UNIVERSITÀ DEGLI STUDI DI BARI

FACOLTÀ DI SCIENZE MM.FF.NN
DIPARTIMENTO DI INFORMATICA

DOTTORATO DI RICERCA IN INFORMATICA — XVI CICLO

Theory of Fuzzy Information Granulation: Contributions to Interpretability Issues

Autore:
Corrado Mencar

Tutor:
Prof. Anna Maria Fanelli

Tesi sottomessa in parziale accoglimento dei requisiti per il conseguimento del titolo di
DOTTORE DI RICERCA IN INFORMATICA

Dicembre 2004

UNIVERSITY OF BARI

FACULTY OF SCIENCE
DEPARTMENT OF INFORMATICS

PHD COURSE IN INFORMATICS — CYCLE XVI

Theory of Fuzzy Information Granulation: Contributions to Interpretability Issues

Author:
Corrado Mencar

Supervisor:
Prof. Anna Maria Fanelli

A thesis submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSPHY IN INFORMATICS

December 2004

Abstract

Granular Computing is an emerging conceptual and computational paradigm for information processing, which concerns representation and processing of complex information entities called "information granules" arising from processes of data abstraction and knowledge derivation. Within Granular Computing, a prominent position is assumed by the "Theory of Fuzzy Information Granulation" (TFIG) whose centrality is motivated by the ability of representing and processing perception-based granular information. A key aspect of TFIG is the process of data granulation in a form that is interpretable by human users, which is achieved by tagging granules with linguistically meaningful (i.e. metaphorical) labels belonging to natural language. However, the process of interpretable information granulation is not trivial and poses a number of theoretical and computational issues that are subject of study in this thesis.

In the first part of the thesis, interpretability is motivated from severa points of view, thus endowing with a robust basis for justifying its study within the TFIG. On the basis of this analysis, the constraint-based approach is recognized as an effective means for characterizing the intuitive notion of interpretability. Interpretability constraints known in literature are hence deeply surveyed with a homogeneous mathematical formalization and critically reviewed from several perspectives encompassing computational, psychological, and linguistic considerations.

In the second part of the thesis some specific issues on interpretability constraints are addressed and novel theoretical contributions are proposed. More specifically, two main results are achieved: the first concerns the quantification of the distinguishability constraint through the possibility measure, while the second regards the formalization of a new measure to quantify information loss when information granules are used to design fuzzy models.

The third part of the thesis is concerned with the development of new algorithms for interpretable information granulation. Such algorithms enable the generation of fuzzy information granules that accurately describe available data and are properly represented both in terms of quantitative and qualitative linguistic labels. These information granules can be used as building blocks for designing neuro-fuzzy models through neural learning. To avoid interpretability loss due to the adaptation process, a new architecture for neuro-fuzzy networks and its learning algorithm are proposed with the specific aim of interpretability protection.

Contents

It is not a case that 'wife' and 'life'
are almost identical words.
And I know the reason.

To Laura, with love.

Acknowledgements

This work is the result of three years of research efforts during my Ph.D. studentship. Of course, most of such work would have not been possible without the support of many people that I wish to thank. First of all, my gratitude goes to my supervisor and tutor, prof. Anna Maria Fanelli, who supported me all the time by guiding my work on the right direction. I am also thankful to her for patiently reviewing and improving my work. Last, but not least, she established a pleasant working environment that I found absolutely beneficial for my work.

I express my gratitude also to prof. Andrzej Bargiela, of Nottingham Trent University (Nottingham, United Kingdom) for his fruitful interaction within my research work, which stimulated the development of new contributions that are part of this thesis. He also shed light on new ideas for future scientific investigation. I am grateful also to my co-tutor, prof. Donato Malerba, for his suggestions that turned out to be useful for this work as well as for future research.

Many thanks also to my colleagues Dr. Giovanna Castellano and Dr. Ciro Castiello of the Department of Informatics, University of Bari. Giovanna gave me useful suggestions for the development of my work and for supporting the review of the final thesis. Ciro enlightened me on many philosophical facets of our common research interests.

Finally, a loveful thank to my wife for always standing by me all the time, especially in the toughest moments.

Preface

Granular Computing is an emerging conceptual and computational paradigm for information processing, which plays a fundamental role within the field of Computational Intelligence. It concerns representing and processing complex information entities called "information granules", which arise in the process of abstraction of data and derivation of knowledge from information. Within Granular Computing, a number of formal frameworks have been developed, among which the "Theory of Fuzzy Information Granulation" (TFIG) assumes a prominent position. The centrality of TFIG is motivated by the ability of representing and processing perception-based granular information, thus providing an effective model of human reasoning and cognition. A key aspect of TFIG is the granulation of data in a form that is readily understandable by human users. Interpretability of the resulting fuzzy information granules is usually achieved by tagging them with linguistically meaningful (i.e. metaphorical) labels, belonging to natural language. However, the semantics of information granules resulting from unconstrained granulation processes hardly matches metaphors of linguistic labels. As a consequence, an interesting research issue arises in the development of fuzzy information granulation methods that ensure interpretability of the derived granules. The definition of such methods cannot be separated from a deep theoretical study about the blurred notion of interpretability itself.

This thesis provides some contributions to the study of interpretability issues in the context of TFIG.

The introductory part of the thesis aims at motivating and defining the notion of "interpretability" within TFIG. Specifically, the First Chapter introduces the notion of interpretability in TFIG from several points of view, including computational, epistemological and linguistic ones. In the attempt to formalize the notion of interpretability in Fuzzy Information Granulation, the constraint-based approach is investigated. As there is not an universally accepted set of constraints that characterize the notion of interpretability, in the Second Chapter a corpus of interpretability constraints is collected, homogeneously treated and critically reviewed. Both the Chapters trace an

introductory path that serves as a basis for novel theoretical contributions in the field of Interpretable Fuzzy Information Granulation, and supplies a formal support to design new algorithms for the extraction and refinement of interpretable information granules.

Based on this preliminary study, some theoretical contributions are proposed in the second part of the thesis, by addressing specific issues on interpretability constraints. The first contribution, introduced in the Third Chapter, is focused on the "Distinguishability" constraint, which is deeply surveyed, with special attention to the measures that quantify its fulfillment. More specifically, two measures, namely "similarity" and "possibility" are compared from a theoretical point of view. As a result, the possibility measure emerges as a valid tool to quantify distinguishability with a clear semantics and a low computational cost.

A second theoretical contribution is presented in the Fourth Chapter, which is concerned with the "Information Equivalence Criterion" constraint. A new mathematical characterization is provided for this constraint, which applies to input and output interfaces of a fuzzy model. It is proved that, for a specific class of input interfaces, the Information Equivalence Criterion is met under mild conditions. For output interfaces, which can hardly fulfill such constraint, a measure is defined to quantify the degree of fulfillment. Such measure can be conveniently used to compare different output interfaces, thus serving as a tool for designing interpretable fuzzy models.

The third part of the thesis is concerned with the development of algorithms for interpretable fuzzy information granulation. The first algorithm, presented in the Fifth Chapter, wraps a fuzzy clustering scheme to achieve interpretable granulation by solving a constrained quadratic programming problem. Key features of such algorithm are a low computational cost and a compact representation of the resulting granules.

The Sixth Chapter introduces an algorithm for information granulation to generate information granules in form of hyper-boxes, thus providing an interval-based quantification of data. The granulation algorithm extends a standard fuzzy clustering algorithm to deal with Minkowski metrics of high order. The algorithm produces boxlike granules whose shape may be distorted by some factors. Such distortion effects are deeply examined and properly quantified, in order to analytically evaluate the effectiveness of the granulation process.

Interval representation of information granules has been adopted also to extend the rule-based formalization of knowledge base of some fuzzy inference systems in order to improve its interpretability. In the Seventh Chapter, such extension is obtained with an algorithm for deriving prediction intervals for fuzzy models acquired from data, and its effectiveness is illustrated on a

real-world complex prediction problem.

All such algorithms are especially useful in applicative contexts where information granulation is necessary to represent fuzzy quantities (i.e. fuzzy numbers or fuzzy vectors). In other applications, qualitative representations of information granules are rather preferred. Qualitative information granules represent information in terms of natural language terms – such as adjectives, adverbs etc. – rather than quantitative labels. Although less precise, qualitative information granules are at the basis of human perception, reasoning and communication; hence it is highly desirable that machine could acquire knowledge expressible in the same terms. This is not a trivial task, however, since interpretability constraints are very stringent and require specifically suited algorithms for qualitative information granulation.

In the Eighth Chapter an algorithmic framework is proposed for qualitative information granulation. It is called "Double Clustering" because it is pivoted on two clustering algorithms that can be chosen on the basis of applicative considerations. The proper combination of the two clustering algorithms – the former aimed at discovering hidden relationships among data and the latter at defining interpretable fuzzy sets – enables the generation of fuzzy information granules that properly describe available data and can be conveniently represented by means of qualitative linguistic terms. Some of these instances are described in detail in the Chapter, and their effectiveness is assessed on some benchmark problems.

Information granules acquired from data can be used in fuzzy models to solve problems of prediction, classification, system identification etc.. A common and effective type of fuzzy model is the "Fuzzy Inference System", which enables the representation of the embodied knowledge in form of a rule base, being each rule explicitly defined by means of fuzzy information granules. The key feature of Fuzzy Inference System is its possibility of being translated into a neural network that can be trained to acquire an unknown functional relationship from data. However, a classical unconstrained neural learning scheme can easily hamper the interpretability of the model's knowledge base, since interpretability constraints ace be easily violated during the training process. To avoid the invalidation of interpretability in adapting information granules, a novel neural architecture is proposed in the Ninth Chapter – together with an appropriate learning scheme based on gradient descent. The resulting neural model enables interpretability protection in all stages of neural learning. The correctness of the proposed architecture in preserving interpretability constraints is proved formally and experimentally verified on a set of benchmark problems.

All the contributions reported in this work finally emerge as a unified framework for extraction and refinement of interpretable fuzzy information

granules. The research direction cast by this work is however by no way exhausted, but it is open to further improvements and extensions in future scientific investigations as highlighted in the conclusive part of the thesis.

Most of the work in this thesis has been presented in peer-reviewed international and national conferences, and has been published, or is under consideration for publication, in international journals. The publications related to this thesis are listed below:

Papers on National and International Journals

- G. Castellano, A.M. Fanelli, C. Mencar, "*A Neuro-Fuzzy Network To Generate Human Understandable Knowledge From Data*" in COGNITIVE SYSTEMS RESEARCH 3(2):125-144, Special Issue on Computational Cognitive Modeling, Elsevier, 2002, ISSN: 1389-0417

- G. Castellano, A.M. Fanelli, C. Mencar, "*Generation Of Interpretable Fuzzy Granules By A Double-Clustering Technique*" in ARCHIVES OF CONTROL SCIENCES 12(4):397-410, Special Issue on Granular Computing, Silesian University of Technology, 2002, ISSN: 1230-2384

- G. Castellano, A.M. Fanelli, C. Mencar "*Fuzzy Information Granulation - A Compact, Transparent and Efficient representation*" in JOURNAL OF ADVANCED COMPUTATIONAL INTELLIGENCE AND INTELLIGENT INFORMATICS (JACIII), vol.7(2):160-168, June 2003, Fuji Technology Press, ISSN: 1343-0130

- G. Castellano, C. Castiello, A.M. Fanelli, C. Mencar, "*Knowledge Discovery by a Neuro-Fuzzy Modeling Framework*", in FUZZY SETS & SYSTEMS, VOL. 149(1):187-207, Special Issue on Fuzzy Sets in Knowledge Discovery, Elsevier, ISSN: 0165-0114

- C. Mencar, G. Castellano, A.M. Fanelli, "*Deriving Prediction Intervals for Neurofuzzy Networks*", in MATHEMATICAL AND COMPUTER MODELLING: AN INTERNATIONAL JOURNAL, Elsevier, in press, ISSN 0895-7177 (**invited paper**)

- G. Castellano, C. Castiello, A.M. Fanelli, C. Mencar, "*Una Metodologia Neuro-Fuzzy per la Predizione delle Proprietà dei Rifiuti Industriali*", in INTELLIGENZA ARTIFICIALE, vol. 1(3):27-35, ISSN: 1724-8035

Chapters in International Books

- C. Mencar, "*Extracting Interpretable Fuzzy Knowledge from Data*" in B. Apolloni, F. Kurfess (eds.), "FROM SYNAPSES TO RULES: DISCOVERING SYMBOLIC RULES FROM NEURAL PROCESSED DATA", pp. 109-116, Kluwer Academic/Plenum Publishers, New York, 2002, ISBN: 0306474026 (**invited paper**)

- G. Castellano, A.M. Fanelli, C. Mencar, "*Design of Transparent Mamdani Fuzzy Inference Systems*", in A. Abraham, M. Köppen, K. Franke (eds.), "DESIGN AND APPLICATION OF HYBRID INTELLIGENT SYSTEMS", pp. 468-476, IOS Press, The Netherlands, 2003, ISBN: 1-58603-394-8

- G. Castellano, A.M. Fanelli, C. Mencar, "*Bi-monotonic Fuzzy Sets Lead to Optimal Fuzzy Interfaces*", in F. Masulli, A. Petrosino (eds.) "LECTURE NOTES IN COMPUTER SCIENCE" (Proceedings of the INTENATIONAL WORKSHOP ON FUZZY LOGIC AND APPLICATIONS (WILF 2003)), Springer-Verlag, 9-11 oct. 2003, Istituto Italiano Studi Filosofici, Naples, Italy, in press, Springer-Verlag, in press

- G. Castellano, A.M. Fanelli, C. Mencar, "*Deriving Prediction Intervals for Neurofuzzy Networks*", in T.E. Simos (ed.), COMPUTATIONAL METHODS IN SCIENCES AND ENGINEERING (ICCMSE 2003), pp. 104-109, World Scientific, Kastoria, Greece, 12-16 sept. 2003, ISBN 981-238-595-9

Papers on National and International Conference Proceedings

- G. Castellano, A.M. Fanelli, C. Mencar, "*Discovering Classification Rules from Neural Processed Data*" in atti del VIII CONVEGNO DELL'ASSOCIAZIONE ITALIANA PER L'INTELLIGENZA ARTIFICIALE (AI*IA), pp. 473-482, Siena, Italy, 10-13 sept. 2002

- G. Castellano, A.M. Fanelli, C. Mencar, "*A Double-Clustering Approach for Interpretable Granulation of Data*", in Proceedings of 2ND IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS (IEEE SMC'02), Yasmine Hammamet, Tunisy, 6-9 oct. 2002., ISBN 2-9512309-4-X

- G. Castellano, A.M. Fanelli, C. Mencar, "*A Compact Gaussian Representation of Fuzzy Information Granules*" in Proceedings of JOINT 1ST INTERNATIONAL CONFERENCE ON SOFT COMPUTING AND INTELLIGENT SYSTEMS AND 3RD INTERNATIONAL SYMPOSIUM ON ADVANCED INTELLIGENT SYSTEMS (SCIS&ISIS 2002), Tsukuba, Japan, 21-25 oct. 2002. **Excellent Presentation Award**.

- G. Castellano, A.M. Fanelli, C. Mencar, "*Discovering human understandable fuzzy diagnostic rules from medical data*" in Proceedings of EUROPEAN SYMPOSIUM ON INTELLIGENT TECHNOLOGIES, HYBRID SYSTEMS AND THEIR IMPLEMENTATION ON SMART ADAPTIVE SYSTEMS (EUNITE 2003), pp. 227-233, Oulu, Finland, 10-12 july 2003

- G. Castellano, A.M. Fanelli, C. Mencar, "*Fuzzy Granulation of Multidimensional Data by a Crisp Double Clustering algorithm*", in Proceedings of the 7TH WORLD MULTI-CONFERENCE ON SYSTEMICS, CYBERNETICS AND INFORMATICS (SCI 2003), pp. 372-377, Orlando, Fl, USA, 27-30 july 2003, ISBN 980-6560-01-9

- G. Castellano, C. Castiello, A.M. Fanelli, C. Mencar, "*Discovering Prediction Rules by a Neuro-Fuzzy Modelling Framework*", in Proceedings of 7TH INTERNATIONAL CONFERENCE ON KNOWLEDGE-BASED INTELLIGENT INFORMATION & ENGINEERING SYSTEMS (KES 2003), vol. 1, pp. 1242-1248, Springer, Oxford, UK, 3-5 sept. 2003, ISBN 3-540-40803-7

- G. Castellano, A.M. Fanelli, C. Mencar, "*DCClass: A Tool to Extract Human Understandable Fuzzy Information Granules for Classification*", in Proceedings of the 4TH INTERNATIONAL SYMPOSIUM ON ADVANCED INTELLIGENT SYSTEMS (SCIS&ISIS 2003), pp. 376-379, Jeju Island, Korea, 25-28 sept. 2003

- G. Castellano, A.M. Fanelli, C. Mencar, "*A Fuzzy Clustering Approach for Mining Diagnostic Rules*", in Proceedings of 2003 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN & CYBERNETICS (IEEE SMC'03), vol. 1, pp. 2007-2012, 5–8 oct. 2003 – Hyatt Regency, Washington, D.C., USA, ISBN 0-7803-7952-7

- C. Mencar, A. Bargiela, G. Castellano, A.M. Fanelli, "*Interpretable Information Granules with Minkowski FCM*", in Proceedings of the CONFERENCE OF NORTH AMERICAN FUZZY INFORMATION SOCIETY (NAFIPS2004), pp. 456-461, 27-30 june 2004, Banff, Alberta, Canada (**Best Student Paper Citation**), ISBN 0-7803-8377-X

- G. Castellano, A.M. Fanelli, C. Mencar, "*Optimality Degree Measurement in Fuzzy System Interfaces*", in Proceedings of EUROPEAN SYMPOSIUM ON INTELLIGENT TECHNOLOGIES, HYBRID SYSTEMS AND THEIR IMPLEMENTATION ON SMART ADAPTIVE SYSTEMS (EUNITE 2004), pp. 443-451, 10-12 june 2004, Aachen, Germany

- G. Castellano, A.M. Fanelli, C. Mencar, "*An Optimality Criterion for Fuzzy Output Interfaces*", in Proceedings of IEEE CONFERENCE ON INTELLIGENT SYSTEMS DESIGN AND APPLICATIONS (ISDA 2004), pp. 601-605, 26-28 aug. 2004, Budapest, Hungary

- C. Mencar, G. Castellano, A. Bargiela, A.M. Fanelli, "*Similarity vs. Possibility in Measuring Fuzzy Sets Distinguishability*", in Proceedings of RASC 2004 (RECENT ADVANCES IN SOFT COMPUTING), pp. 354-359, 16-18 December 2004, Nottingham, UK, ISBN 1-84233-110-8

Papers under consideration for publication

- G. Castellano, A.M. Fanelli, C. Mencar, "*Interface Optimality in Fuzzy Inference Systems*", submitted to INTERNATIONAL JOURNAL OF APPROXIMATE REASONING, Elsevier (**invited paper**)

Part I

Interpretability Issues in Fuzzy Information Granulation

Chapter 1

Granular Computing

The eternal mystery of the world
is its comprehensibility
(A. Einstein)

1.1 The emergence of Granular Computing

The last fourty years will be certainly reminded as the age of the "Information Revolution", a phenomenon that refers to the dramatic social changes in which information-based jobs and tasks become more common than jobs and tasks in manufacturing or agriculture. The information revolution – which ultimately moulded our society into an Information Society – would not have happened without computers and related technologies. Computers[1] – and Information Technology in general – are the fundamental tool for storing, retrieving, processing and presenting information. Recently, the rapid advances of Information Technology have ensured that large sections of the world population have been able to gain access to computers on account of falling costs worldwide, and their use is now commonplace in all walks of life.

Government agencies, scientific, business and commercial organizations are routinely using computers not just for computational purposes but also for storage, in massive databases, of the immense volume of data they routinely generate, or require from other sources. Furthermore, large-scale networks, emblematized by the Internet, has ensured that such data has become accessible to more and more people. All of this has led to an information explosion, and a consequent urgent need for methodologies that organize

[1] In 1983, Time magazine picked the computer as its "Man of the Year", actually listing it as "Machine of the Year". This perhaps shows how influential the computer has become in our society.

such high volumes of information and ultimately synthesize it into useful knowledge. Traditional statistical data summarization and database management techniques do not appear sufficiently adequate for handling data on this scale as well as for extracting knowledge that may be useful for exploring the phenomena responsible for the data, and for providing support to decision-making processes (Pal, 2004).

The quest for developing systems that perform intelligent analysis of data, and for intelligent systems in a wider sense, has manifested in different ways and has been realized in a variety of conceptual frameworks, each of them focusing on some philosophical, methodological and algorithmic aspects (e.g. pattern recognition, symbolic processing, evolutionary computing, etc.). One of the recent developments concerns *Computational Intelligence* as a novel methodology for designing intelligent systems.

The definition of Computational Intelligence has evolved during the years from its early introduction as a property of intelligent systems in (Bezdek, 1992):

> [...] a system is *computationally intelligent* when it: deals only with numerical (low-level) data; has a pattern recognition component; does not use knowledge in the AI [Artificial Intelligence] sense [...]

Successively, Computational Intelligence has been recognized as a methodology in (Karplus, 1996):

> [...] CI [Computational Intelligence] substitutes intensive computation for insight into how the system works. NNs [Neural Networks], FSs [Fuzzy Sets] and EC [Evolutionary Computation] were all shunned by classical system and control theorists. CI umbrellas and unifies these and other revolutionary methods.

To a wider extent, Computational Intelligence is ‹‹*a methodology for the design, the application, and the development of biologically and linguistically motivated computational paradigms emphasizing neural networks, connectionist systems, genetic algorithms, evolutionary programming, fuzzy systems, and hybrid intelligent systems in which these paradigms are contained*››[2].

In (Pedrycz, 1997), an interesting definition of Computational Intelligence is provided:

[2]This is actually the scope of The IEEE Computational Intelligence Society, as published at http://ieee-cis.org/

> Computational Intelligence is a research endeavor aimed at conceptual and algorithmic integration of technologies of granular computing, neural networks and evolutionary computing.

The last definition is enhanced by the inclusion of the "*Granular Computing*" as a conceptual backbone of Computational Intelligence.

Granular Computing is an emerging computing paradigm of information processing. It deals with representing and processing of information in form of "*information granules*". Information granules are complex information entities that arise in the process – called "*information granulation*" – of abstraction of data and derivation of knowledge from information (Bargiela and Pedrycz, 2003a). Generally speaking, information granules are collection of entities, usually originating at the numeric level, that are arranged together due to their similarity, functional adjacency, indistinguishability, coherency or alike (Pedrycz, 2001; Bargiela, 2001; Pedrycz and Bargiela, 2002; Zadeh, 1979; Zadeh, 1997; Zadeh and Kacprzyk, 1999; Pedrycz and Vukovi, 1999; Pedrycz and Smith, 1999; Pedrycz et al., 2000).

The notions of information granule and information granulation are highly pervasive and can be applied to a wide range of phenomena. Human perceptions are intrinsically granular: time granules (e.g. years, days, seconds, etc.), image granules (objects, shapes, etc.), auditory granules, etc. are the basis for human cognition. In addition, methodologies and technologies like qualitative modelling, knowledge-based systems, hierarchical systems, etc. all exploit the notion of information granulation. Granular Computing establishes therefore a sound research agenda that promotes synergies between new and already established technologies, and appears especially suited for modelling and understanding intelligent behavior.

1.1.1 Defining Granular Computing

Although it is difficult to give a precise and uncontroversial definition of Granular Computing, it could be described from several perspectives. Granular Computing can be conceived as a label of theories, methodologies, techniques and tools that make use of information granules in the process of problem solving (Yao, 2000a). In this sense, Granular Computing is used as an umbrella term to cover topics that have been studied in various fields in isolation. By examining existing studies in the unified framework of Granular Computing and extracting their commonalities, it could be able to develop a general theory for problem solving. Under such perspective, there is a fast growing interest in Granular Computing (Inuiguchi et al., 2003; Lin, 1998; Lin, 2001; Lin et al., 2002; Pedrycz, 2001; Polkowski and Skowron, 1998; Skowron,

2001; Skowron and Stepaniuk, 1998; Skowron and Stepaniuk, 2001; Yager and Filev, 1998; Yao, 2000b; Yao, 2004; Yao and Zhong, 2002; Zadeh, 1998; Zhong et al., 1999).

In a more philosophical perspective, Granular Computing can be intended as a way of thinking that relies on the human ability to perceive the real world under various levels of granularity, in order to abstract and consider only those things that serve a specific interest, and to switch among different levels of granularity. By focusing on different levels of granularities, one can obtain different levels of knowledge, as well as inherent knowledge structure. Granular computing is thus essential in human problem solving, and hence has a very significant impact on the design and implementation of intelligent systems (Yao, 2004).

The ideas of Granular Computing have been investigated in Artificial Intelligence through the notions of "granularity" and "abstraction". Hobbs (Hobbs, 1985) proposed a theory of granulation observing that:

> We look at the world under various grain sizes and abstract from it only those things that serve our present interests. [...] Our ability to conceptualize the world at different granularities and to switch among these granularities is fundamental to our intelligence and flexibility. It enables us to map the complexities of the world around us into simpler theories that are computationally tractable to reason in.

The notions of granularity and abstraction are used in many fields of Artificial Intelligence. As an example, the granulation of time and space plays an important role in spatial and temporal reasoning (Bettini and Montanari, 2000; Bettini and Montanari, 2002; Euzenat, 2001; Hornsby, 2001; Stell and Worboys, 1998). Furthermore, based on such notions, many authors studied some fundamental topics of Artificial Intelligence, such as knowledge representation (Zhang and Zhang, 1992), theorem proving (Giunchigalia and Walsh, 1992), search (Zhang and Zhang, 1992; Zhang and B., 2003), planning (Knoblock, 1993), natural language understanding (Mani, 1998), intelligent tutoring systems (McCalla et al., 1992), machine learning (Saitta and Zucker, 1998) and data mining (Han et al., 1993).

1.1.2 Information Granulation

A fundamental task of Granular Computing is the construction of information granules, a process that is called "*information granulation*". According to Zadeh, granulation is one of three main tasks that underlie human cognition:

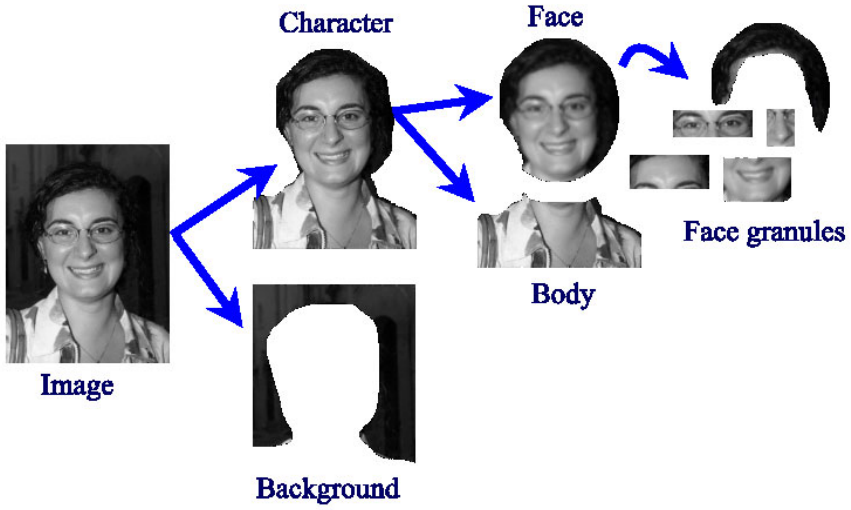

Figure 1.1: Example of a granulation process

granulation, organization and causation (Zadeh, 1997). Specifically, granulation is governed by the following "*granulation principle*" (Zadeh, 2000):

> to exploit the tolerance for imprecision, employ the coarsest level of granulation[3] which is consistent with the allowable level of imprecision.

In a broad sense, granulation involves decomposition of the whole into parts (see fig. 1.1), while organization involves integration of parts into whole and causation relates to the association of causes with effects.

Independently on the technology involved for information granulation, there are several essential factors that drive all the pursuits of information granulation (Bargiela and Pedrycz, 2003a). Such factors include:

- The need to split a problem into more tractable sub-problems, according to the well-known "*divide et impera*" strategy. In such context, granulation serves as an efficient vehicle to modularize the problem;

- The need to comprehend a problem by providing a better insight into its essence rather than being overwhelmed with all unnecessary details. Here, information granulation serves as an abstraction mechanism that

[3]Roughly speaking, the level of granulation is the number of objects in the granule related to the total number of granules employed in a context

hides unnecessary information. By changing the level of granularity, it is possible to hide or reveal details according to the required specificity during a certain design phase;

- The need for processing information in a human-centric modality. Actually, information granules do not exist as tangible physical entities but they are conceptual entities that emerge from information as a direct consequence of the continuous quest for abstraction, summarization and condensation of information by human beings.

The process of information granulation and the nature of information granules imply the definition of a formalism that is well-suited to represent the problem at hand. There are a number of formal frameworks in which information granules can be defined. Such frameworks are well-known and thoroughly investigated both from a theoretical and applicative standpoint. A short list of such frameworks, which includes the most common used within Granular Computing, is the following:

- **Set Theory and Interval Analysis** (Bargiela, 2001; Hansen, 1975; Jaulin et al., 2001; Kearfott and Kreinovich, 1996; Moore, 1962; Morse, 1965; Sunaga, 1958; Warmus, 1956)

- **Fuzzy Set Theory** (Dubois and Prade, 1980; Dubois and Prade, 1997; Kandel, 1982; Klir and Folger, 1988; Klir, 2001; Pedrycz and Gomide, 1998; Zadeh, 1965; Zadeh, 1975a; Zadeh, 1975b; Zadeh, 1978; Zimmermann, 2001)

- **Rough Set Theory** (Lin and Cercone, 1997; Pal and Skowron, 1999; Pawlak, 1982; Pawlak, 1999; Nguyen and Skowron, 2001)

- **Probabilistic (Random) Set Theory** (Butz and Lingras, 2004; Cressie, 1993; Matheron, 1975; Stoyan et al., 1995; Zadeh, 2002; Zadeh, 2003)

- **Dempster-Shafer Belief Theory** (Dempster, 1966; Dubois and Prade, 1992; Fagin and Halpern, 1989; Klir and Ramer, 1990; Nguyen, 1978; Ruspini et al., 1992; Shafer, 1976; Shafer, 1987)

- etc.

1.2 Theory of Fuzzy Information Granulation

Among all formal frameworks for Granular Computing, Fuzzy Set Theory has undoubtedly a prominent position[4]. The key feature of Fuzzy Set Theory stands in the possibility of formally express concepts of continuous boundaries. Such blurred concepts are at the core of human perception processes, which often end up with linguistic terms belonging to natural language. It is evident that while these concepts are useful in the context of a certain problem, as well as convenient in any communication realized in natural language, their set-based formal model will lead to serious representation drawbacks. Such "epistemic divide" between concepts and their formalization is primarily due to the vagueness property of conceptual representations, as well-pointed by Russel (Russell, 1923):

> [...] Vagueness and precision alike are characteristics which can only belong to a representation, of which language is an example. They have to do with the relation between a representation and that which it represents. Apart from representation, whether cognitive or mechanical, there can be no such thing as vagueness or precision; things are what they are, and there is an end of it. [...] law of excluded middle is true when precise symbols are employed, but it is not true when symbols are vague, as, in fact, all symbols are. [...] All traditional logic habitually assumes that precise symbols are being employed. It is therefore not applicable to this terrestrial life, but only to an imagined celestial existence.

Vagueness cannot be removed from human cognition, as already pointed out by the philosopher Arthur Schopenhauer (1788–1860), whose position could be synthesized with the words of Tarrazo (Tarrazo, 2004):

> we work with (symbolic) representations of ourselves, our environment, and the events we notice; these representations are approximate at best because we simply do not know enough; what matters to us is to make better decisions using these subjective representations. We employ those representations to solve important problems, which are unfailingly interdisciplinary and forward-looking. Anything interdisciplinary is bound to be only

[4]Granular Computing is often subsumed by Fuzzy Set Theory, especially in the context of Computational Intelligence and Soft Computing. However, it is thought that a distinction of the two terms acknowledges the credits of all non-fuzzy, yet granular-based, frameworks.

> imperfectly represented. Further, the problems we must address are important (e.g. career decisions, house purchases, retirement financing, etc.) because their effects are often felt substantially and during many years into the future. Very importantly, our strategies to deal with them must be played out in a shifting future where many elements can only be approximated.

Fuzzy Sets are the contribution of Lotfi Zadeh (Zadeh, 1965). They can be conceived as generalizations of ordinary sets that admit partial membership of elements. Based on such extended assumption, Fuzzy Set Theory – and the corresponding Fuzzy Logic – has been formalized in several ways (Bellman and Giertz, 1973; Goguen, 1967; Gottwald, 1979; Hájek, 1998; Mendel and John, 2002; Novák et al., 1999; Ramot et al., 2002). In the simplest setting, a fuzzy set is described by a membership function, which maps the elements of a Universe of Discourse U into the unit interval $[0, 1]$. Membership functions quantify the notion of partial membership and define the semantics of a fuzzy set. Dubois and Prade (Dubois and Prade, 1997) discussed three points of interpretation of fuzzy set semantics:

Similarity The membership value of an element of the Universe of Discourse is interpreted as a degree of proximity to a prototype element. This is an interpretation that is particularly useful in the field of Pattern Recognition (Bellman et al., 1966). This is also an appropriate interpretation in the context of Granular Computing, since information can be granulated based on the similarity of object to some prototypes.

Preference A fuzzy set represent a granule of more or less preferred objects and the membership values indicate their respective degree of desiderability, or – in a decision making context – the degree of feasibility (Saaty, 1980). This point of view is deeply rooted in the realm of decision analysis with approximate reasoning.

Uncertainty The membership value determines the degree of possibility that some variable assumes a certain value. The values encompassed by the support of the fuzzy sets are mutually exclusive and the membership grades rank these values in terms of their plausibility (Zadeh, 1978). Possibility can have either a physical or an epistemic meaning. The physical meaning of possibility is related to the concept of feasibility, while its epistemic meaning is associated to the degree of "unsurprisingness" (Dubois and Prade, 1988).

From their introduction, fuzzy sets gained numerous achievements, especially in applicative frameworks. According to Zadeh, fuzzy sets gained

success because they ‹‹*exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost*›› (Zadeh, 1994). Fuzzy Set Theory is not the only theoretical framework that allows such kind of exploitation: actually several methodologies are inspired to the tolerance for imprecision and uncertainty, and they fall in the wider field of Soft Computing[5]. However, the key feature of Fuzzy Set Theory stands in the possibility of giving a formal and procedural representation of linguistic terms used to express human-centered concepts (Ostasiewicz, 1999). This facet is in line with the basic principles of Granular Computing and it led to the development of the Theory of Fuzzy Information Granulation, whose principal aim is to give a mathematical foundation to model the cognitive tasks of humans in granulating information and reason with it (Zadeh, 1997).

1.2.1 Granulation as a cognitive task

Within the Theory of Fuzzy Information Granulation, a special emphasis is given to the cognitive tasks of granulation (intended as dividing a whole into parts) and fuzzification, that replaces a crisp set with a fuzzy set. The combination of granulation and fuzzification (called "f.g-granulation" or "f.g-generalization") leads to the basic concept of linguistic variable, fuzzy rule set and fuzzy graph.

The necessity for f.g-generalization is given by several motivations, including:

1. Unavailability of precise (i.e. crisp or fine-grained) information. This occurs in everyday decision making, as well as economic systems, etc.

2. High cost for precise information, e.g. in diagnostic systems, quality control, decision analysis, etc.

3. Unnecessity of precise information. This occurs in most human activities, like parking, cooking, art crafting, etc.

4. Low-cost systems, like consumer or throw-away products, where the trade-off between quality and cost is critical.

The Theory of Fuzzy Information Granulation is based on the concept of generalized constraint, according to which an information granule formally represents an elastic constraint on the possible values of a modelled

[5]The terms "Soft Computing" and "Computational Intelligence" assumed over the years a convergent meaning, so that it is common nowadays to refer to both terms as synonyms.

attribute. This theory provides a basis for Computing with Words, which is a methodology where words are used in place of numbers for computing and reasoning. Fuzzy Sets plays a pivotal role in Computing with Words in the sense that words are viewed as labels of fuzzy granules, which in turn model soft constraints on physical variables (Zadeh, 1996a). The inferential machinery that enables information processing within Computing with Words is the Approximate Reasoning, which is accomplished by means of the underlying Fuzzy Set Theory (Yager, 1999). Computing with Words provides a powerful tool for human-centric – yet automated – information processing. It allows the setting of a highly interpretable knowledge base as well as a highly transparent inference process, thus capturing the transparency benefits of symbolic processing without the inferential rigidity of classical expert systems. The Theory of Fuzzy Information Granulation and Computing with Words shed new lights in classical Artificial Intelligence issues, as exemplified by the Precisiated Natural Language and Semiotical Cognitive Information Processing for natural language processing and understanding (Rieger, 2003; Zadeh, 2004).

The Theory of Fuzzy Information Granulation can be also cast into the realm of Cognitive Science, whose objectives include the understanding of cognition and intelligence (Rapaport, 2003). In such context, the Theory of Fuzzy Information Granulation appears as a promising conceptual tool to bridge the historical gap between the "Physical Symbol System Hypothesis" – which states that cognition can be understood as symbolical sequential computations that use mental representation as data (Fodor, 1975) – and the connectionist approach, according to which intelligence and cognition emerge from a system with an appropriate organized complexity, without being necessarily coded into symbols (Knight, 1990). The Theory of Fuzzy Information Granulation offers a double layer of knowledge representation: a first, symbolic layer of linguistic terms eventually structured within knowledge representation schemes, like fuzzy rule, fuzzy graphs, etc.; and a numerical level for expressing the semantics of the linguistic terms, which can emerge from perception-based sub-systems (such as neural networks) and can manipulated by means of Approximate Reasoning.

1.2.2 Symbolic representation of fuzzy information granules

The key advantage of a symbolically represented knowledge is the possibility of communicating it among agents. Communication can have a continuous or discrete (analog) nature. Usually, analog communication is used to express

emotive information among natural agents. Music, visive arts, but also odors, voice tones and animal sounds are all forms of analog communications. However, communication of knowledge is always discrete and symbolical[6]. The reasons behind the discrete nature of knowledge communication are the same the justify the overwhelming success of digital computers over the analog devices: simplicity and reliability.

As a discrete process, knowledge communication needs a language of symbols to exist. Generally speaking, a language can be regarded as a system consisting of a representation (symbols) along with metaphor and some kind of grammar. Metaphor is a relation between representation and semantics and is implicitly shared among all communicating actors. If a metaphor is not shared among actors, communication cannot take place. As an example, the word (symbol) TALL, in the context of human heights, has a shared metaphor among all English-literate people and hence can be used in knowledge communication; on the other hand, the word AXDRW cannot be used in communication until its meaning is fully specified.

For the purposes of this study, knowledge communication can be categorized according to the type of actors involved. A first type of communication takes place when all actors are human. Here, natural language is almost always used[7], which is often integrated with analog emotive forms of communications (facial expressions, tones, gestures, etc.). Natural language is very rich, but it is ambiguous, imprecise and has a strong pragmatic component that is hard to model in any computational paradigm. It is nevertheless the language that allowed the evolution of Science until the present days, even if in the last two centuries it has been strongly supported by mathematics. Nevertheless, the strict precision of the mathematical formalism has been challenged in recent years by the development of alternative scientific approaches, like qualitative modelling (IEE, 1988; Kuipers, 1994).

The knowledge objects exchanged in a natural language communication are sentences made of linguistic labels. Such labels *denote* real-world objects by a composite cognitive task of *perception* and *speech*. Perception is understood as a process of translating sensory stimulation into an organized experience. Though the perceptual process cannot be directly observed, relations

[6] It is interesting to note that communication of knowledge involves a process of recognition of symbolical structures from analog media. Examples of such process are the recognition of an electric/magnetic state of a digital medium as well as the recognition of utterances, words or gestures. All such processes are different forms of information granulation.

[7] With exception to mathematical communication, where a formal language is preferred. However, it is very common that mathematical communication is enriched with several natural language notes to help understanding the involved concepts.

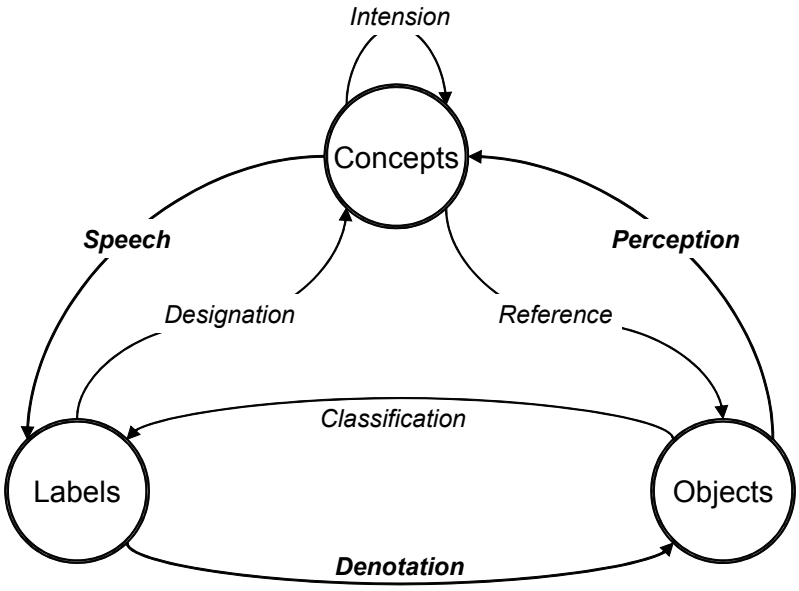

Figure 1.2: Basic epistemic relationships between labels, concepts and objects

can be found between the various type of stimulation and their associated experiences or concepts[8]. The latter are complex constructions of simple elements joined through association and are basic for human experience, which however, is of organized wholes rater than collections of elements. In fig. 1.2, the relationships between linguistic labels, objects and concepts are depicted. Such relations correspond to a cognitive task, and include other important tasks such as denotation, reference, intension and classification. This model, introduced in (Toth, 1999) highlights the prominent importance of the perception task within human cognition.

According to H. Toth (Toth, 1997), fuzziness is a property of perceptions, i.e. concepts originated from perceptive tasks. Perceptions, as vaguely defined entities, stand opposite to measurement-based objects, which are the basic knowledge granules of classical Science. However, even though classical Science has led to brilliant successes during time, it failed to solve problems in which humans are mostly adequate. Agile robots, text summarization,

[8]Strictly speaking, perception is a cognitive process whose results are called percepts. However, since there is not a uniform terminology in literature, hereafter the term "perception" will be used to indicate both the process and the results. The latter will also be denoted as "concepts".

language translation, car driving, etc. are only few examples of problems that are not – and maybe cannot be – solved by measurement-based science. Based on such consideration, novel perspectives can be envisioned that take perceptions as central objects for scientific exploitation (Barsalou, 1999; Zadeh, 2000).

Opposite to human communication is communication among computers. In the most usual case, a formal language is adopted so as to avoid any possible information loss due to the adoption of imprecise languages. This common choice is motivated by the necessity of exact information processing that can replace humans in the most tedious and mechanical tasks. Such formal languages are inherently measurement-based and are specifically suited for symbolic information processing. However, with the advent of intelligent agents that act in complex environments, precise knowledge communication has been challenged and novel research directions, like "communication between granular worlds", have been opened (Bargiela and Pedrycz, 2003a).

When communication involves both humans and computers, two significant distinctions must be made, depending on the direction of communication. In the case of human-to-computer communication, the languages used are almost always formal in order to avoid ambiguity and indeterminacy in computer information processing. Most of such kind of languages are universal, i.e. they can express anything (or almost anything) that can be computable in the Turing sense. The distinction between all such languages stands in the specific purposes the language is design for (e.g. programming languages, query languages, declarative languages, etc.). The principal metaphors of such languages are usually simple, well-defined and measurement-based. For example, the metaphors of a declarative programming language are: predicates, terms and connectives.

In the classical case, perceptions can be communicated to computers in forms of facts and predicates, but their meaning must be exhaustively specified with a huge corpus of axioms and statements that – apart from trivial cases – only approximate the real significance of the perceptions. As a consequence of the simple metaphors involved in such formal languages, the computer does not understand the meaning of perceptions but can derive facts based on syntactical derivation rules. Such possibility of reasoning and acting based only on formal manipulation of symbols is at the center of the philosophical debate on the possibility of a real intelligence in computer systems (Dreyfus and Dreyfus, 1986; Searle, 1980). Furthermore, most computer languages do not support vagueness in representing information, thus disabling any possibility of representing perceptions in a human-centered fashion. In such a situation, Granular Computing, and especially the Theory of Fuzzy Information Granulation with derived paradigms like Precisiated

Natural Language, promise to be effective tools for humans to communicate perception-based information also to computer systems.

When computers communicate knowledge to humans, they can use several means. Usually, text-based communication provides for measurements and precise information (conditions, states, facts, etc.), while graphical – and multimedia – communication easily conveys perception-based information. For example, charts provide graphical objects that are perceived by humans to form judgements, which could be subsequently used to take decisions. *Intuitive* communication of information is a research issue within the field of Human-Computer Interaction (Raskin, 1994; Shneiderman and Plaisant, 2004). However, it should be observed that computers still maintain measurement-based information, which is rendered in visual form to stimulate perceptions by humans; as a consequence, computers still cannot perform perception-based reasoning and actions. An interesting question concerns the possibility for computers of dealing with perception-based information both for reasoning and communication. The Theory of Fuzzy Information Granulation provides a way for representing perception-based information, and Computing with Words – along with the Logic of Perception – enables reasoning and communication of perception-based knowledge. However, such theories and methodologies do not tackle the problem of *learning* perception-based knowledge, or stated differently, forming information granules from data.

1.2.3 Interpretability of fuzzy information granules

The automatic generation of information granules from data is an extremely important task, since it gives to machines the ability of adapting their behavior dynamically. Granules generation (or extraction) can be accomplished through the application of learning techniques especially suited for the formal framwork in which information granules are defined.

Within the Theory of Fuzzy Information Granulation, the extraction of information granules can be achieved through learning techniques that acquire fuzzy sets from data. Learning techniques that make use of Fuzzy Set Theory are numerous and embrace several approaches. Fuzzy Clustering (Baraldi and Blonda, 1999), Fuzzy Decision Trees (Janikow, 1998), Fuzzy Classifiers (Kuncheva, 2000), Fuzzy Cognitive Maps (Kosko, 1986), Neuro-Fuzzy Networks (Nauck et al., 1997), etc. are just few examples of methods and models that are aimed to acquire knowledge from data defined by means of fuzzy sets. It could be stated that almost all learning techniques classically belonging to Machine Learning, Explorative Data Analysis and Data Mining have been adapted to acquire knowledge represented by fuzzy sets or

fuzzy relations. Furthermore, the existence of universal approximation models based on Fuzzy Set Theory has been proved (Wang, 1992). However, such models and techniques do not take into account the perceptive nature of the acquired knowledge. The communication of such knowledge to a human user must be carried out only by means of a formal language, which associates arbitrary symbols to fuzzy sets and provides for each of them a functional form to represent their semantics. This language for communication is very far from natural language, where symbols carry a meaning *ipso facto*, on the basis of an implicit metaphor shared by the actors of the communication. As a consequence of the language gap, communication of the knowledge acquired by adaptive models can be only made by means of measurements (e.g. the parameters of the fuzzy membership functions), thus losing the potential benefits of a perception-based framework for reasoning and communicating.

To obtain perception-based models, the learning algorithms used to acquire knowledge from data must yield fuzzy information granules that can be naturally labelled by linguistic terms, i.e. symbols that belong to the natural language. The attachment of such terms to fuzzy sets must be done by taking into account the metaphor carried by each linguistic term. As an example, the linguistic term TALL can be assigned to a fuzzy set only if its semantics actually embraces the class of heights that people perceive as tall (provided a tolerance for the vague definition off "being tall") and does not include (or includes with low membership) heights that people do not perceive as tall. Information granules that can be attached to a linguistic term are called *interpretable* information granules, and models based on interpretable information granules are called interpretable fuzzy models.

In literature, the term interpretability is often assimilated as a synonymous of *transparency*, so that the two words are used interchangeably. In addition, in the attempt to give a distinct definition of the two terms, some confusion exists. As an example, in (Riid and Rüstern, 2003) the definitions of transparency and interpretability are exchanged w.r.t. definitions given in (Johansson et al., 2004)[9]. The two terms actually refer to two distinct properties of models. Hereafter, the term "transparency" is meant to refer to an inherent property of a model whose behavior can be explained in terms of its components and their relations. On the other hand, a model is interpretable if its behavior is *intelligible*, i.e. it can be easily perceived and understood by a user. Transparency is a metamathematical concept[10] while interpretability has a more cognitive aspect and its study permeates several disciplines, including Machine Learning (IJCAI, 1995). Based on such definitions, fuzzy

[9] Here, the term *comptrehensibility* is used instead of interpretability

[10] Transparency can be related with "white-box" models

models can be considered inherently transparent, because their behavior is clearly explained in terms of the involved information granules and the inference process, but their interpretability cannot be guaranteed until further insights are made.

Interpretability in Fuzzy Information Granulation is an intuitive property, because it deals with the association of fuzzy information granules – which are mathematical entities – with metaphorical symbols (the linguistic terms), which conversely belong to the blurry realm of human linguistics. As a consequence, research on interpretable information granules and interpretable fuzzy models does not lead to a unique direction. Furthermore, interpretability issues have been investigated only in recent years since formerly accuracy was the main concern of fuzzy modelling.

Interpretability issues in fuzzy modelling has been tackled in several ways, including two main approaches. A first approach concerns the complexity reduction of the knowledge structure, by balancing the trade-off between accuracy and simplicity. The strategy of reducing the complexity of the knowledge structure is coherent to the notorious Occam Razor's Principle[11], which is restated in Information Theory as the Minimum Description Length principle (Rissanen, 1978). Complexity reduction can be achieved by reducing the dimensionality of the problem, e.g. by feature selection (Cordón et al., 2003; Tikk et al., 2003; Vanhoucke and Silipo, 2003), while feature extraction (e.g. Principal Component Analysis) is not recommended because the extracted features do not have any physical – hence interpretable – meaning. Other methods to reduce complexity include the selection of an appropriate level of granularity of the knowledge base (Cordón et al., 2003; Guillaume and Charnomordic, 2003), the fusion of similar information granules (Espinosa and Vandewalle, 2003; Abonyi et al., 2003; Sudkamp et al., 2003), elimination of low-relevant information granules (Baranyi et al., 2003; Setnes, 2003), and hierarchical structuring of the knowledge base (Tikk and Baranyi, 2003a).

The Minimum Description Length principle has been effectively adopted in Artificial Intelligence for inductive learning of hypotheses (Gao et al., 2000). However, in interpretable fuzzy modelling it could be insufficient to guarantee interpretable information granules since labelling with linguistic terms could be still hard if further properties, other than simplicity, are not satisfied. Based on such consideration, some authors propose the adoption of several constraints that are imposed on the knowledge based and on the

[11] ‹‹*Pluralitas non est ponenda sine neccesitate*››, which translates literally into English as "Plurality should not be posited without necessity". Occam's Razor is nowadays usually stated as "of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred"

model to achieve interpretability. These interpretability constraints actually define the mathematical characterization of the notion of interpretability. Several authors propose a constraint-oriented approach to interpretable fuzzy modelling, including the pioneering works of de Oliveira (Valente de Oliveira, 1999b) and Chow et al. (Chow et al., 1999).

Interpretability constraints force the process of information granulation to (totally or partially) satisfy a set of properties that are judged necessary to allow the attachment of linguistic labels to the information granules constituting the knowledge base. A preliminary survey of methods to protect interpretability in fuzzy modelling is given in (Guillaume, 2001; Casillas et al., 2003). Constrained information granulation of data can be achieved in several ways, which can be roughly grouped in three main categories

Regularized learning The learning algorithms are aimed to extract information granules so as to optimize an objective function that promotes the definition of an accurate knowledge base but penalizes those solutions that violate interpretability constraints. The objective function must be encoded in an appropriate way so as to be applied to classical constrained optimization techniques (e.g. Lagrangian multipliers method).

Genetic algorithms Information granules are properly encoded into a population of individuals that evolve according to an evolutionary cycle, which involves a selection process that fosters the survival of accurate and interpretable information granules. Genetic algorithms are especially useful when the interpretability constraints cannot be formalized as simple mathematical functions that can be optimized according to classical optimization techniques (e.g. gradient descent, least square methods, etc.). Moreover, the multi-objective genetic algorithms are capable to deal with several objective functions simultaneously (e.g. one objective function that evaluates accuracy and another to assess interpretability). However, the drawback of genetic algorithms is their inherent inefficiency that restricts their applicability.

Ad-hoc algorithms Interpretable information granules are extracted according to learning algorithms that encode interpretability constraints within the extraction procedure. Such type of algorithms are usually more efficient than others but, due to their variable design, are difficult to develop.

Scientific investigation in interpretable information granulation is still open. Despite the numerous techniques for interpretable fuzzy modelling,

there in no agreement on a definitive set of constraints that characterize interpretability. As a consequence, different sets of constraints are proposed, which involve the definition of methods and technique that are sometimes incomparable. Such disagreement is due to the blurry nature of the notion of interpretability, which is subjective and in some cases application oriented. Moreover, each single interpretability constraint can have a psychological support, a common sense justification or even it has no justifications at all. For such reason, a comprehensive study of all interpretability constraints proposed in literature would lead to a number of benefits, including:

- An homogenous description, which helps the selection of interpretability constraints for specific applicative purposes;
- The identification of potential different notions of interpretability, related to the nature of the information granules (e.g. information granules describing quantities vs. those describing qualities);
- A critical review of interpretability constraints, which may help to discard those constraints that are not strongly justified by objective or experimental supports, or to subsume different constraints in more generalized ones, from which novel properties may emerge;
- A common notation, which helps the encoding of different interpretability constraints within the same framework or algorithm;
- The possibility of identifying new methods for interpretable information granulation.

In the next Chapter, interpretability constraints known in literature are deeply surveyed with a homogeneous mathematical formalization and critically reviewed from several perspectives encompassing computational, psychological, and linguistic considerations.

Chapter 2

Interpretability Constraints for Fuzzy Information Granules

2.1 Introduction

The aim of this Chapter is to deeply survey the mathematical characterization of the intuitive notion of interpretability within the context of the Theory of Fuzzy Information Granulation. The survey is the result of a systematic insight of scientific works concerning interpretable fuzzy modelling.

A common approach used to define interpretability involves the definition of a set of constraints to be verified on information granules. Unfortunately, there is not a set of constraints universally accepted to characterize the notion of interpretability. On the contrary, every author adopts a customized set of interpretability constraints. Such variability is due to the subjective judgement motivating the inclusion of each single constraint in the attempt to formalize interpretability. Indeed, while some constraints are psychologically motivated (also with experimental support), others are justified only by common sense or are application-specific. As a consequence, the choice of including a formal constraint in defining interpretability could depend on subjective decision. To achieve a better agreement on which constraints are strictly necessary to define interpretability and which are more superfluous, an analytical systematization of all formal constraints could be a valid help.

In this Chapter, interpretability constraints are analytically described by providing, for each one, a formal definition – where possible – and a justification of its use in interpretable fuzzy modelling. As the number of interpretability constraints is quite high, they have been classified according to several categories. The chosen taxonomy is only one of several possible categorizations. As an example, in (Peña-Reyes and Sipper, 2003) a divi-

sion between "syntactic" and "semantic" constraints can be found. Since this work is aimed in analyzing interpretability within the context of Granular Computing, a different taxonomy has been preferred. Following the representation of fuzzy information granules as linguistically labelled multidimensional fuzzy sets, the following taxonomy of interpretability constraints has been used:

- Constraints for (one-dimensional) fuzzy sets;
- Constraints for frames of cognition, i.e. families of one-dimensional fuzzy sets defined on the same Universe of Discourse;
- Constraints for fuzzy information granules;
- Constraints for fuzzy rules;
- Constraints for fuzzy models;
- Constraints for learning algorithms;

The last three categories of constraints have been included because the most frequent use of fuzzy information granules is within rule-based fuzzy models, which are often equipped with learning capabilities to refine their predictive accuracy.

In the following, interpretability constraints are described according to the aforementioned taxonomy. It should be noted that some constraints have a rigid mathematical definition, while others have a more fuzzy description. Such variability of formulations should not be considered negatively but as a direct consequence of the inherent blurriness of the "interpretability" concept.

2.2 Constraints on Fuzzy sets

The first level of interpretability analysis concerns each fuzzy set involved in the representation of an information granule. In order to be labelled by a semantically sound linguistic term, each fuzzy set must be shaped so as to satisfy some requirements justified by common sense.

Hereafter, a fuzzy set is A is fully defined by its membership function[1] μ_A defined as:

$$\mu_A : U \rightarrow [0, 1] \tag{2.1}$$

[1]Note that on the semantic level, the symbols A and μ_A denote the same object. However the use of two symbols helps to distinguish the variable A from its value μ_A

where the domain U of the membership function is called *Universe of Discourse* and includes all admissible values that may belong to a fuzzy set. The set of all possible fuzzy sets definable on U is denoted with $\mathcal{F}(U)$.

Before introducing interpretability constraints, some considerations about the Universe of Discourse U are noteworthy. In this work, it is assumed that U is a closed interval of the real line. Such definition has some important consequences:

1. The Universe of Discourse is numerical and real[2]. The choice of numerical universe is necessary when physical systems have to be represented by fuzzy models, though categorical universes may be chosen when it is necessary to represent more abstract data types (e.g. conceptual categories). For fuzzy sets defined on categorical universes, however, many interpretability constraints are not applicable while other constraints – though formally applicable – loose significance. Real numbers are adopted because they generalize natural and rational numbers. The adjoint value of including all real elements instead of only rational ones stands in facilitating model analysis, which is often accomplished with mathematical Calculus. More structured domains, such as complex numbers or mathematical/symbolical structures are more application specific and hence are not included in interpretability analysis.

2. The Universe of the Discourse is one-dimensional, hence fuzzy sets model single attributes of a complex system and composite attributes are represented by more complex information granules. Unidimensionality is often required in interpretable fuzzy modelling and is implicitly assumed by most authors in the field (Babuška, 1999; Ishibuchi and Yamamoto, 2002; Setnes et al., 1998b). Unidimensional fuzzy sets help to better understand complex systems in terms of composite relations (e.g. "SEPAL IS WIDE, PETAL IS LONG AND COLOR IS PURPLE"), while they could be unable to model complex relations that can hardly be represented in terms of composition of single attributes (e.g. how to represent the knowledge that "JOHN IS NICE" in terms of eyes' color, nose dimension, etc.?). This latter form of knowledge is "non-representable"[3] and hence it is out of scope for interpretable fuzzy in-

[2] i.e. subset of the real line

[3] The representability of knowledge is a long disputed theme in Cognitive Science. From the strong assumptions of the Representational Theory of Mind (Fodor, 1975), several arguments have been arisen so much to deny the representational nature of human knowledge in favour of emergent phenomena (Norman, 1981). More recent developments try to relate and fruitfully exploit both points of view and hybrid systems (such as neuro-fuzzy networks) are concrete realizations of this new school of thougth.

formation granulation. In such case, black box models, such as neural networks, are more suitable.

3. As a closed interval, the Universe of Discourse has an infimum $m_U = \inf U$ and a supremum $M_U = \sup U$ and all numbers between the two extreme values are legitimate elements of the Universe. This definition of Universe of Discourse simplifies modelling and is not particularly limiting when representing physical attributes, which are almost always expressible as values in a range. In quantitative modelling, either m_U or M_U may be infinite. In such cases, interpretability analysis must be adapted to enable infinite support domains[4]. In other contexts, many authors assume the Universe of Discourse to be the unitary interval $[0, 1]$ or $[-1, 1]$. This restriction is useful when elements of different Universes of Discourse appear in the same computation (e.g. in clustering algorithms) to avoid problems due to different scales. Such rescaling is obviously not limiting since it is always possible to find bijective mappings that transform these specific intervals in any required closed interval, being it limited or not. However, this mapping may hamper the physical meaning of the attribute, so the rescaling should be limited only within the processing steps that require it, and the original universe of discourse must be resorted when knowledge has to be represented in an interpretable form.

In the following, a number of interpretability constraints are defined and described. Such constraints apply to each fuzzy set singularly, which is assumed to be characterized by a membership function defined over a Universe of Discourse U defined as a limited closed interval $[m_U, M_U] \subset \mathbb{R}$.

Normality

CONSTRAINT 1 *A fuzzy set A must be **normal**, i.e. there exists at least one element (called **prototype**) with full membership:*

$$\exists x \in U : \mu_A(x) = 1 \tag{2.2}$$

A fuzzy set is sub-normal if it is not normal. In fig. 2.1, normality and sub-normality are graphically depicted for two fuzzy sets. The normality

[4]The adaptation of interpretability analysis to infinite support domains is straightforward in most cases, but may require adjoint mathematical burden that does not convey useful information in analyzing the key interpretability issues of fuzzy information granulation. For such reason, in this work the Universe of discourse is assumed to be limited, i.e. $M_U < +\infty$ and $m_U > -\infty$.

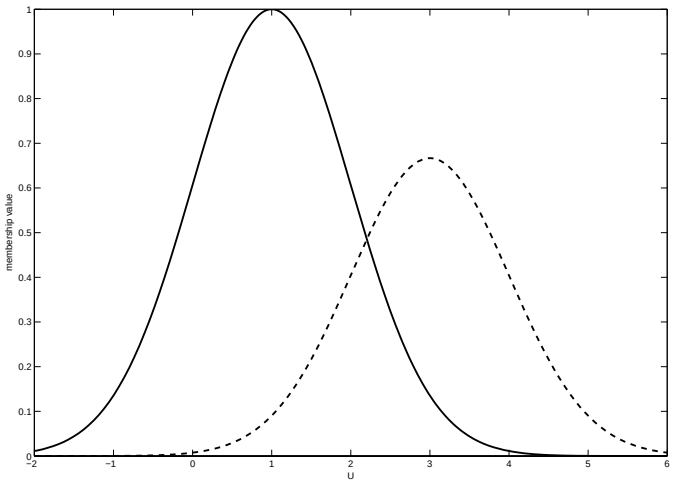

Figure 2.1: Example of a nornal fuzzy set (solid) and a sub-normal fuzzy set (dashed)

requirement is very frequent. Indeed, while few authors require it explicitly, see (Chow et al., 1999; Roubos and Setnes, 2001; Jin et al., 1998a; Setnes and Roubos, 1999), it is implicitly assumed in almost all literature concerning interpretable fuzzy modelling[5].

In an epistemic perspective, the normality constraint implies that at least one element of the Universe of Discourse should exhibit full matching with the concept semantically represented by the fuzzy set (Espinosa and Vandewalle, 2000; Peña-Reyes and Sipper, 2003; Valente de Oliveira, 1999a). Moreover, in fuzzy sets expressing vague quantities, the normality requirement is necessary for modeling existing but not precisely measurable values (Hermann, 1997).

On the logical level, subnormality is an indication of inconsistency of a fuzzy set (Pedrycz and Gomide, 1998). Indeed, if a fuzzy set is sub-normal, then the degree of inclusion of A within the empty set $\varnothing$ is

$$\nu(\varnothing, A) = 1 - \pi(A, U) = 1 - \sup \mu_A > 0 \tag{2.3}$$

where ν is the necessity measure and π the corresponding possibility measure, defined as:

$$\pi(A, B) = \sup_{x \in U} \min \{\mu_A(x), \mu_B(x)\} \tag{2.4}$$

[5] An interesting study concerning linguistic approximation of subnormal fuzzy sets can be found in (Kowalczyk, 1998)

Inconsistent fuzzy sets are problematic in inference systems because they may produce unexpected and unsatisfactory results[6]. As a consequence, the normality constraint should be verified for all fuzzy sets involved in a knowledge-based model.

Convexity

CONSTRAINT 2 *A fuzzy set A must be* ***convex***[7] *i.e. the membership values of elements belonging to any interval are not lower than the membership values at the interval's extremes:*

$$\forall a, b, x \in U : a \leq x \leq b \rightarrow \mu_A(x) \geq \min\{\mu_A(a), \mu_A(b)\} \tag{2.5}$$

Alternatively, α-cuts *may be used to define convexity:*

$$\forall \alpha \in [0,1] \, \exists m_\alpha, M_\alpha : [A]_\alpha = [m_\alpha, M_\alpha] \tag{2.6}$$

being each α*-cut defined as:*

$$[A]_\alpha = \{x \in U : \mu_A(x) \geq \alpha\} \tag{2.7}$$

The fuzzy set A is ***strictly convex*** *if inequalities in (2.5) are strict, i.e. without equality signs.*

Convexity is an important requirement that is often implicitly assumed in interpretable fuzzy modelling (convexity is explicitly required in (Chow et al., 1999)). It can be considered as a completion of the normality requirement. Indeed, while normality constraints imposes that at least one element (a prototype) of the Universe of Discourse is fully represented by the fuzzy set, convexity assures that the concept represented by the fuzzy set gradually loses its evidence as elements of fuzzy sets become far from the prototypes. In this way, the concept represented by the fuzzy set can be conceived as elementary, i.e. it is related to a single specific property of a perceived object (see fig. 2.2).

Convexity is a defining property of fuzzy numbers and fuzzy vectors[8]. A fuzzy number is a semantic representation of vague quantities labelled, e.g., with "ABOUT x" where x is a rational number. Fuzzy numbers can

[6] This could be viewed as a generalization of the pseudo-Scoto meta-theorem, which states that from a contradiction every formula (including its negation) can be derived.

[7] It should be remarked that the notion of convexity of fuzzy sets is different from the notion of convexity of functions. Actually, convex fuzzy sets are often characterized by membership functions that are not convex

[8] A fuzzy vector is a vector of fuzzy numbers

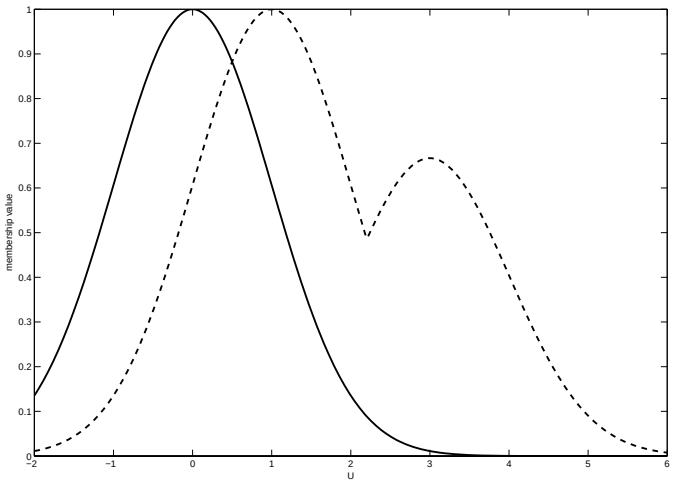

Figure 2.2: A convex fuzzy set (solid) and a non-convex fuzzy set (dashed). While the convex fuzzy set may be associated to a semantically sound lingusitic label, the non-convex fuzzy set conveys represents the semantic of a complex concept that is hard to represent by a lingustic label.

be processed within the Fuzzy Arithmetic, which is an extension of Interval Arithmetic that manipulates with vaguely defined quantities. Convexity is necessary for fuzzy numbers since the perceived evidence of a quantity of being a specific number (or belonging to a specific interval) decreases monotonically as the distance of the quantity from the number (respectively, interval) increases. In addition, convexity is a pivotal property that enables to easily extend arithmetic operations on fuzzy number without excessive computational burden (Kaufmann and Gupta, 1985).

Convexity plays an important role in other interpretability requirements, such as in the Error Free Reconstruction and in Distinguishability analysis of fuzzy sets, as will be clear forth. Nevertheless, many well-known fuzzy clustering procedures underlying information granulation processes do not yield convex fuzzy sets. As an example, in fig. 2.3 it is shown a typical membership function derived from application of the notorious Fuzzy C-Means algorithm (Bezdek, 1981). To avoid interpretability loss due to non-convex fuzzy sets derived from clustering algorithm, a successive Chapter is devoted at describing a proposal of deriving convex fuzzy information granules from clustering results. The proposed strategy augments the clustering process with a highly efficient procedure, which finds fuzzy information granules represented by Gaussian fuzzy sets that optimal in a sense. In this way, the

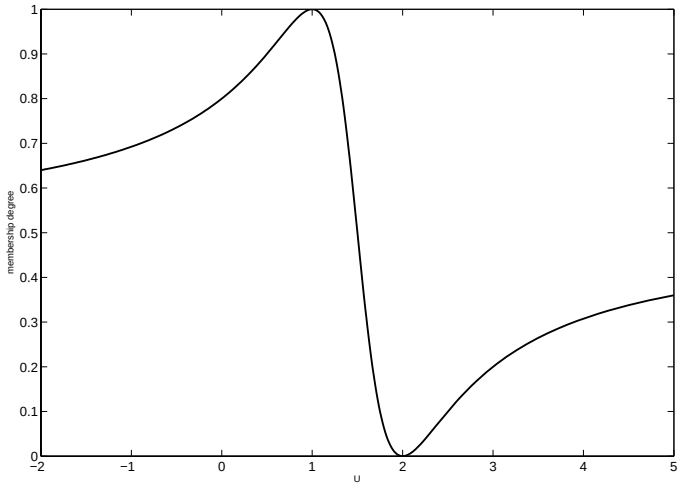

Figure 2.3: An example of one-dimensional membership function that can be derived by application of Fuzzy C-Means algorithm.

explorative features of clustering algorithms is exploited, and interpretability of the resulting information granules is protected.

Unimodality

CONSTRAINT 3 *A fuzzy set A should be* ***unimodal*** *i.e. there exists only one element with maximum membership function, which is called* prototype*:*

$$\exists p \in U : \left(\mu_A(p) = \max_{x \in U} \mu_A(x) \wedge \forall q \neq p : \mu_A(q) < \mu_A(p) \right) \tag{2.8}$$

Unimodality is a direct consequence of strict convexity, as stated in the following proposition:

PROPOSITION 2.1 *A strictly convex fuzzy set A is also unimodal.*
Proof.*Let M be the maximum membership value of the fuzzy set A:*

$$M = \max_{xeU} \mu_A(x) \tag{2.9}$$

Suppose that there exist two distinct prototypes $p \neq q$ with maximum membership:

$$\mu_A(p) = \mu_A(q) = M \tag{2.10}$$

Because of strict convexity, any element between p and q could have a membership value greater than M, which is absurd. Thus two distinct prototypes cannot exist, hence $p = q$. ■

In quantitative modelling, unimodality is a needed requirement, which must be satisfied by fuzzy numbers and can be relaxed only when representing fuzzy intervals. In qualitative modelling, unimodality is not a mandatory requirement, as long as it is semantically significant to have a *set* of prototypes (which is always an interval in convex fuzzy sets) instead of just one. However, attention must be paid when unimodality is not considered in fuzzy models, because prototypes become indistinguishable and may lead to oversimplified models, as illustrated in the following example.

EXAMPLE 2.1 (OVERSIMPLIFIED MODEL) *Consider the fuzzy set labeled* TALL*, which is characterized by the following membership function:*

$$\mu_{\text{TALL}}(x) = \begin{cases} 0 & \text{if } x \leq 150 \\ \frac{x-150}{30} & \text{if } 150 < x < 180 \\ 1 & \text{if } x \geq 180 \end{cases} \tag{2.11}$$

A subject of height 180 cm *and a subject of height 200* cm *are considered equally tall by the system, without any possibility to comparing them in terms of height. This is an oversimplified model that is very far from common sense perception of height.*

Continuity

CONSTRAINT 4 *A fuzzy set A should be* ***continuous*** *i.e. its membership function μ_A is continuous in the universe of discourse.*

Continuity is motivated by the common assumption that is well caught by the words of Leibniz, "*Natura non facit saltus*„[9] (Gottfried Wilhelm Leibniz, *Nouveaux Essais*, IV, 16.). Since fuzzy sets are mathematical characterization of perceptions, it is expected that their membership functions are continuous.

Actually, this constraint is always met in interpretable fuzzy information granulation, but it is rarely explicated in literature. However, neither the Fuzzy Set Theory nor the Theory of Fuzzy Information Granulation guarantee continuity of fuzzy sets *ipso facto*, hence this constraint should be included in interpretability analysis.

[9] "Nature does not make anything with leaps„. This postulate is evidently true in macro-physics, while it loses its validity in micro-physics, where quantum physics seems to better justify behaviour of nature by discrete states. Quantum physics is the basis of a theoretical approach to study intelligence (Penrose, 1989).

2.3 Constraints on Frames of Cognition

A Frame of Cognition (briefly, frame) is defined as a finite collection of fuzzy sets defined over the same Universe of Discourse (Cios et al., 1998). It is convenient to provide a total ordering $\preceq$ to the collection, hence a frame can be provisionally formalized as:

$$\mathbf{F} = \langle U, \mathbf{F}, \preceq \rangle \tag{2.12}$$

where:

- $\mathbf{F} = \{A_1, A_2, ..., A_n : A_i \in \mathcal{F}(U), n > 1\}$ is a collection of fuzzy sets[10];
- $\preceq$ is an ordering on $\mathbf{F}$. For sake of simplicity, it is assumed that $i \leq j \longleftrightarrow A_i \preceq A_j$

In fuzzy modelling, frames define all linguistic terms that can be used to describe a physical attribute of the modelled system. For interpretable modelling, the frame must be semantically sound, so that the fuzzy sets belonging the frame are labelled by an assignment:

$$A \in \mathbf{F} \mapsto \mathcal{L}A = string \tag{2.13}$$

The assignment of a label of each fuzzy set must reflect the metaphor of the assigned label. This is a delicate, yet not automatically controllable point that influences the overall interpretability of the fuzzy model. If labels are meaningless (e.g. $\mathcal{L}A$ =A12), the frame of cognition is hard to understand by human users. Similarly, if labels are meaningful but they are not properly assigned (e.g. $\mathcal{L}A$ =Low and $\mathcal{L}B$ =High but $B \preceq A$), the frame of cognition generates confusion in any attempt to understand the semantics it conveys (see fig. 2.4 for an example).The mapping $\mathcal{L}$ must be invertible, so that it is possible to retrieve a fuzzy set if only the label is given. Such possibility is fundamental in in inference process.

For a complete definition of a Frame of Cognition, a symbolic variable v is attached so as to denotate the physical attribute modelled by the frame. The resulting augmented structure of the frame of cognition is the following:

$$\mathbf{F} = \langle U, \mathbf{F}, \preceq, \mathcal{L}, v \rangle \tag{2.14}$$

Frames of Cognition are similar – yet simpler – to linguistic variables. Linguistic variables are formal structures that provide for the syntax and the

[10]With a slight abuse of notation, the set of fuzzy sets and the frame of cognition are denoted with the same symbol. This avoids the introduction of too many symbols. Amiguity is solved by the context of the discussion.

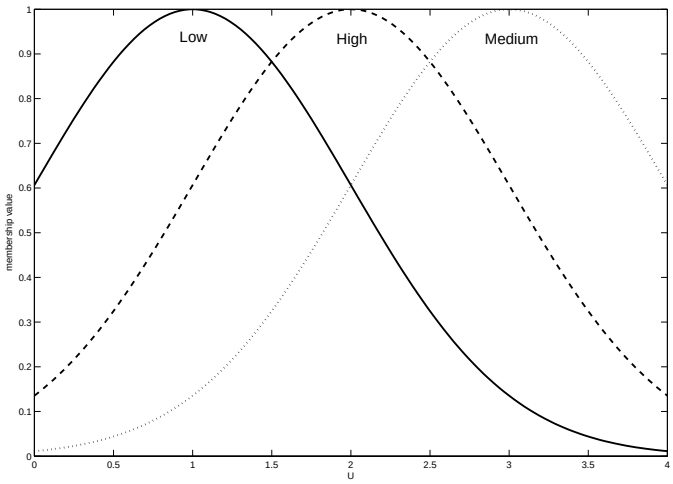

Figure 2.4: A Frame of Cognition with three fuzzy sets. While each individual fuzzy sets has a well defined semantics, they are not properly labelled thus hampering their interpretability

semantics of a set of linguistic edges that can be associated to a variable representing a physical attributes (Zadeh, 1975a). The set of all admissible linguistic edges are derived by a generative grammar that provides for a rich language. However, when Linguistic Variables are used to represent information granules automatically acquired from data, the language richness may be disadvantageous[11]. For such reason, a simpler structure is adopted in this work, though the adoption of linguistic variables for interpretable fuzzy information granulation belongs to the research agenda.

Semantic soundness of Frames of Cognition translates into a number of properties, which are described in the following subsections.

[11]In (Marìn-Blàzquez and Shen, 2002) an approach is proposed that generate fuzzy information granules represented by linguistic variables with a rich language. In their experimental results, the authors report induced rules such as "Sepal Width is Greatly Medium And Petal Length is Very Medium [...]" or "Petal Width is Lower Greatly Wide". These examples clearly show that the use of tortuous linguistic edges does not enhance interpretability of the acquired information granules

Proper ordering

CONSTRAINT 5 *A frame of cognition* $\mathbf{F} = \langle U, \mathbf{F}, \preceq, \mathcal{L}, v \rangle$ *should be* ***properly ordered*** *i.e. the order of fuzzy sets* $\preceq$ *reflects the order of membership values:*

$$\forall A, B \in \mathbf{F} : A \preceq B \rightarrow \exists t \in U \forall x \in U : \\ (x \leq t \rightarrow \mu_A(x) \geq \mu_B(x)) \wedge (x \geq t \rightarrow \mu_A(x) \leq \mu_B(x)) \quad (2.15)$$

According to the previous definition, if two fuzzy sets A and B are related as $A \preceq B$, then there exists a midpoint t such that each point less than t has membership to A greater than to B, and each point greater than t has membership to B greater than to A. Thus A better represents elements of the Universe of Discourse that are less than the points represented by B. In this sense, the ordering of fuzzy sets reflects the semantics formalized by their membership functions.

It should be noted that in general relation (2.15) induces a partial ordering of fuzzy sets. Indeed, if a fuzzy set A in the frame is a subset of B (meaning that $\forall x : \mu_A(x) \leq \mu_B(x)$), then $A \npreceq B$ and $B \npreceq A$. Also, if A is not subset of B, ordering may not be established by (2.15) as can be seen by the following example:

EXAMPLE 2.2 *Let* A *and* B *be two fuzzy sets with the following membership functions:*

$$\mu_A(x) = \exp\left(-\frac{(x-\omega_A)^2}{2\sigma_A^2}\right) \quad (2.16)$$

and:

$$\mu_B(x) = \exp\left(-\frac{(x-\omega_B)^2}{2\sigma_B^2}\right) \quad (2.17)$$

The points where $\mu_A(x) = \mu_B(x)$ *are:*

$$\mu_A(x) = \mu_B(x) \iff x = \begin{cases} \frac{\sigma_B\omega_A \mp \sigma_A\omega_B}{\sigma_B \mp \sigma_A} & \text{if } \sigma_A \neq \sigma_B \\ \frac{\omega_A + \omega_B}{2} & \text{if } \sigma_A = \sigma_B \end{cases} \quad (2.18)$$

Hence there are at most two intersection points, which can be denoted by x_1 *and* x_2 *assuming* $x_1 \leq x_2$.
If $\sigma_A \neq \sigma_B$ *then there are two distinct intersection points. As can be seen from fig. 2.5, the existence of two distinct intersection points violates the proper ordering constraint.*

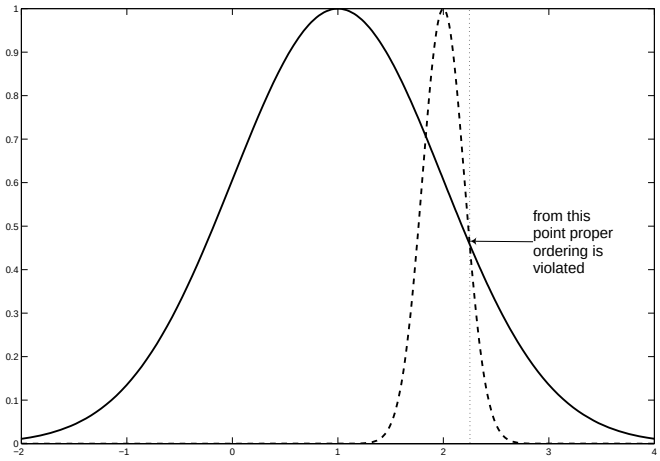

Figure 2.5: An example of violation of proper ordering

From the previous example, it could be observed that if $\omega_A \leq \omega_B$, then A and B could be considered ordered in some sense (e.g. if A and B represent the fuzzy numbers 'ABOUT ω_A' and 'ABOUT ω_B', then A and B are intuitively ordered as $A \preceq B$). This weaker kind of order can be formalized with the following weaker condition:

CONSTRAINT 6 *A frame of cognition* $\mathbf{F} = \langle U, \mathbf{F}, \preceq, \mathcal{L}, v \rangle$ *should be* ***weakly properly ordered*** *i.e. the order of fuzzy sets* $\preceq$ *reflects the order of prototypes:*

$$\forall A, B \in \mathbf{F} : A \preceq B \Leftrightarrow \min \operatorname{core} A \leq \min \operatorname{core} B \tag{2.19}$$

where $\operatorname{core} A = \arg\max \mu_A$.

This new definition is weaker than that in (2.15) , in the sense that it may not reflect the intuitive idea of ordering in some special cases (e.g. when one fuzzy set is subset of another fuzzy set). However, for normal, unimodal and convex fuzzy sets, weakly proper ordering can be considered as an adequate constraint for interpretability analysis. In summary, weakly proper ordering can be considered as a valid constraint for frames consisting of normal, unimodal and convex fuzzy sets, while (strict) proper ordering can be assured when using specific shapes for fuzzy sets (e.g. isotope[12] Gaussian fuzzy sets).

[12]Two Gaussian fuzzy sets are isotope if they have the same width. In such case, the interection point between the two fuzzy sets is unique

Justifiable number of elements

CONSTRAINT 7 *The number of fuzzy sets in the frame should not be too high, preferably less than* 7 ± 2

This criterion is motivated by a psychological experiment reported in (Miller, 1956) and considered in a number of papers concerning interpretable fuzzy modelling, e.g. in (Espinosa and Vandewalle, 2000; Jamei et al., 2001; Jiménez et al., 2001; Jin, 2000; Meesad and Yen, 2002; Valente de Oliveira, 1999a; Paiva and Dourado, 2001; Peña-Reyes and Sipper, 2003; Setnes et al., 1998a). In his test battery, Miller proved that the span of absolute judgment[13] and the span of immediate memory[14] impose a severe limitation on the amount of information that a human being is able to perceive, process and remember. Experimentally, the author found that the number of entities that can be clearly remembered for a short time is around 7, plus or minus 2, depending on the subject under examination. Interestingly, and quite surprisingly from a cognitive standpoint, complex information can be mentally organized into unitary chunks of information that are treated as single entities. In this way, the human brain is capable of perceive, process or remember in its short term memory both simple perceptions (e.g. tones) or more complex entities (e.g. faces), provided that the number of entities is less than "*the magical number seven, plus or minus two*„.

The transposition of such important psychological result into interpretability analysis leads to the criterion of reducing the number of fuzzy sets in a Frame of Cognition to a justifiable number of elements (see fig. 2.6). It also gives a scientific explanation of the commonly adopted criterion of reducing the complexity of a fuzzy model to make it more interpretable. Moreover, the criterion can be applied also to other element constituting a fuzzy model, e.g. the number of attributes (for the entire model or for each information granule) and the number of information granules. However, the psychologically imposed limit of elements can be surpassed when it is necessary to define more accurate models, which may require a greater number of elements (fuzzy sets, attributes or rules) to describe the behavior of a system.

Distinguishability

CONSTRAINT 8 *Any fuzzy set in the Frame of Cognition must be well distinguishable from the remaining fuzzy sets of the frame*

[13] Absolute judgment is the task of assigning numbers to stimuli of different perceived magnitudes. An example of absolute judgment is the assignment to different numbers to acustic impulses based on perceived tones.

[14] The number of events that are reminded in the short term.

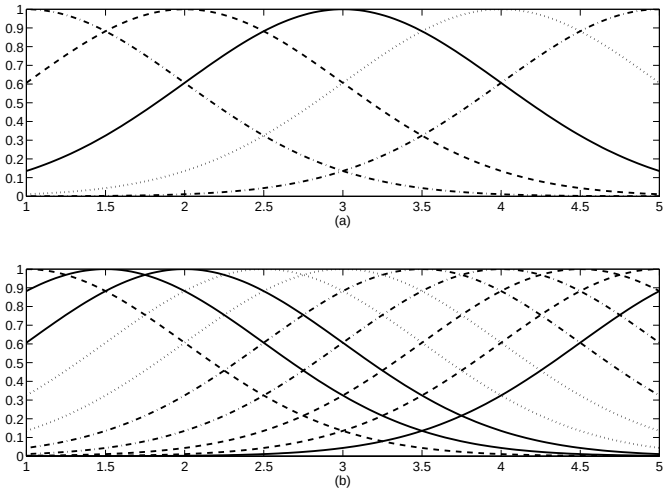

Figure 2.6: Example of Frame of Cognition with a five fuzzy sets (a) and another with ten fuzzy sets (b). The association of liguistic labels to fuzzy sets of the first frame is clearly easier than ti fuzzy sets of the second frame

Distinguishability is one of the most common interpretability constraints adopted in fuzzy modelling literature (Babuška, 1999; Chow et al., 1999; Espinosa and Vandewalle, 2000; Guillaume and Charnomordic, 2004; Jamei et al., 2001; Jiménez et al., 2001; Jin and Sendhoff, 2003; Meesad and Yen, 2002; Nauck and Kruse, 1998; Valente de Oliveira, 1999a; Paiva and Dourado, 2001; Peña-Reyes and Sipper, 2003; Roubos and Setnes, 2001; Setnes et al., 1998a).

Formally, Distinguishability is a relation between fuzzy sets defined on the same Universe of Discourse. Roughly speaking, distinguishable fuzzy sets are well disjunct so they represent distinct concepts and can be assigned to semantically different linguistic labels. Well distinguishable fuzzy sets provide the following advantages in fuzzy modelling:

- Obviate the subjective establishment of membership-function/linguistic term association (Valente de Oliveira, 1998);
- Avoids potential inconsistencies in fuzzy models (Valente de Oliveira, 1999b);
- Reduce model's redundancy and consequently computational complexity (Setnes et al., 1998a);

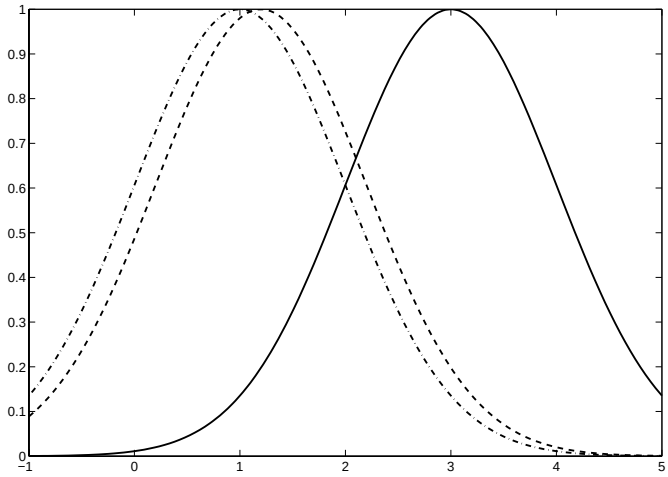

Figure 2.7: Example of non-distinguishable fuzzy sets (dash vs. dash-dots) and distinguishable fuzzy sets (solid vs. dash or solid vs. dash-dots)

- Linguistic interpretation of the fuzzy model is easier because fuzzy sets represent well separated concepts (Setnes et al., 1998a);

Completely disjunct fuzzy sets[15] are maximally distinguishable. However, in interpretable fuzzy modelling, different constraints may require overlapping fuzzy sets. As a consequence, distinguishability must be balanced with other constraints in designing a fuzzy model.

Distinguishability can be formalized in different ways. The most adopted formalization is by means of *similarity* measures. In (Setnes et al., 1998a) similarity measures are deeply discussed in the context of fuzzy modelling. There, similarity is interpreted as a fuzzy relation defined over fuzzy sets and corresponds to the "degree to which two fuzzy sets are equal". Such interpretation is then formally characterized by a set of axioms. Similarity measures well characterize the notion of distinguishability, but their calculation is usually computationally intensive. As a consequence, most strategies that adopt similarity for interpretability are based on massive search algorithms such as Genetic Algorithms (Roubos and Setnes, 2001; Meesad and Yen, 2002; Setnes et al., 1998b), Evolution Strategies (Jin, 2000), Symbiotic Evolution (Jamei et al., 2001), Coevolution (Peña-Reyes and Sipper, 2003), Multi-Objective Genetic Optimization (Jiménez et al., 2001). Alternatively, distinguishability improvement is realized in separate learning stages, often

[15]Two fuzzy sets A and B are completely disjunct if $\forall x \in U : \mu_A(x) \cdot \mu_B(x) = 0$

after some data driven procedure like clustering, in which similar fuzzy sets are usually merged together (Paiva and Dourado, 2001; Setnes et al., 1998a).

When distinguishability constraint must be included in less time consuming learning procedures, like in neural learning, similarity measures are no longer used. In (Paiva and Dourado, 2001), after a merging stage that uses similarity, fine tuning is achieved by simply imposing some heuristic constraints on centers and width of membership functions. In (Jin and Sendhoff, 2003), a distance function between fuzzy sets (restricted to the Gaussian shape) is used in the regularization part of the RBF cost function. In (Guillaume and Charnomordic, 2004) a sophisticated distance function is used to merge fuzzy sets. In (Nauck and Kruse, 1998; Espinosa and Vandewalle, 2000; Castellano et al., 2002) the possibility measure is adopted to evaluate distinguishability.

The possibility measure has some attracting features that promote a deeper investigation in the context of distinguishability assessment. Although it is not a similarity measure, it has a clear and well-established semantics since it can be interpreted as the degree to which the flexible constraint "X is A" is satisfied (Zadeh, 1978). Moreover, the possibility measure between fuzzy sets can be often analytically defined in terms of fuzzy sets' parameters. This makes possibility evaluation very efficient, thus allowing the employment of this measure in computationally inexpensive learning schemes. For this reason, a successive Chapter is devoted to a discussion concerning the theoretical justification of the adoption of the possibility measure in the context of distinguishability analysis.

Completeness (coverage)

CONSTRAINT 9 *A Frame of Cognition* $\mathbf{F}$ *must be* ***complete*** *i.e. each element of the Universe of Discourse* U *belongs at least to one fuzzy set of the Frame of Cognition:*

$$\forall x \in U \exists A \in \mathbf{F} : \mu_A(x) > 0 \tag{2.20}$$

A more general definition of the completeness constraint is given in (Jang, 1992):

CONSTRAINT 10 *A Frame of Cognition* $\mathbf{F}$ *is* $\boldsymbol{\alpha}-$***complete*** *if each element of the Universe of Discourse* U *belongs at least to one fuzzy set of the frame with membership not lower than* α*:*

$$\forall x \in U \exists A \in \mathbf{F} : \mu_A(x) \geq \alpha \tag{2.21}$$

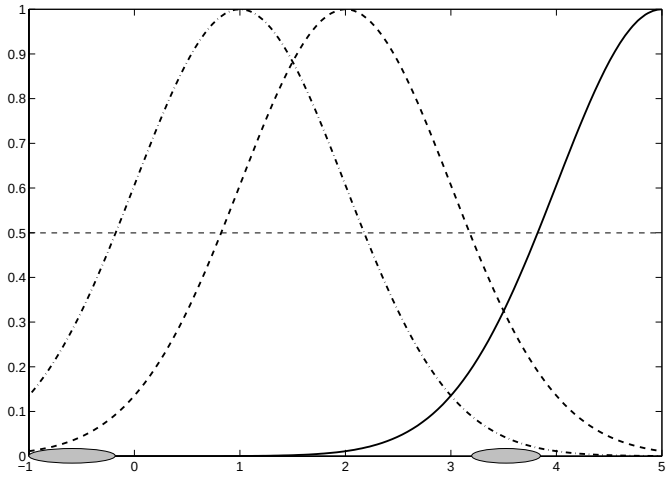

Figure 2.8: Example of completeness violation. In the highlighted regions of the Universe of Discourse (ellipses) 0.5-completeness is not verified.

Completeness is a property of deductive systems that has been used in the context of Artificial Intelligence to indicate that the knowledge representation scheme can represent every entity within the intended domain (Valente de Oliveira, 1999b). When applied to fuzzy models, completeness states that the fuzzy model should be able to infer a proper conclusion for every input (Lee and Lee, 1975). Hermann (Hermann, 1997) justifies completeness (there called 'cover full range') by the fact that in human reasoning there will never be a gap of description within the range of the variable. On the contrary, incompleteness may be a consequence of model adaption from data and can be considered a symptom of overfitting (Jin et al., 1999). Generally, $\alpha-$completeness is preferable in interpretable fuzzy modelling, because it assures that each element of the Universe of Discourse is well represented by the frame of cognition, i.e. each element possess a property (represented by a fuzzy set) with a minimum degree α (see also fig. 2.8).

In an attempt to assess $\alpha-$completeness (and $\alpha-$incompleteness), in (Meesad and Yen, 2002) the following measure is proposed (here adapted to continuous Universe of Discourse):

$$CD_\alpha = \frac{\int_{\left\{x \in U : \max\limits_{X \in \mathbf{F}} \mu_X(x) > \alpha\right\}} dx}{\int_U dx} \tag{2.22}$$

$$ID_\alpha = 1 - CD_\alpha \tag{2.23}$$

Incompleteness Degree (ID_α) is hence the complementary of the Completeness Degree (CD_α). These to measures, which take values in the range $[0, 1]$, can be effectively used in genetic optimization process in order to maximize completeness while improving the accuracy of the model. When interpretability is of major concern, however, techniques must be adopted so that the completeness criterion is always fulfilled (i.e. $CD_\alpha = 1$ for a given α).

When using triangular fuzzy sets, 0.5-completeness is easy to guarantee by a proper placement of fuzzy sets parameters (Espinosa and Vandewalle, 2000; Guillaume and Charnomordic, 2004; Kruse et al., 1994). More specifically, given two triangular fuzzy sets of membership functions $T[a_1, b_1, c_1]$ and $T[a_2, b_2, c_2]$, where:

$$T[a, b, c](x) = \max\left\{0, \min\left\{\frac{x-a}{b-a}, \frac{x-c}{b-c}\right\}\right\} \tag{2.24}$$

0.5-completeness in the interval $[b_1, b_2]$ is guaranteed if $b_1 < b_2$ and:

$$a_2 = b_1 \wedge b_2 = c_1 \tag{2.25}$$

Indeed, with such constraints on the parameters the membership functions can be simplified as:

$$\forall x \in [b_1, b_2] : \begin{cases} T[a_1, b_1, c_1](x) = \frac{x-c_1}{b_1-c_1} \\ T[a_1, b_1, c_1](x) = T[b_1, c_1, c_2](x) = \frac{x-b_1}{c_1-b_1} \end{cases} \tag{2.26}$$

Hence, if $T[a_1, b_1, c_1](x) \leq 0.5$, then:

$$\frac{x-b_1}{c_1-b_1} \leq \frac{1}{2} \implies -x \geq -\frac{c_1+b_1}{2} \implies$$
$$T[a_1, b_1, c_1](x) = \frac{x-c_1}{b_1-c_1} \geq \frac{1}{2} \tag{2.27}$$

From this result, a useful theorem can be settled:

PROPOSITION 2.2 *Let a Frame of Cognition* $\mathbf{F} = \langle U, \mathbf{F}, \preceq, \mathcal{L}, v \rangle$ *with*

$$\mathbf{F} = \{A_1, A_2, ..., A_n : A_i \preceq A_{i+1}\} \tag{2.28}$$

be defined an the Universe of Discourse $U = [m_U, M_U]$ *and assume that each fuzzy set* A_i *is triangular with membership function* $T[a_i, b_i, c_i]$ *such that* $a_{i+1} = b_i$, $b_{i+1} = c_i$ *and* $b_1 = a_1 = m_U$ *and* $b_n = c_n = M_U$. *Then,* $0.5-$*completeness in the sense of 2.21 is verified.*

Proof. As previously observed, each couple of consecutive fuzzy set guarantees 0.5-coverage in the interval $[b_i, b_{i+1}]$. Since:

$$U = \bigcup_{i=1}^{n-1} [b_i, b_{i+1}] \tag{2.29}$$

completeness is verified in the entire Universe of Discourse. ■

The results on triangular fuzzy sets can be easily extended to trapezoidal fuzzy sets by taking care of the interval of prototypes that belong to their core. When using Gaussian fuzzy sets, coverage can be guaranteed by proper constraints on the parameters of the membership functions, as described in the following proposition.

PROPOSITION 2.3 *Let a Frame of Cognition* $\langle U, \mathbf{F}, \preceq, \mathcal{L}, v \rangle$ *with*

$$\mathbf{F} = \{A_1, A_2, ..., A_n : A_i \preceq A_{i+1}\} \tag{2.30}$$

be defined on the Universe of Discourse $U = [m_U, M_U]$ *and assume that each fuzzy set* A_i *is Gaussian with membership function* $G[\omega_i, \sigma_i]$ *defined as:*

$$\mu_{A_i}(x) = G[\omega_i, \sigma_i](x) = \exp\left(-\frac{(x-\omega_i)^2}{2\sigma_i^2}\right) \tag{2.31}$$

Let $t_1, t_2, ..., t_{n-1}$ *a sequence of points such that* $m_U < t_1 < t_2 < ... < t_{n-1} < M_U$ *and let* $t_0 = 2m_U - t_1$ *and* $t_n = 2M_U - t_{n-1}$. *If the centers and the widths of the membership functions in (2.31) are such that:*

$$\omega_i = \frac{t_{i-1} + t_i}{2} \tag{2.32}$$

and:

$$\sigma_i = \frac{t_i - t_{i-1}}{2\sqrt{-2\ln\alpha}} \tag{2.33}$$

then $\alpha-$*completeness in the sense of 2.21 is verified.*

Proof.*The membership values of* t_i *to the fuzzy sets* A_i *and* A_{i+1} *are the following:*

$$\mu_{A_i}(t_i) = \exp\left(-\frac{\left(t_i - \frac{t_{i-1}+t_i}{2}\right)^2}{2\left(\frac{t_i - t_{i-1}}{2\sqrt{-2\ln\alpha}}\right)^2}\right) = \exp\ln\alpha = \alpha \tag{2.34}$$

$$\mu_{A_{i+1}}(t_i) = \exp\left(-\frac{\left(t_i - \frac{t_i+t_{i+1}}{2}\right)^2}{2\left(\frac{t_{i+1} - t_i}{2\sqrt{-2\ln\alpha}}\right)^2}\right) = \exp\ln\alpha = \alpha \tag{2.35}$$

Hence, t_i is the intersection point of the two fuzzy sets A_i and A_{i+1}. Since Gaussian fuzzy sets are strictly convex, then:

$$\forall x \in [\omega_i, t_i] : \mu_{A_i}(x) \geq \alpha \tag{2.36}$$

$$\forall x \in [t_i, \omega_{i+1}] : \mu_{A_{i+1}}(x) \geq \alpha \tag{2.37}$$

As a consequence, in each interval $[\omega_i, \omega_{i+1}]$ $\alpha-$completeness is verified. Moreover, since $\omega_1 = (t_0 + t_1)/2 = m_U$ and $\omega_n = (t_{n-1} + t_n)/2 = M_U$, it follows that $U = [\omega_1, \omega_2] \cup [\omega_2, \omega_3] \cup \cdots [\omega_{n-1}, \omega_n]$ hence $\alpha-$completeness is verified in all the Universe of Discourse. ■

In (Valente de Oliveira, 1999a) a more general criterion is given to guarantee $\alpha-$completeness. The criterion states that for all elements of the Universe of Discourse, the sigma-count (Kosko, 1992) of all membership values must be greater than α:

$$\forall x \in U : \sqrt[p]{\sum_{X \in \mathbf{F}} \mu_X^p(x)} \geq \alpha \tag{2.38}$$

The authors claim that the proposed criterion can be applied to fuzzy sets of any shape. However, no proof is provided to guarantee that constraint (2.38) guarantees $\alpha-$completeness, without assuming specific shapes to fuzzy sets. Indeed, it is trivial to find simple examples of frames of cognition where (2.38) is true but $\alpha-$completeness is not assured:

EXAMPLE 2.3 *Consider a frame of two fuzzy sets A and B such that there exists an element of the universe of discourse where:*

$$\mu_A(x) = \mu_B(x) = \frac{\sqrt{2}}{4} \tag{2.39}$$

Then, for $p = 2$:

$$\sqrt{\mu_A^2(x) + \mu_B^2(x)} = 0.5 \tag{2.40}$$

However, $\max\{\mu_A(x), \mu_B(x)\} = \frac{\sqrt{2}}{4} < 0.5$. Moreover, it can be easily proved that for $p \leq 2$, inequality (2.38) is verified but $0.5-$completeness is not assured.

EXAMPLE 2.4 *In a more general settlement, consider a frame of two fuzzy sets A and B such that there exists an element of the universe of discourse where:*

$$\mu_A(x) = \mu_B(x) = \beta < \alpha \tag{2.41}$$

Then, for a generic $p \geq 1$:

$$\sqrt[p]{\mu_A^p(x) + \mu_B^p(x)} \geq \alpha \longleftrightarrow \beta \geq \frac{\alpha}{2^{\frac{1}{p}}} \tag{2.42}$$

Since, for each $p \geq 1$, $2^{\frac{1}{p}} > 1$, then it is always possible to find β such that $\beta < \alpha$ and inequality (2.38) is verified. Actually, it can be easily proved that equivalence between (2.38) and (2.21) is verified only for $p \to +\infty$, since $2^{\frac{1}{p}} \underset{p\to\infty}{\to} 1$.

The choice of the coverage level α is subjective and application specific. However, a value $\alpha \geq 0.5$ since preferable since in this way it is assured that each element of U is well represented by at least one concept semantically defined by one fuzzy set of the frame[16].

Complementarity ($\Sigma 1$-criterion)

CONSTRAINT 11 *For each element of the Universe of Discourse, all membership values of the frame must sum up to one:*

$$\forall x \in U : \sum_{X \in \mathbf{F}} \mu_X(x) = 1 \tag{2.43}$$

The Complementarity criterion assures that the Universe of Discourse is divided into a Ruspini partition (Ruspini, 1969), also called Bezdek partition (see fig. 2.9 for an example).

The adoption of Ruspini partitions is very common in fuzzy modelling (especially in fuzzy clustering (Bezdek, 1981)) but it is controversial in interpretability analysis. In (Peña-Reyes and Sipper, 2003), the authors state that complementarity "guarantees uniform distribution of meaning among element. However, such statement is unclear and does not shed light on any motivation to adopt Ruspini partitions in interpretable fuzzy information granulation.

Herrmann (Hermann, 1997) motivates complementarity with the following properties, which are direct consequences of the adoption of a Ruspini partition:

Full acceptance Whenever a membership value decreases from 1, another value increases from zero:

$$\forall x \in U \forall A \in \mathbf{F} \exists B \in \mathbf{F} \setminus \{A\} : \mu_A(x) < 1 \to \mu_B(x) > 0 \tag{2.44}$$

[16] In (Pedrycz and Gomide, 1998) it is also proved that the best crisp approximation of a fuzzy set is its 0.5-cut. Thus, membership values greater than this treshold denote a better evidence of the truth of the proposition semantically interpreted by a fuzzy set.

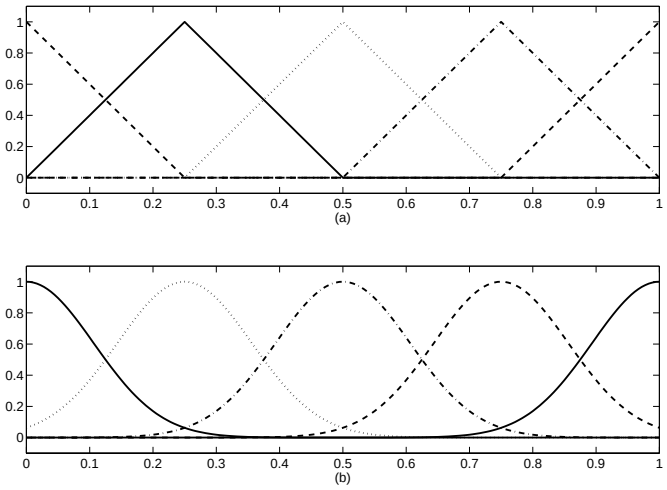

Figure 2.9: Example of two frames of cognition with complementarity constraint verified (a) and violated (b).

From a semantic perspective, full acceptance implies that when a concept does not fully represent an element, then there exists another concept that (at least partially) represents the element.

Negations The negation of a fuzzy set is expressed as the sum of the remaining membership values:

$$\forall x \in U \forall A \in \mathbf{F} : \mu_{\neg A}(x) = \sum_{X \in \mathbf{F} \setminus \{A\}} \mu_X(x) \tag{2.45}$$

However, this property has not any explanation in terms of interpretability analysis, because the summation operator is not a t-conorm[17]. Moreover, negation in fuzzy modelling has a far more complex semantics, as pointed out in (Ferri et al., 1999). As a consequence, this property can be considered as a only mere technical consequence of the complementarity criterion.

No three a time For any element of the Universe of Discourse, no more than two fuzzy sets have non-zero membership degree:

$$\forall x \in U : |\{X \in \mathbf{F} : \mu_X(x) > 0\}| \leq 2 \tag{2.46}$$

[17] A t-conorm is a diadic function that is used to define the semantics of the disjunction logical operator.

This property is very important in the inference context, because it assures that the number of elements involved in the inference process is limited, thus economizing the resources (time and space) involved in the inference. However, such property appears quite stringent in interpretability analysis, since it precludes that more than two concepts partially represent the same element (e.g. if x is LOW with high degree and MEDIUM with small degree, it could also be expected that x cannot be also HIGH, though with a small degree). Actually, limited support fuzzy sets (e.g. triangular or trapezoidal) allow economic inference without satisfying the property (2.46). Moreover, infinite support fuzzy sets (e.g. Gaussian) can be used in efficient inference by discarding fuzzy sets whose membership degrees are lower than a threshold. Hence the property (2.46) does not appear significant in interpretability analysis, nor in efficient inference processes.

The most significant contribution of complementarity to interpretability appears to be full acceptance, which can be also seen as a consequence of the completeness constraint. Furthermore, complementarity has been criticized by some authors because it is not recognized that degrees of truth (or degrees of membership) have an additive nature. In some cases, fuzzy clustering algorithms based on Ruspini partitions have been directly related to probabilistic clustering schemes (Wang et al., 2004). In (Krishnapuram and Keller, 1993), Ruspini partitions have been considered as expressing a "degree of sharing„ and possibilistic clustering has been proposed to better represent the intuitive idea of cluster membership. From the above consideration, with the additional difficulty of preserving complementarity in non-piecewise linear fuzzy sets, complementarity does not appear as a valid constraint for interpretable fuzzy modelling.

Uniform granulation

CONSTRAINT 12 *The cardinalities of all fuzzy sets*[18] *of the frame must be almost the same*

$$\forall A, B \in \mathbf{F} : |A| \approx |B| \tag{2.47}$$

Uniform granulation is a criterion required in (Lotfi et al., 1996) as a discriminating property for distinguishing understandable models – like those proposed in (Glorennec, 1993; Ishibuchi et al., 1994; Kong and Kosko, 1992) –

[18] The simplest definition of the cardinality of a fuzzy set is $|A| = \int_U \mu_A(x)\, dx$. A deeper discussion on fuzzy set cardinalities can be found in (Maciej, 2000).

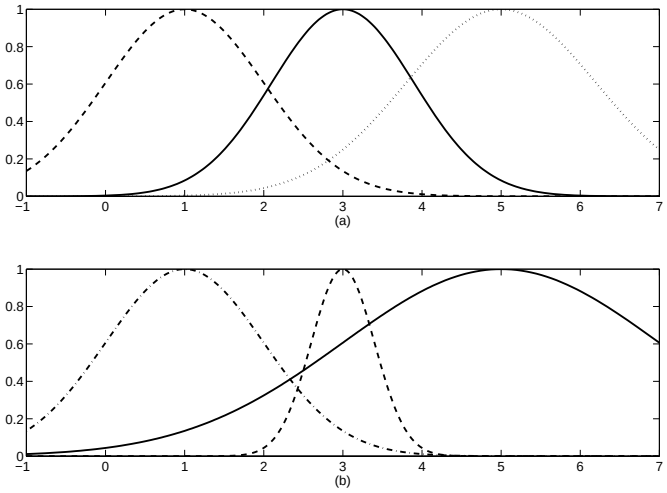

Figure 2.10: Comparative example of two Frames of Congition where Uniform Granulation is verified (a) and violated (b)

from those models where linguistic understandability is not of major concern, as in (Berenji and Khedkar, 1993; Sun, 1994). The criterion can be formalized by fixing a class of membership functions (e.g. Gaussian) characterized by a set of parameters (e.g. center and width). Uniform granulation is then formalized either by restricting the range of variability of such parameters (e.g. the width must vary only in a predefined range) or by assigning a penalty value depending on the actual values of each parameter, e.g.:

$$\Re(p) = \frac{1}{1 + \exp\left(-\left(p - p_{\min}\right)/\sigma_p\right)} - \frac{1}{1 + \exp\left(\left(p - p_{\max}\right)/\sigma_p\right)} \qquad (2.48)$$

where p is a parameter, $p_{\min}$ and $p_{\max}$ are the minimum/maximum values allowed for p and σ_p is the dispersion index. In fig. 2.10 a Frame of Cognition with Uniform Granulation is graphically compared with a frame that violates the constraint.

While it is recognizable that uniform granulation appears as a desirable property for interpretable fuzzy models, it should be noted that such constraint is not often used in specialized literature. This can be motivated by the necessity of specifying a high number of hyper-parameters (i.e. range of variability and the dispersion index), especially when fuzzy models must be acquired from data. An additional problem may arise when data to be mined for model induction have not a uniform distribution. In such a case, the uniform granulation may oppose to an accurate – yet understandable –

description of data. Finally, it should be observed that human beings do not perceive stimuli with a uniform scale (e.g. the intensity of perceived sound is logarithmically related to the actual intensity of sound). Such nonlinearities, which translate into non uniformity of scales, should be taken into account in interpretable fuzzy modelling.

Leftmost/Rightmost fuzzy sets

CONSTRAINT 13 *The minimum (resp. maximum) element of the universe of discourse must be prototypes for some fuzzy set of the Frame of Cognition:*

$$\exists A, B \in \mathbf{F} : \mu_A(\min U) = 1 \wedge \mu_B(\max U) = 1 \tag{2.49}$$

This constraint has been explicitly required in (Chow et al., 1999), but it is extensively used – without mention – in several works concerning interpretable fuzzy modelling. The constraints states that there exist two extreme fuzzy sets that fully represent the limit values of the Universe of Discourse. An obvious consequence of frames that have leftmost/rightmost fuzzy sets and satisfy the weak proper ordering criterion (2.19) is that the leftmost fuzzy set is the first fuzzy set and the rightmost fuzzy set is the last, according to the ordering relation $\preceq$ specified within the frame. In fig. 2.11 an example of two Frames of Cognition, one verifying leftmost/rightmost fuzzy sets and the other violating the constraint, is depicted.

In a semiotic perspective, the constraint of leftmost/rightmost fuzzy sets is extremely important because it could hamper the linguistic interpretability of the frame if violated. The following example illustrates the problem:

EXAMPLE 2.5 *Consider a frame of cognition of convex normal fuzzy sets, where the first fuzzy set (in the weakly properly order sense), labelled as* LOW*, has not its prototype in* $\min U$*, i.e.* $\mu_{\text{LOW}}(\min U) < 1$*. Since the fuzzy set is normal, there exists a prototype, which could be denoted as* $\min U + \varepsilon$*, being* $\varepsilon > 0$*. Each element of the Universe of Discourse belonging to* $[\min U, \min U + \varepsilon[$ *has a degree of membership to the fuzzy set* LOW *that is lower than membership degree of element* $\min U + \varepsilon$*. On a semiotic level, this means that elements in* $[\min U, \min U + \varepsilon[$ *have an evidence of "lowness" that is smaller than element* $\min U + \varepsilon$*, but this opposes to the metaphorical semantic of the concept "being low". This paradox can be solved by introducing another fuzzy set, which may be labelled as* VERY LOW*, which has its prototype in* $\min U$*. In this way, the leftmost fuzzy set constraint is verified.*

Leftmost/rightmost fuzzy sets are hence important in those frames that express qualities on the Universe of Discourse (e.g. LOW, MEDIUM, HIGH),

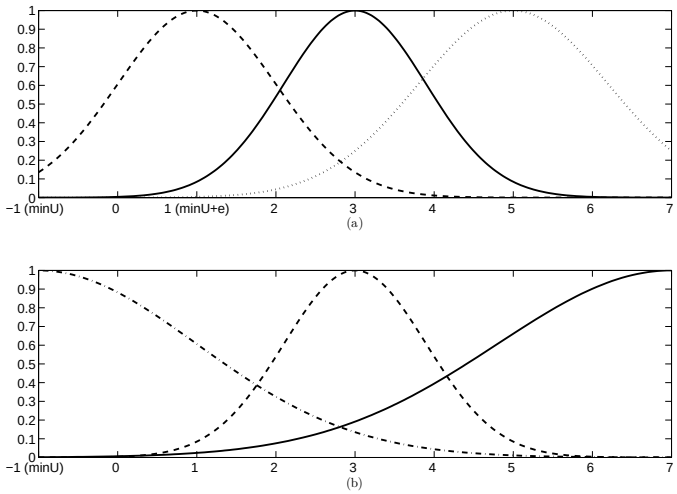

Figure 2.11: Example of a Frame of Cognition violating leftmost/rightmost fuzzy sets (a) and a frame veryfing such constraint (b).

so as to adhere to human intuition. However, such extreme fuzzy sets may not exist in frames that express fuzzy quantities (e.g. ABOUT X, PLUS OR MINUS Y, etc.), since these labels do not convey any meaning of extremity. It is perhaps for this reason that many interpretable fuzzy information granules, designed by fuzzy clustering techniques, do not have leftmost/rightmost membership values. However, when results of fuzzy clustering are used to define qualitatively labelled fuzzy sets, this constraint should be taken into account.

Natural zero positioning

CONSTRAINT 14 *If the number 0 belongs to the Universe of Discourse, it must be prototype of a fuzzy set in the Frame of Cognition:*

$$0 \in U \rightarrow \exists A \in \mathbf{F} : \mu_A(0) = 1 \tag{2.50}$$

This property states that the number 0, if included in the Universe of Discourse, should be fully represented by a fuzzy set of the Frame. Stated differently, this constraint specifies that the Frame must include a fuzzy set that represents the concept "ABOUT 0" or similar. However, as stated by the same authors that proposed this constraint, Natural Zero Positioning is a problem-specific constraint, especially useful in control problems (Valente de Oliveira, 1999a; Valente de Oliveira, 1999b). Generally speaking,

natural zero positioning makes sense – and its application should be evaluated – to Universes of Discourse of rational scale, i.e. sets where division between elements of the universe is a semantically correct operation (e.g. length/area/volume/weight measures lay on rational scales, but temperatures are not measured in rational scales).

More interestingly, Natural Zero Positioning calls for a more general criterion, which may be called "Natural S Positioning", where S is a set of elements in the Universe of Discourse which should be prototype of some fuzzy sets because they have privileged semantics. The constraint of Leftmost/rightmost fuzzy sets falls in this more general criterion, having $\{\min U, \max U\} \subseteq S$. Another example can be taken from the domain of accelerations, where the value $9.8\,\mathrm{m/s^2}$ (i.e. the gravitational acceleration) should be prototype of a fuzzy set. In Celsius temperatures, both $0\,^{\circ}\mathrm{C}$ and $100\,^{\circ}\mathrm{C}$ may be considered prototypes of two fuzzy sets, eventually labelled as IcingPoint and BoilingPoint respectively, if the frame is referred to water temperatures, while the value $37\,^{\circ}\mathrm{C}$ can be prototype for a fuzzy set if the frame refers to human body temperatures.

Error-free reconstruction

Constraint 15 *Given a frame of cognition*

$$\mathbf{F} = \{A_1, A_2, \ldots, A_n\} \tag{2.51}$$

and a function

$$D\colon [0,1]^n \to \mathbb{R} \tag{2.52}$$

each element of the Universe of the Discourse x must be perfectly reconstructed by applying the D to its membership degrees:

$$\forall x \in U : D\left(\mu_{A_1}(x), \mu_{A_2}(x), ..., \mu_{A_n}(x)\right) = x \tag{2.53}$$

Error-free reconstruction guarantees that the 'internal representation' of an element x of the Universe of Discourse – which is given by the sequence of membership degrees $\mu_{A_1}(x), \ldots, \mu_{A_n}(x)$ is a perfect representation of x w.r.t. the function D (Espinosa and Vandewalle, 2000; Valente de Oliveira, 1999a). For interpretability analysis, this property is of great importance because it assures that the processing carried out by the model is coherent with the measured value transformed by the frame, and different elements should lead to different processing tasks.

Put more formally, error free reconstruction requires that the mapping operated by the frame to the element must be invertible. In this sense,

however, even if such mapping is actually invertible, a wrong choice of the function D may hide bijection. For such reason, in interpretability analysis a weaker form of error-free reconstruction appears more adequate:

CONSTRAINT 16 *Given a frame of cognition*

$$\mathbf{F} = \{A_1, A_2, \ldots, A_n\} \tag{2.54}$$

there exists a function $D: [0,1]^n \rightarrow \mathbb{R}$ *such that each element of the Universe of the Discourse* x *must be perfectly reconstructed by applying* D *to its membership degrees:*

$$\exists D : [0,1]^n \rightarrow \mathbb{R}, \forall x \in U : D\left(\mu_{A_1}(x), \mu_{A_2}(x), ..., \mu_{A_n}(x)\right) = x \tag{2.55}$$

In fuzzy modelling, a dual problem should be also addressed, which concern the perfect reconstruction of the sequence of membership degrees when a single numerical value is given. This problem arises when the fuzzy model provides a single numerical value by means of a defuzzification procedure and one asks whether the defuzzified value is coherent with the membership degrees provided after processing. This is a far more complex problem than that concerning error free reconstruction stated in (2.55) because it addresses bijection of a mapping that transforms a multi-dimensional space into a one-dimensional space.

The issues related to error-free reconstruction are analyzed in the context of *fuzzy interfaces*, which are frames of cognition analyzed in a pure functional viewpoint, i.e. without explicitly addressing the interpretability issues. Nevertheless, theoretical results on fuzzy interfaces are of great importance in interpretability analysis. Theoretical issues on fuzzy interfaces are analyzed in a successive Chapter, which focus on characterizing the class of fuzzy sets that guarantee error-free reconstruction (in the weak sense), and provides theoretical results concerning the assessment of fuzzy interface when error-free reconstruction is practically impossible.

2.4 Constraints on Fuzzy Information Granules

A fuzzy information granule is defined as a linguistically labelled multidimensional fuzzy set. A convenient linguistic representation of an information granule is by means of soft-constraints, proposed by Zadeh in (Zadeh, 1997). In his original work, Zadeh proposed a general framework to represent information granules deriving from different paradigms, including the Theory of

Probability or the Rough Sets Theory. Here, only the Fuzzy Set Theory is considered for representing information granules, which can be formalized as described in the following.

An atomic soft constraint is defined as a formal expression of the form:

$$v \text{ IS } FL_v \tag{2.56}$$

where v is a formal variable, "IS" is a copula and FL_v is a label associated to a fuzzy set. Atomic soft constraints can be directly derived by a Frame of Cognition, which can be conveniently denoted as

$$\mathbf{F}_v = \langle U_v, \mathbf{F}_v, \preceq_v, \mathcal{L}_v, v \rangle \tag{2.57}$$

provided that the label of the soft constraint is assigned to some fuzzy set of the frame:

$$\exists A \in \mathbf{F}_v : \mathcal{L}_v(A) = FL_v \tag{2.58}$$

The semantics of a soft constraint is straightforward. The formal variable is evaluated according to an assignment function that associates the variable v to a fuzzy set in $\mathcal{F}(U_v)$:

$$\Im(v)[x] = x \in \mathcal{F}(U_v) \tag{2.59}$$

In most cases, however, the evaluation of the variable maps to a single numerical value:

$$\Im(v)[x] = \{x\} \tag{2.60}$$

The evaluation of the label FL_v is achieved by retrieving the fuzzy set labelled by FL_v:

$$\Im(FL_v) = \mathcal{L}_v^{-1}(FL_v) \in \mathbf{F}_v \tag{2.61}$$

Finally, the interpretation of the soft constraint "v IS FL_v", denoted by FC_v for convenience, is defined as

$$\Im(FC_v)[x] = \Im(v \text{ IS } FL_v)[x] = \pi\left(\mathcal{L}_v^{-1}(FL_v), \Im(v)[x]\right) \tag{2.62}$$

where π is the possibility function, defined as:

$$\pi(A, B) = \sup \min_{x \in U_v} \{\mu_A(x), \mu_B(x)\} \tag{2.63}$$

Computation is simplified if the variable assignment associates a single scalar value to the variable as in (2.60). In such a case the evaluation of a fuzzy constraint reduces to:

$$\Im\left(FC_v\right)[x] = \mu_A(x) \tag{2.64}$$

being

$$A = \mathcal{L}_v^{-1}\left(FL_v\right) \tag{2.65}$$

More complex soft constraints can be composed by conjunction, disjunction or negation of several atomic soft constraints, eventually defined in different Universes of Discourse. In the most common case, conjunctive soft constraints are used, i.e. forms of the type

$$FC_{v_1} \text{ AND } FC_{v_2} \text{ AND } \ldots \text{ AND } FC_{v_{n_V}} \tag{2.66}$$

The interpretation of such type of complex soft constraint (denoted as $\mathbf{G}$) depends on the interpretation of the operator AND. Usually, a *t-norm* $\otimes$ is associated, i.e. a function

$$\otimes : [0,1]^2 \rightarrow [0,1] \tag{2.67}$$

such that the following properties are verified:

Commutativity $a \otimes b = b \otimes a$

Associativity $a \otimes (b \otimes c) = (a \otimes b) \otimes c = a \otimes b \otimes c$

Monotonicity $a \leq c \wedge b \leq d \rightarrow a \otimes b \leq c \otimes d$

1-unity $a \otimes 1 = a$

The evaluation of the complex soft constraint is hence defined as:

$$\begin{aligned}\Im\left(\mathbf{G}\right)[x_1, x_2, \ldots, x_{n_V}] = \\ \Im\left(FC_{v_1}\right)[x_1] \otimes \Im\left(FC_{v_2}\right)[x_2] \otimes \cdots \otimes \Im\left(FC_{v_{n_V}}\right)[x_{n_V}]\end{aligned} \tag{2.68}$$

Fuzzy information granules are represented as complex soft constraints, usually in the conjunctive form. The semantic of a fuzzy information granule is given by the interpretation of the soft constraint, thus resulting in a multidimensional fuzzy set defined on the Cartesian product of one-dimensional Universes of Discourse. The resulting Universe of Discourse is hence defined as:

$$\mathbf{U_G} = U_{v_1} \times U_{v_2} \times \cdots \times U_{v_{n_V}} \tag{2.69}$$

A fuzzy information granule defined over interpretable frames of cognition is itself interpretable. However, the augmented structure provided by conjunction of atomic soft constraints may require the definition of additional interpretability constraints.

Description length

CONSTRAINT 17 *The description length of the information granule should be as small as possible.*

The description length of the information granule is the number of soft constraints included in its description. If such number is too high, the information granule description is difficult to read and understand. The maximum number of soft constraints is suggested by psychology and should not exceed seven, plus or minus two (a deeper discussion of such limit is portrayed in the "Justifiable Number of Elements" constraint).

If the number of variables is higher than such limit, it is advisable to apply some simplification procedure, e.g. by selecting a subset of the variables to be included in the rule. The selection of an appropriate subset of variables is usually performed by means of genetic algorithms (Jiménez et al., 2001; Jin et al., 1999; Meesad and Yen, 2002; Peña-Reyes and Sipper, 2003; Ishibuchi and Yamamoto, 2002). However, other approaches may be effectively used (Shen and Chouchoulas, 2001)

In modelling contexts, an information granule can be described by a subset of the variables actually used in a model. For implementation pursuits, in such situation the remaining variables can be represented in the granule by "don't care" soft constraints, i.e. soft constraints with constant unitary truth value (Bonarini, 1997; Nauck et al., 1997). For instance, the following representations are semantically equivalent:

> TEMPERATURE IS LOW AND PRESSURE IS HIGH
>
> TEMPERATURE IS LOW AND PRESSURE IS HIGH AND HUMIDITY IS DONTCARE

since, by definition:

$$\forall x \in U_{\text{HUMIDITY}} : \Im\left(\text{HUMIDITY IS DONTCARE}\right)[x] = 1 \tag{2.70}$$

The first description is evidently more readable than the second, which in turn comprehends all the variables of the model. The latter type of description can be used internally in the model, e.g. to allow a matrix representation of an entire collection of information granules.

Attribute correlation

CONSTRAINT 18 *Information granules should be represented in terms of correlation between variable pairs*

Gaweda and Zurada (Gaweda and Zurada, 2001) propose an approach to find interpretable fuzzy models based on the representation of empirical correlation between input variables. Such correlations are expressed in linguistic terms such as:

> TEMPERATURE OUTSIDE IS MORE OR LESS THE SAME AS TEMPERATURE INSIDE

which may be used as descriptions of information granules. However, when granules are acquired from data, the general schema of the resulting description is more general and looks like the following:

> L_{i1} v_{i1} IS POSITIVELY/NEGATIVELY CORRELATED WITH L_{j1} v_{j1} AND ... AND L_{in} v_{in} IS POSITIVELY/NEGATIVELY CORRELATED WITH L_{jn} v_{jn}.

This type of formalization, while proved effective by the authors in system identification, becomes hard to understand when the number of correlations in a single information granule is high. Moreover, the notation "L_{i1} v_{i1} IS POSITIVELY/NEGATIVELY CORRELATED WITH L_{j1} v_{j1}", which states that correlation between the two variables v_{i1} and v_{j1} subsists in the fuzzy region

$$\mathcal{L}_{v_{i1}}^{-1}(L_{i1}) \times \mathcal{L}_{v_{j1}}^{-1}(L_{j1}) \tag{2.71}$$

can be difficult to read and understand in many applicative contexts. As a consequence, even though this approach shows a promising way to express understandable knowledge, there is room for research aimed at acquiring more readable knowledge from data.

2.5 Constraints on Rules

A rule is a formal structure that expresses a piece of knowledge. In the context of Granular Computing, rules define a relationship between two information granules. Their formal description follows the schema:

$$\text{IF } Ant \text{ THEN } Cons \tag{2.72}$$

where *Ant* and *Cons* are labels representing information granules, while the symbols IF/THEN formalize the direction of the relation. It must be mentioned, however, that the above mentioned rule schema is only one possible formalization for rules. A deeper investigation on fuzzy rules can be found in (Dubois and Prade, 1996).

The semantics of a rule is directly derived by the semantics of the involved information granules. More specifically, a rule can be interpreted as follows:

$$\Im\left(Rule\right)\left[\mathbf{x},\mathbf{y}\right]=\Im_{\text{THEN}}\left(\Im\left(Ant\right)\left[\mathbf{x}\right],\Im\left(Cons\right)\left[\mathbf{y}\right]\right) \tag{2.73}$$

where $\Im_{\text{THEN}}$ is a function $\odot:\left[0,1\right]^{2}\rightarrow\left[0,1\right]$. By varying the assignments of the variables in the antecedent and in the consequents, the rule defines a fuzzy multidimensional relation. Such relation can be used to infer the values of the consequent variables when the antecedent variables are properly assigned. More specifically, for a given $\mathbf{x}$ in the (multi-dimensional) Universe of Discourse, the fuzzy set $Y_{Rule}\subseteq\mathcal{F}\left(\mathbf{U}_{Cons}\right)$ with the following membership function can be inferred:

$$\mu_{Y_{Rule}}\left(\mathbf{y}\right)=\Im\left(Rule\right)\left[\mathbf{x},\mathbf{y}\right] \tag{2.74}$$

Usually, the consequent of the rule is reduced to an atomic soft constraint, which leads to a one-dimensional information granule. In such case, the rule is named "Mamdani type".

A specific case of Mamdani type rule is the "Takagi Sugeno" type, in which consequents are scalar numbers described as

$$v\ \text{IS}\ f\left(v_{1},v_{2},\ldots,v_{n_{V}}\right) \tag{2.75}$$

where v is the variable occurring in the consequent soft constraint, v_1,v_2, ..., v_{n_V} are all the variables occurring in the antecedent and f is a form denoting a function of n_V variables. Usually f denotes a polynomial of order k. In such case, the rule is called "Takagi-Sugeno of order k". The consequent information granule in a Takagi-Sugeno rule is reduced to a singleton set. Given a numerical assignment for each variable $\Im\left(v_{1}\right)\left[x_{1}\right]$, $\Im\left(v_{2}\right)\left[x_{2}\right]$, ..., $\Im\left(v_{n_{V}}\right)\left[x_{n_{V}}\right]$, the interpretation of the consequent is defined as:

$$\Im\left(v\ \text{IS}\ f\left(v_{1},v_{2},\ldots,v_{n_{V}}\right)\right)\left[y\right]=\begin{cases}1 & \text{if}\quad y=f\left(x_{1},x_{2},\ldots,x_{n_{V}}\right)\\ 0 & \text{otherwise}\end{cases} \tag{2.76}$$

Such a simplified interpretation allows for a faster inference process, since, for $\mathbf{x}=\left(x_{1},x_{2},\ldots,x_{n_{V}}\right)$:

$$\begin{aligned}\mu_{Y_{Rule}}\left(y\right)=\Im\left(Rule\right)\left[\mathbf{x},y\right]=\Im_{\text{THEN}}\left(\Im\left(Ant\right)\left[\mathbf{x}\right],\Im\left(Cons\right)\left[y\right]\right)=\\ =\begin{cases}\Im\left(Ant\right)\left[\mathbf{x}\right]\odot1 & \text{if}\quad y=f\left(x_{1},x_{2},\ldots,x_{n_{V}}\right)\\ \Im\left(Ant\right)\left[\mathbf{x}\right]\odot0 & \text{otherwise}\end{cases}\end{aligned} \tag{2.77}$$

The operator $\odot$ can have different realizations. In expert systems, it is defined as an implication while in many fuzzy models it is defined as a

conjunction. Some theoretical discussion about similarities and differences of the two realization are portrayed next in the Chapter.

Takagi-Sugeno rules are widely used in fuzzy modelling because of the computational efficiency of the inference process. However, it is undisputed that Mamdani type of rules offer a better interpretability as they offer a granulated information about consequents that is more adequate in describing complex systems.

Zero order rules are especially used in fuzzy modelling since they assign a constant scalar value to the output variable, which is hence independent from the values assigned in the antecedent variables. This kind of rules offer an interesting interpretation in classification tasks. Indeed, rules of the Takagi-Sugeno type of order 0 can express knowledge oriented to pattern recognition, in the form:

> IF *Ant* THEN OBJECT BELONGS TO CLASS κ WITH MEMBERSHIP ν

Given an object represented by the feature vector $\mathbf{x}$, the previous rule assigns the object to class κ with membership degree:

$$\mu_{Class\_\kappa}(\mathbf{x}) = \Im(ANT)[\mathbf{x}] \odot \nu \tag{2.78}$$

Usually, the membership degree is fixed to $\nu = 1$ and the rule schema is further simplified to

> IF *Ant* THEN OBJECT BELONGS TO CLASS κ

In other cases, multiple classes are simultaneously present in the same rule leading to rules of the form:

> IF *Ant* THEN OBJECT BELONGS TO CLASS κ_1 WITH MEMBERSHIP ν_1, CLASS κ_2 WITH MEMBERSHIP ν_2, ..., CLASS κ_{n_K} WITH MEMBERSHIP ν_{n_K}

In all cases, rules are formal modifications of 0^{th} order Takagi Sugeno rules, which hence appear as a useful formalization for complex classification tasks (Nauck et al., 1996; Castellano et al., 2001).

The structure of a fuzzy rule leads to a very interpretable form of knowledge, because it is close to expert rules that are usually applied in human reasoning and decision making. Such closeness enables a tight integration of expert rules and rules acquired from data within the same modelling framework. This kind of integration is extensively studied within the so-called "knowledge intensive learning", where benefits are shown both from empirical studies and theoretical considerations. Notwithstanding the clear structure of a fuzzy rule, some additional constraints, further described in detail, can be useful for further improving its interpretability.

Rule length

CONSTRAINT 19 *The length of the rule should be as small as possible*

The length of the rule is the sum of the length of the antecedent and the consequent information granules. With the assumption of a single consequent in each rule, the length of the rule is totally related to the length of the antecedent information granule. A discussion of interpretability related to the length of an information granule is portrayed in Section 2.4.

High-order consequents

CONSTRAINT 20 (MAMDANI TYPE OF RULES) *Each consequent of the rule is a fuzzy set describing the value of the function and its derivative*

CONSTRAINT 21 (TAKAGI-SUGENO TYPE OF RULES) *Each consequent of the rule is a truncated Taylor series centered on the antecedent prototype.*

Consequents are very crucial in the final understanding of a rule. When rules are used to identify unknown functions, consequents may be more readable if they provide a description of the shape of the function locally represented by the rule. In (Höppner and Klawonn, 2000), Mamdani consequents are used to describe the values of the function and its derivative in the local neighborhood covered by each rule. An example of such rules is the following:

> IF X IS NEAR 0 THEN G(X) HAS A STEEP LOCAL MINIMUM NEAR 0

Such rule richly describes the shape of the function $g(x)$ in the neighborhood of zero by providing a fuzzy quantification of the value of the function ("NEAR 0"), its first derivative ("LOCAL MINIMUM") and its second derivative ("STEEP" local minimum). It should be noted, however, that readability can be seriously hampered if the number of dimensions is high (more than two). This is because such rules offer an effective way of building an imaginary geometrical representation of the identified function, but human beings can hardly imagine function graphs with more than two dimensions.

In (Bikdash, 1999), more precise interpolation is provided by first order Takagi-Sugeno rules. To preserve interpretability, the linear coefficients of the consequent function are constrained to represent the coefficient of the truncated Taylor expansion of the identified function, e.g.:

> IF v_1 IS FL_{v_1} AND...AND v_1 IS FL_{v_1} THEN Y=B0 + B1(v_1-r_1) +...+ Bn(v_n-r_n)

where each r_i is the prototype of the convex unimodal fuzzy set $L_{v_i}^{-1}(FL_{v_i})$. Eventually, second order derivatives can be accommodated by using second-order Takagi-Sugeno rules.

Both the approaches offer an undisputed improvement of interpretability of the rule, even though their usefulness is especially evident in system identification, while in other applications different forms of consequent representation may be more significative.

Improved consequents

CONSTRAINT 22 (MAMDANI RULES) *Consequents are expressed with two fuzzy sets in the form*

$$\text{IF } \ldots \text{ THEN } v \text{ IS BETWEEN } FL_{v,1} \text{ AND } FL_{v,2} \tag{2.79}$$

CONSTRAINT 23 (0TH ORDER TAKAGI-SUGENO RULES) *Scalar consequents must be replaced by a suitable fuzzy set belonging to a reference frame of cognition.*

Cordón and Herrera reported a number of problems that arise when using classical Mamdani rules (Cordón and Herrera, 2000):

- Lack of flexibility due to the rigid partitioning of input and output spaces;
- The homogeneous partitioning of these spaces when the input/output mapping varies in complexity is inefficient and does not scale to high-dimensional spaces;
- Dependent input variables are very hard to partition;
- The small size of the rule-base lead to model inaccuracies.

Such problems come up especially when fuzzy sets are not automatically acquired from data but are initially established before model design. The authors coped with them through the development of an extended version of fuzzy rules that associates two fuzzy sets for each variable in the consequent. Each rule is internally elaborated as a couple of rules, with the same antecedent but each with one of the two consequents. A dangerous side effect of such internal representation consists in the introduction of potential inconsistencies (see Section 2.6), which is one of the most required constraints for interpretability.

In Takagi-Sugeno rules, issues are different because the output space is not partitioned but only singleton values are included in rules. Input/output mapping is then realized by linear interpolation that uses the rule activation strengths[19] as weight factors. Takagi-Sugeno rules are widely used because of their efficiency in numerical mappings, but they lose readability when compared with Mamdani systems. To improve readability, Setnes et al. propose a simple method to convert 0th order Takagi-Sugeno rules (eventually acquired from data) into Mamdani rules (Setnes et al., 1998b). They fix a reference frame of cognition consisting of a collection of fuzzy sets defined on the output universe of discourse and then they associate each numerical consequent with the most possible fuzzy set of the reference frame. In this way, the rule becomes of Mamdani type and gains in interpretability. The new Mamdani rules are however used only for representational pursuits, while the rules are internally stored in the Takagi-Sugeno form, in order to avoid the accuracy loss that afflicts Mamdani rules.

2.6 Constraints on Fuzzy Models

A fuzzy rule-based model[20] is a complex structure

$$\mathcal{M} = \langle \mathbb{F}_I, \mathbb{F}_O, \mathbf{R}, \prec, \otimes, \odot, \oplus, D \rangle \tag{2.80}$$

where:

- $\mathbb{F}_I$ and $\mathbb{F}_O$ are sets of input and output frames of cognition, as defined in (2.14). It is assumed that $\mathbb{F}_O$ consists of only one frame of cognition, while $\mathbb{F}_I$ consists of $n_I > 0$ frames of cognition. $\mathbb{F}_I$ and $\mathbb{F}_O$ constitute the *database* of the model;
- $\mathbf{R}$ is a finite set of n_R rules defined as in Section 2.5. It is the *knowledge base* of the model. If all rules are of Mamdani form, the model is said of Mamdani type. Similarly, if all rules are in Takagi-Sugeno form (of order o), the model is said of Takagi-Sugeno type (of order o). Hybrid models are theoretically possible, since Takagi-Sugeno rules are actually Mamdani rules with degenerate consequents, but they are generally not used;

[19] The activation sternght of a rule is defined as the membership degree of the antecedent information granule, eventually multiplied by a weighting factor. This latter factor has not been included in this discussion, since it does not appear useful in interpretability analysis.

[20] Hereafter, the term "fuzzy model" actually indicates the more exact specification "fuzzy rule-based model".

- $\prec$ defines an ordering of the input frames of cognitions. For sake of simplicity, each frame of cognition is denoted by $\mathbf{F}_i = \langle U_i, \mathbf{F}_i, \preceq_i, \mathcal{L}_i, v_i \rangle$ and it is assumed that $i < j \rightarrow \mathbf{F}_i \prec \mathbf{F}_j$;
- $\otimes$ is a t-norm;
- $\odot$ is an operator used to interpret the "Then" part of each rule
- $\oplus$ is an operator used to aggregate all fuzzy sets coming from each rule
- $\mathcal{D}$ is a defuzzificator, i.e. a function $\mathcal{D} : \mathcal{F}(U_O) \rightarrow U_O$ where U_O is the Universe of Discourse of the output frame of cognition.

An inference $\Im_{\mathcal{M}}(x_1, x_2, \ldots, x_{n_I})$ of the model is an effective procedure that:

1. Associates the value $x_i \in U_i$ to the variable v_i, i.e. $\Im(v_i)[x_i] = x_i$
2. Interprets each rule $R_j \in \mathbf{R}$ according to (2.73), by associating the t-norm $\otimes$ to the symbol AND and $\odot$ to the symbol THEN.
3. The resulting interpretations of each rule provides an ordered set of R fuzzy sets, which are further unified into one fuzzy set according to the operator $\oplus$, attaining the final fuzzy set B with membership function:

$$\mu_B(y) = \Im(R_1)[y] \oplus \ldots \oplus \Im(R_R)[y] \tag{2.81}$$

The inference $\Im_{\mathcal{M}}$ is concluded by applying the defuzzificator $\mathcal{D}$ to the resulting fuzzy set B. Hence the inference function is a mapping from the global universe of discourse $\mathbf{U} = U_1 \times U_2 \times \cdots \times U_{n_I}$ to a value of the output Universe of Discourse U_O:

$$\Im_{\mathcal{M}} : \mathbf{U} \rightarrow U_O \tag{2.82}$$

For sake of completeness it should be mentioned that the described inference procedure is only one of two possible inference strategies, which are, namely:

FITA (First Infer Then Aggregate) The output of each rule is first inferred, then, all outputs are aggregated by union. This is the strategy described previously and adopted hereafter.

FATI (First Aggregate Then Infer) All antecedents of the rules are aggregated to form a multidimensional fuzzy relation. Via the composition principle the output fuzzy set is derived.

However, it can be proved that the FITA and FATI approaches are equivalent if inputs are scalar values (Czogala and Leski, 2000). For this reason, only the FITA approach will be considered hereafter.

The inference $\Im_{\mathcal{M}}$ is a mapping from a numerical space $\mathbf{U}$ to a numerical space U_O. This kind of inference is considered as default because it is the most widely used in literature. However, another form of inference could be considered, which takes fuzzy inputs and provides a fuzzy output defined by (2.81). In such a case, the defuzzificator $\mathcal{D}$ is not applied. This kind of fuzzy mapping will be indicated as:

$$\Im_{\mathcal{M}} : \mathcal{F}(\mathbf{U}) \rightarrow \mathcal{F}(U_O) \tag{2.83}$$

A number of interpretability issues arise when considering a model as a whole system consisting of a formal knowledge base and an inference mechanism. In the following, some interpretability constraints are portrayed. For sake of simplicity, the following notation will be used:

$$\Im_R = \langle \mathbf{A}, B \rangle \text{ or } \Im_R = \langle \mathbf{A}, f \rangle \tag{2.84}$$

to indicate that the rule R in the rule set $\mathbf{R}$ has all fuzzy sets in the antecedent part whose Cartesian product is $\mathbf{A}$ and in the consequent part has a fuzzy set B (Mamdani rule) or a function f (Takagi-Sugeno rule).

Implication and aggregation

An important interpretability issue for fuzzy models concerns the inference process, which depends on the specific choices of the operators $\odot$ (inference from each rule) and $\oplus$ (aggregation of results inferred from each rule). More specifically, two forms of interpretations can be adopted:

Conjunctive Interpretation The "Then" part of a rule is actually a conjunction, hence it is interpreted with a t-norm. Rules are consequently disjointed and the aggregation of rules is made by t-conorms. This type of rules defines a *fuzzy graph*, as indicated by Zadeh in (Zadeh, 1996b);

Implicative Interpretation The "Then" part of a rule is actually an implication, which may be interpreted in several ways (Trillas et al., 2000). Rules are consequently conjoined and the aggregation is realized by t-norms.

In Boolean logics the two types of inference are formally equivalent, provided that some weak conditions on the antecedents and consequents hold

(Mundici, 2002). In fuzzy logics such equivalence is lost due to the invalidity of some logical tautologies (e.g. the distributivity property). As a consequence, the choice of the type of inference is a crucial step when designing fuzzy models.

The first form of interpretation is generally used for representing case-defined functions, like:

> EITHER v IS $FL_{v,1}$ AND THEN u IS $FL_{u,1}$ OR v IS $FL_{v,2}$ AND THEN u IS $FL_{u,2}$ OR...OR v IS $FL_{v,R}$ AND THEN u IS $FL_{u,R}$

Most fuzzy models are designed to represent case-defined functions, being each case attained by an automatic granulation procedure. For this reason, most rule-based fuzzy models are interpreted according to the first type of inference, even though the representation schema is in the "IF...THEN" formalism.

The second type of inference is usually adopted in fuzzy expert systems. The main difference between this type of inference and the first type is that consequents are still valid even if antecedents are false. This feature is convenient in automated reasoning but it may create some difficulties in engineering applications where the first type of inference is preferred. Nevertheless, only fuzzy models in which the second type of inference is adopted the "IF...THEN" formalization is legitimate, while it is abused (though widely accepted) when the first type of inference is used. From the interpretability analysis standpoint, the correct choice of the rule schema is an important design point for an interpretable knowledge base.

Number of rules (compactness)

CONSTRAINT 24 *The total number of rules in a model must not be too high.*

This constraint is similar to the "Justifiable number of elements" for Frames of Cognition, as well as the "Rule length" criterion, and finds the same psychological motivations. The principle of reducing the complexity of the model can be considered as an application of the well-known "Occam's Razor" principle[21] – widely used in symbolic Artificial Intelligence – which states that among all possible correct hypotheses, the one that reflects the reality is most probably the simplest. Occam's Razor finds its justification in philosophy and metaphysics, but can be motivated by our need to comprehend phenomena with the simplest possible explanation of it. This principle translates practically in fuzzy modelling by keeping the number of fuzzy sets,

[21] See note 11, p. 18

frames of cognition and rules as small as possible. In learning theory, a model with a small number of rules is a candidate for endowing a knowledge base that is general enough, while complex fuzzy models are more prone to overfitting[22]. However, if the model is too simple, it may be useless since the space of mappable functions is too poor (Tikk and Baranyi, 2003b).

Compactness is a widely required constraint for interpretable fuzzy modelling. This requirement is needed because the upper bound of fuzzy rules is exponentially related to the input dimensionality. Indeed, if each $i = 1, 2, \ldots, n_I$, the i-th input Frame of Cognition $\mathbf{F}_i$ consists of K_i fuzzy sets, then the number of all possible rules follows the number of all possible combinations of fuzzy sets (one fuzzy set for each frame), resulting in the upper bound:

$$R_{\max} = \prod_{i=1}^{n_I} K_i \tag{2.85}$$

In (Combs and Andrews, 1998) the authors propose a different way to combine fuzzy sets, claiming that their method eliminates the problem of combinatorial rule explosion. More specifically, the authors use a different approach in building antecedents by using disjunct expressions instead of conjunctions. However such method has been severely criticized in (Dick and Kandel, 1999; Mendel and Liang, 1999). The main flaw of the technique is its validity only for monotonic case-based functions, but in this circumstance also models with conjunctive antecedents do not suffer of combinatorial rule explosion.

To avoid combinatorial rule explosion, several techniques have been proposed, such as: tree partition of the Universe of Discourse (Fischer et al., 1998); model pruning (Hermann, 1997); rule merging (Jiménez et al., 2001; Paiva and Dourado, 2001), etc.. Other techniques are aimed at reducing the number of fuzzy sets (Jin et al., 1999; Jin and Sendhoff, 2003), e.g. by windowing the dataset from which the fuzzy model is generated by using the results of B. Kosko (Kosko, 1995). Genetic approaches are also used (Peña-Reyes and Sipper, 2003), but the most widely used technique to generate parsimonious fuzzy models is cluster analysis, see e.g. (Angelov, 2004; Auephanwiriyakul and Keller, 2002; Chepoi and Dumitrescu, 1999; Corsini et al., 2004; Chiang et al., 2004; Groenen and Jajuga, 2001; Klawonn and Keller, 1997; Liao et al., 2003; Pedrycz, 1996a; Pedrycz, 2002; Roubos and Setnes, 2001; Setnes et al., 1998b; Yao et al., 2000; Zahid et al., 2001; Zhang and Leung, 2004).

[22] This problem has been faced within the field of neurocomputation under the name of "Bias/Variance Dilemma" (Geman et al., 1992)

Clustering is widely used because it allows to discover hidden relationships among multidimensional data. Cluster analysis, both crisp and fuzzy, has a long tradition (Baraldi and Blonda, 1999; Jain et al., 1999), but its application on interpretable fuzzy modeling is problematic because interpretability constraints are easily violated if the clustering process is not properly guided.

In (Guillaume, 2001) a number of clustering algorithms for fuzzy information granulation are surveyed with special emphasis on interpretability issues. There, clustering techniques emerge as viable solutions for compact models, especially for high-dimensional Universes of Discourse, but they are put in contrast to "shared partitioning" techniques (e.g. grid or hierarchical partitioning) for the interpretability of the derived information granules. To exploit both the features of clustering techniques and shared-partitioning techniques, in a successive Chapter a new information granulation framework is proposed, which has been called "Double Clustering". The application of Double Clustering enables the generation of fuzzy information granules that – similarly to clustering techniques – well capture hidden relationships among data, and – similarly to shared partitioning – are defined by shared fuzzy sets to improve their interpretability.

Number of firing rules

CONSTRAINT 25 *The number of rules simultaneously activated (firing) by an input should be small*

A rule in a model is said to be activated if the antecedent of the rule has nonzero truth degree (also called 'activation strength'). To improve interpretability, in (Peña-Reyes and Sipper, 2003) it is suggested to limit the number of rules simultaneous activated by a single input , in order to provide a simple local view of the model behavior. Moreover, a limited number of firing rule assures an efficient inference process since the complexity of the inference procedure is linearly related to the number of rules, as can be seen in (2.81).

When limited support fuzzy sets are used in the model, such as triangular or trapezoidal fuzzy sets, the number of fuzzy rules is limited, but it is still upper bounded by a number exponentially related to the dimensionality of the Universe of Discourse.

PROPOSITION 2.4 *If for each dimension the number of fuzzy sets with nonzero membership is at most k, then the number of firing rules is at most k^n, where n is the dimension of the Universe of Discourse.*

Proof. *The proof is by induction. For $n = 1$ the statement is trivially true. Suppose that it is also true for $n - 1 \geq 1$. Consider a model $\mathcal{M}_{n-1}$ defined on an*

input Universe of Discourse $\mathbf{U}_{n-1}$ *of dimension* $n-1$*. It consists of* R_{n-1} *rules interpreted as* $\langle \mathbf{A}_i, B_i \rangle$ *for* $i = 1, 2, \ldots, R_{n-1}$*. A Frame of Cognition consisting of* K_n *fuzzy sets can be considered, and a new set of rules can be defined. For each rule* $\langle \mathbf{A}_i, B_i \rangle$ *a set of* K_n *rules are defined with the following interpretation* $\langle \mathbf{A}_i \times C_j, B_i \rangle$*, for* $i = 1, 2, \ldots, R_{n-1}$ *and* $j = 1, 2, \ldots, K_n$*. The total number of new rules is hence* $R_n = K_n \cdot R_{n-1}$*. Such new rules are the basis for a new model* $\mathcal{M}_n$ *defined on a* n*-dimensional Universe of Discourse and having* $\mathcal{M}_{n-1}$ *as a sub-model. Consider an input* $\mathbf{x}_n$ *defined on the* n*-th Universe of Discourse* $\mathbf{U}_n = \mathbf{U}_{n-1} \times U$*. The input can be conveniently represented as* $\mathbf{x}_n = [\mathbf{x}_{n-1}, x]$ *where* $\mathbf{x}_{n-1} \in \mathbf{U}_{n-1}$ *and* $x \in U$*. The sub-vector* $\mathbf{x}_{n-1}$ *is a suitable input for* $\mathcal{M}_{n-1}$ *which activates at most* k^{n-1} *rules, by induction hypothesis. Furthermore, by hypothesis the scalar value* x *activates at most* k *fuzzy sets of the Frame of Cognition defined on the* n*-th dimension. As a consequence, the number of firing rules in* $\mathcal{M}_n$ *is bounded by the number of rules firing in* $\mathcal{M}_{n-1}$ *times the number of fuzzy sets with non-zero membership degree on the* n*-th dimension. Such upper bound is hence* $k \cdot k^{n-1} = k^n$*.* ■

As a corollary of the previous proposition, it should be clear that, even though the adoption of limited support fuzzy sets helps in significantly reducing the number of firing rules, its upper bound *is exponentially related to the number of dimensions*. As a consequence, the adoption of limited support fuzzy sets may not be enough for efficient and interpretable inference process. The upper bound of firing rules should also be lowered by reducing the total number of rules in the knowledge base. For this reason clustering techniques are widely used to design fuzzy models, since the number of generated rules is usually linearly related to the number of dimensions, thus limiting combinatorial explosion. When a low number of rules defines the knowledge base of the fuzzy model, even infinite support fuzzy sets may be used, like Gaussian fuzzy sets. Eventually, further efficiency optimization can be attained by dropping from inference fuzzy rules activation strength smaller than a proper threshold.

Shared fuzzy sets

Constraint 26 *Fuzzy sets should be shared among rules.*

If the distinguishability constraint is taking into account when designing interpretable fuzzy models, it may be necessary that different information granules share the same fuzzy sets, as in the following example:

R_1 : If x is Small and y is High Then z is Small

R_2 : If x is Small and y is Medium Then z is Medium

R_3 : If x is Medium and y is High Then z is Medium

In the example, R_1 and R_2 share the same fuzzy set for the input variable x ("Small"), while R_2 and R_3 share the same fuzzy set for the output variable z ("Medium"). Shared fuzzy sets are helpful for representing the knowledge base with a small number of different symbols.

To enable fuzzy set sharing, several methods have been proposed. In the simplest case, all possible combinations of fuzzy sets coming from different frames of cognition are used to define rules (Ishibuchi et al., 1994). However, such technique provides for an exponential number of rules that poorly describe the hidden input/output relationship in data. To overcome such problem, techniques for partition refinement (Rojas et al., 2000) or genetic algorithms are used (Ishibuchi et al., 1995). More sophisticated techniques use Neuro-Fuzzy decision trees (Ichihashi et al., 1996) or fuzzy clustering followed by similarity-based merging techniques to enable sharing well distinguishable fuzzy sets among several rules (Babuška, 1998).

Rule locality

Constraint 27 *The mapping realized by the Takagi-Sugeno fuzzy model is a composition of independent local models, being each local model defined by a rule of the knowledge base:*

$$\forall \mathbf{x} \in \mathbf{U} \forall R \in \mathbf{R} : \Im_R = \langle \mathbf{A}, f \rangle \wedge \mu_{\mathbf{A}}(\mathbf{x}) = 1 \rightarrow \Im_{\mathcal{M}}(\mathbf{x}) = f(\mathbf{x}) \wedge \bigtriangledown \Im_{\mathcal{M}}(\mathbf{x}) = \bigtriangledown f(\mathbf{x}) \quad (2.86)$$

Rule locality is an interpretability constraint especially required when fuzzy models are used in control applications (Johansen et al., 2000; Lotfi et al., 1996; Fischer et al., 1998; Setnes et al., 1998b; Yen et al., 1998). Actually, rule locality is a desirable in all applications of Takagi-Sugeno fuzzy models because it requires that each rule well represents a localized region of the domain $\mathbf{U}$ and hence corresponds to a meaningful piece of knowledge. Rule locality is not a specific requirement for fuzzy models, but also for models defined by crisp regression rules (Malerba et al., 2004).

Rule locality is hard to achieve because often models are tuned so as to achieve a good accuracy on an available dataset. The tuning procedure, however adapts the model by considering the output provided by the model, which results by the aggregation of an entire set of rules. In this way, the tuning process can improve the overall quality of the model, but the locality of each rule can be seriously hampered.

On the other hand, strategies that acquire fuzzy models by independently generating each single rule may not achieve good accuracy results

when all rules are considered simultaneously in the inference process. Combined global/local learning is a promising approach to reach rule locality while maintaining good global accuracy of the model (Yen et al., 1998). However, it should be noted that even in combined learning, the rule locality constraint is only approximately satisfied.

Modus-Ponens

CONSTRAINT 28 (FUZZY MAPPING) *If an input matches the antecedent of a rule, the output of the system must coincide with that rule consequent:*

$$\forall \mathbf{X} \in \mathcal{F}(\mathbf{U}) \forall R \in \mathbf{R} : \Im_R = (\mathbf{A}, B) \wedge \pi(\mathbf{A}, \mathbf{X}) = 1 \rightarrow \Im_{\mathcal{M}}(\mathbf{X}) = B \quad (2.87)$$

CONSTRAINT 29 (NUMERICAL MAPPING) *If a rule in the system has full activation strength for a given input, the output of the system must belong to the core*[23] *of the rule consequent:*

$$\forall \mathbf{x} \in \mathbf{U} \forall R \in \mathbf{R} : \Im_R = (\mathbf{A}, B) \wedge \mu_{\mathbf{A}}(\mathbf{x}) = 1 \rightarrow \mu_B(\Im_{\mathcal{M}}(\mathbf{x})) = 1 \quad (2.88)$$

In classical logics, if a rule base contains a rule in the form $A \implies B$ and the antecedent A is verified by the input, than the consequent B must be verified by the output. This inference meta-rule, called *Modus Ponens*, is fundamental of any reasoning scheme. As a consequence, it is desirable that Modus Ponens could be verified also in fuzzy inference systems. Unfortunately, Modus Ponens satisfaction is not verified *ipso facto* but must be guaranteed by an appropriate design of the fuzzy model. As an example, in (Lazzerini and Marcelloni, 2000), a study is made where proper interpretations of logical connectives verify Modus Ponens in Mamdani fuzzy models. In (Riid and Rüstern, 2000), it is proved that the adoption of constrained triangular fuzzy sets allows for Modus Ponens verification in 0th order Takagi-Sugeno fuzzy models. It should be noted, however, that such conditions are only sufficient, and the existence of other design configurations that enable Modus Ponens verification are possible.

Consistency

CONSTRAINT 30 *If two or more rules have simultaneously high activation strength, (e.g. antecedents are very similar), the consequents must be very similar*

[23] The core of a fuzzy set is the set of elements with full membership, $\operatorname{core} A = \{x \in U : \mu_A(x) = 1\}$

Consistency if one of the most crucial requirements of any expert system (Russell and Norvig, 1995). The knowledge base of an expert system can be viewed as a formal theory in which rules are the axiomatic postulates. A theory is consistent if it is not possible to derive a statement and its negation. If the theory is inconsistent, then any statement can be derived from the theory, be it true or false in the world modeled by the expert system. As a consequence, an expert system with inconsistent knowledge base is useless.

In rule-based expert systems, inconsistency comes up when there are two rules in the form $A \rightarrow B$ and $A \rightarrow C$ where B and C are mutually exclusive concepts, i.e. either $B = \neg C$ or it is possible to derive $B \rightarrow \neg C$ (or, equivalently $C \rightarrow \neg B$) from the knowledge base. In fuzzy logic, consistency is a less stringent requirement because the 'non contradictoriness' law is no more valid, i.e. the statement $A \wedge \neg A$ can be true with a certain degree greater than zero. This is considered a point of strength of fuzzy logic since it is still possible to make inference even in presence of inconsistency, as often occurs in real world. In fuzzy logic (in the broad sense), consistency is a matter of degree; hence in interpretable fuzzy modelling inconsistence could be tolerated as long as its degree is acceptably small.

Jin and Sendhoff extend the concept of inconsistency to cover all situations that may generate confusion in understanding the knowledge base (Jin and Sendhoff, 2003). Such inconsistency conditions are exemplified as follows:

CONDITION 2.1 *Consequents are the same, but consequents are very different*

R1: IF x1 IS A1 AND x2 IS A2 THEN y IS PositiveLarge

R2: IF x1 IS A1 AND x2 IS A2 THEN y IS NegativeLarge

This is the classical inconsistency situation, where two quasi mutually exclusive consequents are derived by the same antecedent. The two consequents can have very low overlap and consequently the inferred fuzzy set results highly subnormal[24] without any clear indication of a suitable numerical output to be provided (see fig. 2.12).

CONDITION 2.2 *Conditions are apparently differently, actually being physically the same*

R1: IF cycletime IS Low THEN y IS PositiveLarge

R2: IF frequency IS High THEN y IS NegativeLarge

[24]Depending on the interpretation of the "Then" symbols, as discussed next in the Chapter

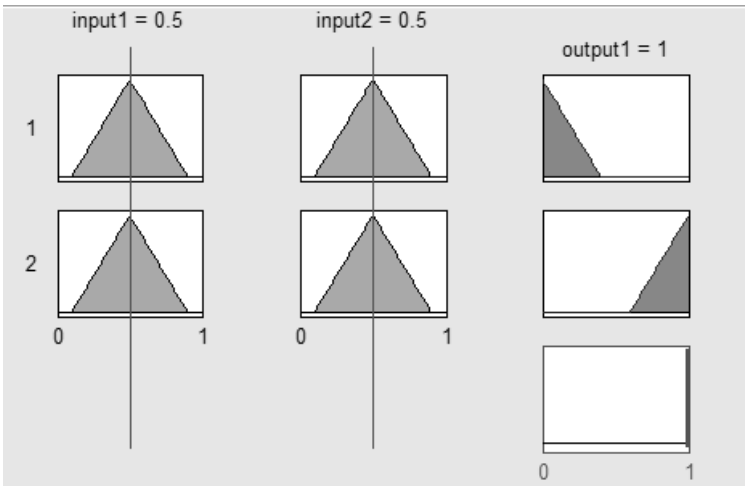

Figure 2.12: Example of inference with rules in implicative interpretation. The inferred fuzzy set (bottom-right) is empty

This is an insidious problem that cannot be checked automatically, unless a complex ontology on the variables and fuzzy set labels is provided. This kind of inconsistency occurs when there exists a (generally unknown) functional relation between two or more variables (in the example, there exists the relation $frequency = 1/cicle\_time$). Usually, detection of such type of inconsistency is left to the domain experts that analyze the knowledge base.

Condition 2.3 *Conditions are contradictory*

R1: IF sun is Bright and rain is Heavy then ...

This situation does not lead to inconsistency in the proper sense because contradicting concepts appear in the antecedents but not in the consequents. However, rules of this type can be considered inconsistent (in a larger extent) since they will never be activated (with high strength) by any input configuration. Rules of this type can be safely removed from the knowledge base because they do not significantly influence the inference process but can generate confusion when reading the knowledge base.

Condition 2.4 *Consequents involve contradictory actions*

R1: IF ... then acceleration is High and brake is On

This kind of inconsistence is similar to the first type (same antecedent, different consequents) but it is more subtle because consequents are defined

on different Universes of Discourse and inconsistency cannot be checked automatically. However, a proper integration of expert knowledge can overcome the problem by inserting rules that exclude contradictory actions. In the example the additional rule "IF ACCELERATION IS GREATER THAN ZERO THEN BRAKE IS OFF" enables the automatic detection of an inconsistency.

When inconsistency can be automatically detected, it can be assessed by some quantitative measurements. If the degree of inconsistence degree is too high, the knowledge base must be modified so as to improve its consistency. To assess consistency, Jin et al. propose the following measure to be used in the fitness function of a genetic algorithm (Jin et al., 1998b):

$$f_{Incons} = \sum_{i=1}^{R} \sum_{\substack{k=1 \\ k \neq i}}^{R} \left(1 - Cons\left(R_i, R_k\right)\right) \tag{2.89}$$

where:

$$Cons\left(R_i, R_k\right) = \exp\left(-\frac{\left(\frac{S(\mathbf{A}_i, \mathbf{A}_k)}{S(B_i, B_k)} - 1\right)^2}{\left(S\left(\mathbf{A}_i, \mathbf{A}_k\right)\right)^{-2}}\right) \tag{2.90}$$

being S a similarity measure, $\Im_{R_i} = \langle \mathbf{A}_i, B_i \rangle$ and $\Im_{R_k} = \langle \mathbf{A}_k, B_k \rangle$. In (Meesad and Yen, 2002), inconsistency is evaluated as:

$$RS = \frac{2}{R\left(R-1\right)} \sum_{i=1}^{R} \sum_{\substack{k=1 \\ k \neq i}}^{R} RS\left(R_i, R_k\right) \tag{2.91}$$

where:

$$RS\left(R_i, R_k\right) = \frac{1}{n+1} \sum_{j=1}^{n} \left(S\left(A_i^j, A_k^j\right) + \mathrm{eq}\left(B_i, B_k\right)\right) \tag{2.92}$$

being $\mathbf{A}_i = A_i^1 \times \cdots \times A_i^n$, $\mathbf{A}_k = A_k^1 \times \cdots \times A_k^n$, and $\mathrm{eq}\left(B_i, B_k\right) = 1$ if $B_i = B_k$, 0 otherwise.

Consistency of rule-based fuzzy models has also been theoretically analyzed in the specific context of implicative interpretation. This analysis is well synthesized by Pedrycz and Gomide (Pedrycz and Gomide, 1998). Briefly speaking, inconsistency is defined at the level of a single fuzzy set and it is measured as "The degree to which set is included in the empty set", i.e.:

$$c\left(A\right) = \nu\left(\varnothing, A\right) = 1 - \pi\left(U, A\right) = 1 - \sup_{x \in U} \mu_A\left(x\right) \tag{2.93}$$

being ν the necessity measure that measures the degree of inclusion between two fuzzy sets (Dubois and Prade, 1980).

Yager and Larsen use such definition to derive a necessary condition for two rules to be inconsistent (Yager and Larsen, 1991). Given two rules R_1 and R_2 such that $\Im_{R_1} = (\mathbf{A}_1, B_1)$ and $\Im_{R_2} = (\mathbf{A}_2, B_2)$, they are $\alpha-$inconsistent if there exist a fuzzy input $\mathbf{A}$ such that the fuzzy set B inferred by the model has a degree of inconsistency α measured by (2.93). If the two rules are $\alpha-$inconsistent, then the following inequalities hold:

$$\pi(B_1, B_2) \leq 1 - \alpha \tag{2.94a}$$
$$\pi(\mathbf{A}_1, \mathbf{A}_2) \geq \alpha \tag{2.94b}$$

The two inequalities formalize the condition of similar antecedents (2.94b) and dissimilar consequents (2.94a).

To check whether two rules are potentially inconsistent, the following function is used:

$$E(x) = \mu_{\neg\mathbf{A}_1}(x) \otimes \mu_{\neg\mathbf{A}_2}(x) \otimes \pi(B_1, B_2) \tag{2.95}$$

If there exists some x for which $E(x) < 1$ then, potential inconsistency exists and the corresponding $\alpha-$level can be calculated. Scalpelli and Gomide extended this method to deal with multiple rules by constructing a matrix of fuzzy sets that is subject of computation to derive potential inconsistencies among couples of rules (Scarpelli and Gomide, 1994). The authors also extend this approach to deal with chaining rules, i.e. rules whose consequent variables are used as antecedent variables in other rules of the knowledge base, by using the idea of Agarwal and Taniru (Agarwal and Taniru, 1992).

Some observation are noteworthy when analyzing consistency of rule-based fuzzy models. As previously stated, consistency is a requirement of classical expert systems that has been extended to fuzzy models. However, as previously discussed, two interpretations exists of a fuzzy knowledge base: implicative interpretation and conjunctive interpretation. Implicative interpretation is a direct generalization of classical rule-based expert system within the fuzzy framework. In such interpretation, consistency issues occur also in fuzzy models and can be analyzed in the ways previously illustrated. But the most common interpretation of the fuzzy knowledge base is conjunctive, which is not equivalent to the implicative interpretation and gives rise to different considerations.

To illustrate the consistency issues in a conjunctive interpretation of fuzzy rules, a simplified setting can be considered. Let be $\mathbf{R}$ a trivial crisp knowledge base consisting of only two rules. If such rules are interpreted in a

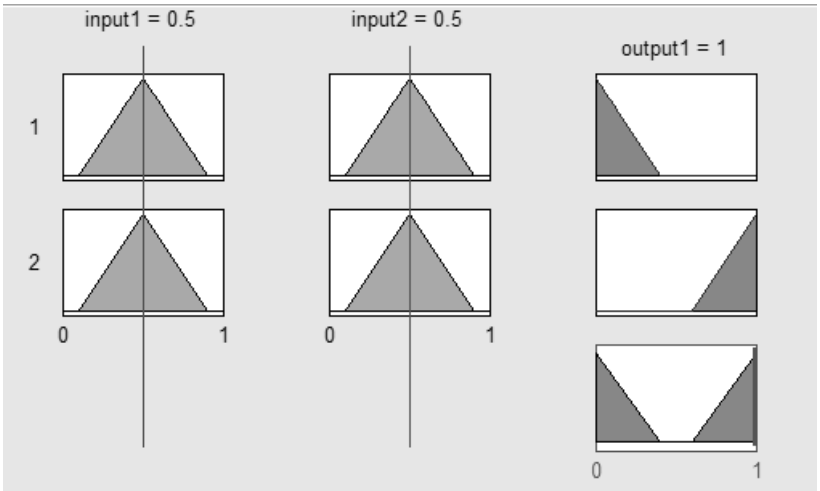

Figure 2.13: Example of inference with rules in conjuctive interpretation. The inferred fuzzy set (bottom right) represents two distinct and equally possible alternatives.

conjunctive modality, the entire knowledge base can be formalized in classical logics as

$$(Ant_1 \wedge Cons_1) \vee (Ant_2 \wedge Cons_2) \tag{2.96}$$

Suppose that $Ant_1 = Ant_2$ while $Cons_1 = \neg Cons_2$. If the common antecedent is verified by an input, then it is possible to derive $Cons_1 \vee \neg Cons_1$ *which is a tautology, not a contradiction.* Actually, contradiction cannot occur because rules are tied by disjunction, so 'non contradictoriness' law cannot apply. In light of such considerations, consistency conditions should be reviewed according to the new interpretation. For example, if two rules have the same antecedent but different consequent, i.e.

$$\Im_{R_1} = (\mathbf{A}_1, B_1) \wedge \Im_{R_2} = (\mathbf{A}_1, B_2) \wedge \pi\left(B_1, B_2\right) \ll 1 \tag{2.97}$$

then, for an input that fully verifies the antecedent $\mathbf{A}_1$, the model infers $B_1 \vee B_2$, meaning that both B_1 and B_2 are possible fuzzy sets in which the output can belong. As an illustrative example, in fig. 2.13 an inference process is graphically depicted, which illustrates the inexistence of inconsistency when rules are interpreted in the conjunctive modality. It is interesting to compare such inference with that depicted in fig. 2.12, where inconsistency degree is maximal.

Model completeness

CONSTRAINT 31 (COMPLETENESS) *For each input, at least one rule must have non-zero activation strength:*

$$\forall \mathbf{x} \in \mathbf{U} \exists R \in \mathbf{R} : \Im_R = (\mathbf{A}, B) : \mu_{\mathbf{A}}(\mathbf{x}) > 0 \tag{2.98}$$

CONSTRAINT 32 (α-COMPLETENESS) *For a given threshold α, and for each input, at least one rule must have activation strength greater than α:*

$$\forall \mathbf{x} \in \mathbf{U} \exists R \in \mathbf{R} : \Im_R = (\mathbf{A}, B) \wedge \mu_{\mathbf{A}}(\mathbf{x}) > \alpha \tag{2.99}$$

Completeness is a property of deductive systems that has been used in the context of Artificial Intelligence to indicate that the knowledge representation scheme can represent every entity within the intended domain. When applied to fuzzy models, this property informally states that the fuzzy model should be able to infer a proper action for every input (Herrera et al., 1998; Lee, 1990). On the other hand, incompleteness is symptom of overfitting (Jin et al., 1999).

As stated in (Stamou and Tzafestas, 1999), a complete fuzzy model can achieve proper operation avoiding undesirable situations. Such undesirable situations can be easily understood by analyzing the behavior of the fuzzy model when no rules are activated by an actual input. Suppose that there exists an input vector that does not activate any rule:

$$\exists \mathbf{x} \in \mathbf{U} \forall R \in \mathbf{R} : \Im_R = (\mathbf{A}, B) \rightarrow \mu_{\mathbf{A}}(\mathbf{x}) = 0 \tag{2.100}$$

If the "Then" part of the rules is interpreted as conjunction, then the fuzzy sets deriving from the interpretation of each rule are all empty:

$$\forall R \in \mathbf{R} \forall y \in U_O : \Im(Rule)[\mathbf{x}, y] = 0 \tag{2.101}$$

As a consequence, the fuzzy set obtained by aggregating rules is the empty set. The inferred fuzzy set is empty, thus excluding any possible numerical outcome from the model. Put differently, no action is decided for the given input. However, many defuzzification procedures always provide a numerical outcome even if the inferred fuzzy set is an empty set. This is necessary when an action is always required from the model (e.g. in control applications). In this case, however, the final outcome is arbitrary and is not a logical consequence of the knowledge base embodied in the model.

If the "Then" part is interpreted as an implication, a completely opposite situation is likely to happen. Since material implication is always verified

when the antecedent is true, the fuzzy set derived for each rule coincides with the output Universe of Discourse U_O, i.e.:

$$\forall R \in \mathbf{R} \forall y \in U_O : \Im\left(Rule\right)\left[\mathbf{x}, y\right] = \Im\left(Ant\right) \odot \mu_{\mathcal{L}_y^{-1}(FL_y)}\left(y\right) = 0 \odot \mu_{\mathcal{L}_y^{-1}(FL_y)}\left(y\right) = 1 \quad (2.102)$$

The conjunctive aggregation of all rules coincides to the output Universe of Discourse, hence the inferred fuzzy set gives full possibility to any element of the Universe of Discourse and the defuzzification procedure can legitimately choose any random element from this Universe. As a consequence, the behavior of the fuzzy model is undefined in input regions where no rules are fired. Undesirable situations may occur even if some rule is fired by the input, but the activation strength is too low. This phenomenon is also called "weak firing". For such reason, stronger α-completeness is preferred over simple completeness to guarantee a proper behavior of the fuzzy model (Pedrycz, 1993).

Between the two forms of completeness, namely completeness of the Frames of Cognition (coverage) and completeness of the model, the former is more often required in interpretable fuzzy modeling even though model incompleteness can lead to undesirable situations as those previously illustrated. Clearly model completeness implies completeness of Frames of Cognition, but the contrary is not true unless the knowledge base is composed of rules defined by all possible combinations of antecedent fuzzy sets. However, all combinations of fuzzy sets lead to a combinatorial explosion of the number of rules, making the model impractical.

In addition, often the input space has nonuniform possibility distribution but the multidimensional Universe of Discourse is defined as Cartesian product of one-dimensional fuzzy sets for interpretability pursuits. As a consequence, the multidimensional Universe of Discourse may be only a superset of the plausible domain in which input data effectively belong, while inputs not belonging to such domain can be considered as outliers. In such situation, model incompleteness can be tolerated since undesirable behavior may result only in regions of the Universe of Discourse where inputs are unlikely to occur.

If outliers have to be properly handled by a fuzzy model, model completeness should be restored but, in order to avoid combinatorial rule explosion, two approaches may be used:

- Infinite support fuzzy sets can be adopted in the design of the Frames of Cognition. In such case, all inputs of the Universe of Discourse activate all rules. This approach is simple but may decrease the inference efficiency and can originate the weak-firing phenomenon;

- A "default rule" can be introduced (Peña-Reyes and Sipper, 2003; Vuorimaa, 1994). A default rule is a special rule that is activated only when no rules of the remaining knowledge base are fired. The default rule may be associated with a special output signaling 'out of range operation' or an indecision status. Default rules may be crisp or fuzzy. In the latter case, the activation strength is defined as the complement of the maximum activation strength among all ordinary rules. Default fuzzy rules can be thus used to fight the weak-firing phenomenon but its semantics must be carefully defined due to its special role in the fuzzy model.

Number of variables

CONSTRAINT 33 *The number of input and output variables should be as small as possible*

The number of input/output variables are strictly related to the dimensionality of the input/output Universes of Discourse. A high number of input/output variables is negative in fuzzy modelling for several reasons, such as:

1. The length of each rule becomes to high, thus hampering interpretability. Unless sophisticated techniques are adopted to reduce the number of variables in each rule (see 2.5), a high number of variables – typically more than seven – reduces the readability of the knowledge base;

2. Fuzzy models are usually acquired from data by means of identification techniques. Like all data-identified models, fuzzy models are subject of the so-called "curse of dimensionality" (Bellman, 1961). According to Bellman, the number of parameters needed for approximating a function to a prescribed degree of accuracy increases exponentially with the dimensionality of the input space (Haykin, 1999)[25]. In fuzzy models, such parameters correspond to the number of fuzzy sets and the number of fuzzy rules. The basic reason for the curse of dimensionality is that a function defined in high dimensional space is likely to be much more complex than a function defined in a lower dimensional space, and those complications are harder to discern (Friedman, 1995).

[25]The complete formulation of the curse of dimensionality also involves a "smoothing factor" that measures the oscillatory nature of the approximating function and is assumed constant as the number of dimensions increases. For the sake of simplicity – motivated by the qualitative level of the discussion – such smoothing factor has not been considered here.

According to the above-mentioned motivations, the number of input variables and that of output variables have a different effect on the model degradation. Specifically, while a high number of output variables hampers the interpretability of the model by lengthening the rules, a high number of input variables damages interpretability and causes the curse of dimensionality. For such reason, input and output variables are treated differently in the attempt of reducing their number.

Input variables The reduction of the number of input variables can improve the readability of the knowledge base and avoids the curse of dimensionality. Clearly, a drastic reduction of the number of input variables reduces the number of free parameters in the model and, consequently, the ability of identifying a large set of functions. For such reason, any technique to reduce the number of input variables must be balanced with the desired accuracy of the final model. In literature, there exists an impressive number of techniques to reduce the number of input variables, which can be classified in two main categories:

Feature extraction The multidimensional Universe of Discourse $\mathbf{U}$ is transformed into another multidimensional Universe of Discourse $\mathbf{U}'$ with a dimensionality lower than $\mathbf{U}$. Such transformation is usually achieved by exploiting the distribution of an available dataset and can be linear on non-linear. There is a wide number of feature extraction techniques and a review of them can be found in (Jain et al., 2000). While effective for black-box models, feature extraction techniques are dangerous in interpretable fuzzy models because the new set of variables coming from the transformed Universe of Discourse $\mathbf{U}'$ cannot be associated to any physical meaning. For such reason, feature extraction techniques should be avoided in interpretable fuzzy modelling.

Feature selection Feature selection techniques aim to extract a small subset of input variables without degrading too much the accuracy of the fuzzy models. Differently to feature extraction techniques, feature selection does not destroy interpretability because the selected input variables are not new but a subset of the original variables. As a consequence, the physical meaning associated to each variable is retained. Feature selection is less effective than feature extraction, but it is necessary when interpretability is the major concern of a fuzzy model. A wide number of feature selection techniques exists in literature, but those that are especially suited for interpretable fuzzy modelling define a ranking of features before selecting a subset of them. In feature ranking, an ordered list of input variables is provided, where the higher is

the position of the variable the higher is its influence in determining the input/output relation captured by the fuzzy model (Linkens and Chen, 1999). Such ordered list is very useful for interpretable fuzzy modeling because it offers an interaction tool with the domain experts, who can decide to force some input variables into the model – perhaps because their physical meaning is essential – or to discard some input variables even though their relative position in the list is high. Such decision are better supported if the ordered list is associated with a correlation analysis, which can ensure that the elimination of some variables does not hamper the overall accuracy of the model, when highly correlated variables still remain in the ranked list.

Output variables Throughout this Chapter the number of variables of the model has been assumed to be only one, but in many applicative problems the number of variables can be very high. Two approaches can be used to deal with a high number of output variables: building a single Multi-Input Multi-Output (MIMO) model or defining several Multi-Input Single-Output (MISO) models.

MIMO Models A MIMO fuzzy model is defined by rules where the consequent is a conjunction of two or more fuzzy constraints. Usually, the inference procedure of a MIMO model operates by inferring the output value of each variable independently. This is possible because of the conjunctive relation of the output fuzzy constraints[26]. When the number of output variables is too high, rules of MIMO models become illegible and must be replaced by models with simpler knowledge representation;

MISO Models A MISO model is defined by rules with a single consequent. When multiple outputs are required, several MISO models are defined, one for each output variables. The main advantage of MISO models is legibility of the rule base, which is simpler than MIMO models. Moreover, since MISO models are identified independently, the resulting knowledge bases are better suited to represent the input/output relationships for each output variable. A consequent benefit is the greater accuracy of MISO models w.r.t. a single MIMO model (this accuracy improvement is also justified by the increased number of free

[26]Some authors generalize the definition of rules by including also disjunctive consequents (Pedrycz and Gomide, 1998). In such case, however, inference cannot be accomplished independently for each output variable. For such reason, disjunctive consequents are very rare in fuzzy modeling.

parameters that characterize a family of MISO models with respect to a single MIMO model). The main drawback of MISO models is that the Frames of Cognition of the same input variable are identified independently for each model. As a result, Frames of Cognition for the same linguistic variable may have the same linguistic labels but different semantics. This situation of ambiguity clearly impedes a clear understanding of the acquired knowledge, hence MISO models should be avoided if interpretability is of major concern.

Granulated output

CONSTRAINT 34 *The output of the fuzzy model should be an information granule rather than a single numerical value*

In most cases, fuzzy models are used to identify an unknown input/output mapping defined on numerical domains. Even though information is processed within the model by means of fuzzy sets manipulation, often the required outcome is a single numerical value that must be used in the environment for decision making, classification, control, etc. In other cases, the outcome of the fuzzy model is just an information for further processing (e.g. in medical diagnosis, the outcome of the fuzzy model can be used by physicians to actuate a proper decision). In situations like this latter, it is favorable that the output of the fuzzy model would be an information granule rather than a single numerical value. Indeed, granular information is richer than numerical information because the former also conveys uncertainty or vagueness information that are valuable in a decision making process.

A first approach to provide granulated information is to avoid the defuzzification procedure and just return the resulting fuzzy set attained by aggregating the outcomes inferred by each rule. Such fuzzy set provides vagueness information about the model output but can be hard to interpret because it generally does not satisfy none of the interpretability constraints for fuzzy sets. An alternative approach consists in enriching the numerical outcome provided by the defuzzificator with uncertainty information about the behavior simulated by the fuzzy model. This new information can be well represented by means of prediction intervals. A successive Chapter is devoted at describing a proposed method for integrating prediction intervals within fuzzy models.

2.7 Constraints on Fuzzy Model Adaption

Fuzzy information granules are often used as building blocks for designing rule-based fuzzy models, which can be effectively used to solve soft-computing problems such as predictions, classifications, system identification, etc. From a functional point of view, fuzzy models are approximative representations of an underlying functional input/output relationship that has to be discovered. To achieve such objective, often a dataset is available, which provides a finite collection of input/output examples from which the model's knowledge has to be built. Information granulation techniques operate on such dataset to properly extract information granules that will be used to define the knowledge base. In such context, information granulation can be accomplished in two main ways:

1. The complete dataset of examples is granulated, involving both input and output information. At the end of the granulation process, the knowledge base of the fuzzy model is complete and can be immediately used for approximating the underlying functional relationship. A typical example of fuzzy model defined by this kind of knowledge base is the Mamdani Fuzzy Inference System;

2. The examples are split in the input and output components, and the granulation process is executed only in the input part. This choice is typical in classification problems, where outputs are class labels that belong to a symbolic domain that cannot be granulated. This form of granulation is also chosen when accuracy has a great impact in the assessment of the model quality.

When the second form of granulation is chosen, the resulting information granules are able to describe the relationships among input data, but they miss to correctly represent the underlying input/output relationship. Therefore the knowledge base of the fuzzy model has to be *refined* (or adapted) in order to better approximate the unknown functional relationship by means of supervised learning schemes[27].

In an extreme synthesis, the adaption of a fuzzy model can be viewed as a transformation

$$\mathcal{L}(\mathcal{M}, D) = \mathcal{M}' \tag{2.103}$$

[27]It should be noted that information granulation on the input space can be considered as an unsupervised learning scheme. The fuzzy model resulting from both unsupervised and supervised learning is then a hybrid model.

that maps an original model $\mathcal{M}$ into a new model $\mathcal{M}'$ with the support of a finite set of examples D. The transformation is aimed at optimizing an objective function $\wp$ such that:

$$\wp(\mathcal{M}') \geq \wp(\mathcal{M}) \tag{2.104}$$

The objective function may be a measure of accuracy, generalization ability, interpretability, etc. or a combination of these[28]. Even though it is not the general case, it is assumed hereafter that the adaption process does not modify the structure of the Frames of Cognition involved in the model, i.e. their number, their cardinality and labelling. What is usually changed by the learning process are the free parameters that characterize input and output fuzzy sets. In some cases, the information granules of the rules are changed by removing useless soft-constraints. The rule structure is admitted to changing, e.g. by removing useless rules or adding new ones. With a slight abuse of notation, hereafter each element X belonging to the model (especially fuzzy sets) that are subject to adaptation by means of $\mathcal{L}$ will be indicated with $\mathcal{L}X$.

There exist several methods to adapt a model to data, but most of them fall in two main paradigms:

Genetic Algorithms The fuzzy model is adapted by evolving its structure and/or its free parameters. Three classical approaches may be adopted to adapt fuzzy models to data:

Michigan Approach (Michalewicz, 1996) Each individual represents a single rule. The knowledge base is represented by the entire population. Since several rules participate in the inference process, the rules are in constant competition for the best action to be proposed, and cooperate to form an efficient fuzzy model. Such cooperative/competitive necessitates an effective credit-assignment policy to ascribe fitness values to individual rules.

Pittsburgh Approach (Michalewicz, 1996) Each individual is a candidate knowledge base. Genetic operations provide new populations of knowledge bases that can be evaluated for their accuracy but also for additional properties such as transparency. The main drawback of this approach is the computational cost.

Iterative Rule Learning Approach (Herrera et al., 1995) This approach is used to both identify and refine fuzzy models by first

[28] A complete and self-contained functional model for supervised learning can be found in (Vapnik, 1999).

providing a model with a single rule, then new rules are added until an appropriate knowledge base is defined.

In recent years, a new promising approach for genetic optimization has been proposed, which appears particularly suitable to find fuzzy models balanced in accuracy and interpretability. These "multi-objective" genetic algorithms allow for searching multiple solutions in parallel with the aim of optimizing several fitness functions simultaneously (Fonseca and Fleming, 1995). Multi-objective genetic algorithms have been successfully used to design and refine fuzzy models from data by simultaneously optimizing conflicting fitness functions such as accuracy and interpretability (Jiménez et al., 2001; Gómez-Skarmeta et al., 1998; Ishibuchi et al., 1997).

Neuro-Fuzzy Networks The fuzzy model is converted into a specific neural network, which is trained according to some neural learning scheme. After learning, the neural network is re-converted back into a fuzzy model. Two approaches for neuro-fuzzy modeling are very common (Nauck and Kruse, 1994):

Cooperative approach The neural network and the fuzzy model are distinct objects that cooperate to form an accurate final model. For example, the neural network can be used to define a proper shape of fuzzy sets, or a proper weight of rules, while the fuzzy model is responsible in making the input/output mapping;

Hybrid approach There exist a bijective mapping from nodes of the neural network and the components of the fuzzy model, so that the two models can be viewed as two facets of the same system, called "neuro-fuzzy network". This kind of architecture is called "translational" in the Knowledge-Based Neurocomputing paradigm (Cloete and Zurada, 2000), because the fuzzy model can be converted into the neural network and vice-versa in every stage of processing. One of the first neuro-fuzzy networks appeared in literature is ANFIS (Adaptive Network-based Fuzzy Inference System) (Jang, 1993). Actually, the hybrid approach has overcome the cooperative approach because of its versatility and efficacy.

In interpretable fuzzy modelling, the adaption process must not destroy the interpretability constraints, i.e. the adapted model $\mathcal{M}'$ must satisfy at least the same interpretability constraints of the original model $\mathcal{M}$. More constraints can be added if the objective of adaption is interpretability improvement. If – as often occurs – the adaption process is iterative, the

intermediate models emerging after each iteration may not satisfy the interpretability constraints. This is usually the case when adaption is carried out by means of regularized learning (Bishop, 1995) or when interpretability constraints are applied after several adaption stages for efficiency pursuits (Altug et al., 1999). However, when the adaption process can be halted at every stage (e.g. to allow user interaction or in real-time applicative contexts) it is more desirable that all intermediate models satisfy interpretability constraints (Nauck et al., 1997).

In addition to the previously discussed interpretability constraints, the adaption process must be designed so as to not twisting the semantics of the fuzzy sets involved in the knowledge base. Indeed, the adaption of the model has the role of refine and not completely modify the meaning conveyed by the knowledge base. For such reason, additional constraints could be required on the adaption process for interpretable fuzzy modelling.

No position exchange

CONSTRAINT 35 *Two fuzzy sets belonging to the same Frame of Cognition must not exchange position after learning:*

$$\forall \mathbf{F} \in \mathbb{F}_I \cup \mathbb{F}_O \; \forall A, B \in \mathbf{F} : A \prec B \rightarrow \mathcal{L}A \prec \mathcal{L}B \tag{2.105}$$

This criterion, elicited by Nauck and Kruse, assures that the learning process does not switch the meaning of two linguistic labels associated to fuzzy sets belonging to a Frame of Cognition (Nauck and Kruse, 1998). This is especially important when linguistic labels express qualities on the Universe of Discourse, like LOW, MEDIUM and HIGH. It would be indeed undesirable that, after learning, fuzzy sets associated to these labels exchange their position such that the final order does not reflect the metaphors of the labels, e.g. when LOW $\prec$ HIGH $\prec$ MEDIUM. Note that a relabelling operation is possible so as to restore the semiotic ordering of linguistic labels, but this operation is against the aims of model adaptation, according to which the model should be only finely tuned to better meet the objective function. In such a viewpoint, any relabelling operation would be seen as a model re-design.

When labels express vague quantities (like ABOUT_2) ore precise quantities (as in the consequents of Takagi-Sugeno rules), position exchange is admissible provided that the label changes in agreement to the fuzzy sets. In such case, re-labelling is an admissible operation because the fuzzy sets express quantities that are susceptible to changes during the tuning process.

No sign change

CONSTRAINT 36 *Fuzzy sets with cores in positive (resp. negative) range must not change sign after learning:*

$$\forall \mathbf{F} \in \mathbb{F}_I \cup \mathbb{F}_O \; \forall A \in \mathbf{F} : \min \operatorname{core} A \geq 0 \rightarrow \min \operatorname{core} \mathcal{L}A \geq 0 \quad (2.106)$$

$$\forall \mathbf{F} \in \mathbb{F}_I \cup \mathbb{F}_O \; \forall A \in \mathbf{F} : \max \operatorname{core} A \leq 0 \rightarrow \max \operatorname{core} \mathcal{L}A \leq 0 \quad (2.107)$$

This criterion is mostly related to fuzzy sets used to express vague quantities (Nauck et al., 1997; Nauck and Kruse, 1998). Informally speaking, the criterion states that if a fuzzy set represents a signed quantity (either positive or negative), then the corresponding fuzzy set resulting from adaption must represent a quantity with the same sign. In this way, the learning process $\mathcal{L}$ does not drastically change the semantics of the knowledge base during fine-tuning. Note that if a fuzzy set is unimodal and satisfies the "Natural Zero Positioning" constraint, i.e. $\operatorname{core} A = \{0\}$, then the "No Sign Change" constraint requires that, after adaption, the refined fuzzy set still represents the number 0, since $\operatorname{core} \mathcal{L}A = \{0\}$. As Natural Zero Positioning can be extended to "Natural S Positioning" (see 2.3), also the following constraints can be formulated:

$$\forall s \in S \; \forall \mathbf{F} \in \mathbb{F}_I \cup \mathbb{F}_O \; \forall A \in \mathbf{F} : \min \operatorname{core} A \geq s \rightarrow \min \operatorname{core} \mathcal{L}A \geq s \quad (2.108)$$

and

$$\forall s \in S \; \forall \mathbf{F} \in \mathbb{F}_I \cup \mathbb{F}_O \; \forall A \in \mathbf{F} : \max \operatorname{core} A \leq s \rightarrow \max \operatorname{core} \mathcal{L}A \leq s \quad (2.109)$$

Similarity preservation

CONSTRAINT 37 *Any adapted fuzzy set should be similar to the original one:*

$$\forall \mathbf{F} \in \mathbb{F}_I \cup \mathbb{F}_O \; \forall A \in \mathbf{F} : S(A, \mathcal{L}A) \approx 1 \quad (2.110)$$

where S is a similarity measure.

This constraint informally states that fuzzy sets subject of adaption must not change too much in order to preserve their original semantics. The formalization of the criterion is a generalization of the work of Lotfi, who uses a rough-fuzzy representation of Gaussian membership functions to restrict the variability of parameters (namely, center and width) into a predefined range (Lotfi et al., 1996). This constraint captures the very nature of fine

tuning, i.e. adaption of the model without excessive modification of the internal knowledge. On the other hand, it should be observed that if the adaption process leads to a strong modification of the model, than a bad initial design of the model could be hypothesized. In such a case, a complete redesign of the model should be considered.

2.8 Final remarks

Thoroughly this Chapter, a number of interpretability constraints have been surveyed. Some of these constraints are mandatory for interpretable fuzzy information granulation, while others can be included or dropped from a design strategy according to subjective judgements. Furthermore, the application of some interpretability constraints depends on the final representation of the information granules, and, in general, of the knowledge base. As an example, some interpretability constraints make sense only for information granules representing qualities on the Universe of Discourse, while other constraints are especially suited for information granules representing fuzzy quantities.

Additional interpretability constraints can be added for specific applicative needs. As an example, in (Johansen et al., 2000) a number of special-purpose interpretability constraints have been proposed for system identification applications. Because of limited interest, special-purpose interpretability constraints have not been considered in this Chapter, rather preferring general purpose constraints because of their wider applicability. In summary, on the engineering standpoint, a number of issues should be taken into account in the selection of interpretability constraints, including:

- Type of representation of information granules (e.g. qualitative vs. quantitative information granules);
- Type of fuzzy model (e.g. Mamdani vs. Takagi-Sugeno models);
- Concerns of modelling: accuracy, efficiency, interpretability;
- Epistemological assumptions (e.g. implicative vs. conjunctive rules);
- Applicative context.

Scientific investigation on interpretability constraints is still open. Further inquiries in the field may include the analysis of interpretability on different types (e.g. discrete) of Universes of Discourse. Also, the integration of feature extraction techniques within interpretable fuzzy information

granulation is of great interest (e.g. to form complex linguistic constraints such as "PRODUCT OF VOLUME AND PRESSURE IS LOW").

Within Cognitive Science, interpretability constraints may provide a valid help in understanding the mechanisms of human behavioral phenomena like conceptual model building and explanation. In this context, the integration of results coming from different disciplines could be highly benefic. As an emblematic example, the psychological experiments of Miller in (Miller, 1956) provide a better understanding of human abilities as well as important guidelines for interpretable fuzzy information granulation and modelling. The example sheds light on the inter-disciplinary character of interpretable fuzzy information granulation, which results a promising theoretical framework for bringing machines nearer to humans with the aid of perception-based computing.

Part II

Theoretical Contributions to Interpretabilty in Fuzzy Information Granulation

Chapter 3

Distinguishability Quantification

3.1 Introduction

Distinguishability of fuzzy sets is one of the most common interpretability constraints adopted in literature. In brief, distinguishability can be viewed as a relation between fuzzy sets defined on the same Universe of Discourse. Distinguishable fuzzy sets are well disjunct so as to represent distinct concepts and can be associated to metaphorically different linguistic labels. Well distinguishable fuzzy sets provide the following advantages within the Theory of Fuzzy Information Granulation:

- Obviate the subjective establishment of membership-function/linguistic term association (Valente de Oliveira, 1998);
- Avoids potential inconsistencies in fuzzy models (Valente de Oliveira, 1999b);
- Reduce model's redundancy and consequently computational complexity (Setnes et al., 1998a);
- Linguistic interpretation of the fuzzy information granules is easier (Setnes et al., 1998a);

Completely disjunct fuzzy sets are maximally distinguishable. However, disjunct fuzzy sets may violate some other interpretability constraints (e.g. coverage), hence partially overlapping fuzzy sets are more preferable for the definition of interpretable fuzzy information granules.

Distinguishability quantification can be realized in different ways. The most adopted characterization is by means of *similarity* measures. In (Setnes et al., 1998a) similarity measures are deeply discussed in the context of fuzzy modelling. There, similarity is interpreted as a fuzzy relation defined over fuzzy sets and corresponds to the "*degree to which two fuzzy sets are equal*". Such interpretation is then formally characterized by a set of axioms.

Similarity measures well capture all the requirements for distinguishable fuzzy sets, but their calculation is usually computationally intensive. As a consequence, most strategies of model building that adopt similarity measures for interpretability enhancement are based on massive search algorithms such as Genetic Algorithms (Roubos and Setnes, 2001; Meesad and Yen, 2002; Setnes et al., 1998b), Evolution Strategies (Jin, 2000), Symbiotic Evolution (Jamei et al., 2001), Coevolution (Peña-Reyes and Sipper, 2003), or Multi-Objective Genetic Optimization (Jiménez et al., 2001). Alternatively, distinguishability improvement is realized in a separate design stage, often after some data driven procedure like clustering, in which similar fuzzy sets are usually merged together (Paiva and Dourado, 2001; Setnes et al., 1998a).

When the distinguishability constraint has to be included in less time consuming learning paradigms, like neural learning, similarity measure is seldom used. In (Paiva and Dourado, 2001), after a merging stage that uses similarity, fine tuning is achieved by simply imposing some heuristic constraints on centers and width of membership functions. In (Jin and Sendhoff, 2003), a distance function between fuzzy sets (restricted to the Gaussian shape) is used in the regularization part of the RBF[1] cost function. In (Guillaume and Charnomordic, 2004) a sophisticated distance function is used to merge fuzzy sets. In (Nauck and Kruse, 1998; Espinosa and Vandewalle, 2000; Castellano et al., 2002) the possibility measure is adopted to evaluate distinguishability.

The possibility measure has been intensively studied within Fuzzy Set Theory and has some attracting features that promote a deeper investigation in the context of distinguishability assessment. Although it is not a similarity measure, it has a clear and well-established semantics since it can be interpreted as the degree to which a flexible constraint "X is A" is satisfied (Zadeh, 1978; Toth, 1999). In addition, the possibility quantification of two fuzzy sets can be often analytically expressed in terms of fuzzy sets' parameters. This makes possibility evaluation very efficient so that it can be effortlessly embodied in computationally inexpensive learning schemes.

The objective pursued in the Chapter is to derive some significant relationships between similarity and possibility measures. Specifically, some sufficient conditions are demonstrated to positively correlate possibility and

[1] Radial Basis Function neural network

similarity *in the worst case*, i.e. the lower is the possibility between two fuzzy sets, the lower is the upper-bound of similarity that can be measured between the same fuzzy sets. As a final result, some theorems formally prove that, under some mild conditions, any transformation aimed to reduce possibility between fuzzy sets actually reduces also their similarity measure and, consequently, improves their distinguishability. In light of such theoretical results, the possibility measure emerges as a good candidate for interpretability analysis as well as for efficient interpretable fuzzy modelling.

3.2 Distinguishability measures

In this Section the most common measures to quantify distinguishability are formalized and briefly described. A deep description is provided for the similarity and the possibility measures, since they are subject of further investigation in successive sections. Other measures proposed in literature are also surveyed, and some theoretical results for them are established.

For the sake of clarity, the adopted formal notation is repeated here. In this context any fuzzy set is denoted with a capital letter (A, B, etc.) and the corresponding membership function with μ_A, μ_B, etc. Each membership function is defined on the same Universe of Discourse U, which is assumed to be a one-dimensional closed interval $[m_U, M_U] \subset \mathbb{R}$. The set of all possible fuzzy (sub-)sets defined over U is denoted with $\mathcal{F}(U)$, while the finite family of fuzzy sets actually involved in a fuzzy model is called "Frame of cognition" and it is denoted with $\mathbf{F}$.

3.2.1 Similarity measure

According to (Setnes et al., 1998a), the similarity measure between two fuzzy sets A and B is a fuzzy relation that expresses the degree to which A and B are equal. Put formally, similarity is a function:

$$S : \mathcal{F}(U) \times \mathcal{F}(U) \rightarrow [0,1] \tag{3.1}$$

such that:

1. Non-overlapping fuzzy sets are totally non-similar and vice versa[2]:

$$S(A,B) = 0 \iff \forall x \in U : \mu_A(x)\,\mu_B(x) = 0 \tag{3.2}$$

[2] In (Setnes et al., 1998a) another definitory property states that overlapping fuzzy sets must have similarity greater than zero. However, this is a direct consequence of property 1 and definition of S.

The latter definition, given in (Setnes et al., 1998a), may not be satisfactory for fuzzy sets that overlap in a finite (or countable) number of points. In such case, similarity can be still zero even if the fuzzy sets do overlap (see, e.g. example 3.1). In a more precise formalization the left-to-right implication would be truncated.

2. Only equal fuzzy sets have maximum similarity:

$$S(A,B) = 1 \iff \forall x \in U : \mu_A(x) = \mu_B(x) \tag{3.3}$$

The same objection for property 1 could be stated for fuzzy sets that are equal everywhere except in a finite (or countable) number of points. Even in this case, the left-to-right implication would be truncated.

3. The similarity measure is invariant under linear scaling of the fuzzy sets, provided that the scaling is the same:

$$S(A,B) = S(A',B') \tag{3.4}$$

where:

$$\exists k \neq 0 \exists l \in \mathbb{R} : \begin{array}{l} \mu_A(x) = \mu_{A'}(kx+l) \\ \mu_B(x) = \mu_{B'}(kx+l) \end{array} \tag{3.5}$$

Property 1 assures that mutually exclusive fuzzy sets are not similar at all, while property 2 assures that equal fuzzy sets are totally similar. Property 3 establishes that similarity between fuzzy sets does not change if the Universe of Discourse changes according to a linear transformation of its values (e.g. if two temperatures are similar when expressed Celsius degrees, they must remain similar also when expressed in Fahrenheit degrees).

For sake of completeness, the following symmetry property is added, which makes similarity more adherent to the informal definition of "degree of equality":

4. Similarity is a symmetrical function:

$$S(A,B) = S(B,A) \tag{3.6}$$

Several similarity measures have been proposed in literature, and some of them can be found in (Cross, 1993; Setnes, 1995). However, in interpretability analysis, the most commonly adopted similarity measure is the following:

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.7}$$

where intersection $\cap$ and union $\cup$ are defined by a proper couple of τ-norm and τ-conorm and $|\cdot|$ is the cardinality of the resulting fuzzy set, usually defined as:

$$|A| = \int_U \mu_A(x)\, dx \tag{3.8}$$

Often, minimum and maximum are used as τ-norm and τ-conorm respectively, hence definition (3.7) can be rewritten as:

$$S(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.9}$$

The definition (3.9) of similarity will be considered hereafter. Distinguishability in a Frame of Cognition is guaranteed by imposing that similarity between any two distinct fuzzy sets in the frame must not exceed a user-given threshold σ:

$$\forall A, B \in \mathbf{F} : A \neq B \rightarrow S(A,B) \leq \sigma \tag{3.10}$$

The evaluation of the similarity measure, either in the general form (3.7) or in the specific form (3.9), can be done analytically only for particular classes of fuzzy sets (e.g. triangular fuzzy sets), but may become computationally intensive for other classes of fuzzy sets (e.g. Gaussian) because of the integration operation (3.8), which is necessary for determining the cardinality of the involved fuzzy sets. For such reason, similarity is not used in some learning schemes, where more efficient measures are instead adopted.

3.2.2 Possibility measure

The possibility measure between two fuzzy sets A and B is defined as the degree of applicability of the soft constraint "A is B". Possibility is evaluated according to the following definition:

$$\Pi(A,B) = \sup_{x \in U} \min\left\{\mu_A(x), \mu_B(x)\right\} \tag{3.11}$$

An useful interpretation of possibility measure is the extent to which A and B overlap (Pedrycz and Gomide, 1998). As for similarity measure, an

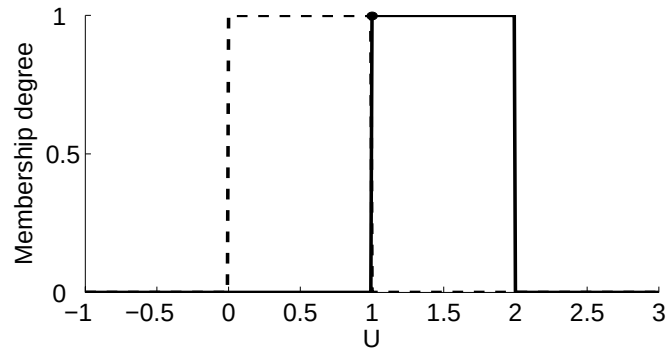

Figure 3.1: Example of fuzzy sets with full possibility but zero similarity

overlapping threshold ϑ can be imposed between any two fuzzy sets in a Frame of Cognition:

$$\forall A, B \in \mathbf{F} : A \neq B \rightarrow \Pi(A, B) \leq \vartheta \tag{3.12}$$

Some important considerations are noteworthy for possibility measure, especially in comparison with the definition of similarity measure. First of all, the possibility measure is not a similarity measure because properties 2 and 3 for similarity measures are not verified. Moreover, in general there is not any monotonic correlation between possibility and similarity as showed in the following two examples.

EXAMPLE 3.1 *Let A and B be two crisp sets defined as closed intervals $[0, 1]$ and $[1, 2]$ respectively (see fig. 3.1). Suppose that intersection and union operations are defined by the minimum and maximum operators respectively. Then:*

$$S(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{1\}|}{|[0, 2]|} = \frac{0}{2} = 0 \tag{3.13}$$

On the other hand, the possibility measure has a very different value:

$$\Pi(A, B) = \sup_{x \in [0,2]} \min\{\mu_A(x), \mu_B(x)\} = \sup_{x \in [0,2]} \begin{cases} 0, & x \neq 1 \\ 1, & x = 1 \end{cases} = 1 \tag{3.14}$$

This result is coherent with the definition of the two measures. Indeed, the two intervals are quasi-disjunct (i.e. disjunct everywhere except in a single point), hence their similarity should be very low (zero using the standard definition (3.9)). On the other hand there is the full possibility that an element of the Universe of Discourse belongs both to A and B, hence the possibility measure is maximal.

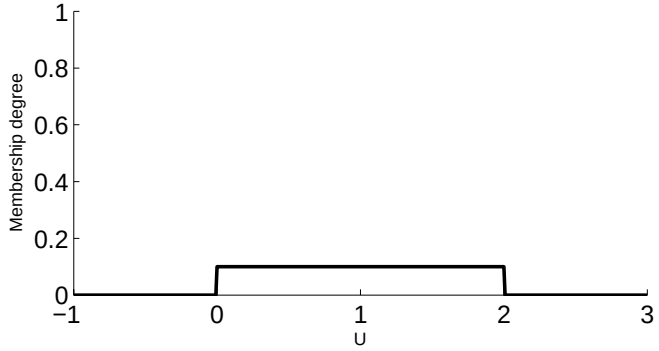

Figure 3.2: Example of fuzzy set with low possibility

EXAMPLE 3.2 *Let A be a fuzzy set defined over the entire Universe of Discourse U with constant membership function $0 < \varepsilon \ll 1$ (see fig. 3.2). Then, by definition:*

$$S(A, A) = 1 \tag{3.15}$$

but:

$$\Pi(A, A) = \sup_{x \in U} \mu_A(x) = \varepsilon \tag{3.16}$$

Here, again, the similarity is maximal as a correct formalization of "degree of equality". On the other hand, the possibility is very low since it is very unlikely (or undesirable, etc.) that an element belongs to the fuzzy set A.

The previous examples show that the possibility measure cannot be used to evaluate similarity between two fuzzy sets. Nevertheless, the possibility measure has features that are important in distinguishability analysis:

1. The possibility threshold ϑ has a clear semantics as it can be interpreted in the context of Possibility Theory (Dubois and Prade, 1988). On the other hand, the similarity threshold σ has a more arbitrary nature (it also depends on the specific definition of intersection and union operators);

2. Numerical integration is not necessary when calculating possibility, in contraposition with similarity calculation. Moreover, although the general definition (3.11) may require a numerical sampling of the Universe of Discourse, the possibility measure can be evaluated analytically for several classes of membership functions (e.g. triangular, Gaussian, bell-shaped, etc., see table 3.1 for an example). This feature enables the adoption of possibility measure in efficient learning schemes;

Table 3.1: Possibility and similarity measures for two common classes of fuzzy sets (special cases not considered).

Shape	**Possibility / Similarity**
Triangular $\mu_i(x) = \max\left\{0, \min\left\{\frac{x-a_i}{b_i-a_i}, \frac{x-c_i}{b_i-c_i}\right\}\right\}$	$\Pi = (c_1 - a_2) / (b_2 - b_1 - a_2 + c_1)$ $S = -(c_1 - a_2)^2 / (\alpha + \beta + \gamma)$ where $\alpha = a_1 (b_2 - b_1 - a_2 + c_1)$ $\beta = -a_2 (b_1 - b_2 - c_1 - c_2)$ $\gamma = b_1 c_2 - b_2 c_1 - b_2 c_2 - c_1 c_2$
Gaussian $\mu_i(x) = \exp\left(-\frac{(x-\omega_i)^2}{2\sigma_i^2}\right)$	$\Pi = \exp\left(-\frac{\omega_1^2 - 2\omega_1\omega_2 + \omega_2^2}{2\sigma_1^2 - 4\sigma_1\sigma_2 + 2\sigma_2^2}\right)$ S not analytically definable

The two aforementioned arguments promote a deeper investigation on possible relationships occurring between similarity and possibility, especially in the context of interpretability analysis.

3.2.3 Other distinguishability measures

In (Chow et al., 1999) a different characterization of distinguishability is used and is called "overlap". Such measure is defined as:

$$D(A, B) = \frac{\operatorname{diam}([A]_\alpha \cap [B]_\alpha)}{\operatorname{diam}([A]_\alpha)} \tag{3.17}$$

being $[A]_\alpha$ and $[B]_\alpha$ the α-cuts[3] of A and B respectively, and "diam" evaluates the extension of the crisp set in argument:

$$\operatorname{diam} X = \max X - \min X \tag{3.18}$$

According to the authors, the overlap constraint is satisfied when

$$L \leq D(A, B) \leq U \tag{3.19}$$

being L and U two user-defined thresholds and A, B two adjacent fuzzy sets in the Frame of Cognition[4]. Such special measure of overlap is not commonly used in literature (especially in interpretability analysis), conceivably for the following drawbacks:

[3]The α-cut of a fuzzy set A is a crisp set defined as $[A]_\alpha = \{x \in U : \mu_A(x) \geq \alpha\}$

[4]In this context, two fuzzy sets in a Frame of Cognition are said adjacent if there is not any other fuzzy set in the frame whose prototype (assumed only one) lies between the prototypes of A and B. Put formally, the two fuzzy sets are adjacent if one is successive to the other according to the proper ordering $\preceq$ defined within the Frame.

1. The measure depends on three parameters (α, L and U). Moreover, the sensibility of the measure w.r.t. such parameters is very high, as admitted by the same authors;

2. The measure is asymmetrical ($D(A,B) \neq D(B,A)$), due to the denominator of (3.17). This property is counter-intuitive, as the distinguishability appears as a symmetrical relation between two fuzzy sets.

In (Valente de Oliveira, 1999a), another measure is adopted to quantify distinguishability. It is not conceived as a measure between fuzzy sets, but a pointwise property that must hold everywhere in the Universe of Discourse. Specifically, the following condition must be verified:

$$\forall x \in U : \sqrt[p]{\sum_{X \in \mathbf{F}} \mu_X(x)^p} \leq 1 \tag{3.20}$$

Such condition assures that any element of U will not have simultaneously high membership grades in different fuzzy sets of the same Frame of Cognition. The parameter p regulates the strength of the condition, which vanishes for $p \to +\infty$. This kind of constraint has been used in regularizing a RBF network together with other interpretability constraints. This measure is strictly related to the possibility measure, as proved here:

PROPOSITION 3.1 *Let $\mathbf{F} = \langle U, \mathbf{F}, \preceq, \mathcal{L}, v \rangle$ be a Frame of Cognition. If the constraint (3.20) holds, then:*

$$\forall A, B \in \mathbf{F} : A \neq B \rightarrow \Pi(A,B) \leq 2^{-\frac{1}{p}} \tag{3.21}$$

Proof. *Consider the* min *function $(x,y) \in [0,1]^2 \mapsto \min(x,y)$ and let $\Phi \subseteq [0,1]^2$. In order to find the supremum of the* min *function within the region Φ, the following line segments can be defined:*

$$u(x) = \left\{(x,y) \in [0,1]^2 : y \geq x\right\} \tag{3.22}$$

and:

$$r(y) = \left\{(x,y) \in [0,1]^2 : x \geq y\right\} \tag{3.23}$$

Let be:

$$\overline{x_1} = \sup\{x \in [0,1] : u(x) \cap \Phi \neq \varnothing\} \tag{3.24}$$

and:

$$\overline{y_2} = \sup\{y \in [0,1] : r(y) \cap \Phi \neq \varnothing\} \tag{3.25}$$

The maximum between $\overline{x_1}$ *and* $\overline{y_2}$ *is the supremum of* min *within* Φ*. Indeed:*

$$\forall (x,y) \in [0,1]^2 : (x,y) \in \Phi \wedge y \geq x \rightarrow x \leq \overline{x_1} \implies \forall (x,y) \in [0,1]^2 : (x,y) \in \Phi \wedge y \geq x : \min(x,y) \leq \overline{x_1} \quad (3.26)$$

and similarly:

$$\forall (x,y) \in [0,1]^2 : (x,y) \in \Phi \wedge x \geq y \rightarrow y \leq \overline{y_2} \implies \forall (x,y) \in [0,1]^2 : (x,y) \in \Phi \wedge x \geq y : \min(x,y) \leq \overline{y_2} \quad (3.27)$$

Since, by definition,

$$\exists y_1 \in [0,1] : (\overline{x_1}, y_1) \in \Phi \quad (3.28)$$

and

$$\exists x_2 \in [0,1] : (x_2, \overline{y_2}) \in \Phi \quad (3.29)$$

then:

$$\forall (x,y) \in \Phi : \min(x,y) \leq \max\{\overline{x_1}, \overline{y_2}\} \quad (3.30)$$

and:

$$\exists (x,y) \in \Phi : \min(x,y) = \max\{\overline{x_1}, \overline{y_2}\} \quad (3.31)$$

As a consequence:

$$\sup_{(x,y)\in\Phi} \min(x,y) = \max\{\overline{x_1}, \overline{y_2}\} \quad (3.32)$$

This general result can be applied to a specific class of sets, defined as follows:

$$\Phi_p = \left\{ (x,y) \in]0,1]^2 : x^p + y^p \leq 1 \right\} \quad (3.33)$$

If the constraint (3.20) is verified by the Frame of Cognition $\mathbf{F}$*, then the following inequality is verified:*

$$\Psi_{A,B} \subseteq \Phi_p \quad (3.34)$$

where:

$$\Psi_{A,B} = \left\{ (\mu_A(u), \mu_B(u)) \in [0,1]^2 : u \in U \right\} \quad (3.35)$$

Figure 3.3 illustrates the relationships between the sets $\Psi_{A,B}$ *and* Φ_p*, as well as the contour of the* min *function. Relation (3.34) is due to the following inequality:*

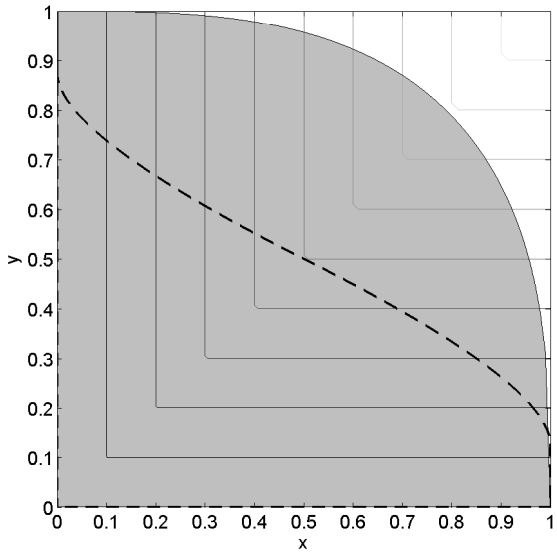

Figure 3.3: The contour of the min function (gray lines), the set Φ_3 (shadowed) and an example of $\Psi_{A,B}$ (dashed line)

$$\mu_A(u)^p + \mu_B(u)^p \leq 1 - \sum_{X \in \mathbf{F} \setminus \{A,B\}} \mu_X(x)^p \leq 1 \tag{3.36}$$

Given two sets A and B in the frame $\mathbf{F}$, the possibility measure between A and B, namely $\Pi(A,B)$, verifies the following equality:

$$\Pi(A,B) = \sup_{(x,y) \in \Psi_{A,B}} \min(x,y) \tag{3.37}$$

Since $\Psi_{A,B} \subseteq \Phi_p$, then:

$$\Pi(A,B) \leq \sup_{(x,y) \in \Phi_p} \min(x,y) \tag{3.38}$$

In order to quantify the right side of (3.38), the values $\overline{x_1}$ and $\overline{y_2}$ should be calculated as in (3.24) and (3.25). It is actually unnecessary to calculate both values, since they coincide due to the symmetry of Φ_p w.r.t. the straight line $y = x$.
It is easy to prove that $\overline{x_1} = \overline{y_2} = 2^{-\frac{1}{p}}$. Indeed, for a given $\overline{x} > 2^{-\frac{1}{p}}$ then

$$\forall (x,y) \in u(\overline{x}) : x^p + y^p > 2\overline{x}^p > 1 \tag{3.39}$$

hence $u(\overline{x}) \cap \Phi_p = \varnothing$. Furthermore, for a given $\overline{x} < 2^{-\frac{1}{p}}$ there exists $\overline{\overline{x}}$ such that $\overline{x} < \overline{\overline{x}} < 2^{-\frac{1}{p}}$ such that $u\left(\overline{\overline{x}}\right) \cap \Phi_p \neq \varnothing$. As a consequence:

$$\sup_{(x,y) \in \Phi_p} \min(x,y) = 2^{-\frac{1}{p}} \tag{3.40}$$

which proves the theorem. ■

Based on the previous proposition, it is possible to affirm that the constraint (3.20) is actually a constraint on the possibility measure, hence its study in relation to distinguishability quantification can be subsumed to the analysis of the possibility measure.

In (Jin and Sendhoff, 2003), the distinguishability is evaluated according to a distance function especially suited for Gaussian membership functions. If A and B are Gaussian fuzzy sets with centers ω_A and ω_B respectively, and widths σ_A and σ_B respectively, then the distance is evaluated according to:

$$d_J(A,B) = \sqrt{(\omega_A - \omega_B)^2 + (\sigma_A - \sigma_B)^2} \tag{3.41}$$

From the distance function d_J, a new similarity measure can be induced according to the set-theoretical approach described in (Zwick et al., 1987). More specifically, the following similarity measure can be defined:

$$S_J(A,B) = \frac{1}{1 + d_J(A,B)} \tag{3.42}$$

Such kind of measure, while inexpensive for its calculation, is strictly limited to Gaussian fuzzy sets. Moreover, it is not a similarity measure since property 3 is not verified. Indeed, the following proposition is proved.

PROPOSITION 3.2 *Given two Gaussian fuzzy sets A, B of centers ω_A, ω_B and widths σ_A, σ_B and a rescaling of factor $k > 0$ and shift l, then*

$$S_J(A',B') \neq S_J(A,B) \tag{3.43}$$

being A', B' the re-scaled fuzzy sets
Proof. *The membership functions of A', B' are*

$$\mu_{A'}(x) = \mu_A(kx + l) = \exp\left(-\frac{(x - \omega'_A)^2}{2\sigma'^2_A}\right) \tag{3.44}$$

and:

$$\mu_{B'}(x) = \mu_B(kx + l) = \exp\left(-\frac{(x - \omega'_B)^2}{2\sigma'^2_B}\right) \tag{3.45}$$

where:

$$\omega_{A'} = \frac{\omega_A - l}{k}, \omega_{B'} = \frac{\omega_B - l}{k} \tag{3.46}$$

and

$$\sigma'_A = \frac{\sigma_A}{k}, \sigma'_B = \frac{\sigma_B}{k} \tag{3.47}$$

As a consequence:

$$d_J\left(A', B'\right) = \frac{d_J\left(A, B\right)}{k} \tag{3.48}$$

and hence

$$S_J\left(A', B'\right) = \frac{k}{k + d_J\left(A, B\right)} \neq S_J\left(A, B\right) \tag{3.49}$$

for $k \neq 1$. ■

This and other issues are faced in (Guillaume and Charnomordic, 2004), where a more sophisticated metric is adopted for fuzzy partitioning with semantic constraints.

3.3 Theorems about Similarity and Possibility measures

On the basis of the considerations discussed in the previous Section, the possibility measure emerges as a potentially good candidate for distinguishability quantification. The main advantage deriving from adopting possibility consists in a more efficient evaluation of distinguishability, which can be used in on-line learning schemes. However, the adoption of possibility for quantifying distinguishability is consistent provided the existence of a monotonic relation between possibility and similarity, i.e. a relation that assures low grades of similarity for small values of possibility and vice versa. Unfortunately, as previously showed in examples 3.1 and 3.2, such relation does not exist unless some restrictions are imposed on the involved fuzzy sets.

To find a relationship between possibility and similarity, in this Section it will be proved that the involved fuzzy sets must verify the following interpretability constraints:

- Normality;
- Convexity;
- Continuity.

The three above-mentioned constraints are widely used in interpretable fuzzy modeling, since usually fuzzy sets used in model design belong to specific classes (such as triangular, Gaussian, trapezoidal, etc.) that fulfill all the three properties. As a consequence, assuming the validity of the above properties does not limit the subsequent analysis to any specific class of interpretable fuzzy model.

Before enunciating the relation between similarity and possibility, two technical lemmas are necessary to establish some useful properties of convex fuzzy sets, which will turn useful in the successive discussion.

LEMMA 3.1 *For any convex fuzzy set A there exists an element $p \in U$ such that the membership function μ_A restricted to the sub-domain*

$$U_L = \{x \in U : x \leq p\} \tag{3.50}$$

is non-increasing. Similarly, the membership function μ_A restricted to the sub-domain

$$U_R = \{x \in U : x \geq p\} \tag{3.51}$$

is non-decreasing.

Proof. *Let $p \in \arg\max_{x \in U} \mu_A(x)$. Consider two points $x_1 < x_2 \leq p$ and the corresponding membership values $\alpha_1 = \mu_A(x_1)$ and $\alpha_2 = \mu_A(x_2)$. By definition of convexity, $\alpha_2 \geq \min\{\mu_A(p), \alpha_1\} = \alpha_1$. Hence, the membership function μ_A is non-decreasing in U_L. With the same procedure, it can be proved that μ_A is non-increasing in U_R.* ■

LEMMA 3.2 *Let A and B be two continuous convex normal fuzzy sets with $p_A \in \arg\max \mu_A$ and $p_B \in \arg\max \mu_B$ such that $p_A < p_B$. Then, there exists a point between p_A and p_B whose membership degree (both on A and B) corresponds to the possibility measure between A and B:*

$$\Pi(A, B) = \vartheta \rightarrow \exists x \in [p_A, p_B] : \mu_A(x) = \mu_B(x) = \vartheta \tag{3.52}$$

Proof. *By definition of possibility (3.11):*

$$\vartheta = \sup_{x \in U} \min\{\mu_A(x), \mu_B(x)\} \tag{3.53}$$

Because of continuity of fuzzy set A, and by Weierstrass theorem, the range of μ_A in the interval $[p_A, p_B]$ is a closed interval $[m_A, 1]$. Similarly, the range of μ_B in the same interval is $[m_B, 1]$. Let be:

$$\widetilde{\vartheta} = \sup_{x \in [p_A, p_B]} \min\{\mu_A(x), \mu_B(x)\} \tag{3.54}$$

Let $\widetilde{x} \in [p_A, p_B]$ *such that* $\min\{\mu_A(\widetilde{x}), \mu_B(\widetilde{x})\} = \widetilde{\vartheta}$. *Without loss of generality, suppose that* $\mu_A(\widetilde{x}) = \widetilde{\vartheta} < 1$. *Because of convexity of A, then* $\forall x : p_A \leq x < \widetilde{x} \rightarrow \mu_A(x) > \widetilde{\vartheta}$. *As a consequence* $\mu_B(\widetilde{x}) \geq \widetilde{\vartheta}$. *If* $\mu_B(\widetilde{x}) > \widetilde{\vartheta}$, *then for any* $\Delta\widetilde{x} > 0$ *sufficiently small,* $\mu_B(\widetilde{x} - \Delta\widetilde{x}) > \widetilde{\vartheta}$ *and* $\mu_A(\widetilde{x} - \Delta\widetilde{x}) > \widetilde{\vartheta}$. *Thus,* $\min\{\mu_A(\widetilde{x} - \Delta\widetilde{x}), \mu_B(\widetilde{x} - \Delta\widetilde{x})\} > \widetilde{\vartheta}$ *and (3.54) would be invalidated. As a consequence,* $\mu_B(\widetilde{x}) = \mu_A(\widetilde{x}) = \widetilde{\vartheta}$. *In the specific case of* $\widetilde{\vartheta} = 1$, *the equality of the two membership values is trivial.*
Suppose now, ab absurdo, that the possibility measure ϑ *is different from* $\widetilde{\vartheta}$, *i.e.* $\vartheta > \widetilde{\vartheta}$ *(the case* $\vartheta < \widetilde{\vartheta}$ *is impossible by definition of possibility measure). Two cases can be considered:*

$$\textit{(i)}\ \exists x_\vartheta < p_A : \min\{\mu_A(x), \mu_B(x)\} = \vartheta \tag{3.55}$$

$$\textit{(ii)}\ \exists x_\vartheta > p_B : \min\{\mu_A(x), \mu_B(x)\} = \vartheta \tag{3.56}$$

We consider first case (i). Then, $x_\vartheta < \widetilde{x} \leq p_B$ *and* $\mu_B(x_\vartheta) = \vartheta > \widetilde{\vartheta} = \mu_B(\widetilde{x})$. *But since B is convex, by Lemma 3.1* $x_\vartheta < \widetilde{x} \leq p_B \rightarrow \mu_B(x_\vartheta) = \vartheta \leq \mu_B(\widetilde{x}) = \widetilde{\vartheta}$. *This is a contradiction, thus,* $\vartheta = \widetilde{\vartheta}$. *Case (ii) is symmetrical to case (i).* ■

The last lemma states that for convex, continuous and normal fuzzy sets, the "intersection point" of the fuzzy membership functions μ_A and μ_B laying between the modal values of A and B determines the possibility value (see fig. 3.4). In this way, the analysis can be limited to the interval between the modal values of the two fuzzy sets, thus simplifying further considerations on relating possibility and similarity.

Now it is possible to establish an important relation between possibility and similarity with the following theorem.

THEOREM 3.1 *Let A and B be two fuzzy sets that are continuous, normal and convex. Let* $p_A \in \arg\max \mu_A$, $p_B \in \arg\max \mu_B$ *and suppose* $p_A < p_B$. *Let* $\vartheta = \Pi(A, B)$ *and* $x_\vartheta \in [p_A, p_B]$ *such that* $\mu_A(x_\vartheta) = \mu_B(x_\vartheta) = \vartheta$. *In addition, suppose that:*

$$\forall x \in]p_A, x_\vartheta[: \frac{d^2\mu_A}{dx^2}(x) \leq 0 \tag{3.57}$$

and:

$$\forall x \in]x_\vartheta, p_B[: \frac{d^2\mu_B}{dx^2}(x) \leq 0 \tag{3.58}$$

Then, the similarity between A and B is upper-bounded by:

$$S(A, B) \leq S_{\max} = \frac{2\vartheta}{r + 2\vartheta - r\vartheta} \tag{3.59}$$

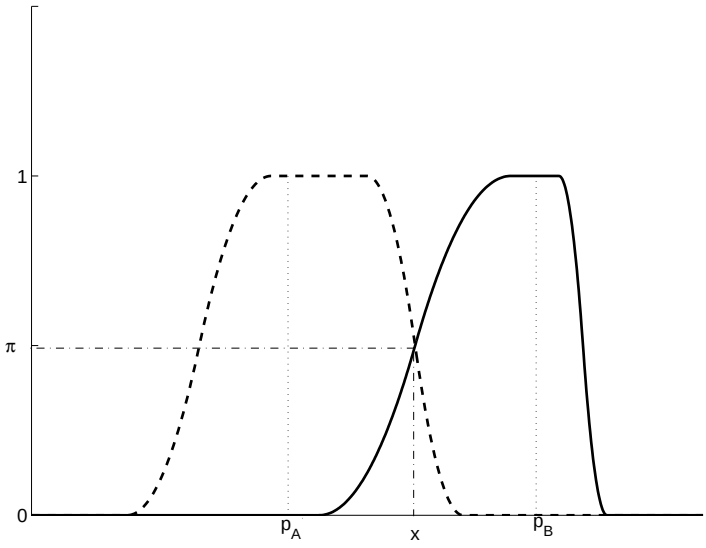

Figure 3.4: The membership degree of intersection point between two membership functions corresponds to the possibility between the two fuzzy sets.

being r the ratio between the distance $p_B - p_A$ and the length of the support[5] *of $A \cup B$:*

$$r = \frac{p_B - p_A}{|\text{supp}\, A \cup B|} \tag{3.60}$$

Proof. *The first objective is to define two normal and convex fuzzy sets that are maximally similar but have possibility measure ϑ. These fuzzy sets must be defined so that the cardinality of their intersection is the highest possible, while the cardinality of their union is the smallest possible. The following two fuzzy sets $A_{\max}$ and $B_{\max}$ satisfy such requirements (see fig. 3.5 for an illustrative example):*

$$\mu_{A_{\max}}(x) = \begin{cases} \vartheta & \text{if} \quad x \in [\min \text{supp}\, A \cup B, p_A[\\ \frac{x(\vartheta-1)+x_\vartheta - p_A\vartheta}{x_\vartheta - p_A} & \text{if} \quad x \in [p_A, x_\vartheta] \\ \vartheta & \text{if} \quad x \in]x_\vartheta, \max \text{supp}\, A \cup B] \\ 0 & \text{elsewhere} \end{cases} \tag{3.61}$$

$$\mu_{B_{\max}}(x) = \begin{cases} \vartheta & \text{if} \quad x \in [\min \text{supp}\, A \cup B, x_\vartheta[\\ \frac{x(\vartheta-1)+x_\vartheta - p_B\vartheta}{x_\vartheta - p_B} & \text{if} \quad x \in [x_\vartheta, p_B] \\ \vartheta & \text{if} \quad x \in]p_B, \max \text{supp}\, A \cup B] \\ 0 & \text{elsewhere} \end{cases} \tag{3.62}$$

[5] The support of a fuzzy set is the (crisp) set of all elements with non-zero membership, i.e. $\text{supp}\, X = \{x \in U : \mu_X(x) > 0\}$. For convex fuzzy sets, the support is an interval.

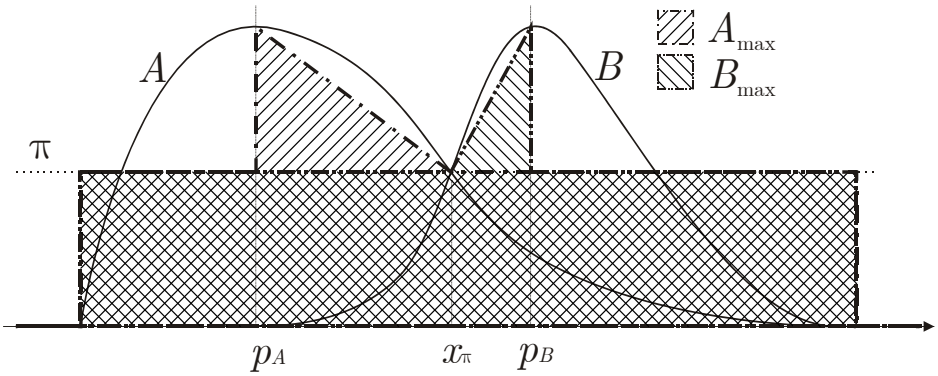

Figure 3.5: Example of fuzzy sets with maximal similarity for a given possibility measure

The membership functions so defined are such that intersection and union coincide in all the support except the interval $[p_A, p_B]$*, while within such interval the membership functions have null second derivative. As a consequence, any couple of fuzzy sets* A *and* B *satisfying the hypothesis will have:*

$$\forall x \in [p_A, x_\vartheta] : \mu_A(x) \geq \mu_{A_{\max}}(x) \tag{3.63}$$

and

$$\forall x \in [x_\vartheta, p_B] : \mu_B(x) \geq \mu_{B_{\max}}(x) \tag{3.64}$$

In this way, the cardinality of the intersection fuzzy set is minimized while the union is maximized. More specifically, the intersection of the two fuzzy sets has the following membership function:

$$\mu_{A_{\max} \cap B_{\max}}(x) = \begin{cases} \vartheta & \text{if} \quad x \in \operatorname{supp} A \cup B \\ 0 & \text{elsewhere} \end{cases} \tag{3.65}$$

The union of the two fuzzy sets has the following membership function:

$$\mu_{A_{\max} \cup B_{\max}}(x) = \begin{cases} \vartheta & \text{if} \quad x \in [\min \operatorname{supp} A \cup B, p_A[\\ \frac{x(\vartheta-1)+x_\vartheta-p_A\vartheta}{x_\vartheta-p_A} & \text{if} \quad x \in [p_A, x_\vartheta] \\ \frac{x(\vartheta-1)+x_\vartheta-p_B\vartheta}{x_\vartheta-p_B} & \text{if} \quad x \in [x_\vartheta, p_B] \\ \vartheta & \text{if} \quad x \in]p_B, \max \operatorname{supp} A \cup B] \\ 0 & \text{elsewhere} \end{cases} \tag{3.66}$$

The similarity of the two fuzzy sets is:

$$S(A_{\max}, B_{\max}) = \frac{|A_{\max} \cap B_{\max}|}{|A_{\max} \cup B_{\max}|} = \frac{\vartheta \, |\operatorname{supp} A \cup B|}{\vartheta \, |\operatorname{supp} A \cup B| + \frac{1}{2}(1-\vartheta)(p_B - p_A)} \tag{3.67}$$

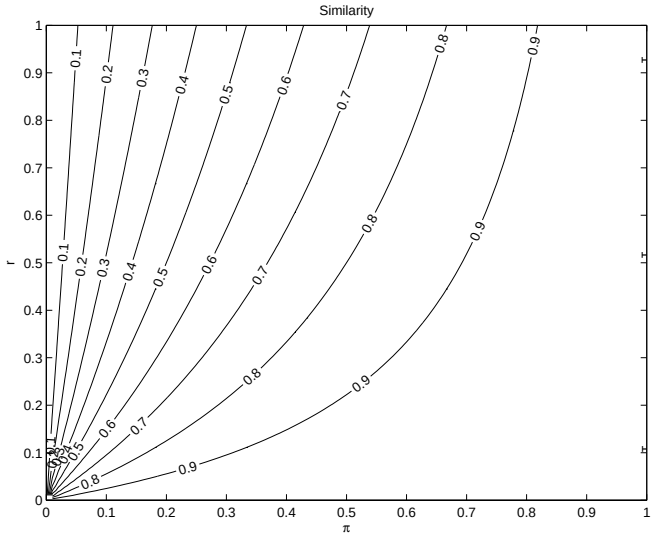

Figure 3.6: Contour plot of maximal similarity $S_{\max}$ with respect to r and π.

By defining r as in (3.60), the similarity is shown to be equal to (3.59). Note that $A_{\max}$ and $B_{\max}$ are not continuous. However, continuous fuzzy sets may be defined so as to be arbitrary similar to $A_{\max}$ and $B_{\max}$. Hence, by defining $S_{\max} = S(A_{\max}, B_{\max})$, the maximal similarity measure is the upper-bound of the actual similarity between the original fuzzy sets A and B. ■

It should be remarked that the additional constraint for the second derivatives, required in the theorem hypothesis, is not particularly limiting, since commonly used fuzzy set shapes (triangular, trapezoidal, Gaussian, bell-shaped, etc.) satisfy such requirement.

It should be also noted that the relationship between possibility and similarity established by the theorem holds only for the upper-bound of the similarity measure, while the actual value is strictly related to the shape of the membership function. Paradoxically, the actual similarity measure between two low-possibility fuzzy sets may be higher than two high-possibility fuzzy sets. However, *relation (3.59) assures that the similarity measure does not exceed a defined threshold that is monotonically related to the possibility measure.* As a consequence, any modelling technique that assures small values of possibility between fuzzy sets, indirectly provides small values of similarity and, hence, good distinguishability between fuzzy sets (see fig. 3.6). Thus, relation (3.59) justifies the adoption of possibility measure in interpretable fuzzy modelling.

A problem arises when Gaussian fuzzy sets are adopted, since the second

derivatives of the respective membership functions may not be negative as requested by the theorem. In order to satisfy the theorem hypothesis, it is necessary that the intersection point between two Gaussian fuzzy set must lay between the prototype and the inflection point of each membership function. To guarantee this condition, the possibility threshold should not be less than[6] $e^{-1/2} \approx 0.606\,53$. However, a specific analysis can be made for Gaussian fuzzy sets, which shows that possibility and similarity are still monotonically related. Indeed, the following proposition can be proved.

PROPOSITION 3.3 *Let A and B be two Gaussian fuzzy sets with same width. If the possibility value between A and B is ϑ then their similarity measure is:*

$$S(A,B) = \frac{1 - \operatorname{erf}\left(\sqrt{-\ln \vartheta}\right)}{1 + \operatorname{erf}\left(\sqrt{-\ln \vartheta}\right)} \tag{3.68}$$

where:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{3.69}$$

Proof. *Let ω_A and ω_B be the centers of the fuzzy sets A and B respectively. Without loss of generality, it can be assumed $\omega_A \leq \omega_B$. Let $\sigma > 0$ be the common width of the two fuzzy sets. By lemma 3.2, it is possible to state that there exists a point $x_\vartheta \in [\omega_A, \omega_B]$ such that:*

$$\mu_A(x_\vartheta) = \vartheta = \mu_B(x_\vartheta) \tag{3.70}$$

Moreover, in example 2.2 (page 32) it has been proved that the intersection point between two fuzzy sets of equal width is unique. As a consequence, the cardinality of the intersection between the fuzzy sets A and B is the area depicted in fig 3.7. The expression for the cardinality of the intersection between A and B is therefore:

$$|A \cap B| = \int_{\mathbb{R}} \min\left\{\mu_A(x), \mu_B(x)\right\} dx = \int_{-\infty}^{x_\vartheta} \mu_B(x)\, dx + \int_{x_\vartheta}^{+\infty} \mu_A(x)\, dx \tag{3.71}$$

For the sake of simplicity, it is assumed that the Universe of Discourse coincided with the real line $\mathbb{R}$. Since each membership function is symmetrical w.r.t. its center, and since the membership function μ_B can be viewed as a horizontal translation of μ_A of length $\omega_B - \omega_A$, it is possible to simplify the previous relation in the following:

$$|A \cap B| = 2 \int_{x_\vartheta}^{+\infty} \mu_A(x)\, dx \tag{3.72}$$

[6] Note that this value is stable against possible variations in the definition of Gaussian membership functions. Indeed, it is easy to prove that the indicated possibility threshold is constant for any membership function of the form $\exp\left(-(x-\omega)^2 / a\sigma^2\right)$ with $a > 0$.

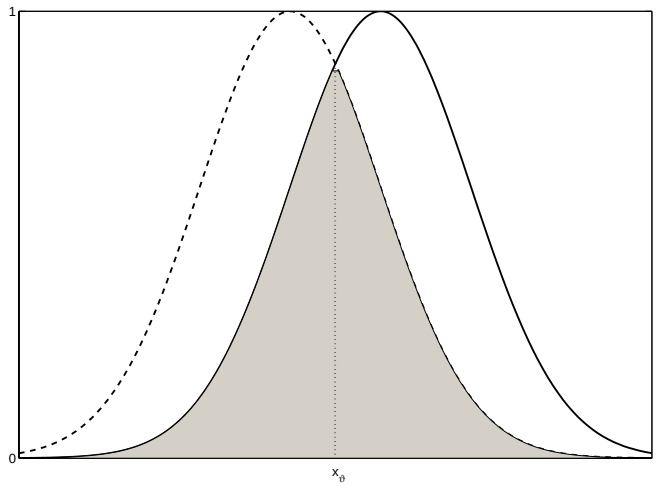

Figure 3.7: The cardinality of the intersection between A and B is the area of the shaded region.

By explicating the definition of Gaussian membership function, equation (3.72) can be rewritten as:

$$|A \cap B| = 2 \int_{x_\vartheta}^{+\infty} \exp\left(-\frac{(x-\omega_A)^2}{2\sigma^2}\right) dx \tag{3.73}$$

The following change of variable can be conveniently defined:

$$t = \frac{x-\omega_A}{\sigma\sqrt{2}} \tag{3.74}$$

In this way, relation (3.73) can be rewritten as:

$$|A \cap B| = 2\sqrt{2}\sigma \int_{\frac{x_\vartheta-\omega_A}{\sigma\sqrt{2}}}^{+\infty} e^{-t^2} dt = \sqrt{2\pi}\sigma \left(1 - \operatorname{erf}\left(\frac{x_\vartheta-\omega_A}{\sigma\sqrt{2}}\right)\right) \tag{3.75}$$

Now the point x_ϑ can be defined in terms of ϑ. More specifically, since $\mu_A(x_\vartheta) = \vartheta$ and $x_\vartheta > \omega_A$, then:

$$x_\vartheta = \omega_A + \sigma\sqrt{-2\ln\vartheta} \tag{3.76}$$

By replacing (3.76) into (3.75), it results:

$$|A \cap B| = \sqrt{2\pi}\sigma \left(1 - \operatorname{erf}\left(\sqrt{-\ln\vartheta}\right)\right) \tag{3.77}$$

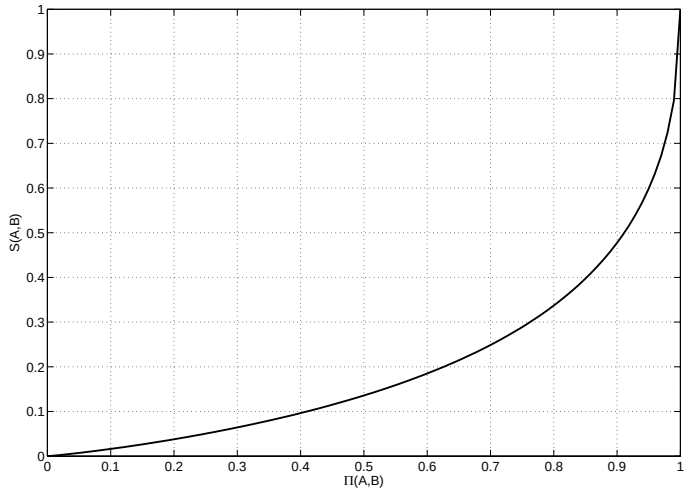

Figure 3.8: Functional relationship between possibility and similarity of two Gaussian fuzzy sets of equal width

The cardinalities of A and B are identical, and can be calculated from (3.72) by taking $x_\vartheta = 0$, resulting in:

$$|A| = |B| = 2\int_0^{+\infty} \mu_A(x)\,dx = \sqrt{2\pi}\sigma \tag{3.78}$$

Finally, the similarity measure between A and B is calculated as:

$$S(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{1 - \operatorname{erf}\left(\sqrt{-\ln \vartheta}\right)}{1 + \operatorname{erf}\left(\sqrt{-\ln \vartheta}\right)} \tag{3.79}$$

■

It is noteworthy that the similarity of two Gaussian membership functions with equal width is related to their possibility measure but it is independent from the value of width. In fig. 3.8 the relationship between possibility and similarity is depicted, showing the monotonic behavior that justifies the adoption of possibility measure for distinguishability quantification even when Gaussian fuzzy sets (of equal width) are adopted.

When Gaussian fuzzy sets have different width, the relation between possibility and similarity becomes more complex and non monotonic[7]. As can be

[7]It can be proved that the relationship between possibility and similarity depends on the specific widths of the Gaussian fuzzy sets. The proof is omitted since it does not convey further arguments.

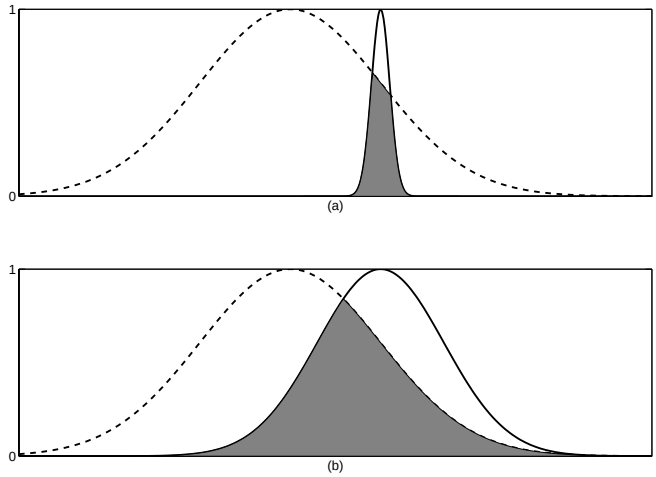

Figure 3.9: Two intersections between Gaussian fuzzy sets with different width: (a) very different widths; (b) similar widths.

seen from fig. 3.9(a), when widths are very different the area delimited by the intersection of two fuzzy sets has a very different shape compared with that depicted in fig. 3.7, where equal width fuzzy sets are intersected. As a consequence, the functional relationship between possibility and similarity does not follow the trend depicted in fig. 3.8. In this case, the sufficient condition provided by theorem 3.1 should be considered. However, if the widths of the Gaussian fuzzy sets are similar, as exemplified in fig 3.9(b), the monotonic relationship between possibility and similarity can be still considered roughly monotonic.

3.4 Theorem about distinguishability improvement

Reducing similarity between fuzzy sets is a classical approach to improve their distinguishability. However, the calculation of similarity calls for computationally intensive methods (e.g. genetic algorithms) or separate stages (e.g. fuzzy sets merging). When efficient learning schemes are necessary, other measures are adopted in substitution of similarity. Hence, an interesting issue concerns how reducing non-similarity measures effectively reduces similarity. Here the analysis is focused on possibility measure as an alternative quantification of distinguishability. The following lemma characterize a

wide class of possibility-reducing procedures.

LEMMA 3.3 *Let A and B two fuzzy sets defined on the Universe of Discourse U, which are continuous normal and convex. Let $p_A \in \arg\max \mu_A$, $p_B \in \arg\max \mu_B$ and suppose $p_A < p_B$. Let $\Phi : \mathcal{F}(U) \to \mathcal{F}(U)$ be a transformation such that $B' = \Phi(B)$ is a continuous, normal and convex fuzzy set, and $\forall x \in U : x \leq p_B \to \mu_{B'}(x) \leq \mu_B(x)$. Then,*

$$\Pi(A, B') \leq \Pi(A, B) \tag{3.80}$$

Conversely, if B' is such that $\forall x \in U : x \leq p_B \to \mu_{B'}(x) \geq \mu_B(x)$, then

$$\Pi(A, B') \geq \Pi(A, B) \tag{3.81}$$

Proof. *We consider only the first thesis (3.80), because the other can be proved similarly. Let ϑ be the possibility measure $\Pi(A, B)$. In force of Lemma 3.2,*

$$\exists x_\vartheta \in [p_A, p_B] : \mu_A(x) = \mu_B(x) = \vartheta \tag{3.82}$$

By hypothesis, $\mu_{B'}(x_\vartheta) \leq \vartheta$, and re-applying Lemma 3.2 to A and B', it can be stated that:

$$\exists x' \in [p_A, p_{B'}] : \mu_A(x') = \mu_{B'}(x') = \Pi(A, B') = \vartheta' \tag{3.83}$$

where $p_{B'} \in \arg\max \mu_{B'}$. If $\mu_{B'}(x_\vartheta) < \mu_B(x_\vartheta)$, then $\forall x'' < x_\vartheta : \mu_A(x'') \geq \vartheta \wedge \mu_{B'}(x'') \leq \mu_{B'}(x_\vartheta) < \vartheta$. Hence, $\forall x'' < x_\vartheta : \mu_A(x'') \neq \mu_B(x'')$. As a consequence, $x' \geq x_\vartheta$. In such case, $\vartheta' = \mu_A(x') \leq \mu_A(x) = \vartheta$, i.e. $\Pi(A, B') \leq \Pi(A, B)$.
■

Two very common examples of transformations satisfying the lemma's hypothesis are the translation ($\mu_{B'}(x) = \mu_B(x - x_0), x_0 > 0$) and the contraction ($\mu_{B'}(x) = \mu_B(x)^p, p > 1$) of the membership function. Such transformations can be effectively used to reduce possibility between two fuzzy sets, but in order to establish whether such transformations also reduce similarity an additional condition must be introduced, as proved in the following theorem.

THEOREM 3.2 *Any transformation $\Phi : \mathcal{F}(U) \to \mathcal{F}(U)$ such that $B' = \Phi(B)$ preserves lemma 3.3 hypothesis and additionally*

$$r' = \frac{p_{B'} - p_A}{|\operatorname{supp} A \cup B'|} \geq \frac{p_B - p_A}{|\operatorname{supp} A \cup B|} = r \tag{3.84}$$

produces a decrease of the maximal similarity $S_{\max}$.

Proof. *The function* $S_{\max}(\vartheta, r) = \frac{2\vartheta}{r+2\vartheta-r\vartheta}$ *given in (3.59) is continuous and derivable for every* $\vartheta \in [0,1]$*. Its derivative is:*

$$\frac{\partial S_{\max}}{\partial \vartheta} = \frac{2r}{(r+2\vartheta-r\vartheta)^2} \tag{3.85}$$

Since the ratio r *is always positive,* $S_{\max}$ *is monotonically decreasing as* ϑ *decreases, for a fixed ratio* r*. However, the new fuzzy set* B' *may determine a different ratio* r' *different from* r*. The directional derivative of* $S_{\max}$ *w.r.t. the direction* $(\Delta\vartheta, \Delta r)$ *is:*

$$\frac{dS_{\max}}{d(\Delta\vartheta, \Delta r)} = \nabla S_{\max} \cdot \frac{(\Delta\vartheta, \Delta r)}{\|(\Delta\vartheta, \Delta r)\|} = \frac{2r\Delta\vartheta + 2\vartheta(\vartheta-1)\Delta r}{\|(\Delta\vartheta, \Delta r)\|(r+2\vartheta-r\vartheta)^2} \tag{3.86}$$

Such derivative is negative (i.e. $S_{\max}$ *is decreasing) when:*

$$\Delta r > \frac{2r\Delta\vartheta}{2\vartheta(1-\vartheta)} \tag{3.87}$$

For $\Delta\vartheta < 0$ *(reduced possibility), the second member of (3.87) is negative, hence any transformation that does not reduce the ratio* ($\Delta r = r' - r \geq 0$) *will effectively reduce the maximal possibility* $S_{\max}$*.* ■

As a corollary of the theorem, every method aimed to reduce possibility actually reduces (maximal) similarity, thus improving distinguishability. In this sense, the adoption of possibility as a measure of distinguishability is fully justified. The additional constraint required in the corollary (the ratio r must not decrease) is always fulfilled by any translation that lengthens the distance between the prototypes, as well as by contractions. However, attention must be paid for those transformation that reduce the support, for which the relation (3.87) must be carefully taken into account.

3.5 Final remarks

In this Chapter, it has been shown that possibility is both an effective and computationally efficient measure to quantify distinguishability, an important constraint for interpretable fuzzy modelling. Indeed, while similarity can be considered as the most representative measure for distinguishability of fuzzy sets, it has been proved that under mild conditions (always satisfied by interpretable fuzzy sets) possibility and similarity are related monotonically, so that small values of possibility imply small values of similarity. The added values of possibility measures are its sound semantic meaning and the computationally efficiency of the calculation procedure. As a matter of fact, in most cases possibility measure can be expressed analytically in terms of

fuzzy sets parameters, so it can be used in many learning schemes without resorting computationally intensive algorithms.

Chapter 4

Interface Optimality

4.1 Introduction

Frames of Cognition are employed in fuzzy modelling to provide a granulated view of a simple (i.e. one-dimensional) attribute by means of interpretable fuzzy sets that can be designated by semantically sound linguistic labels. Frames of Cognition are at the basis for the definition of more complex information granules that relate different attributes so as to express knowledge and perform approximate reasoning.

From a functional point of view, Frames of Cognition can be regarded as mappings that convert an external representation of information to an internal one (or vice versa), which is suited to the information processing that is carried out within a fuzzy model. For such reason, Frames of Cognition are parts of the so-called "*Input Interfaces*" and "*Output Interfaces*", which are – along with the "*Processing Module*" – the main functional blocks most fuzzy models.

The role of the Input Interface consists in the conversion of information coming from the environment in an internal format acceptable by the Processing Module. Symmetrically, the Output Interface transforms information coming from the processing module into a suited external representation to be used in the environment. The information transformation from the external to the internal representation can be carried out through a matching procedure employing a family of referential fuzzy sets that belong to some Frame of Cognition. Specifically, the Input Interface provides a mapping from numerical data into a structure of fuzzy membership values, while the Output Interface transforms such structures into numerical results. Although the roles of the input and output interfaces are quite different, both of them share the same framework as being based on Frames of Cognition

(de Oliveira, 1993; de Oliveira, 1995; Pedrycz and de Oliveira, 1996).

As it is widely recognized that the impact of the performance of fuzzy models heavily depends on the input/output interfaces, a careful design of such functional blocks is crucial for the overall quality of the fuzzy models. With this purpose, several criteria have been proposed in order to guarantee well-designed interfaces in terms of two main factors: the semantic integrity of the modeled linguistic terms and the precise representation of uncertain input data (Valente de Oliveira, 1999a; Pedrycz, 1996b). In particular, precise representation can be guaranteed by the so-called "*Information Equivalence Criterion*" (Pedrycz and de Oliveira, 1996), also called "*Anti-fuzzy-aliasing Condition*" (de Oliveira, 1996) or "*Error-free Reconstruction*" (Pedrycz, 1994). According to the Information Equivalence Criterion, an interface should preserve the information coming into (or outcoming from) the processing block by converting it into a relevant internal representation appropriate for further processing. An interface that satisfies the Information Equivalence Criterion is also called "Optimal Interface" (de Oliveira, 1996). Using optimal interfaces, it can be proved that each external datum has its own internal representation, and referential fuzzy sets cover the entire Universe of Discourse (Valente de Oliveira, 1998).

In interpretable fuzzy modelling, optimal Input Interfaces are important because they provide for a precise internal representation of the information coming from the environment. More specifically, for different inputs, an optimal Input Interface returns different representations that are further processed by the Processing Module. In case of sub-optimality, there would exists two or more inputs for which the same representation is provided by the Input Interface. In this way, the inputs would processed exactly in the same way: they would be indistinguishable even though they are different. But often fuzzy models are used because their internal knowledge is defined by formal concepts (i.e. the information granules) that are always verified *to a degree*. Therefore, it could be expected that different inputs verify concepts to different degrees, in the same way humans perceive different stimuli with different intensities. As a consequence, sub-optimality of Input Interfaces could make fuzzy models too approximative to be acceptable in certain applicative contexts

Conversely, optimal Output Interfaces provide for an invertible representation of the inference process carried out by the Processing Module. In the context of interpretable fuzzy modelling, this means that the output of the model could be re-fuzzified to yield the *same* fuzzy representation provided by the Processing Module. Such representation is a structure of truth degrees, usually consequence degrees of a list of rules. Therefore, the output of the model would verify the consequent of each rule with the same degree

provided by the Processing Module. This is exactly what is required by a fully interpretable inference process: if the strength of a rule is high (resp. low), it is expected that the inferred output verifies the rule consequent with a high (resp. low) degree[1]. In this sense, optimality of the Output Interface is a necessary condition to guarantee that the output provided by the model is in strict correspondence with the conclusion inferred by the Processing Module.

Optimal Interfaces have been extensively studied in the context of Frames of Cognition defined by triangular fuzzy sets (de Oliveira, 1993; de Oliveira, 1995; Pedrycz, 1994; Pedrycz and de Oliveira, 1996), but little attention has been paid to interfaces based on fuzzy sets of different shapes, such as trapezoidal, Gaussian, bell-shaped, etc. (Zimmermann, 2001). Moreover, some authors are skeptical concerning the real effectiveness of triangular fuzzy sets in solving modeling problems (see for example (Mitaim and Kosko, 2001)). Conversely, convex fuzzy sets are regarded as a promising class of fuzzy sets to guarantee interface optimality (de Oliveira, 1995). Hence, the issue of optimality of interfaces based on different shapes of fuzzy sets should be addressed.

In this Chapter, it is proved that optimality of Input Interfaces can be guaranteed for a wide class of fuzzy sets, provided that mild conditions are satisfied. In particular, after the definition of two main classes of fuzzy sets, namely "strictly bi-monotonic" fuzzy sets and "loosely bi-monotonic" fuzzy sets, two theorems are proved. The first theorem proves that optimal interfaces are always guaranteed for strictly bi-monotonic fuzzy sets (e.g. Gaussian fuzzy sets), provided that weak conditions hold. The second theorem states that optimality is guaranteed for loosely bi-monotonic fuzzy sets (e.g. triangular and convex fuzzy sets in general) if stronger conditions hold. The theorems provide sufficient conditions for optimality for a wide class of fuzzy sets, including convex fuzzy sets, so that their results can be useful in a broad range of fuzzy modeling contexts.

Such theoretical results, derived for Input Interfaces, are not applicable to Output Interfaces. Indeed, while Input Interfaces are usually implemented by a matching mechanism that derives membership values from numerical values, Output Interfaces are generally defined by an aggregation operation that can easily hamper the optimality condition. Nevertheless, the Information Equivalence Criterion should be taken into account in the design of Output Interfaces, since it is a necessary condition to define transparent and accurate models. For such reason, a different optimality criterion is proposed, which leads to the definition of an Optimality Degree that measures

[1] Assuming a conjunctive interpretation of rules, see §2.6

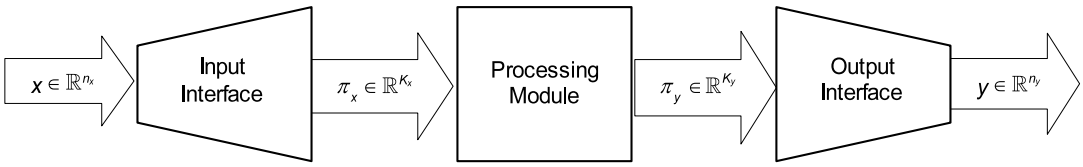

Figure 4.1: The three functional blocks of a Fuzzy Model

the quality of a fuzzy Output Interface. According to such criterion, different Output Interfaces can be compared and their quality ranked according to their optimality.

The Chapter is organized as follows. Section 4.2 provides a formal description of Input and Output Interfaces. In Section 4.3, the optimality of Input Interfaces is addressed: the class of bi-monotonic fuzzy sets is defined and two theorems on optimality conditions for bi-monotonic fuzzy sets are proved. In Section 4.4, the optimality of Output Interfaces is discussed, and the Optimality Degree is introduced. Also, an illustrative example is presented to show how the defined optimality degree can be conveniently used to highlight quality variations attained by different output interfaces. Finally, in Section 4.5 some conclusions are drawn.

4.2 Fuzzy Interfaces

A Fuzzy Model is typically composed of three main functional blocks (fig. 4.1), namely the Input and the Output Interface communicating between the modeling environment, and the processing module dedicated for fuzzy computations. In the following,he role of each functional block is briefly described along with the basic terminology.

4.2.1 Input Interface

The role of the Input Interface is to transform the data about a certain application from the usual numerical level into the linguistic level. In other words, an Input Interface provides a fuzzy discretization of external data by means of a family of K_x reference fuzzy sets. For n_x-dimensional data of domain $\mathbf{X} \subseteq \mathbb{R}^{n_x}$, the (multidimensional) Input Interface can be formalized as follows:

$$\Im : \mathbf{X} \subset \mathbb{R}^{n_x} \rightarrow [0,1]^{K_x} \tag{4.1}$$

such that:

$$\forall x \in \mathbf{X} : \Im(x) = [A_i(x)]_{i=1,2,\ldots K_x} \tag{4.2}$$

where $A_i(x)$ is the membership degree of the i-th reference fuzzy set. It should be remarked that the definition of Input Interfaces can be extended to deal with fuzzy inputs, as remarked in (Pedrycz and de Oliveira, 1996), but in such case it is impossible to perfectly reconstruct inputs when only membership values are given (see (Pedrycz and Gomide, 1998) for a complete proof).

Usually, in interpretable fuzzy models, an Input Interface is composed of several one-dimensional Input Interfaces defined by Frames of Cognition. In such a case, the (multidimensional) Input Interface is defined as follows:

$$\forall x \in \mathbf{X} \subseteq \mathbb{R}^{n_x} : \Im(x) = \left[\Im^{(j)}(x_j)\right]_{j=1,2,\ldots n_x} \tag{4.3}$$

being $x = [x_1, x_2, \ldots, x_{n_x}]$ and $\Im^{(j)}$ the j-th one-dimensional Input Interface which is in the form:

$$\Im^{(j)} : X^{(j)} \subseteq \mathbb{R} \rightarrow [0,1]^{K_x^{(j)}}, \quad j = 1, 2, \ldots n_x \tag{4.4}$$

Each one-dimensional Input Interface can be defined as the functional counterpart of a Frame of Cognition. Specifically, if for the j-th attribute the following Frame of Cognition is defined:

$$\mathbf{F}^{(j)} = \left\langle U^{(j)}, \mathbf{F}^{(j)}, \preceq^{(j)}, \mathcal{L}^{(j)}, v^{(j)} \right\rangle \tag{4.5}$$

being the family of fuzzy sets:

$$\mathbf{F}^{(j)} = \left\{ A_1^{(j)}, A_2^{(j)}, \ldots, A_{K_x^{(j)}}^{(j)} \right\} \tag{4.6}$$

then, the j-th input interface is defined as:

$$\Im^{(j)}(x) = \left(A_1^{(j)}(x), A_2^{(j)}(x), \ldots, A_{K_x^{(j)}}^{(j)}(x) \right) \tag{4.7}$$

with the assumption that $A_i^{(j)} \preceq^{(j)} A_{i+1}^{(j)}$.

4.2.2 Processing Module

The outcome of the Input Interface is transformed into a structure of output membership degrees by the processing module, which is dedicated to all computations at a fuzzy level. From a functional point of view, the processing module can be formalized as follows:

$$\mathfrak{P} : [0,1]^{K_x} \rightarrow [0,1]^{K_y} \tag{4.8}$$

The processing module can be implemented in several forms, including fuzzy relational calculus, fuzzy regression models and fuzzy neural networks (Pedrycz, 1995).

4.2.3 Output Interface

The role of the Output Interface (Output Interface) consists in transforming the results produced by the processing module into a form acceptable by the modeling environment, which often calls for numerical values. Hence, such transformation completes a mapping from the linguistic to the numerical level. From a functional point of view, the Output Interface can be formalized as:

$$\mathfrak{O} : [0,1]^{K_y} \rightarrow \mathbb{R}^{n_y} \tag{4.9}$$

For simplicity, it is assumed that $n_y = 1$, i.e. the Output Interface provides only a one-dimensional numerical value. The extension to multidimensional outputs is straightforward by considering n_y Output Interfaces independently.

Generally, an Output Interface is characterized by a family of referential fuzzy sets belonging to a related Frame of Cognition and the final numerical output is calculated by means of a defuzzification algorithm, which operates by aggregating membership degrees coming from the processing module.

Depending on the type of the fuzzy model, the Output Interface can be very simple or quite complex. In Takagi-Sugeno Fuzzy Inference Systems with rules of the following type:

$$\text{IF } x \text{ is } \mathbf{A}^{(r)} \text{ THEN } y = w^{(r)} \tag{4.10}$$

the Output Interface is defined as a weighted average formula:

$$\mathfrak{O}_{TS}\left(\pi_y\right) = \frac{\sum_{r=1}^{K_y} \pi_y^{(r)} \cdot w^{(r)}}{\sum_{r=1}^{K_y} \pi_y^{(r)}} \tag{4.11}$$

being $\pi_y = \left[\pi_y^{(1)}, \pi_y^{(2)}, \ldots, \pi_y^{(K_y)}\right]$ the vector of membership degrees provided by the Processing Module, with $\pi_y^{(r)} = \mathbf{A}^{(r)}\left(x\right)$. In a Mamdani Fuzzy Inference System with rules of the following type:

$$\text{IF } x \text{ is } \mathbf{A}^{(r)} \text{ THEN } y \text{ is } B^{(r)} \tag{4.12}$$

the Output Interface derives, for each rule, a new fuzzy set as follows:

$$\bar{B}^{(r)}\left(y\right) = B^{(r)}\left(y\right) \otimes \pi_y^{(r)} \tag{4.13}$$

being $\otimes$ a t-norm like the minimum or the product operator. Then, all such fuzzy sets are aggregated to form a unique fuzzy set:

$$B = \bigcup_{r=1}^{K_y} \bar{B}^{(r)} \tag{4.14}$$

centroid	$\widetilde{y} = \frac{\int_{\mathbb{R}} B(y) \cdot y \cdot dy}{\int_{\mathbb{R}} B(y) \cdot dy}$
bisector	$\widetilde{y}\ s.t.\ \int_{-\infty}^{\widetilde{y}} B\left(y\right) dy = \int_{\widetilde{y}}^{+\infty} B\left(y\right) dy$
mean of maxima (MOM)	$\widetilde{y} = \frac{\sum_{\phi e M} \phi}{\lvert M \rvert}, M = \arg\max B$
sup of maxima (SOM)	$\widetilde{y} = \sup\left(\arg\max B\right)$
inf of maxima (LOM)	$\widetilde{y} = \inf\left(\arg\max B\right)$

Table 4.1: Some common defuzzification formulas

where the union operation can be computed by means of a t-conorm, like the maximum or the probabilistic sum. Finally the aggregate fuzzy set B is reduced to a single numerical value $\tilde{y}$ according to a defuzzification method, like those illustrated in table 4.1 (for a more complete review of defuzzification methods, see (Lee, 1990)). Hence, in case of Mamdani Fuzzy Inference Systems, the Output Interface is implicitly defined by one of the above defuzzification formulas.

4.3 Theorems about optimality for Input Interfaces

The idea behind the concept of optimal interface states that an error-free conversion should exist when the values of a given numerical variable are transformed in the internal representation, and vice versa. Hence, an Input Interface is defined as *optimal* if it is invertible, i.e. there exists a way to exactly determine the input value when the vector of membership degrees is given. Formally:

$$\text{opt}\left(\Im, \mathbf{X}\right) \Leftrightarrow \exists \Im^{-1}\ s.t. \forall x \in \mathbf{X} : \Im^{-1}\left(\Im\left(x\right)\right) = x \tag{4.15}$$

The optimality condition of a multidimensional Input Interface defined as in (4.3) is met when all its composing one-dimensional Input Interfaces are optimal.

It should be remarked that this definition is slightly different from those found in literature to characterize optimality (see, e.g., (Valente de Oliveira, 1998; Valente de Oliveira, 1999a; Pedrycz, 1994)). Indeed, other authors provide a definition of optimality for a given couple of I/O interfaces, being the Output Interface the inverse function of the Input Interface. This inverse function is usually defined by a defuzzification formula. Conversely, in (4.15) the inverse function of the Input Interface is not explicitly given, but it is simply required to exist. In other words, such inverse function may not coincide

with any known defuzzification method, but the definition still guarantees that no ambiguity is introduced in the conversion process performed by the Input Interface.

In the design of an Input Interface, the optimality condition can be easily guaranteed if the referential fuzzy sets are properly chosen. Here, a family of fuzzy sets which guarantee optimality condition is characterized. This family embraces two classes of fuzzy sets, called bi-monotonic, that cover most types of fuzzy sets, including convex fuzzy sets. In the following, after the definition of the two classes of bi-monotonic fuzzy sets, namely strictly bi-monotonic fuzzy sets and loosely bi-monotonic fuzzy sets, two theorems about optimality condition are formulated and proved.

4.3.1 Bi-monotonic fuzzy sets

DEFINITION 4.1 (BI-MONOTONIC FUZZY SETS) *A one-dimensional fuzzy set A defined over the Universe of Discourse $X \subseteq \mathbb{R}$ is strictly bi-monotonic if there exists a prototype $p \in X$ (i.e. $A(p) = 1$) such that the restrictions of the membership function*[2] *$A : X \to [0,1]$ to the sub-domains $X_L = X \cap]-\infty, p]$ and $X_R = X \cap [p, +\infty[$ are strictly monotonic. The fuzzy set A is (loosely) bi-monotonic if such restrictions are loosely monotonic. Of course, strictly bi-monotonic fuzzy sets are also loosely bi-monotonic. We will call the two above mentioned restrictions as A_L (left restriction) and A_R (right restriction).*

LEMMA 4.1 *If A is (strictly/loosely) bi-monotonic, then the complement of A (i.e. $\overline{A}(x) = 1 - A(x)$) is bi-monotonic.*

LEMMA 4.2 *If A is strictly bi-monotonic, then $\forall \pi \in [0,1] : |A^{-1}(\pi)| \leq 2$, where $A^{-1}(\pi) = \{x \in X | A(x) = \pi\}$.*

Proof. *If the range of the membership function of A is $\rho(A) \subseteq [0,1]$, then $\forall \pi \in [0,1] - \rho(A) : A^{-1}(\pi) = \emptyset$. Let p be a prototype of A. If $\pi \in \rho(A)$, let:*

$$A_L^{-1}(\pi) = \{x \in X \cap]-\infty, p] \,|\, A(x) = \pi\} \tag{4.16}$$

$$A_R^{-1}(\pi) = \{x \in X \cap [p, +\infty[\,|\, A(x) = \pi\} \tag{4.17}$$

Then $\left|A_L^{-1}(\pi)\right| \leq 1$ because if $\exists x_1, x_2 : x_1 < x_2 \wedge x_1 \in A_L^{-1}(\pi) \wedge x_2 \in A_L^{-1}(\pi)$, then $\exists x_1, x_2 : x_1 < x_2 \wedge A_L(x_1) = A_L(x_2)$ that is, the left restriction A_L would not be strictly monotonic. The same is true for A_R. As a consequence, $A^{-1}(\pi) = A_L^{-1}(\pi) \cup A_R^{-1}(\pi) \Rightarrow \left|A^{-1}(\pi)\right| \leq \left|A_L^{-1}(\pi)\right| + \left|A_R^{-1}(\pi)\right| \leq 2$. ■

[2] Hereafter, the membership function of a fuzzy set A is denoted with the same symbol A and not by μ_A, as in other chapters. This notation has been preferred for ease of reading.

COROLLARY 4.1 *If p is a prototype of A, then $|A^{-1}(A(p))| = 1$*
Proof. *Note that $p \in A_L^{-1}(A(p)) \wedge p \in A_R^{-1}(A(p))$. If, ad absurdum, $\exists q$ s.t. $p \neq q \wedge q \in A^{-1}(A(p))$, then $q \in A_L^{-1}(A(p)) \vee q \in A_R^{-1}(A(p))$, that is, $\left|A_L^{-1}(A(p))\right| > 1 \vee \left|A_R^{-1}(A(p))\right| > 1$. This is absurd.* ■

COROLLARY 4.2 *The support of a strictly bi-monotonic fuzzy set A defined over X is $X - O$, where O is a set with at most two elements, and can be eventually empty.*

LEMMA 4.3 *Convex one-dimensional fuzzy sets are loosely bi-monotonic.*
Proof. *Let A be a convex fuzzy set and $p \in \arg\max_{x \in X} A(x)$. Consider two points $x_1, x_2 \in X$ such that $x_1 < x_2 < p$ and their respective membership degrees $\alpha_1 = A(x_1), \alpha_2 = A(x_2)$. Because of convexity of A, $\alpha_2 = A(x_2) \leq \min\{A(x_1), A(p)\} = A(x_1) = \alpha_1$, hence $\alpha_1 \leqslant \alpha_2$. As a consequence, the restriction A_L is loosely monotonic. Similarly, A_R is loosely monotonic, so A is loosely bi-monotonic.* ■

COROLLARY 4.3 *Gaussian membership functions (and their complements) are strictly bi-monotonic. Triangular and trapezoidal membership functions (with their complements) are loosely bi-monotonic but, in general, not strictly bi-monotonic.*

4.3.2 Optimality conditions

Two sufficient conditions of optimality are provided, for Input Interfaces based on the two classes of bi-monotonic fuzzy sets (strictly and loosely). Such conditions are fulfilled if some mild constraints are satisfied.

THEOREM 4.1 *Let $\Phi = \{A_1, A_2, \ldots, A_n\}$ be a family of one-dimensional strictly bi-monotonic fuzzy sets, with respective distinct prototypes $p_1 < p_2 < \cdots < p_n$ such that $\forall x \in X \exists p_i, p_j : p_i \leq x \leq p_j$.. Then, the corresponding Input Interface $\Im(x) = (A_1(x), A_2(x), \ldots, A_n(x))$ is optimal.*
Proof. *Let $x \in X$ and $\Im(x) = (\pi_1, \pi_2, \ldots, \pi_n)$. Let π_i, π_j be the values of $(\pi_1, \pi_2, \ldots, \pi_n)$ corresponding to two fuzzy sets $A_i, A_j \in \Phi$ with prototypes p_i, p_j such that $p_i \leq x \leq p_j$. Then, $A_i^{-1}(\pi_i) \cap A_j^{-1}(\pi_j) = \{x\}$. Indeed, suppose, ad absurdum, that $A_i^{-1}(\pi_i) \cap A_j^{-1}(\pi_j) = \{y, z\}$, with $y < z$ (note that $x = y \vee x = z$, since x will always compare in the intersection by construction of the defined sets). By Lemma 4.2, we have $y \in A_{iL}^{-1}(\pi_i) \wedge z \in A_{iR}^{-1}(\pi_i)$ and, symmetrically, $y \in A_{jL}^{-1}(\pi_j) \wedge z \in A_{jR}^{-1}(\pi_j)$. As a consequence, the prototypes p_i, p_j are such that $y \leq p_i < p_j \leq z$, that is $x \leq p_i < p_j \vee p_i < p_j \leq x$. This is absurd. Since any two fuzzy sets are sufficient in uniquely determining the original*

input value x, provided that x is included between the corresponding prototypes, the function:

$$\mathfrak{N}(\pi_1, \pi_2, \ldots, \pi_n) = x \Leftrightarrow A_1^{-1}(\pi_1) \cap A_n^{-1}(\pi_n) = \{x\} \tag{4.18}$$

is the inverse function of $\Im$, which is hence optimal. ■

Note that definition of $\mathfrak{N}$ requires only the two extreme components π_1, π_n. This is theoretically acceptable, but in numerical simulations, more numerically robust solutions can be adopted, depending on the shape of the membership functions. As an example, if Gaussian membership functions are adopted, more robust numerical results can be obtained by selecting the two highest values of $(\pi_1, \pi_2, \ldots, \pi_n)$ and proceeding analytically to find the value of x.

THEOREM 4.2 *Let $\Phi = \{A_1, A_2, \ldots, A_n\}$ be a family of one-dimensional loosely bi-monotonic fuzzy sets, such that*

$$\forall A \in \Phi \forall x \in X \exists B \in \Phi : \left| A^{-1}(A(x)) \cap B^{-1}(B(x)) \right| = 1 \tag{4.19}$$

with respective distinct prototypes $p_1 < p_2 < \cdots < p_n$ such that $\forall x \in X \exists p_i, p_j : p_i \leq x \leq p_j$. Then, the associated Input Interface $\Im$ is optimal.

Proof. *The proof is similar to the previous theorem, where the condition $A_i^{-1}(\pi_i) \cap A_j^{-1}(\pi_j) = \{x\}$ is true by hypothesis.* ■

The theorems here proved guarantee optimality of Input Interfaces provided that mild conditions hold. As shown, while for strictly bi-monotonic fuzzy sets such constraints are easy to satisfy, for loosely bi-monotonic the constraints are more stringent and require careful design.

For strictly bi-monotonic fuzzy sets (like Gaussian fuzzy sets), optimality of Input Interfaces can be easily guaranteed if the prototypes of fuzzy sets are distinct, with the leftmost and rightmost prototypes coinciding with minimum and maximum values of the universe of discourse. Moreover, if two fuzzy sets share the same prototype (e.g. one fuzzy set is the subset of the other) only one can be retained for the definition of the interface, thus preserving optimality. Finally, it is noteworthy that the derivation of the inverse function of the fuzzy interface requires - at least theoretically - only the two fuzzy sets corresponding to the leftmost and the rightmost prototypes. Such derivation is independent on the shape of the intermediate fuzzy sets. As a consequence, interface optimality becomes a trivial condition that is always held when strictly bi-monotonic fuzzy sets are used.

For loosely bi-monotonic fuzzy sets, including all convex fuzzy sets in general, interface optimality is guaranteed provided that a stronger condition

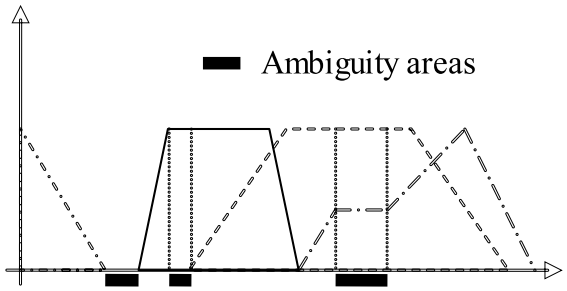

Figure 4.2: Ambiguity areas in loosely bi-monotonic fuzzy sets

holds. In particular, it is required that for each fuzzy set there exists another fuzzy set which eliminates any possible ambiguity when an input is given. While such condition trivially holds for strictly bi-monotonic fuzzy sets, it can be violated for fuzzy sets with "flat" areas (i.e. areas where membership value is constant), like triangular or trapezoidal fuzzy sets (see fig. 4.2 for an example). As a consequence, this condition implies a careful design of the fuzzy interface, since it can be easily violated for particular choices of fuzzy sets parameters. Moreover, in data-driven design of fuzzy systems, the involved learning algorithms must be constrained so as to not violate the optimality conditions. Since such constraints may affect the accuracy of the resulting fuzzy model, strictly bi-monotonic fuzzy sets are recommended in all such modeling scenarios where accuracy is the primarily objective to achieve.

For multi-dimensional bi-monotonic fuzzy sets, interface optimality heavily depends on interface optimality of one-dimensional projections. However, many techniques that automatically generate fuzzy interfaces, like clustering methods, do not assure the fulfillment of conditions that guarantee one-dimensional interface optimality. To overcome such drawback, interpretability - oriented granulation algorithms, like those in (Liao et al., 2003; Abonyi et al., 2002), as well as those presented in successive Chapters, can be effectively applied.

4.4 A measure of optimality for Output Interfaces

Similarly to Input Interfaces, the optimality condition for an Output Interface is satisfied if the interface is invertible, that is:

$$\text{opt}\left(\mathfrak{O}, [0,1]^{K_y}\right) \iff \exists \mathfrak{O}^{-1} \forall \pi_y \in [0,1]^{K_y} : \mathfrak{O}^{-1}\left(\mathfrak{O}\left(\pi_y\right)\right) = \pi_y \quad (4.20)$$

Since the Processing Module cannot provide all possible elements of $[0,1]^{K_y}$ but only a small subset, the optimality condition of an Output Interface can be conveniently redefined as:

$$\text{opt}'\left(\mathfrak{O}, [0,1]^{K_y}\right) \iff \exists \mathfrak{O}^{-1} \forall x \in X : \pi_y = \mathfrak{P}\left(\Im\left(x\right)\right) \rightarrow \mathfrak{O}^{-1}\left(\mathfrak{O}\left(\pi_y\right)\right) = \pi_y \quad (4.21)$$

The last definition restricts optimality condition only for those values of π_y that can be effectively returned by the processing module.

Information Equivalence Criterion is of great importance for the interpretability of the inference process carried out by the fuzzy model. Indeed, in an interpretable inference process, the inferred numerical value should belong to the fuzzy sets of the output Frame of Cognition with the membership degrees coinciding to those provided by the Processing module. This is the sense formalized by the optimality condition 4.21.

Unfortunately, apart from trivial cases optimality of Output Interfaces is hard to achieve. This is essentially due to the aggregation operation of the defuzzification procedure, which is adopted to reduce a highly dimensional structure of membership degrees into low dimensional numerical values (one dimension only in case of single output). However, if an Output Interface is designed so that the inferred values belong to the fuzzy sets of the Output Interface with membership degrees very close to those provided by the Processing Module, it could be still valuable in interpretable fuzzy modelling. This calls for a measure to assess the quality of such interfaces even in case of sub-optimality.

4.4.1 Optimality Degree

To deal with optimality of Output Interfaces, here a measure of optimality for Output Interfaces is introduced, which extends the classical optimality condition based on the Information Equivalence Criterion. The new condition provides an optimality degree of a fuzzy Output Interface ranging from

0 (excluded) to 1. The higher is such degree, the better is the quality of the interface. The maximal optimality degree corresponds to an optimal interface in the classical sense. Conversely, the lower is such degree the less precise is the conversion process performed by the fuzzy interface.

To define the Optimality Degree of an Output Interface as in (4.9), a related Input Interface is considered:

$$\overleftarrow{\mathfrak{O}} : \mathbb{R}^{n_y} \rightarrow [0,1]^{K_y} \tag{4.22}$$

The definition of $\overleftarrow{\mathfrak{O}}$ depends on the form of $\mathfrak{O}$. For Mamdani rules as in (4.12), $\overleftarrow{\mathfrak{O}}$ is defined as follows:

$$\forall y \in \mathbb{R}^{n_y} : \overleftarrow{\mathfrak{O}}(y) = \left[B^{(r)}(y)\right]_{r=1,2,\ldots K_y} \tag{4.23}$$

Once the Input Interface $\overleftarrow{\mathfrak{O}}$ has been derived, its optimality (in the classical sense) can be easily checked. If the related Input Interface $\overleftarrow{\mathfrak{O}}$ is optimal, then $\mathfrak{O}$ is defined as *reversely optimal*. Assuming that the Output Interface is reversely optimal, its optimality degree is computed as follows.

1. When an input x is presented to the fuzzy model, a structure of membership degrees π_y is computed by the Processing Module.

2. The vector π_y is further processed by the Output Interface $\mathfrak{O}$, which provides a numerical value y.

3. Now the related Input Interface $\overleftarrow{\mathfrak{O}}$ is considered, and the structure of membership values $\tilde{\pi}_y = \overleftarrow{\mathfrak{O}}(y)$ is computed for the numerical value y.

4. If $\pi_y = \tilde{\pi}_y$, then the Output Interface is optimal for x. Conversely, the more π_y is different from $\tilde{\pi}_y$, the less optimal is the O-interface.

Given a distance measure $d : [0,1]^{K_y} \times [0,1]^{K_y} \rightarrow \mathbb{R}$ (e.g. the Euclidean distance), the Optimality Degree is hence defined as:

$$od(x) = \exp\left(-d\left(\pi_y, \tilde{\pi}_y\right)\right) \tag{4.24}$$

A value $od(x) \approx 1$ means that the membership values provided by the Processing Module can be reconstructed almost perfectly. When $od(x) \approx 1$, the output y produced by the Output Interface belongs to the fuzzy sets of the output Frame of Cognition with membership degrees very close to those provided by the Processing Module. Conversely, if $od(x) \ll 1$, the inferred

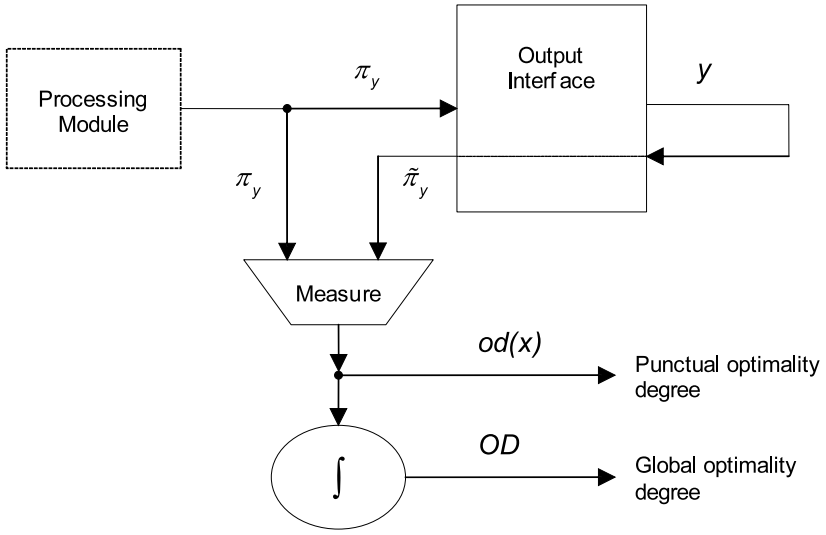

Figure 4.3: The functional block diagram to calculate the optimality degree

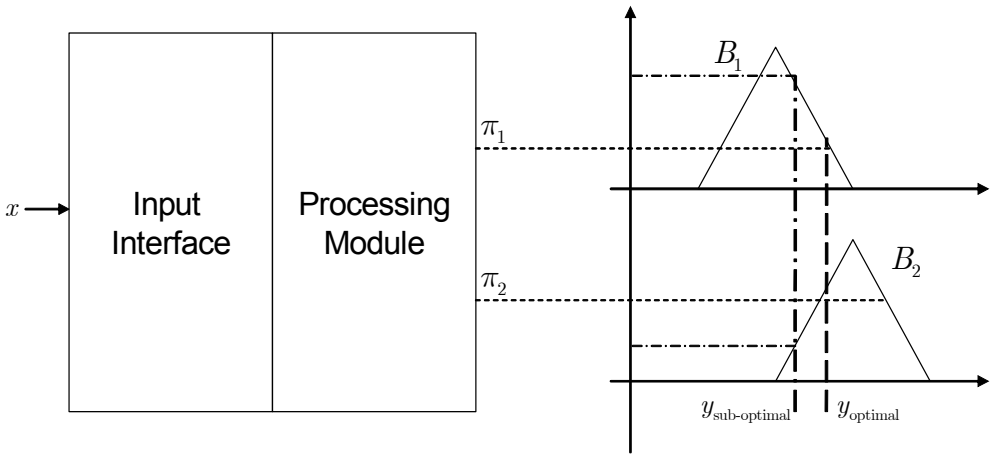

Figure 4.4: Example of optimality degrees of two inferred outputs. The value y_{optimal} has full optimality degree since its membership degrees w.r.t. output fuzzy sets coincide to those provided byt the Processing Module for a given input x. In this sense, the value $y_{\text{sub-optimal}}$ has a very small optimality degree.

output value does not reflect the knowledge expressed in the model's rule base.

The definition of optimal degree in (4.24) is pointwise, so it can be integrated to provide a Global Optimality Degree as follows:

$$OD = \frac{\int_X od\,(x)\,dx}{\int_X dx} \tag{4.25}$$

provided that $\int_X dx$ exists and is finite. Such global measure can be used as a tool to compare different defuzzification methods so as to choose the most suitable defuzzification strategy when designing a fuzzy model. Furthermore, new implementations for output interfaces can be devised with the specific aim of maximizing the optimality degree.

As a final remark, the optimality degree of an Output Interface is influenced by several factors concerning the design of a FIS, such as the choice of the output fuzzy sets, the defuzzification procedure, the Processing Module and the Input Interface. Hence, the optimality degree can be also regarded as a quality index of the fuzzy model under consideration.

4.4.2 Illustrative examples

To show how the proposed optimality degree can be used to evaluate the quality of different Output Interfaces, in this Section several fuzzy models are considered, which differ in the defuzzification procedures.

As a first illustrative example, a very simple Mamdami fuzzy inference system is considered, whose knowledge base is defined by the following two rules:

IF x IS LOW THEN y IS LOW

IF x IS HIGH THEN y IS HIGH

The fuzzy sets associated to the labels LOW and HIGH are identical for the input Universe of Discourse and the output Universe of Discourse, and their membership functions is depicted in fig. 4.5.

The expected behavior of the fuzzy model above defined is a monotonic increasing mapping. More specifically, the inference process would be maximally understandable (interpretable) if, for a given input x, the system responds with the same value, i.e. $y = x$. However, the actual mapping realized by the fuzzy model heavily depends on the defuzzification procedure chosen

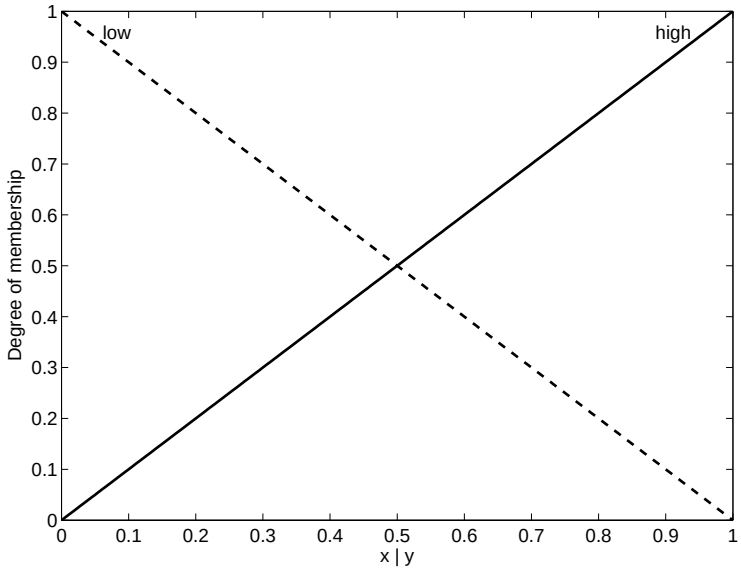

Figure 4.5: The membership functions of the input/output fuzzy sets for the illustrative fuzzy model

to derive a numerical value after the fuzzy inference. In fig. 4.6 it is illustrated how the choice of the defuzzification procedure influences the behavior of the model.

As can be observed, the actual model behavior can be very far from the ideal one. More specifically, the input/output mapping defined by the model is polluted with *spurious nonlinearities* that may heavily hamper the full understanding of the inference process. To evaluate the quality of the output interface in providing a mapping that is more or less close to the ideal behavior, the Optimality Degree can be used. In fig. 4.7 the Optimality Degrees for the Mamdani fuzzy model with different defuzzification procedures are depicted.

As it can be observed from the figures, the Optimality Degree is strictly related to the capacity of the model in providing the expected output. As an example, if the centroid method is used as defuzzification procedure, the mapping realized by the model is a very flat 'S'-shaped graph that is very far from the expected mapping, especially at the extremes of the Universe of Discourse. This behavior is well reflected by the graph of the Optimality Degree, which is maximal at the center of the Universe of Discourse and decreases as the input reaches the limits of its domain.

A very similar behavior is attained when bisector is used as a defuzzi-

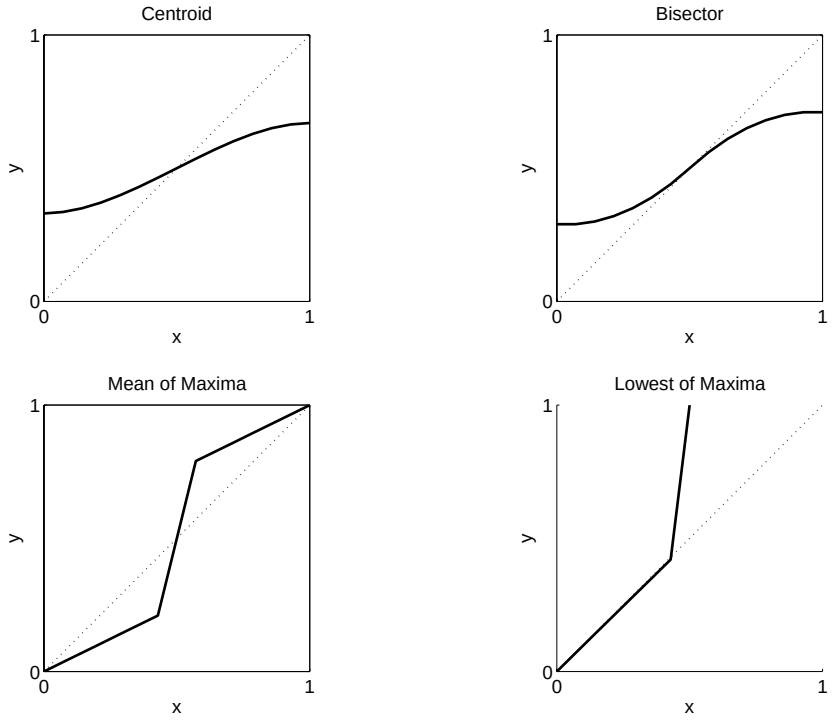

Figure 4.6: The behavior of the fuzzy model according to different choices of defuzzification methods

fication procedure. However, the slope of the graph appears a little closer to the ideal mapping. This slight improvement is captured by the Global Optimality Degree, which is slightly higher for bisector than for centroid.

A completely symmetrical behavior is shown when mean of maxima is used as a defuzzification procedure. In this case, the model mapping is closer to the ideal behavior at the extremes of the Universe of Discourse, while it deteriorates at it center. However, on the overall the Global Optimality Degree shows a better behavior w.r.t. centroid but worse than bisector.

If the lowest of maxima is used as a defuzzification method, an asymmetrical behavior is observed. In the leftmost side of the Universe of Discourse, the mapping carried out by the model is adherent to the ideal behavior, while on the rightmost side the mapping is unacceptably constant with value 1. This behavior is also captured by the trend of the optimality degree. The behavior of the model with supremum of maxima defuzzification method is not shown because it is symmetrical to that emerging from the adoption of lowest of maxima.

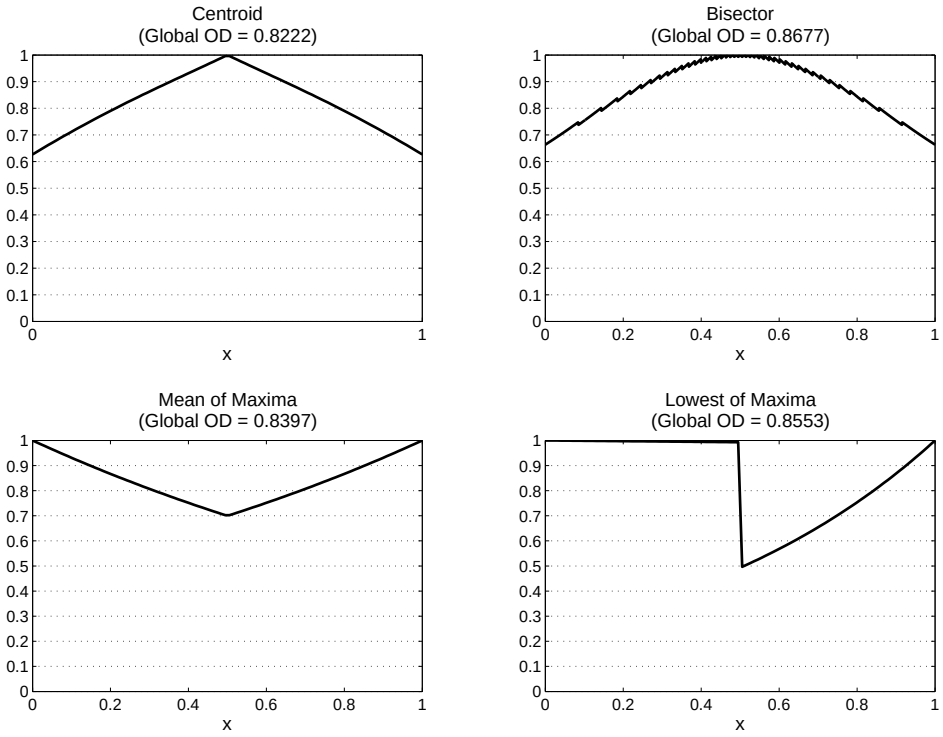

Figure 4.7: Optimality Degrees for the illustrative fuzzy model according to different defuzzification procedures.

In all cases, the quality of the model mapping in terms of its interpretability w.r.t. the knowledge base is well reflected by the graph of optimality degree. It is noteworthy observing that the calculation of the optimality degree does not require the explication of the ideal behavior. As a consequence, it could also be applied to more complex mappings so as to assess, either from a qualitative point of view (through an insight of the optimality degree) or from a quantitative standpoint (through the Global Optimality Degree), the interpretability of the inference process carried out by the model.

An interesting issue concerns the relationship between optimality degree and accuracy of a fuzzy model. To evaluate this relationship, the well-known McKey-Glass chaotic time series identification problem has been considered. The ideal mapping is defined by the following differential equation

$$\frac{dx}{dt} = \frac{0.2x(t-17)}{1+x(t-17)^{10}} - 0.1x(t) \tag{4.26}$$

The identification problem requires the prediction of $x(t+6)$, when $x(t-18)$, $x(t-12)$, $x(t-6)$ and $x(t)$ are given. For this purpose, a dataset of

Table 4.2: Comparison of Optimality Degree and Mean Squared Error for a fuzzy model predicting the McKey Glass time series with different defuzzification methods

Defuzzification Method	**Optimality Degree**	**Mean Squared Error**
Centroid	0.4823	0.0212
Bisector	0.4608	0.0232
MOM/LOM/SOM	0.3832	0.0647

1000 samples has been generated for $t = 118...1117$, with the initial condition $x(0) = 1.2$. The data have been granulated through the application of the Crisp Double Clustering technique, described in Chapter 8. Based on the resulting information granules, a Mamdani fuzzy model has been designed, which consists of three rules with three inputs and one output.

As before, different defuzzification methods are considered for such system. For each defuzzification procedure, the Global Optimality Degree and the Mean Squared Error have been registered as reported in table 4.2. As it can be seen, the higher is the global optimality degree, the more accurate results the fuzzy model. This confirms that the optimality degree can be also regarded as a quality index of the model under consideration.

4.5 Final remarks

In this Chapter, the issue of interface optimality has been addressed. After a formal specification of Input and Output Interfaces, the optimality condition has been analyzed for both types of interfaces. For Input Interfaces, optimality condition can be easily guaranteed if a proper choice of reference fuzzy set is made. In particular, if such fuzzy sets belong to the class of strictly bi-monotonic fuzzy sets (like Gaussian fuzzy sets), optimality condition is easy to guarantee, even if the fuzzy interface is designed by means of data driven methods. On the other hand, Input Interfaces based on loosely bi-monotonic fuzzy sets (e.g. triangular shaped) must be designed carefully if optimality condition must be met.

Differently to Input Interfaces, optimality is hardly satisfied in Output Interfaces, due to the defuzzification methods that implement them. To still deal with the so-called "Information Equivalence Criterion" in Output Interfaces, a measure of optimality for fuzzy interfaces has been proposed. The measure, called Optimality Degree, is defined under a particular condition, i.e. reverse optimality of the Output Interface. The pointwise definition of

the optimality degree highlights the variation of quality of the defuzzification method implementing the Output Interface. The global definition of the optimality degree provides a quantitative estimate of the quality of the FIS, which may be affected by different design choices.

Experimental results have shown that defuzzification methods can be compared in terms of optimality degrees and a high value of such measure is a necessary condition for accurate fuzzy models. As a design tool, the optimality degree can be used to compare different output interfaces (or different defuzzification methods), especially when interpretability if of major concern in the design of the model. Furthermore, Optimality Degree could be used as an objective function that would suggest the investigation of new implementations of Output Interfaces that are specifically aimed at optimizing it.

Part III

Algorithms for Deriving Interpretable Fuzzy Information Granules

Chapter 5

Gaussian Information Granulation

5.1 Introduction

The term "Information Granulation" refers to any task aimed at automatically generating information granules on the basis of a set of observed examples. When the information granules are semantically represented as multidimensional fuzzy sets, the task is called "fuzzy information granulation" and the generated granules "fuzzy information granules"[1]. As already mentioned in the first Chapter, the granulation process is fundamental to any intelligent system with adaptive behavior.

Within the context of interpretable fuzzy modelling, it is necessary that the information granulation process yields information granules that satisfy a set of interpretability requirements. As already observed in the previous chapters, the required interpretability constraints depends on the applicative context, on the desired formal representation of granules and, ultimately, on subjective judgements.

In many applicative contexts, especially in control and system identification, it is highly desirable that information granules provide a vague representation of quantities (e.g. "About 10", "More or Less 2", "Nearly 0", etc.), as well as spacial information (e.g. cities, lands, etc.). To admit this kind of representation, information granules must satisfy a set of interpretability constraints, which include:

- Normality;

[1] Hereafter, as only fuzzy information granulation tasks are analyzed, the adjective "fuzzy" may be dropped for ease of reading.

- Convexity;
- Continuity;

In some cases, the Unimodality constraint may be required if vague quantities represent fuzzy numbers or fuzzy points, while this constraint may be dropped when fuzzy intervals or fuzzy regions are allowed. Other interpretability constraints may be included according to applicative needs and designer judgements, but the three aforementioned constraints constitute the minimal set that is necessary for a proper representation of vague quantities.

The automatic generation of fuzzy information granules can be carried out by means of fuzzy clustering of numerical observations. Fuzzy clustering algorithms are extremely useful in finding hidden relationships among data and hence are widely used for unsupervised learning of empirical knowledge. Unfortunately, in the interpretability context, the adoption of fuzzy clustering algorithms may yield to severe drawbacks that must be taken into account. Indeed, many fuzzy clustering algorithms return a set of prototypes and a partition matrix that contains the membership values of each observation to each cluster (Bezdek, 1981). Such partition does not convey any direct information about fuzzy memberships on the entire Universe of Discourse, since a column of the partition matrix delineates a discontinuous fuzzy set defined only on the available observations.

Furthermore, many clustering algorithms may not produce convex fuzzy sets, thus the association of linguistic labels to the resulting information granules is hard unless further processing is carried out. As an additional difficulty, the multi-dimensional clusters cannot be directly projected onto each dimension without information loss. This drawback further reduces the utility of fuzzy clustering algorithms in interpretable fuzzy modelling. Finally the partition matrix is a structure that may need large memory requirements, since its space complexity is linear in the number of observations and in the number of clusters. This clashes with the simplified representation of knowledge with highly meaningful linguistic terms.

A common approach for the definition of interpretable fuzzy information granules is to define them by interpretable fuzzy sets whose parameters are determined by the information provided by the results of a clustering process. In a simplified setting, however, only the derived prototypes are considered for determining such parameters, while the information provided by the partition matrix is partially or totally ignored. As an example, if interpretable information granules are represented by Gaussian fuzzy sets, the centers of the Gaussian functions are usually made to coincide with the prototypes calculated by the clustering algorithm, while usually no analytical process is adopted to define their widths. Often, heuristic techniques

are applied to define the Gaussian widths (Haykin, 1999), but most of them require the introduction of some user-defined parameters and do not exploit all the information about the fuzzy clusters discovered by the clustering algorithm. Also, the widths are frequently chosen by trial-and-error, and strict assumptions are formulated to simplify the search (e.g. isotropic membership functions with equal widths in all dimensions). The consequence is a large waste of time that sums up with unexploited useful information provided by the clustering process.

In this Chapter, a method to induce fuzzy information granules from data and to represent them by Gaussian functional forms is proposed. First, information granules are extracted from data by a fuzzy clustering algorithm. Then, they are properly represented by Gaussian functions whose widths are determined by solving a constrained quadratic programming problem on membership values returned by the clustering algorithm. The method allows computation of the widths of Gaussian functions by exploiting the information conveyed by the partition matrix of the clustering algorithm. The key advantage of the proposed approach is the ability to automatically find Gaussian representations of fuzzy granules which approximate the membership values in the partition matrix with a small Mean Squared Error. In the proposed approach, any fuzzy clustering algorithm that returns a set of prototypes and the corresponding partition matrix can be adopted. Also, the approach does not require trial-and-error procedures or strong constraints, such as imposing the same width for all the granules (i.e. isotropic Gaussian functions).

In the next Section, the proposed method for representation of information granules induced by fuzzy clustering is introduced. In particular, a Gaussian representation is derived by solving a constrained quadratic programming problem. Then, a rule-based fuzzy model for descriptive modeling is described, as an inference framework in which the derived fuzzy information granules can be applied. Finally, a real-world information granulation problem is considered to validate the proposed method. Also, the section includes the development of a descriptive fuzzy model for MPG (miles per Gallon) prediction benchmark. The Chapter ends with some conclusive remarks.

5.2 A method of Information Granulation with Gaussian fuzzy sets

Gaussian fuzzy sets are especially useful in representing interpretable information granules. Indeed, they satisfy several interpretability requirements – included those considered in this study – and have the added value of a differentiable functional form that is valuable in many adaptive models. An interesting issue concerns the generation of fuzzy information granules semantically represented by Gaussian fuzzy sets by exploiting the results of a clustering process. In this Section, this issue is pursued and an efficient method is proposed.

On a formal level, a fuzzy clustering algorithm can be described as a function that accepts a set of observations and returns a set of prototypes together with a partition matrix. The number of clusters (i.e. information granules) may be predefined or determined by the algorithm. Hence, a generic fuzzy clustering algorithm may be formalized as:

$$fc : \mathbf{U}^N \rightarrow \mathbf{U}^c \times [0,1]^{m \times c} \tag{5.1}$$

for $\mathbf{U} \subseteq \mathbb{R}^n,\ N > 1,\ c \geq 1$ and such that:

$$fc\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right) = \langle P, M \rangle \tag{5.2}$$

where:

$$P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_c] \tag{5.3}$$

is the matrix of all prototypes (one for each column), and:

$$M = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_c] = [m_{ij}]_{\substack{i=1,2,\ldots,N \\ j=1,2,\ldots,c}} \tag{5.4}$$

is the partition matrix, that contains the membership value of each observation to each cluster.

5.2.1 Problem formulation

Given the result of a fuzzy clustering algorithm, the objective of the proposed method is to find a set of Gaussian representations of the discovered clusters, corresponding to the following functional form:

$$\mu_{[\omega,C]}\left(\mathbf{x}\right) = \exp\left(-\left(\mathbf{x}-\omega\right)C^{-1}\left(\mathbf{x}-\omega\right)^T\right) \tag{5.5}$$

where ω is the center and C^{-1} is the inverse of the width matrix. Matrix C^{-1} should be symmetric positive definite (s.p.d.) in order to have a convex shape centered on ω of the function graph. Nevertheless, to achieve interpretability of the resulting granule, the membership function should be projected onto each axis without any loss of information. This can be attained if the amplitude matrix has non-zero elements only in its principal diagonal. Indeed, if C^{-1} is a diagonal matrix, that is:

$$C^{-1} = \operatorname{diag} \mathbf{c} = \operatorname{diag} (c_1, c_2, \ldots, c_n), \ c_i > 0 \tag{5.6}$$

then the fuzzy granule can be represented as product of independent scalar exponential functions[2]:

$$\mu_{[\omega,C]} (\mathbf{x}) = \prod_{i=1}^{n} \mu_{[\omega_i,c_i]} (x_i) = \prod_{i=1}^{n} \exp \left(-c_i (x_i - \omega_i)^2\right) \tag{5.7}$$

where the product can be conveniently interpreted as the conjunction operation of fuzzy sets.

The problem of finding membership parameters can be decomposed into c independent sub-problems that find the best representation for each cluster discovered by the clustering algorithm. Hence, in the following the analysis is focused on a single cluster and the cluster index j is dropped – when unnecessary – for ease of reading.

Generally, there is no an exact solution to the problem, i.e. there is not a pair $\langle \omega, C \rangle$ such that:

$$\forall i : \mu_{[\omega,C]} (\mathbf{x}_i) = m_i \tag{5.8}$$

with the constraint:

$$\forall \mathbf{x} \neq \mathbf{0} : \mathbf{x}^T C^{-1} \mathbf{x} > 0 \tag{5.9}$$

In order to choose the "best" Gaussian representation, some error function has to be defined. Because of the nonlinearity of the equations in (5.8), it is not possible to apply general linear systems theory. On the other hand, the equation system in (5.8) is equivalent to the following:

$$\forall i : -(\mathbf{x}_i - \omega)^T C^{-1} (\mathbf{x}_i - \omega) = \log m_i, \ C^{-1} s.p.d. \tag{5.10}$$

The system (5.10) can be rewritten as:

$$\forall i : \widehat{\mathbf{x}}_i^T C^{-1} \widehat{\mathbf{x}}_i = -\log m_i, \ C^{-1} s.p.d. \tag{5.11}$$

[2]Each unidimensional Gaussian fuzzy set can be reconducted to the canonical form with the definition $\sigma_i = 1/2c_i$

where the center of the Gaussian membership function is put equal to the cluster prototype:

$$\omega = \mathbf{p}_j \tag{5.12}$$

and the following change of variables is done:

$$\widehat{\mathbf{x}_i} = \mathbf{x}_i - \omega \tag{5.13}$$

By imposing C to be positive diagonal, the system can be further simplified as:

$$\forall i : \sum_{k=1}^{n} \widehat{x_{ik}}^2 c_k = -\log m_i,\ c_i > 0 \tag{5.14}$$

where

$$\widehat{\mathbf{x}_i} = [\widehat{x_{ik}}]_{k=1,2,\ldots,n} \tag{5.15}$$

The equations in (5.14) form a constrained linear system; generally, it has not an exact solution, so a constrained least squared error minimization problem can be formulated as follows:

$$\begin{array}{l} \text{minimize: } f(\mathbf{c}) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{n}\widehat{x_{ik}}^2 c_k + \log m_i\right)^2 \\ \text{subject to: } \mathbf{c} > 0 \end{array} \tag{5.16}$$

To solve the minimization problem (5.16) the following matrix is defined:

$$H = \left[\widehat{x_{ik}}^2\right]_{\substack{i=1,2,\ldots,N \\ k=1,2,\ldots,n}} \tag{5.17}$$

then, excluding the constant terms, the problem (5.16) can be restated as:

$$\begin{array}{l} \text{minimize: } f'(\mathbf{c}) = \frac{1}{2}\mathbf{c}^T G \mathbf{c} + \mathbf{g}^T \mathbf{c} \\ \text{subject to: } \mathbf{c} > 0 \end{array} \tag{5.18}$$

where:

$$G = 2H^T H \tag{5.19}$$

and:

$$\mathbf{g} = 2H^T \log \mathbf{m} \tag{5.20}$$

The problem (5.18) can be solved with classical constrained quadratic programming techniques. Usually, quadratic programming algorithms only accept constraints in the form:

$$A\mathbf{c} \geq \mathbf{b} \tag{5.21}$$

In this case, it is useful to express the constraints of the objective function in the form:

$$\mathbf{c} \geq \mathbf{c}_{\min} \tag{5.22}$$

where the vector $\mathbf{c}_{\min}$ defines the maximum admissible amplitudes. If $\mathbf{c}_{\min} = \mathbf{0}$, then all possible amplitudes are admissible, even infinite.

5.2.2 Analysis of the solution

In order to analyze the optimal solution of (5.18) with respect to the original problem setting (5.8), it is useful to rewrite the objective function f as:

$$f(\mathbf{c}) = \frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp\left(-\widehat{\mathbf{x}}_i^T \cdot \operatorname{diag}\mathbf{c} \cdot \widehat{\mathbf{x}}_i\right)}{m_i}\right)^2 \tag{5.23}$$

which is the mean squared log-ratio between the Gaussian membership approximation and the actual membership value assigned by the clustering algorithm. The squared log-ratio is a concave positive function with global minimum in 1 with value 0 (see fig. 5.1).

By expanding the Taylor series of the squared log-ratio with center in 1, it is possible to observe that in a sufficiently small neighbor of point 1, the function is equal to:

$$(\log\xi)^2 = (\xi-1)^2 + O\left((\xi-1)^3\right) \tag{5.24}$$

In such neighborhood, the following approximation can be done:

$$\varepsilon = \left(\log\frac{\mu_{[\omega,C]}(\mathbf{x}_i)}{m_i}\right)^2 \approx \left(\frac{\mu_{[\omega,C]}(\mathbf{x}_i)}{m_i} - 1\right)^2 \tag{5.25}$$

This implies that:

$$\left(\mu_{[\omega,C]}(\mathbf{x}_i) - m_i\right)^2 \approx m_i^2\varepsilon \leq \varepsilon \tag{5.26}$$

As a consequence, if the log-ratio assumes small values, the resulting Gaussian membership function approximates the partition matrix with a small Mean Squared Error. This property validates the proposed approach.

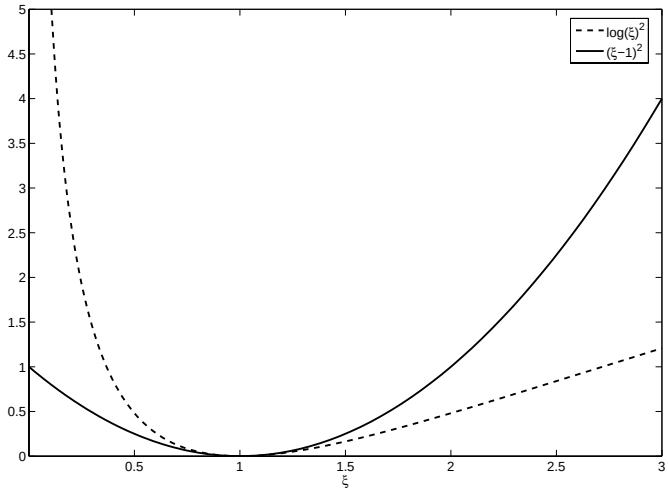

Figure 5.1: The function $(\xi - 1)^2$ (solid) is a second order approximation of $\log^2 \xi$ (dashed) in the neighborhood of 1.

The space complexity of the derived representation is $O(nc)$, while the memory required for storing the partition matrix is $O((M + n)c)$. In this sense, the proposed approach leads to a compact representation of fuzzy granules.

Furthermore, the membership function estimation method is quite general and does not depend on the specific fuzzy clustering algorithm adopted. In the simulations presented forth, the well-known Fuzzy C-Means (FCM) (Bezdek, 1981) is used as basic clustering algorithm, but any other clustering scheme that provides a set of prototypes and a partition matrix can be employed.

5.2.3 Fuzzy model design with Gaussian granules

Information granules represented through the above described method can be properly employed as building blocks of a Fuzzy Inference System, which is a rule-based fuzzy model that can be used to perform inference on a working environment. This is aimed at the characterization of both the descriptive and predictive capability of the model determined on the basis of the identified fuzzy granules.

Hereafter, the classical first-order Takagi-Sugeno fuzzy model is considered (Takagi and Sugeno, 1993). Its rule base is in the form:

$$R_i : \text{IF } Ant \text{ THEN } \mathrm{Y}{=}\mathrm{A}_i^{\mathrm{T}}\mathrm{X}{+}\mathrm{B}_i,\ i = 1, 2, \ldots, c \tag{5.27}$$

where Ant denotes a multi-dimensional fuzzy granule on the input domain, and $\mathrm{A}_i^{\mathrm{T}}$ and B_i denote the the parameters of the local linear model. If the fuzzy granule in the antecedent is defined by a Gaussian membership function as in (8.6) and the amplitude matrix is diagonal, then it can be expressed in an interpretable form, as a conjunction of linguistic labels associated to one-dimensional fuzzy sets defined on the individual components of the input domain:

$$Ant_i \equiv v_1 \text{ IS } FL_{i,v_1} \text{ AND...AND } v_1 \text{ IS } FL_{i,v_1} \tag{5.28}$$

where the membership function of each univariate fuzzy set is in the form:

$$\mu_{A_{ij}}(x_j) = \exp\left(-\left(x_j - \omega_j^i\right)^2 c_j^i\right) \tag{5.29}$$

being $A_{ij} = L_{v_j}^{-1}\left(FL_{i,v_j}\right)$ a Gaussian fuzzy set with center ω_j^i and width $1/2\sigma_j^i$, and c_j^i is the j-th diagonal element of the i-th amplitude matrix.

Given a set of rules as defined in (5.27), the model output is calculated as:

$$\widehat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{c}\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)\left(\mathbf{a}_i^T\mathbf{x} + b_i\right)}{\sum_{i=1}^{c}\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)} \tag{5.30}$$

The model output is linear with respect to linear coefficients $\mathbf{a}_i^T$ and b_i; hence such coefficients can be estimated by a least-square technique when a set of input-output data $T = \{\langle \mathbf{x}_k, y_k\rangle : k = 1, 2, \ldots, N\}$ is given.

The integration of fuzzy granules in a Takagi-Sugeno fuzzy inference model is useful to obtain a fuzzy descriptive model, that can describe experimental data in the language of well-defined, semantically sound and user-oriented information granules.

5.3 Illustrative examples

In order to examine the performance of the proposed granule representation method, two different examples are presented in this section. The first example concerns an information granulation problem and aims to compare the information granules generated by the proposed method with those discovered by the well-known Fuzzy C-Means algorithm. The second example considered is the MPG (miles per gallon) prediction problem, which is a benchmark from the literature. This simulation is performed to verify how much fuzzy granules identified from data through the proposed method are useful in providing good mapping properties when employed as building blocks of fuzzy rules-based models.

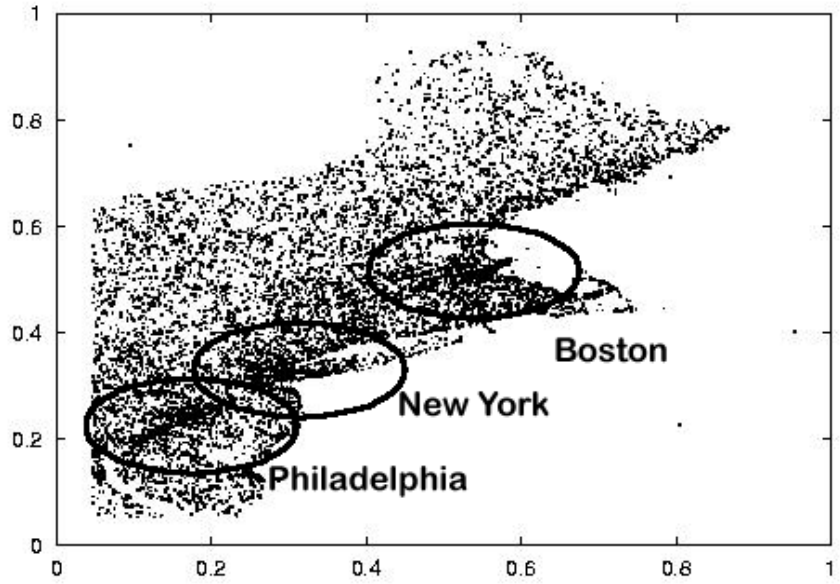

Figure 5.2: The North-East Dataset

5.3.1 Information granulation of postal addresses

As an information granulation problem, the North East dataset is chosen, containing 123,593 postal addresses (represented as normalized coordinates, see fig. 5.2), which represent the three metropolitan areas of New York, Philadelphia and Boston. (Kollios et al., 2003). The dataset can be grouped into three clusters, with a lot of noise, in the form of uniformly distributed rural areas and smaller population centers.

The FCM algorithm has been used to generate three fuzzy clusters from the dataset. Successively, the prototype vector and the partition matrix returned by FCM were used by the proposed method to obtain a Gaussian representation of the three clusters. For FCM and quadratic programming, the Matlab® R11.1 Fuzzy toolbox and Optimization toolbox have been used respectively.

Centers and widths of the derived Gaussian functions are reported in table 5.1. Figures 5.3, 5.4 and 5.5, depict for each cluster both membership values in the partition matrix as grey-levels, and the radial contours of the corresponding Gaussian function.

As it can be seen in the figures, Gaussian granules obtained by the proposed approach properly model some qualitative concepts about the available

Table 5.1: Parameters of the Gaussian Information Granules and Mean Squared Error

	Boston	**New York**	**Philadelphia**
Center	$(0.6027, 0.6782)$	$(0.3858, 0.4870)$	$(0.1729, 0.2604)$
Amplitudes	$(0.0906, 0.1027)$	$(0.0580, 0.0606)$	$(0.1013, 0.1151)$
MSE	0.0360	0.0203	0.0347

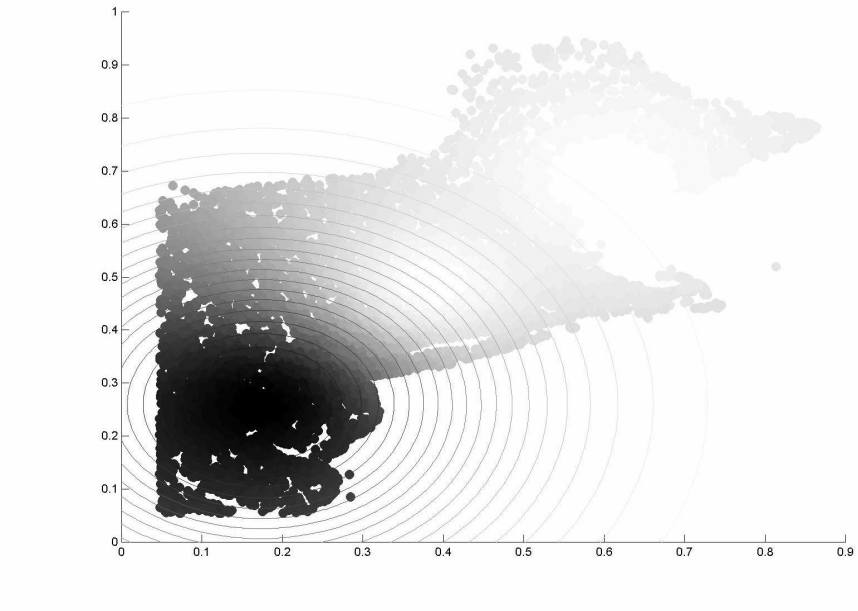

Figure 5.3: Fuzzy granule for Philadelphia city and its Gaussian representation

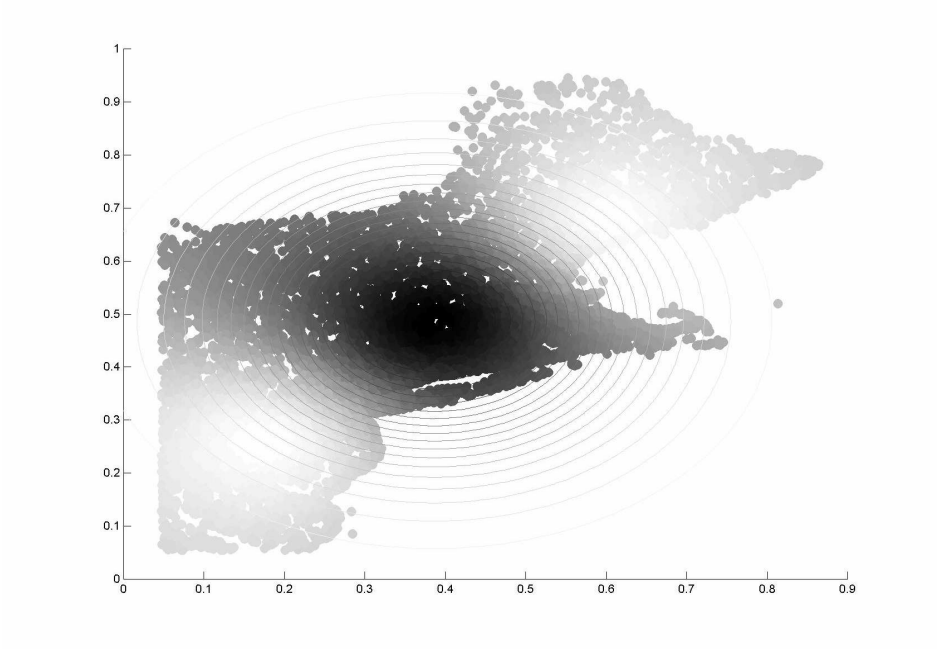

Figure 5.4: Fuzzy granule for New York city and its Gaussian representation

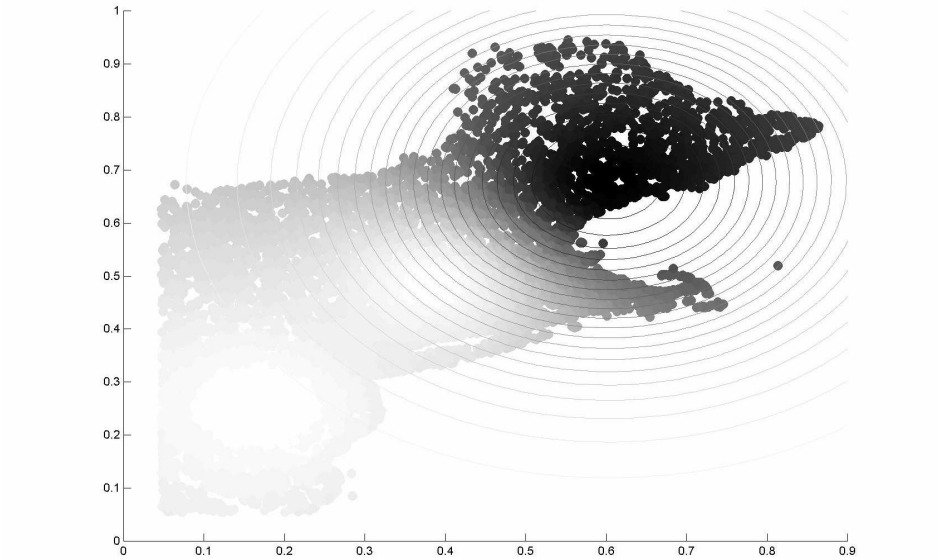

Figure 5.5: Fuzzy Granule for Boston city and its Gaussian representation

data. Specifically, regarding each granule as one of the three metropolitan areas (Boston, New York, Philadelphia), membership values of postal addresses can be interpreted as the degree of closeness to one city (cluster prototype). Such concept is not easily captured with clusters discovered by FCM alone, since, as the figures illustrate, the membership values of the addresses do not always decrease as the distances from the prototype cluster increase.

Also, table 5.1 reports the Mean Squared Errors (MSE) between Gaussian granules and fuzzy clusters, defined as:

$$\varepsilon_j = \frac{1}{N} \sum_{i=1}^{N} \left(\mu_{[\omega_j, C_j]} (\mathbf{x}_i) - m_i \right)^2 \tag{5.31}$$

The low values of MSE for each granule, demonstrate how well the resulting Gaussian membership functions approximate the partition matrix of FCM.

In order to evaluate quantitatively the derived Gaussian information granules, the Xie Beni index has been used as compactness and separation validity measure (Xie and Beni, 1991). Such measure is defined as:

$$S = \frac{\sum_{j=1}^{c} \sum_{i=1}^{N} \vartheta_{ij}^2 \left\| \mathbf{p}_j - \mathbf{x}_i \right\|^2}{N \min\limits_{i,j} \left\| \mathbf{p}_j - \mathbf{p}_i \right\|^2} \tag{5.32}$$

where:

$$\vartheta_{ij} = \begin{cases} m_{ij}, \text{ for FCM clusters} \\ \mu_{[\omega_j, C_j]} (\mathbf{x}_j), \text{ for Gaussian granules} \end{cases} \tag{5.33}$$

In other words, the Xie-Beni index for the FCM clusters has been directly computed on the partition matrix returned by the clustering algorithm. Conversely, for Gaussian granules the measure has been computed by re-calculating the membership values of each observation of the dataset with the derived Gaussian membership functions.

Table 5.2 summarizes a comparison between fuzzy granules extracted by FCM alone, and those obtained by the proposed approach, in terms of Xie-Beni index, number of floating point operations (FLOPS) and time/memory requirements on a desktop workstation[3].

As it can be seen, the Xie-Beni index values for the Gaussian granules and FCM clusters are comparable. The slight difference is due to the nature of the proposed method that generates convex (Gaussian) approximations

[3]The CPU was an Intel™ Pentium® III @500MHz and the memory capacity 128MB.

Table 5.2: Performace Measurements

Measure	**FCM**	**Gaussian Granulation**
Xie-Beni index	0.1665	0.2687
FLOPS	792M	14.1M
Time required	138.1 s	14.7 s.
Memory required	2,966,280 B	144 B

for the partition matrix, which is generally not convex, i.e. it assumes high values even for points very distant from the prototype (see figs. 5.3-5.5).

The time required for representing granules with Gaussian functional forms is negligible compared to the time required for FCM, hence the total computational cost of the proposed method (FCM + Gaussian representation) is comparable with FCM alone. More important, the method provides a compact representation of the granules. Indeed, each Gaussian granule is fully described only with a prototype vector and a diagonal width matrix. As a consequence, once granules have been represented by Gaussian functions, the partition matrix can be discarded, thus saving a large amount of memory.

5.3.2 Prediction of automobile fuel consumption

The previous case study shows that an efficient and compact representation of information granules can be obtained by the proposed method. However, the real advantage of the granulation method, i.e. interpretability, was not clearly shown. This will be done through the following problem.

A fuzzy model for the prediction of the fuel consumption of an automobile has been constructed as proposed in the previous section on the basis of the data provided in the Automobile MPG dataset (Blake and Merx, 1998). The original dataset has been cleaned, by removing all examples with missing values and excluding the attributes in discrete domains, which is common with other studies. The resulting dataset consists of 392 samples described by the following attributes: 1) displacement; 2) horsepower; 3) weight; 4) acceleration; 5) model year. In the simulation, a 10-fold cross validation was carried out, i.e. ten different splitting of the dataset were considered, in order to attain statistically significant results.

For each fold, FCM has been used to generate two fuzzy clusters from the training data. Successively, the prototype vector and the partition matrix returned by FCM were used to obtain a Gaussian representation of the two clusters. Moreover, the clustering process has been repeated ten times, for each fold, with different initial settings of FCM. This leads to 100 different

Table 5.3: Average RMSE of the 10 Takagi-Sugeno models identified from each fold. The total mean of RMSE is reported in the last row

Fold	**Average Error on Training set**	**Average Error on Test set**
1	3.05	2.55
2	2.58	3.06
3	2.81	2.82
4	2.86	2.83
5	2.71	2.92
6	2.71	3.01
7	2.67	2.96
8	2.66	2.94
9	2.90	2.75
10	2.88	2.74
Mean	*2.78*	*2.86*

granulations of the data in total.

Then, using the information granules resulting from each granulation process, a Takagi-Sugeno model with two rules has been built as previously described. The prediction ability of the identified fuzzy models has been evaluated both on the training set and the test set in terms of root mean squared error (RMSE):

$$\mathrm{RMSE}\,(T) = \sqrt{\frac{1}{|T|}\sum_{\langle \mathbf{x},y\rangle \in T} \left(y - \widehat{y}\,(\mathbf{x})\right)^2} \tag{5.34}$$

Results of such experiments are summarized in table 5.3 which reports, for each fold, the average RMSE of the 10 TS models identified. Also, the total mean of RMSE is reported, as an estimate of the prediction capability of the fuzzy inference model based on information granules derived by the proposed approach. This value shows a good predictive accuracy for such a fuzzy model as compared with previously reported results. Indeed, very similar results are reported by Abonyi, Babuška and Szeifert (Abonyi et al., 2002), where a modified version of Gath-Geva clustering that extracts Gaussian clusters with diagonal amplitude matrices is proposed. With a two-rule model, such a method achieved RMSE value of 2.72 and 2.85 for training and test data. However, it should be noted that these results derive from a single partition of the dataset, and hence they are not statistically significant.

For a further comparison, fuzzy models with only two input variables, namely the Weight and the Year, as suggested by Jang (Jang, 1996), were

Table 5.4: Comparison of the performance (RMSE) of Takagi-Sugeno models with two input variables

	2 rules		**4 rules**	
Method	**train**	**test**	**train**	**test**
Gaussian Granulation	*2.94*	*3.18*	*2.81*	*2.97*
ANFIS	2.67	2.95	2.37	3.05
FMID	2.96	2.98	2.84	2.94
EM-NI	2.97	2.95	2.90	2.96

also identified with two and four rules. In table 5.4 report the RMSE is reported on training and test data averaged on the 100 runs of 10-fold CV using 10 different partitions of the dataset into 10 subsets. Also, the table shows the results reported by Abonyi et al. (Abonyi et al., 2002)for fuzzy models with the same number of rules and inputs obtained by the Matlab fuzzy toolbox (ANFIS model (Jang, 1993)), the fuzzy model identification toolbox FMID (Babuška, 1998) and the method EM-NI introduced in Abonyi et al. (Abonyi et al., 2002). Again, it should be noted that only the results on Gaussian granulation were obtained from 10-fold cross validation, while results of other methods were produced from a single run on a random partition of the data, thus providing a less feasible estimate of the prediction accuracy. Therefore, results given in table 5.4 are only roughly comparable. Moreover, unlike the proposed method, both ANFIS and FMID pursue only accuracy as ultimate goal and take no care about the interpretability of the representation. The issue of interpretability is addressed by EM-NI, which produces clusters that can be projected and decomposed into easily interpretable membership functions defined on the individual input variables. However, as the authors state in their work, this constraint reduces the flexibility of the model produced by EM-NI, which can result in slightly worse prediction performance. The proposed approach gives good results in terms of predictive accuracy while preserving the descriptive property of the derived granules.

This last property can be illustrated graphically for models with only two inputs. In figs. 5.6 and 5.7, the derived Gaussian representation for two fuzzy granules is depicted. Particularly, the figures show the clusters discovered by FCM form the data, with a scatter graph where each point corresponds to a training set example. The brightness of each point is proportional to the membership value of the sample in the partition matrix: the darker the point, the higher the membership value. For each granule, continuous lines represent contour levels of the derived Gaussian representation. The gray level of each contour line represents an average of the membership values

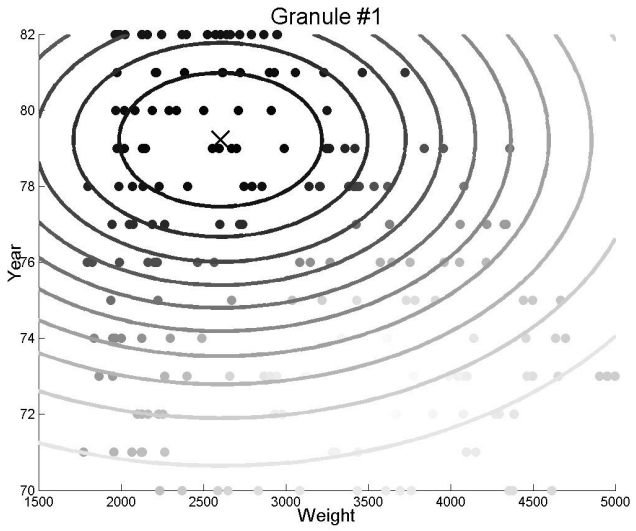

Figure 5.6: First granule representation

relative to the points laying in the contours neighbors.

These fuzzy information granules can be easily projected onto each axis, yielding interpretable Frames of Cognition defined on each input feature, as depicted in figs. 5.8 and 5.9. As it can be seen, for each variable, the two fuzzy sets are represented by very distinct membership functions that turn out to be nicely interpretable. As a consequence, each one-dimensional fuzzy set can be easily associated to a quantitative linguistic label, such as "About 2500 Kg", "About 3500 Kg", "About 1973", "About 1979". Such linguistic labels can be used in the formulation of fuzzy rules, as the followings:

> R1: IF weight is About 2500 Kg AND year is About 1973 THEN mpg = f1(weight, year)
>
> R2: IF weight is About 3500 Kg AND year is About 1979 THEN mpg = f2(weight, year)

where:

$$f_1(w,y) = \frac{-4.124}{1000}w + \frac{27.24}{100}y + 12.62 \tag{5.35a}$$

$$f_2(w,y) = \frac{-10.53}{1000}w + \frac{60.68}{100}y + 9.069 \tag{5.35b}$$

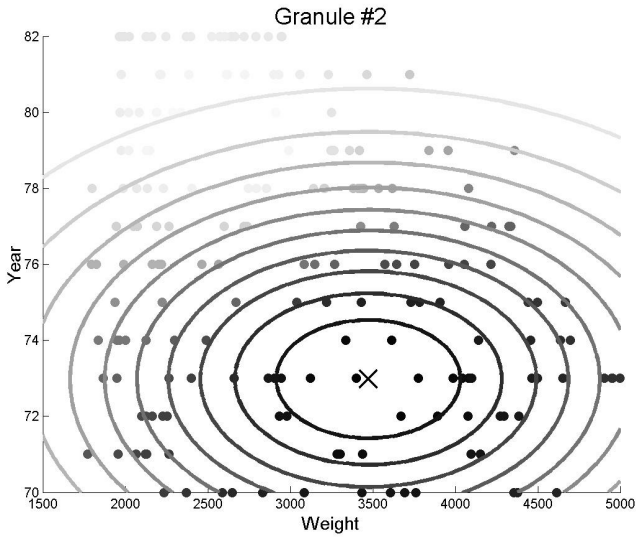

Figure 5.7: Second granule representation

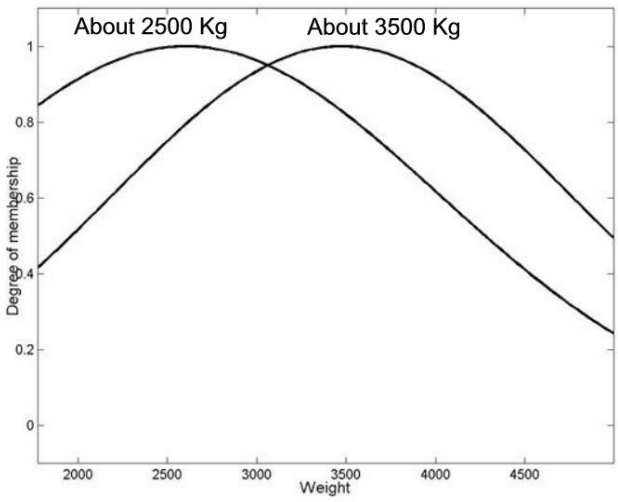

Figure 5.8: The Frame of Cognition for the "Weight" variable

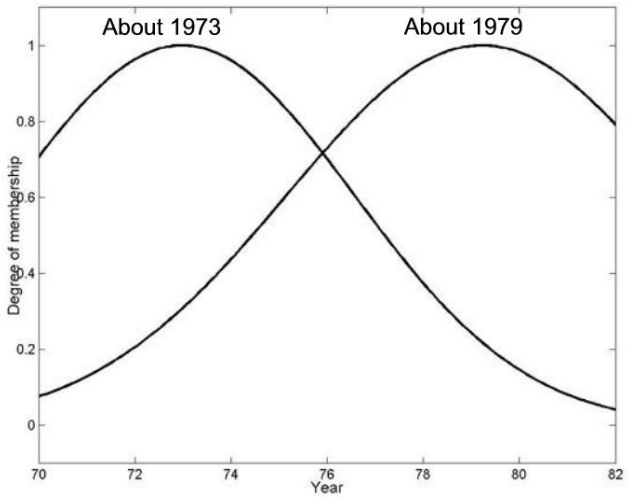

Figure 5.9: The Frame of Cognition for the "Year" variable

5.4 Final remarks

The method proposed in this Chapter is able to efficiently generate Information granules represented by Gaussian fuzzy sets by exploiting the results of a generic fuzzy clustering algorithm. Its key features of efficiency and effectiveness suggest its use as a wrapper procedure that is able to provide interpretable information granules when a fuzzy clustering process is available. The derived granules have good features in terms of fine approximation and compact representation, providing a sound a comprehensible description of the experimental data. Moreover, they are very robust against noise, as application of the method to real-world examples showed. Finally, the information granules derived by the proposed approach can be usefully integrated in most rule-based fuzzy models to perform fuzzy reasoning about the working environment, as showed in the prediction problem, for which Takagi-Sugeno fuzzy inference systems are built.

In table **??** a number of applicable interpretability constraints are listed, showing their fulfillment when Gaussian Granulation is applied. As shown, information granules resulting from the application of the proposed method verify a number of interpretability constraints that are useful in representing quantitative information. Also, the representation of information granules by means of Gaussian fuzzy sets enables the fulfillment of interpretability constraints, such as Error Free Reconstruction and Number of Rules, which

Table 5.5: Table of fufillment of interpretability constraints by Gaussian Granulation (only applicable constraints are shown).

Interpretability constraints fulfillment	
Normality	Yes
Convexity	Yes
Unimodality	Yes
Continuity	Yes
Proper ordering	Weak
Justifiable no. of elements	Yes
Distinguishability	No
Completeness	Yes
Uniform granulation	Yes (*)
Leftmost/Rightmost fuzzy sets	No
Natural zero positioning	No
Error Free Reconstruction	Yes
Description length	No
Rule length	No
Number of rules	Low
Number of firing rules	High
Shared fuzzy sets	No
Model completeness	Yes
(*) The fulfillment of this constraint depends on the fuzzy clustering algorithm. When FCM is used, the derived clusters have roughly a spherical shape with similar radius, which involves uniform granulation.	

are useful in modelling applications. However, the proposed method is not convenient for deriving fuzzy granules representing qualitative information, because some essential interpretability constraints – like Leftmost/Rightmost fuzzy sets, Distinguishability and Shared fuzzy sets– are not guaranteed. Therefore, for qualitative information granulation a different approach is needed, like the one described in Chapter 8.

Chapter 6

Minkowski Information Granulation

6.1 Introduction

In the previous Chapter, a method has been proposed with the specific aim of generating information granules representing vague numbers or vectors. In other applicative contexts, however, a different type of granular information may be required, as one representing intervals or hyper-boxes[1] (either crisp or fuzzy). In this case, Gaussian granulation may not be convenient since the generated granules represent fuzzy vectors that have an elliptical – rather than boxed – distribution of membership degrees (see fig. 6.1).

To generate box-shaped fuzzy information granules, a different approach for information granulation is proposed in this Chapter. The described approach is based on that presented in (Bargiela and Pedrycz, 2003b) and consists in decomposing complex topologies of data structures through hierarchical partitioning of the structures into core and residual parts. The cores are represented as granular prototypes characterized by a transparent boxlike shape that can be interpreted as a decomposable relation among data features.

To generate the granular prototypes, the Minkowski Fuzzy C-Means has been investigated as a tool that allows a degree of control over the geometry of the information granules identified through clustering. The transition from hyper-ellipsoids to hyper-boxes, prompted by the change of the order of the Minkowski distance from 2 to ∞, is somewhat distorted by two concurrent factors subject of study: the Minkowski order and the interactions among clusters.

[1] A hyper-box (or hyper-interval) is the Cartesian product of one or more intervals.

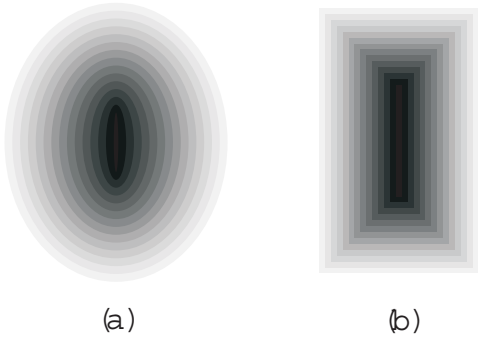

Figure 6.1: Two distributions of membership degrees: (a) elliptical distribution; (b) boxed distribution.

The first deformation factor is due to the estimation error introduced when approximating the Tchebychev distance – which characterizes boxlike granules – with Minkowski distances. To investigate such phenomenon, a first distortion measure is formalized to quantify the approximation error in relation to the Minkowski order. Experimental results show that such distortion measure quickly decreases as the Minkowski order increases, but calculi with too high orders are hampered by perturbation errors involved by finite precision arithmetic. As a consequence the distortion measure can be used as a tool to estimate the best Minkowski order that allows the approximation to the Tchebychev on a given machine.

The second deformation effect is due to interactions among clusters, which distort the boxlike shape of clusters to meet the constraints required by Fuzzy C-Means. To deal with such problem, a second deformation measure has been formalized. This measure evaluates the distortion of each cluster α-cut with an associated "ideal" hyper-box. With the aid of such measure, boxlike granular prototypes can be generated within an accepted deformation threshold.

In the next section, the Minkowski Fuzzy C-Means is described in detail. Then, the approach for extracting interpretable information granules is formalized, and a study on the deformation effects is presented. Finally, some illustrative material is portrayed so as to exemplify the key results of the study from a practical viewpoint.

6.2 A method for Information Granulation through Minkowski fuzzy clustering

In this Section, the proposed method for information granulation through Minkowski fuzzy clustering is described. First, the Minkowski Fuzzy C-Means is formalized and its algorithmic issues are analyzed. Then, the method for information granulation is described with a number of theoretical considerations.

6.2.1 Minkowski Fuzzy C-Means

Let be a set of data patterns to be clustered into fuzzy partitions by minimizing the following objective function:

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} m_{ik}^2 d_{ik}^{(p)} \tag{6.1}$$

where $[m_{ik}]_{i=1,2,\ldots,c}^{k=1,2,\ldots,N}$ is the partition matrix, which is subject to:

$$\forall k \;:\; \sum_{i=1}^{c} m_{ik} = 1 \tag{6.2a}$$

$$\forall i,k \;:\; 0 \leq m_{ik} \leq 1 \tag{6.2b}$$

$$\forall i \;:\; 0 < \sum_{k=1}^{N} m_{ik} < N \tag{6.2c}$$

The distance between a pattern $\mathbf{x}_k$ and a cluster prototype $\mathbf{v}_i$ is denoted by $d_{ik}^{(p)}$ and is defined as:

$$d_{ik}^{(p)} = \sqrt[p]{\sum_{j=1}^{n} |x_{kj} - v_{ij}|^p} \tag{6.3}$$

being p the order of the Minkowski distance $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kn})$, and $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{in})$. The objective of Minkowski Fuzzy C-Means is to find a proper choice of the partition matrix $[m_{ik}]$ and prototypes $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c$ such that (6.1) is minimal while satisfying constraints in (6.2).

Such minimization problem has been tackled specifically for Minkowski distances in (Groenen and Jajuga, 2001). There, the authors propose a gradient-free minimization method called "iterative majorization" (Groenen et al., 1995). However, such method cannot be applied for $p > 2$, it is

not suitable to generate boxlike information granules. As a consequence, an extension of the classical Fuzzy C-Means procedure has been devised to fit Minkowski distance. Fuzzy C-Means is standard to a high extent and is based on alternative optimization of the partition matrix $[m_{ik}]$ and the set of prototypes $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c$ (Bezdek, 1981). These two steps are repeated iteratively until no significant change of the objective function is registered. The update formula of the partition matrix does not depend on the chosen distance function and is derived by means of Lagrange multipliers. After some straightforward calculations, the following expression can be obtained:

$$m_{st}\left[\tau+1\right] = \frac{1}{\sum_{j=1}^{c} \frac{d_{st}^{(p)}[\tau]}{d_{jt}^{(p)}[\tau]}} \tag{6.4}$$

Here, the variable τ denotes the iteration number. Differently from the partition matrix, the derivation of the update formula for the cluster prototypes depends on the analytical form of the distance function. Specifically, the update formula is determined by solving the following equation system:

$$\frac{\partial Q}{\partial v_{st}} = 0, \quad \begin{matrix} s = 1, 2, \ldots, c \\ t = 1, 2, \ldots, n \end{matrix} \tag{6.5}$$

Such equation system has a simple solution for $p = 2$ while cannot be solved analytically for a generic order p. For such reason, the update formula for each prototypes follows the gradient-descent scheme:

$$v_{st}\left[\tau+1\right] = v_{st}\left[\tau\right] - \alpha \frac{\partial Q}{\partial v_{st}} \tag{6.6}$$

Such scheme introduces a novel hyper-parameter α whose value is critical for the convergence of the iterations. However, based on empirical observations, the following choice always leads to convergence:

$$\alpha \propto \frac{1}{N} \tag{6.7}$$

The derivatives of the objective function with respect to each prototype are calculated as follows:

$$\frac{\partial Q}{\partial v_{st}} = \sum_{k=1}^{N} m_{sk}^2 \frac{\partial}{\partial v_{st}} d_{sk}^{(p)} \tag{6.8}$$

where:

$$\frac{\partial}{\partial v_{st}} d_{sk}^{(p)} = -\left(d_{sk}^{(p)}\right)^{1-p} \cdot \left|x_{kt} - v_{st}\right| \cdot \operatorname{sgn}\left(x_{kt} - v_{st}\right) \tag{6.9}$$

being sgn the sign function.

Although the case of $p = +\infty$ is not included in such computation, it should be remarked that the Tchebychev distance can be handled through a fuzzy logic approximation of the 'max' operation as discussed in (Castellano et al., 2004). In particular, for $p \gg 1$ the corresponding Minkowski distance quickly approaches the Tchebychev distance, hence the Minkowski Fuzzy C-Means can be applied to derive box-shaped information granules by using a gradient descent scheme.

In summary, the clustering algorithm can be described as a sequence of the following steps:

1. `Randomly define prototypes`
2. `Repeat`
   (a) `Compute the partition matrix according to (4)`
   (b) `Compute prototypes according to (6)`
3. `Until a termination criterion is satisfied`

The most commonly adopted termination criterion entails the variation of the objective function. Specifically, it is met when:

$$Q[\tau + 1] - Q[\tau] < \varepsilon \tag{6.10}$$

being ε a user-defined threshold. Note that while the update formula of the partition matrix is the same as for the standard Fuzzy C-Means, the update of the prototypes is more time consuming.

6.2.2 Generation of Information Granules

The Minkowski order p used for clustering is user defined and is chosen so as to achieve a control over the shape of the information granules. It should be remarked that as $p \to +\infty$, the distance function approaches the Tchebychev distance defined as:

$$d_{ik}^{(\infty)} = \max_{j=1,2,\ldots,n} |x_{kj} - v_{ij}| \tag{6.11}$$

Such distance function is well suited to identify box-shaped information granules that can be represented as a decomposable relation $\mathbf{R}$ identified as a Cartesian product of one-dimensional information granules:

$$\mathbf{R} = R_1 \times R_2 \times \cdots \times R_n \tag{6.12}$$

Such decomposability leads to transparent information granules and can be effectively adopted in data mining contexts. On the other hand, for $p = 2$, the Minkowski Fuzzy C-Means reduces to the classical Fuzzy C-Means based on Euclidean distance, which is used to extract ellipsoidal-shaped information granules that are not decomposable in the aforementioned sense. As a consequence, an interesting study concerns the quantification of the deformation effects of information granules in relation to the Minkowski order.

Once the final partition matrix has been computed, c multidimensional fuzzy sets $\mathbf{G}_1, \ldots, \mathbf{G}_c$ can be defined by the following membership functions:

$$m_i(\mathbf{x}) = \begin{cases} \delta_{ij}, \text{ if } \mathbf{x} = \mathbf{v}_j \text{ for some } j \\ \left(\sum_{j=1}^{c} \frac{d^{(p)}(\mathbf{x}, \mathbf{v}_i)}{d^{(p)}(\mathbf{x}, \mathbf{v}_j)} \right)^{-1}, \text{ otherwise} \end{cases} \tag{6.13}$$

being:

$$d^{(p)}(\mathbf{x}, \mathbf{v}) = \sqrt[p]{\sum_{j=1}^{n} |x_j - v_j|^p} \tag{6.14}$$

and δ_{ij} the Kronecker's symbol. Definition (6.13) is a direct extension of equation (6.4) in a functional form.

In a small neighborhood of $\mathbf{v}_i$, the contour levels of the i-th fuzzy set have approximately the same shape of the distance function. Indeed, if for $i \neq j : d^{(p)}(\mathbf{x}, \mathbf{v}_i) \ll d^{(p)}(\mathbf{x}, \mathbf{v}_j)$, then:

$$m_i(\mathbf{x}) \approx 1 - \beta d^{(p)}(\mathbf{x}, \mathbf{v}_i) \tag{6.15}$$

being β a positive constant. As the distance between $\mathbf{x}$ and increases $\mathbf{v}_i$, the interactions amongst the different clusters become stronger, and the shape of the contour levels will distort significantly.

By virtue of (6.15), for $p \gg 1$ the contour levels of the fuzzy sets assume the shape of a hyper-box in the neighborhood of each prototype. To assess the distortion involved in approximating the Tchebychev distance with a Minkowski distance, the following measure can be adopted:

$$\Delta^{(p)} = \sum_{i=1}^{c} \int_{\mathbf{U}} |m_i(\mathbf{x}) - \overline{m_i}(\mathbf{x})| \tag{6.16}$$

being:

$$\overline{m_i}(\mathbf{x}) = \begin{cases} \delta_{ij}, \text{ if } \mathbf{x} = \mathbf{v}_j \text{ for some } j \\ \left(\sum_{j=1}^{c} \frac{d^{(\infty)}(\mathbf{x}, \mathbf{v}_i)}{d^{(\infty)}(\mathbf{x}, \mathbf{v}_j)} \right)^{-1}, \text{ otherwise} \end{cases} \tag{6.17}$$

The measure $\Delta^{(p)}$ is theoretical and cannot be evaluated in any analytical way. Nevertheless, it can be estimated as:

$$D^{(p)} = \frac{1}{N} \sum_{i=1}^{c} \sum_{k=1}^{N} \left| m_i \left(\mathbf{x}_k \right) - \overline{m_i} \left(\mathbf{x}_k \right) \right| \tag{6.18}$$

From a theoretical point of view, it is expected that:

$$\lim_{p \to \infty} D^{(p)} = 0 \tag{6.19}$$

However, when the Minkowski Fuzzy C-Means is executed on a digital computer, it is possible that for very high values of p numerical errors significantly perturb the convergence (6.19). As a consequence, the measure (6.18) can be used to evaluate the maximum value of p that allows the best approximation of the Tchebychev distance on a given digital computer.

Once the most suitable value p of has been selected, the geometric properties of the information granules can be studied. As an example, in (Bargiela and Pedrycz, 2003b) an approach is proposed to evaluate the deformation of membership grades due to cluster interactions when box-shaped information granules are required. To quantify such deformation, the γ-cuts of each fuzzy set are considered:

$$\left[\mathbf{G}_i \right]_\gamma = \left\{ \mathbf{x} : m_i \left(\mathbf{x} \right) \geq \gamma \right\} \tag{6.20}$$

For each γ-cut, an "ideal" hyper-box is built according to[2]:

$$\left[\Gamma \right]_\gamma = \left[\mathbf{v}_i - \mathbf{l}_i^{(\gamma)}, \mathbf{v}_i + \mathbf{s}_i^{(\gamma)} \right] \tag{6.21}$$

being $\mathbf{l}_i^{(\gamma)} = \left(l_{i1}^{(\gamma)}, l_{i2}^{(\gamma)}, \ldots, l_{in}^{(\gamma)} \right)$ and $\mathbf{s}_i^{(\gamma)} = \left(s_{i1}^{(\gamma)}, s_{i2}^{(\gamma)}, \ldots, s_{in}^{(\gamma)} \right)$ such that:

$$\overline{m_i} \left(\mathbf{v}_i - l_{ij}^{(\gamma)} \mathbf{e}_j \right) = \overline{m_i} \left(\mathbf{v}_i + s_{ij}^{(\gamma)} \mathbf{e}_j \right) = \gamma \tag{6.22}$$

for each $j = 1, 2, \ldots, n$, and:

$$\mathbf{e}_j = \left(\underbrace{0, 0, \ldots, 0}_{j-1}, 1, \underbrace{0, 0, \ldots, 0}_{n-j} \right) \tag{6.23}$$

[2] A hyper-box is represented as $[\mathbf{a}, \mathbf{b}]$ and is defined as the Cartesian product $[a_1, b_1] \times \cdots \times [a_n, b_n]$

To avoid unlimited hyper-boxes, the value of γ must belong to the open interval $\left]\frac{1}{c}, 1\right]$. Based on the definition of $[\Gamma]_\gamma$, the deformation is measured as follows:

$$\phi_i(\gamma) = \sum_{\mathbf{m} \in \mathbf{V}_j^{(\gamma)}} |\gamma - m_i(\mathbf{m})| \tag{6.24}$$

being $\mathbf{V}_j^{(\gamma)}$ the set of all vertices of the hyper-interval $[\Gamma]_\gamma$:

$$\mathbf{V}_j^{(\gamma)} = \left\{ \mathbf{v}_i + (o_1, o_2, \ldots, o_n) : o_j = -l_{ij}^{(\gamma)} \vee o_j = s_{ij}^{(\gamma)} \right\} \tag{6.25}$$

The analysis of the functions ϕ_i helps in the definition of box-shaped granular prototypes. As an example, a threshold $\phi_{\max}$ can be fixed and the corresponding value $\gamma_{\max}^{(i)}$ such that:

$$\phi_i\left(\gamma_{\max}^{(i)}\right) = \phi_{\max} \tag{6.26}$$

can be considered. The value $\gamma_{\max}^{(i)}$ can be then used to define the granular prototypes as:

$$\mathbf{B}_i = \left[\mathbf{v}_i - \mathbf{l}_i^{\left(\gamma_{\max}^{(i)}\right)}, \mathbf{v}_i + \mathbf{s}_i^{\left(\gamma_{\max}^{(i)}\right)} \right] = \left[\mathbf{b}^{\min}, \mathbf{b}^{\max} \right] \tag{6.27}$$

The derived granular prototypes serve as "focal points" of the structure underlying data and can be employed as building blocks for further information processing. Indeed, the following relation holds:

$$[0,1]^n = \Re \cup \bigcup_{i=1}^{c} \mathbf{B}_i \tag{6.28}$$

where $\Re$ is the residual part of the data domain that is not covered by any information granule. By virtue of relation (6.28), the input domain is decomposed in a "core" part, defined by transparent granular prototypes of clear semantics, and a residual part of more complex topology that does not carry an evident pattern of regularity.

As an optional step, crisp hyper-boxes $\mathbf{B}_i$ can be fuzzified in order to soften their boundaries. Among several possibilities to achieve fuzzification, one that is coherent with the pursuits of the granulation process is to define trapezoidal fuzzy sets (see fig. 6.2). Indeed, trapezoidal fuzzy sets are well suited to represent fuzzy intervals that can be labelled in a linguistic form.

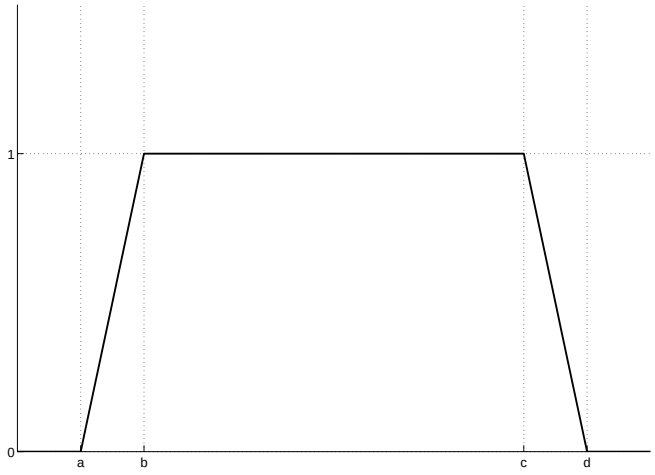

Figure 6.2: Shape of trapezoidal fuzzy set. The fuzzy set is well suited to represent the fuzzy interval with core $[b, c]$

The parametric definition of a one dimensional trapezoidal fuzzy set is the following:

$$\mu_{trap}\left(x; [a, b, c, d]\right) = \begin{cases} 0 & \text{if} \quad x < a \vee x > d \\ \dfrac{x-a}{b-a} & \text{if} \quad b \leq x \leq c \\ 1 & \text{if} \quad b < x < c \\ \dfrac{x-d}{c-d} & \text{if} \quad c \leq x \leq d \end{cases} \tag{6.29}$$

Let $\mathbf{B}_i$ be a granular prototype defined as in (6.27). It can be conveniently represented as a Cartesian product of its one-dimensional projections:

$$\mathbf{B}_i = B_{i1} \times \cdots \times B_{in} \tag{6.30}$$

being each B_{ij} defined as:

$$B_{ij} = \left[b_{ij}^{\min}, b_{ij}^{\max}\right] \tag{6.31}$$

Each interval B_{ij} can be fuzzified by defining a proper trapezoidal fuzzy set $\widetilde{B}_{ij}$ with membership function

$$\mu_{\widetilde{B}_{ij}}\left(x_j\right) = \mu_{trap}\left(x_j; \left[a_{ij}, b_{ij}^{\min}, b_{ij}^{\max}, d_{ij}\right]\right) \tag{6.32}$$

The parameters a_{ij} and d_{ij} can be determined according to different approaches e.g. by setting them so that the support of each $\widetilde{B}_{ij}$ includes all

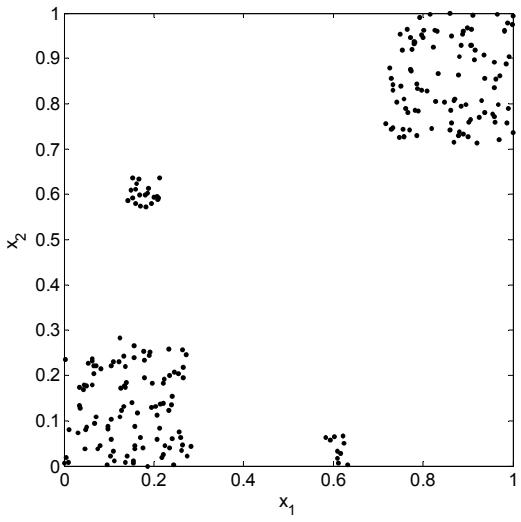

Figure 6.3: Two-dimensional synthetic dataset with four visible clusters of unequal size

elements of the i-th cluster according to a crisp assignment. Each resulting fuzzy set can be assigned to a linguistic label denoting fuzzy intervals, e.g. "ABOUT $[b_{ij}^{\min}, b_{ij}^{\max}]$".

When the fuzzy sets $\widetilde{B}_{ij}$ are fully determined, the multi-dimensional fuzzy prototype can be defined as a conjunction of the fuzzified projection, resulting in a membership function that is a t-norm combination of its components:

$$\mu_{\widetilde{\mathbf{B}}_i}(\mathbf{x}) = \mu_{\widetilde{B}_{i1}}(x_1) \otimes \cdots \otimes \mu_{\widetilde{B}_{in}}(x_n) \tag{6.33}$$

6.3 Illustrative example

As an illustrative example, a synthetic dataset is considered, which involves four clusters, as depicted in fig. 6.3. The two larger data groupings consist of 100 data-points and the two smaller ones have 20 and 10 data-points respectively.

The dataset has been clustered according to four Minkowski distances corresponding to $p = 2, 4, 6, 50$. After carrying out the Minkowski Fuzzy C-Means procedure, four multidimensional fuzzy sets have been defined according to (6.13). As it may be observed from figs. 6.4-6.7 the shape of clusters can be controlled by choosing an appropriate Minkowski distance. For $p = 2$ the fuzzy sets have the typical spherical representation, which

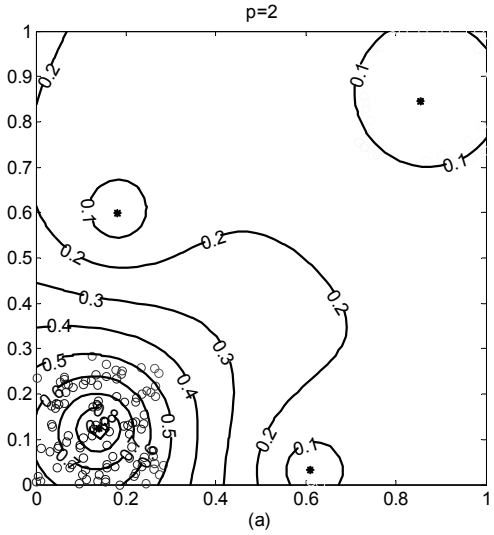

Figure 6.4: Contour levels of a generated information granule, for $p = 2$

gradually becomes boxlike for increasing values of p. For $p \gg 1$, the contour levels of each fuzzy membership function appear neatly sharp, resulting in a close approximation of the clusters attainable with Tchebychev distance.

In fig. 6.8, the distortion estimation $D^{(p)}$ defined in (6.18) is plotted for $p = 2, 3, \ldots, 100$. The plot can be conveniently divided in two parts. In the first part ($p \leq 62$) it is noticed a quick decrease of the distortion error that is coherent with the theoretical relationship between Minkowski and Tchebychev distance (6.19). For $p > 62$, the error perturbations due to the finite precision arithmetic[3] become increasingly significant so as to cause a random trend. As a consequence, the Minkowski distance for $p = 62$ appears as the best approximation of the Tchebychev distance for the problem at hand. It is remarkable, however, that good approximations (i.e. $D^{(p)} < 10^{-3}$) can be achieved for $p > 10$.

To evaluate the best representation of granular prototypes for different Minkowski distances, the plots of the functions $\phi_i(\gamma)$ defined in (6.26) are portrayed in figs. 6.9-**??**. Each graph consists of four plots for the same fuzzy set corresponding to $p = 2, 4, 6, 50$. As it may be easily observed, the adoption of the Euclidean distance ($p = 2$) clearly clashes with the requirement of box-shaped information granules because distortion is evident by high values of $\phi_i(\gamma)$ even for high values of γ. Conversely, higher values of

[3] All simulations were run on a IEEE754 double-precision computer

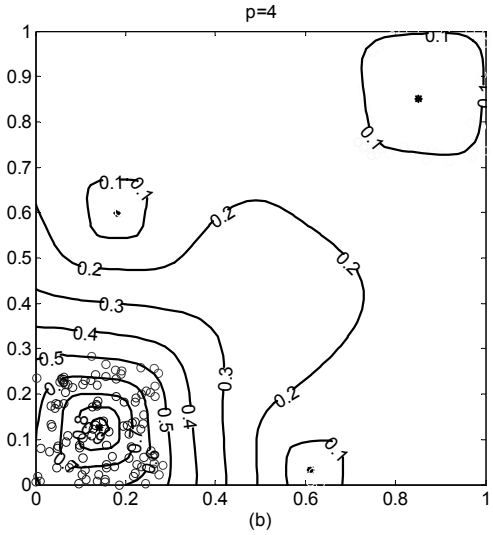

Figure 6.5: Contour levels of a generated information granule, for $p = 4$

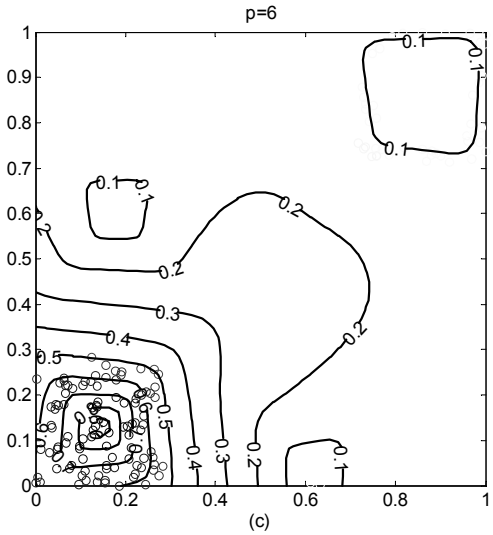

Figure 6.6: Contour levels of a generated information granule, for $p = 6$

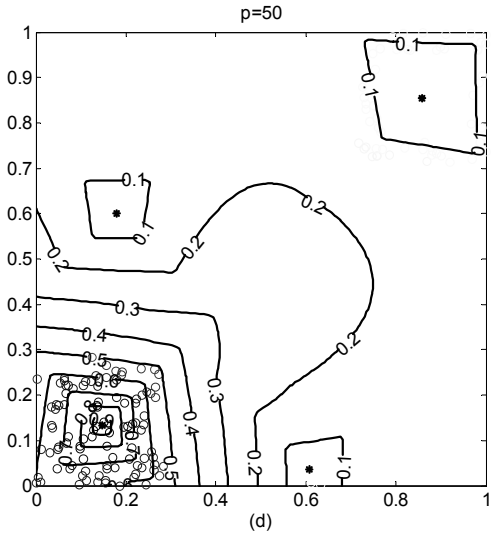

Figure 6.7: Contour levels of a generated information granule, for $p = 50$

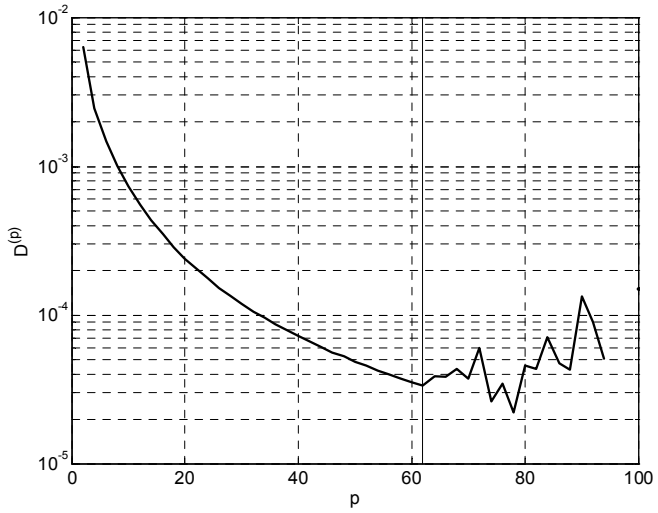

Figure 6.8: Distortion evaluation for Minkowski distance in approximating Tchebychev distance

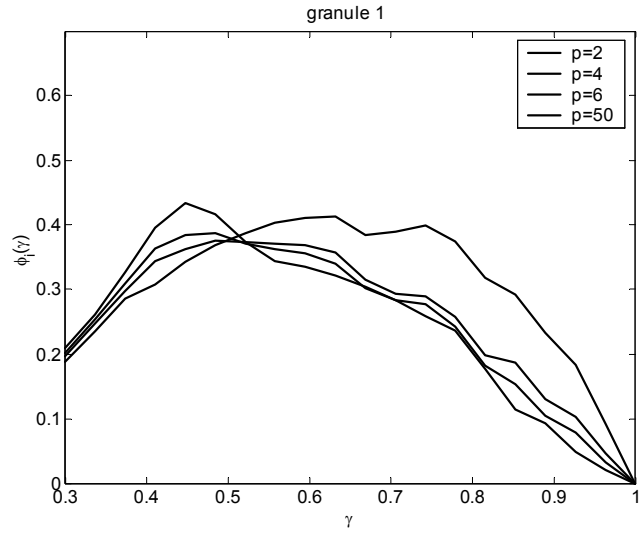

Figure 6.9: The plot of ϕ_1 vs. γ

p determine very similar behavior, which is roughly characterized by a convex trend. Indeed, for $\gamma \approx 1$, interactions amongst clusters is not relevant, thus the distortion value $\phi_i(\gamma)$ is close to zero. As γ decreases, the influence of each information granule becomes significant so that the distortion value raises consequently. For small values of γ, i.e. $\gamma \approx \frac{1}{c}$, the distortion measure becomes small since the membership value of each fuzzy set is asymptotically constant for large distances. Such behavior suggests to select the value of γ approximately within the range $[0.7, 1]$ to extract meaningful granular prototypes.

To derive crisp granular prototypes, the maximum distortion measure has been set to $\phi_{\max} = 0.1$. Based on such threshold, the granular prototypes corresponding to each cluster have been derived for $p = 2$ and $p = 50$ and represented in fig. 6.13. As it may be observed, hyper-boxes derived for $p = 2$ appear too narrow, especially for the two biggest clusters, thus excluding a significant number of patterns. Conversely, hyper-boxes derived for $p = 50$ better capture the underlying structure of data. As an optional step, the resulting hyper-boxes can be fuzzified in order to provide a description of data in terms of soft intervals.

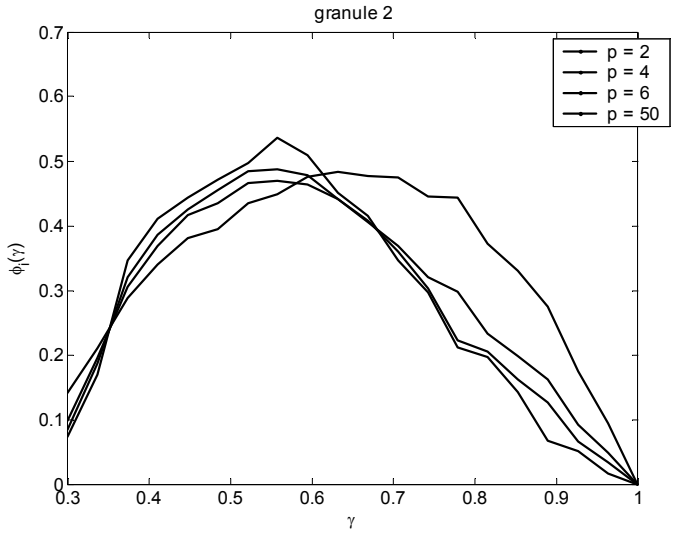

Figure 6.10: The plot of ϕ_2 vs. γ

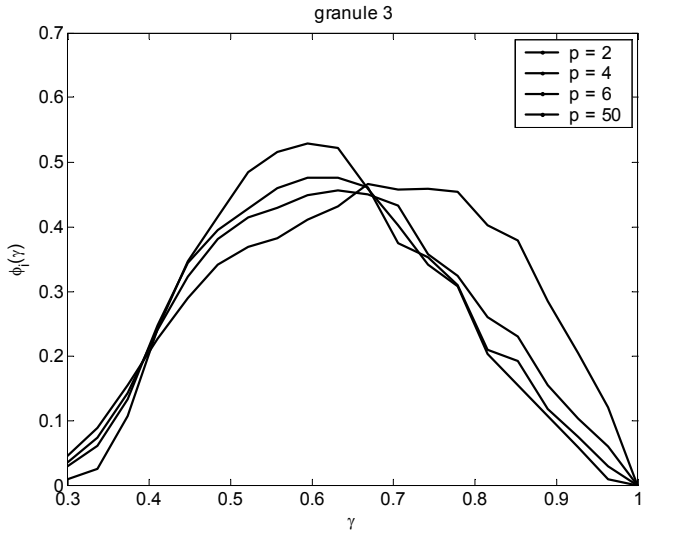

Figure 6.11: The plot of ϕ_3 vs. γ

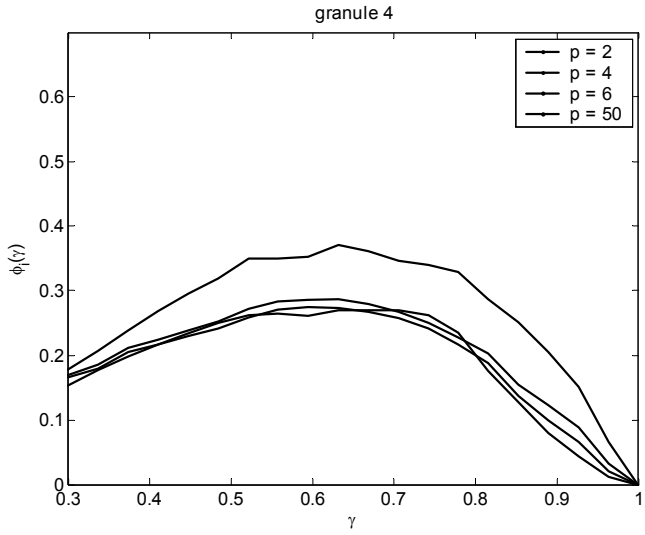

Figure 6.12: The plot of ϕ_4 vs. γ

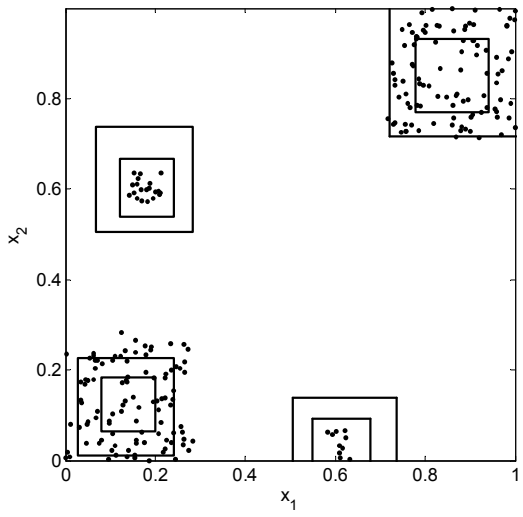

Figure 6.13: Granular prototypes generated for $p = 2$ (dashed boxes) and $p = 50$ (solid boxes) with $\phi_{\max} = 0.1$

6.4 Final remarks

The approach described in this Chapter for data granulation consists in the decomposition of complex topologies described by data into a regular structure and a residual part. The regular structure is defined in terms of granular prototypes, whose interpretability is determined by their representation in form of hyper-boxes.

To generate the granular prototypes, the Minkowski Fuzzy C-Means has been investigated. Specifically, two deformation effects that perturb the desired boxlike shape are analyzed. The first deformation effect is due to the estimation error introduced when approximating the Tchebychev distance – which characterizes boxlike granules – with Minkowski distances. Experimental results show that such approximation error quickly decreases as the Minkowski order increases, but calculi with too high orders are hampered by perturbation errors involved by finite precision arithmetic. As a consequence, it has been found that the Minkowski order should be chosen between 10 and 60 to avoid both rough approximation and perturbation errors.

The second deformation effect is due to interactions among clusters, that distort the boxlike shape of clusters to meet the constraints required by Fuzzy C-Means . To deal with such problem, a deformation measure has been formalized. With the aid of such measure, boxlike granular prototypes can be generated within an accepted deformation threshold.

When granular prototypes are fuzzified, they can be labelled as linguistic terms denoting imprecise ranges. The interpretability constraints fulfilled by this kind of representation are shown in table 6.1. As shown, the information granules resulting from Minkowski granulation are suitable for representing fuzzy quantities (because of normality, convexity and continuity of fuzzy sets) but not fuzzy numbers, since unimodality is not verified. Qualitative labels cannot be assigned to such information granules because some fundamental properties (such as Leftmost/rightmost fuzzy sets, Distinguishability, etc.) are not fulfilled. Moreover, the semantic representation of information granules by means of trapezoidal fuzzy set, while useful in efficient inference (because of the low number of firing rules) can invalidate some important properties like Error Free Reconstruction or Completeness. When such constraints are of fundamental importance in a modelling context, more sophisticated representations should be adopted.

Table 6.1: Table of fufillment of interpretability constraints by Minkowski Granulation (only applicable constraints are shown).

Interpretability constraints fulfillment	
Normality	Yes
Convexity	Yes
Unimodality	No
Continuity	Yes
Proper ordering	No
Justifiable no. of elements	Yes
Distinguishability	No
Completeness	No
Uniform granulation	No
Leftmost/Rightmost fuzzy sets	No
Natural zero positioning	No
Error Free Reconstruction	No
Description length	No
Rule length	No
Number of rules	Low
Number of firing rules	Low
Shared fuzzy sets	No
Model completeness	No

Chapter 7

Information Granulation through Prediction Intervals

7.1 Introduction

Fuzzy information granulation is one of the basic tasks in designing fuzzy models for solving predictive tasks. From a functional point of view, a fuzzy model defines a pointwise input/output mapping that provides a prediction of the output value when an input is presented to the system. Differently from other predictive models, fuzzy models have the added value of an interpretable knowledge base that explains the functional mapping by means of cause/effect rules. For this particularly appreciable feature, fuzzy models have met considerable success in control (Farinwata et al., 2000), engineering (Ross, 1997) as well as other theoretical and applicative areas. However, as already observed in Chapter 2, interpretability is not granted by the simple adoption of a fuzzy information granulation process to design fuzzy models, but a number of interpretability constraints have to be fulfilled.

A fuzzy information granulation process that verifies all the required interpretability constraints may not be sufficient to design an interpretable fuzzy model. Indeed, there are some constraints that do not depend on the granulation technique because they involve the representation of each single rule or the behavior of the entire model. In addition, interpretability must be balanced with accuracy as they are conflicting properties of a fuzzy model that must be equilibrated on the basis of practical issues.

One of the most important interpretability constraints is the compactness of the model, which is directly related to the number of rules that constitute its knowledge base. However, in many real-world problems, data are corrupted by noise or follow complex relationships that are hardly discovered

by simple models; in such cases, a model that provides a granular output - i.e. an aggregate of values rather than a single numerical value – is more appreciable. Prediction intervals as outputs are desirable since they provide a range of values that most likely include the true value to be predicted. Moreover, prediction intervals allow the user to perceive the accuracy of the estimation provided by the model, thus deciding to keep or reject the result.

The integration of prediction intervals within function approximation models has been extensively analyzed in the neural paradigm. In (Dybowski and Roberts, 2001; Papadopoulos et al., 2001; Ungar et al., 1996) some techniques are critically reviewed. Unfortunately, most of the proposed approaches have features that are not suitable for fuzzy models. In particular, many techniques (Nix and Weigen, 1994; Heskes, 1997; Efron and Tibshirani, 1993) yield prediction intervals with length related to the input variables. While this feature can be desirable for black-box models like classical neural networks, it can hamper the interpretability of the knowledge base in a fuzzy model. Moreover, as such techniques are based on the model's gradient, they are dependent on specific architecture and must be completely redefined whenever the architecture of the network changes. This is undesirable in fuzzy models – eventually encoded as neuro-fuzzy networks[1] (Jang and Sun, 1995)– since they are susceptible to architecture changes due to some modification of the represented rules, e.g. by adding/deleting new rules or new input variables. Finally, in other approaches like in (MacKay, 1992; Neal, 1996), the Bayesian framework is considered and the free parameters of the network are estimated by Monte Carlo methods. However such approach is inefficient and requires strict assumptions on the free parameters that are often impractical to solve real-world problems.

In this Chapter, an approach is proposed to derive prediction intervals for fuzzy models that in this way provide an estimate of the uncertainty associated with predicted output values. The derived intervals are constant throughout the input domain and do not require any strict assumption on the unknown distribution of data. Prediction intervals can be attached to each rule so as to provide a granulated output that enhances their interpretability. The resulting prediction intervals, even if expectably wider than those ob-

[1] A neuro-fuzzy network is a fuzzy model that uses a learning algorithm derived from or inspired by neural network theory to determine its parameters (fuzzy sets and fuzzy rules) by processing data samples (Nauck et al., 1997). The key advantage of neuro-fuzzy networks consists in their ability to acquire knowledge directly from data and represent it in form of linguistically sound fuzzy rules. Compared with classical neural networks, neuro-fuzzy systems exhibit similar performances in terms of accuracy and learning speed, but their added value lies in the representation of the acquired knowledge in terms of interpretable fuzzy rules.

tained with classical methods, provide a better understanding on the model's accuracy due to their integration in the knowledge base. In addition, as the mode structure changes, prediction intervals can be recomputed effortlessly.

In the next section, a method of information granulation through prediction interval derivation is introduced, together with a description of their integration in the knowledge base of a referential fuzzy model. In Section 7.3, some experimental results are drawn, both from a synthetic problem and from a real-world problem of ash combustion prediction. The Chapter ends with some final remarks.

7.2 A method of Information Granulation through Prediction Intervals

In this Section, the proposed method of information granulation through prediction is described in detail. First, a fuzzy model is established as a reference for the formalization of the method. Then, the derivation of prediction intervals is formalized and analyzed in the context of fuzzy modelling.

7.2.1 Reference Model

The pilot fuzzy model for prediction interval derivation is defined by 0th order Takagi-Sugeno rules, which have the following schema:

$$R_i : \text{IF } Ant_i \text{ THEN } \text{Y}{=}a_i \tag{7.1}$$

being $i = 1, 2, \ldots, R$. Here, Ant_i is an information granule resulting from an information granulation process and is semantically represented by a multidimensional fuzzy set $\mathbf{A}_i \in \mathcal{F}(\mathbf{U})$, being $\mathbf{U} \subseteq \mathbb{R}^n$. Given a numerical input $\mathbf{x} \in \mathbf{U}$, the output of the model is computed according to the following inference formula:

$$\tilde{y} = \frac{\sum_{i=1}^{R} a_i \mu_{\mathbf{A}_i}(\mathbf{x})}{\sum_{i=1}^{R} \mu_{\mathbf{A}_i}(\mathbf{x})} \tag{7.2}$$

Usually, the consequent parameters a_i are not evaluated by the the information granulation procedure, but through several possible techniques, including neural learning (Jang, 1993). In this latter case, the training process, may also involve the adaption of the fuzzy sets in the rule antecedents so as to minimize a given loss function, usually the Mean Squared Error.

7.2.2 Prediction interval derivation

It is supposed that there exists a supervisor that provides, for each input $\mathbf{x}$, the value $\bar{y}$ of an underlying unknown function:

$$\bar{y} = f(\mathbf{x}) + \varepsilon \tag{7.3}$$

where $f(\mathbf{x})$ is a deterministic input/output relationship and ε is a random variation. Random variation can be due to noise or inadequacy of the input variable $\mathbf{x}$ in fully describing the input/output relationship. Unlike other works known in literature, it is not assumed a zero mean value for ε. Data provided by the supervisor are limited to a finite training set of independent and identically distributed samples:

$$T = \left\{(\mathbf{x}_p, \bar{y}_p) \in \mathbb{R}^{n+1}, p = 1, 2, \ldots, N\right\} \tag{7.4}$$

It is also assumed that the supervisor is memory-less, that is, the returned value is independent on the previously calculated values. A fuzzy model is designed and trained so as to provide an approximated value $\tilde{y}$ for the same input $\mathbf{x}$. The following error function can be hence defined as:

$$e(\mathbf{x}) = \tilde{y}(\mathbf{x}) - \bar{y}(\mathbf{x}) \tag{7.5}$$

After training, a finite number of errors $e_1, e_2, \ldots, e_N$ are available. Such errors can be considered as independent and identically distributed univariate random variables with mean value:

$$\overline{e} = \frac{1}{N}\sum_{i=1}^{N} e_i \tag{7.6}$$

For sampled errors, prediction intervals can be calculated. A prediction interval $[L_\alpha, U_\alpha]$ of confidence level α represents an interval that will include the error e_{new} of a newly drawn example $\mathbf{x}_{new}$, with probability greater than $1 - \alpha$ (Neter et al., 1985). Formally:

$$\mathcal{P}(e_{new} \in [L_\alpha, U_\alpha]) \geqslant 1 - \alpha \tag{7.7}$$

With simple formal transformations, it is possible to derive the following inequality:

$$\mathcal{P}(\bar{y} \in [\tilde{y} - U_\alpha, \tilde{y} - L_\alpha]) \geqslant 1 - \alpha \tag{7.8}$$

Relation (7.8) defines a statistical method to estimate the true value of the underlying function approximated by the fuzzy model, with a desirable

confidence level. Formally, a prediction interval is defined by the following relations:

$$L_\alpha = \overline{e} - t_{\frac{\alpha}{2},[N-1]} \left(s\sqrt{\frac{1}{N}+1} \right) \tag{7.9}$$

$$U_\alpha = \overline{e} + t_{\frac{\alpha}{2},[N-1]} \left(s\sqrt{\frac{1}{N}+1} \right) \tag{7.10}$$

where $t_{\frac{\alpha}{2},[N-1]}$ is the value of the Student distribution with $N-1$ degrees of freedom corresponding to the critical value $\alpha/2$, and s is the sampled standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (e_i - \overline{e})^2}{N-1}} \tag{7.11}$$

The width of the prediction interval is directly related to the model accuracy. As a consequence, the less accurate is the model, or the smaller is the training set cardinality, the wider is the prediction interval.

In the proposed approach, the prediction intervals are derived on the basis of the approximated outputs inferred by the fuzzy model. However, for representation pursuits, prediction intervals can be calculated for each rule consequent, resulting in the following representation:

$$\text{IF } \mathbf{x} \text{ is } \mathbf{A}_i \text{ THEN } y \text{ is in } [a_i - U_{i,\alpha}, a_i - L_{i,\alpha}] \ (1-\alpha)\,\% \tag{7.12}$$

The individual prediction intervals are calculated by assuming each rule to be a simple fuzzy model with input/output mapping:

$$\tilde{y}_i(\mathbf{x}) = \mu_{\mathbf{A}_i}(\mathbf{x}) \cdot a_i \tag{7.13}$$

Each prediction interval is derived according to (7.5) through (7.10). As a consequence, the knowledge base is enriched with granular outputs that convey information on the accuracy of each rule. Moreover, the calculation of prediction intervals for the fuzzy rules leads to a more interpretable knowledge base, since it helps users to perceive an entire range of validity of each rule, instead of a single numerical value. It should be noted, however, that such intervals are used only for rules representation, while the derivation of the prediction interval for the model output must follow relation (7.8). This is due to the normalization effect existing in (7.2), while direct computation of the final prediction interval from the prediction intervals of each rule is possible only if the constraint $\forall \mathbf{x} : \sum_{i=1}^{R} \mu_{\mathbf{A}_i}(\mathbf{x}) = 1$ is added in the design of the fuzzy model.

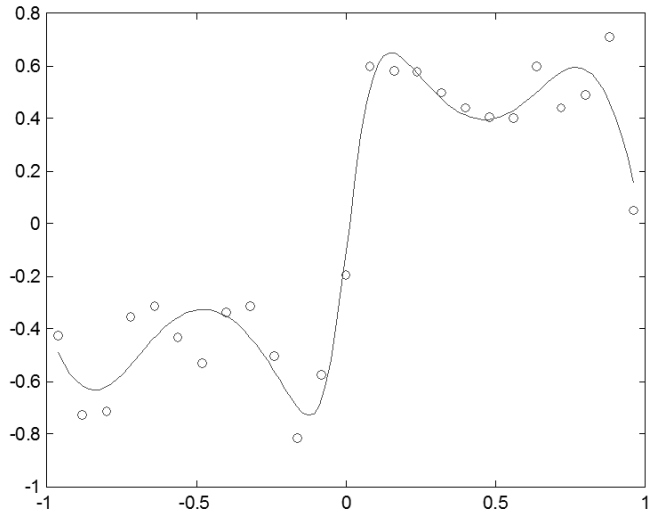

Figure 7.1: The training dataset (circles) and the input/output mapping defined by the neuro-fuzzy system (line)

7.3 Illustrative examples

In order to validate the proposed method, two experimental settings have been considered. The first illustrative example is aimed at showing the effectiveness of the proposed approach in a synthetic problem of nonlinear function approximation. The approach is then applied in a real world experimental setting to illustrate the added value of predictive model with the introduction of prediction intervals.

7.3.1 Nonlinear function approximation

As an illustrative example, a simple fuzzy model has been designed on the basis of two-dimensional synthetic data. The dataset consisted 51 input/output pairs in the range $[-1, 1]^2$. Because of the simplicity of the problem, the adopted granulation process of the input domain simply provided a Frame of Cognition with five Gaussian fuzzy sets of equidistant centers and same width. Such width has been calculated so as to guarantee the fulfillment of the 0.5-coverage of the Frame of Cognition. The structure of the model has been implemented in the neural paradigm as an ANFIS neural network that has been trained on a subset of 25 examples of the dataset. The remaining 26 examples constituted the test set for evaluating the final model in terms of accuracy. The test points and the pointwise functional mapping realized by the fuzzy model are depicted in fig. 7.1.

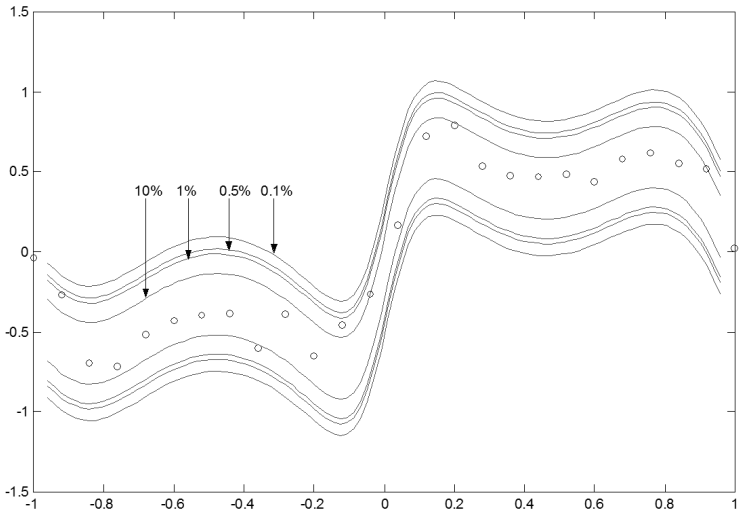

Figure 7.2: Prediction intervals of the input/output mapping for four confidence levels (10%, 1%, 0.5%, 0.1%)

Table 7.1: Prediction errors with different confidence levels

Confidence level	**Examples outside the prediction interval**
10%	2 (7.69%)
1%	1 (3.85%)
0.5%	0
0.1%	0

The proposed approach for deriving prediction intervals has been applied on the fuzzy model acquired through ANFIS learning. More specifically, four prediction intervals have been calculated in correspondence to the confidence levels $\alpha = 10\%,\ 1\%,\ 0.5\%,\ 0.1\%$. The interval mappings provided by the resulting four fuzzy models are illustrated in fig. 7.2.

For each value of the confidence level, the test set has been used to check whether the desired outputs fall within the prediction interval. The results are reported in table 7.1. As it can be seen, the number of desired outputs that fall outside the prediction interval is coherent with the confidence level. In the design phase, the choice of the appropriate value of the confidence level should be a trade-off between precision of the mapping (narrow intervals) and good predictive estimation (large intervals).

7.3.2 Ash combustion properties prediction

The proposed approach has been also adopted in a real-world experimental setting to illustrate the added value of predictive models obtained with the introduction of prediction intervals. The prediction problem consists in the estimation of amounts of chemical elements present after the combustion process in a thermoelectric generator. The dataset and the problem setting have been provided by "ENEL Produzione & Ricerca S.p.A." (Italy), and are fully described in (Castellano et al., 2003a). Briefly speaking, the problem consists in estimating the amount of 22 chemicals after combustion, when a set of 32 measurements of chemicals before combustion are given. The relations between input and output variables appear highly non-linear, due to the chaotic nature of the combustion process and the noisy measurements. In addition, the available dataset is very small in cardinality, thus hardening the complexity of the prediction problem.

For such kind of problem, any model that does not yield granular estimates would not convey sufficient information to understand the process and would provide almost arbitrary estimates. To solve such prediction problem, several neuro-fuzzy networks have been designed and trained according to the framework presented in (Castellano et al., 2003b). Specifically, a neuro-fuzzy network for each output variable to be predicted has been derived from the available dataset of 54 samples, which has been split into a training set of 31 samples and a test set with the remaining 23 samples[2]. Then, prediction intervals of 95% confidence have been derived, as depicted in fig. 7.3 for a selection of 8 models. Moreover, prediction intervals have been integrated in the knowledge base, as exemplified in the rule set in table 7.2 for the model estimating "Vanadium" (V). As it can be seen, prediction intervals in rule consequents convey much more information than scalar values, since the representation of a range of values also expresses the accuracy of each rule. In addition, the adoption of prediction intervals in the inference process provides the user with the range of most probable amounts of each chemical, thus enabling more appropriate decisions.

7.4 Final remarks

In this Chapter a method has been proposed for information granulation through derivation of prediction intervals. The method is applied on fuzzy models to generate a granular output that represents an estimate of the

[2]This split has been suggested by a domain expert, on the basis of physical considerations

If Cu is in $[240, 540]_{0.5}$, V is in $[0, 1100]_{0.5}$, DRAWING is in {S1, S2, S3}, Then V is in [0, 711] (95%);
If Cu is in $[220, 350]_{0.5}$, V is in $[1700, 3400]_{0.5}$, DRAWING is in {S2, S3, S4, S5}, Then V is in [2720, 3890] (95%);
If Cu is in $[260, 540]_{0.5}$, V is in $[1600, 3300]_{0.5}$, DRAWING is in {S2, S3}, Then V is in [1640, 2800] (95%);
If Cu is in $[260, 540]_{0.5}$, V is in $[1600, 3300]_{0.5}$, DRAWING is in {S2, S3}, Then V is in [1640, 2800] (95%);
If Cu is in $[0, 210]_{0.5}$, V is in $[1100, 2900]_{0.5}$, DRAWING is in {S1, S2}, Then V is in [1140, 2310] (95%);
If Cu is in $[0, 260]_{0.5}$, V is in $[0, 1400]_{0.5}$, DRAWING is in {S2, S3, S4, S5}, Then V is in [0, 877] (95%);
If Cu is in $[290, 590]_{0.5}$, V is in $[950, 1710]_{0.5}$, DRAWING is in {S4, S5}, Then V is in [1590, 2750] (95%);
If Cu is in $[0, 290]_{0.5}$, V is in $[0, 1400]_{0.5}$, DRAWING is in {S1, S2, S3}, Then V is in [0, 690] (95%)

Table 7.2: The ruleset generated for predicting Vanadium after the combustion process, on the basis of the quantitis of Copper (Cu) and Vanadium (V) before combustion, and the drawing source. For ease of reading, the fuzzy sets in the rules antecedents have been represented by their respective 0.5-cuts

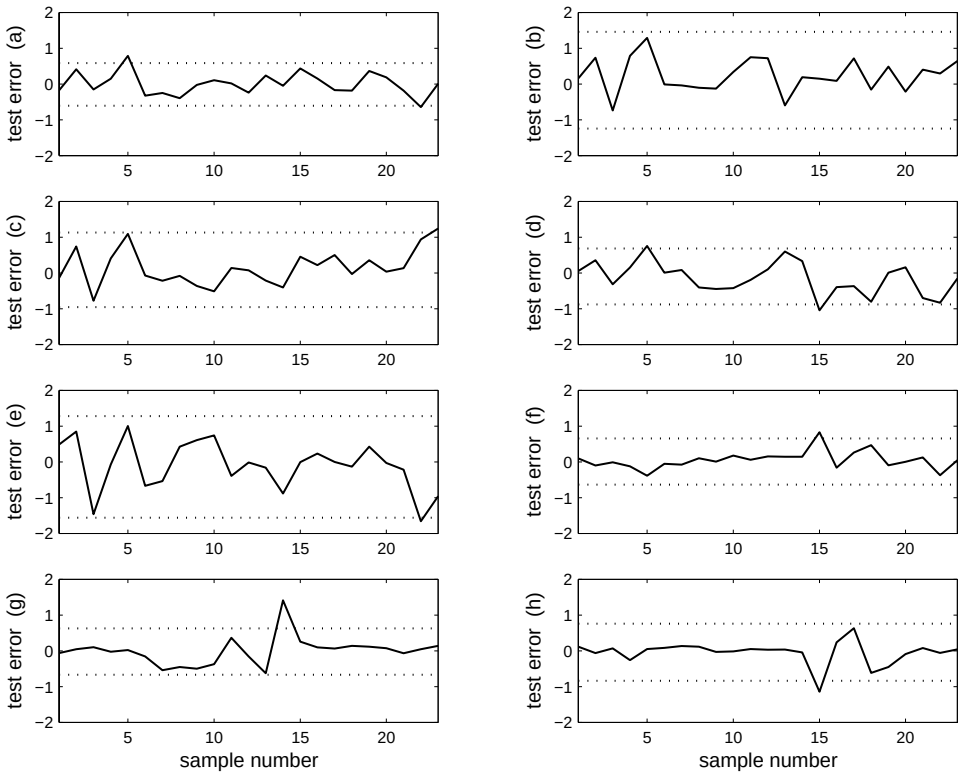

Figure 7.3: Estimation errors (solid line) and relative prediction intervals (dotted lines) for 8 models: Al (a), Ca (b), Mg (c), Ti (d), Ba (e), Co (f), V (g), Zi (h)

predicted numerical output. Such granulated output improves the interpretability of the model since it conveys useful information concerning the accuracy of the embodied knowledge base. Moreover, prediction intervals can be included in the rule base of the fuzzy model, thus improving their representation in terms of interpretability.

The method has been applied on a 0th order Takagi-Sugeno fuzzy model, which is well known for its fast inference process, due to the absence of the defuzzification procedure. On the other hand, the numerical knowledge included in the rule base of such models negatively influences the interpretability of the knowledge base. The integration of prediction intervals within such models allows for an improvement of the interpretability of the knowledge base without sacrificing the inference speed.

Chapter 8

Information Granulation through Double Clustering

8.1 Introduction

The key feature of information granulation consists in forming meaningful pieces of information that exhibit some functional and descriptive representation of data gathered into each information granule. Such a representation could take different forms, depending on the nature of data as well as on applicative needs.

In the previous chapters, some techniques for information granulation have been proposed, with the main objective of representing quantitative information. This type of representation usually takes the form of fuzzy number, fuzzy vector or fuzzy/probabilistic interval, and could be also used as a basis for further arithmetic processing.

In different applicative areas, however, quantitative representation of data may not be necessary, nor helpful. In some cases, a qualitative description of data, which makes use of linguistic terms like "small", "cold", "bright", etc., could be more advantageous. Indeed, qualitative linguistic terms are often used in common speech and are very helpful in conveying information in an economic yet effective way. As a consequence, the adoption of a qualitative representation of information granules could significantly boost the interpretability of a knowledge base.

Information granulation of data with qualitative representation requires a higher number of interpretability constraints than in the case of quantitative representation since the metaphors associated to linguistic terms are related together more tightly than for linguistic quantities. As a consequence, a number of additional constraints is required, especially those defined at the

level of Frame of Cognition[1].

The interpretability constraints that are essential in a sound qualitative representation of information granules are:

- Normality, convexity, continuity, as for quantitatively represented information granules;
- Proper ordering, so as to reflect the ordering of linguistic labels (e.g. "small" recalls a concept that precedes "large" w.r.t. a suitable scale);
- Justifiable number of elements, because the number of linguistic labels used to describe an attribute is usually very limited;
- Distinguishability, because different linguistic labels designate well distinct metaphors;
- Completeness, because each element of the Universe of Discourse could be denoted by at least one linguistic label;
- Leftmost/Rightmost fuzzy sets, which properly refer to limit cases into a Universe of Discourse.

The process of automatically extracting information granules from data with qualitative representation is rather difficult because it implies two major problems: (1) the discover of hidden relationships among multi-dimensional data, and (2) the representation of such relationship in terms in a linguistically interpretable fashion. Usually, the problem of discovering hidden relationships among data is tackled with cluster analysis techniques, which however may fail in providing a qualitative description of data in terms of interpretable fuzzy information granules. On the other hand, the a-priori definition of interpretable fuzzy sets is relatively simple, but it is rather hard to guarantee that information granules based on such definition adequately represent the relationships underlying the distribution of data.

In this Chapter, an algorithmic framework is proposed, called "Double Clustering Framework", which enables the extraction of qualitatively described information granules with through efficient clustering techniques. The framework is mainly centered on two clustering steps: the first, applied on the multidimensional data, is aimed at discovering the hidden relationships among data, while the second, which operates at the level of each dimension, enables the definition of interpretable fuzzy sets that can be labelled

[1]See Chapter 2 for a deeper discussion.

with linguistic terms. The framework does not fully specify the clustering algorithms, which could be chosen according to different applicative needs.

The Double Clustering Framework is described in detail in Section 8.2, along with some implementations that include specific clustering algorithms. Then, some illustrative examples are portrayed and conclusive remarks are drawn.

8.2 The Double Clustering Framework

The Double Clustering Framework is an algorithmic scheme that enables the extraction of fuzzy information granules representable in terms of qualitative linguistic labels. It is made of three main steps:

1. **Data Clustering**. Clustering is performed in the multi-dimensional space of experimental data – assumed to belong to a numerical domain – to embrace similar points into granules. At this stage, the information granules are described by single multi-dimensional prototypes defined on the same numerical level of experimental data;

2. **Prototype Clustering**. The prototypes obtained from the first clustering stage are further clustered along each dimension of the numerical domain, so as to obtain a number of one-dimensional prototypes on each dimension;

3. **Granule Fuzzification**. Multidimensional and one-dimensional prototypes provide useful information to derive information granules that can be conveniently represents by fuzzy sets. Moreover, such fuzzy sets can be built in accord to the interpretability constraints that allow a qualitative description of the information granules.

The framework tries to exploit the features of both multi-dimensional and one-dimensional clustering. Precisely, the multi-dimensional clustering captures the granularity of the data in the multi-dimensional space, but the fuzzification of the resulting granules typically results in fuzzy sets that cannot be associated with qualitative linguistic labels. Conversely, the one-dimensional clustering can provide interpretable fuzzy sets but may loose information about the granularity of the multi-dimensional data. The combined application of both one-dimensional and multi-dimensional clustering exploits the benefits of the two approaches and enables an interpretable granulation of information. A schema of the process carried out within the Double Clustering Framework is portrayed in fig. 8.1.

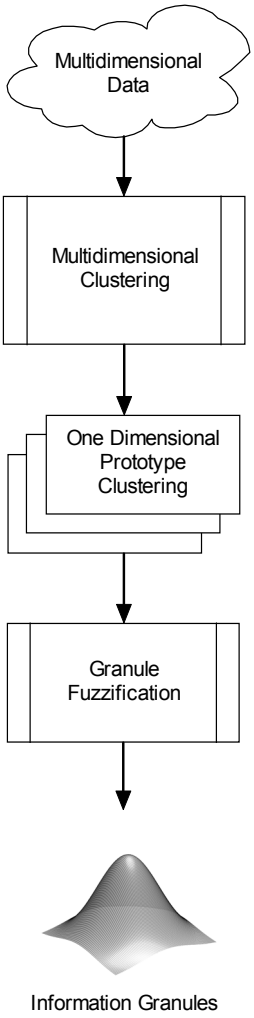

Figure 8.1: The three steps of the Double Clustering Framework

Formally, the Double Clustering Framework can be described as follows. Let

$$X = [m_1, M_1] \times \cdots \times [m_n, M_n] \subseteq \mathbb{R}^n \tag{8.1}$$

be the n-dimensional Universe of Discourse in which an available set of numerical data

$$D = \{\mathbf{x}_i \in X : i = 1, 2, \ldots, N\} \tag{8.2}$$

is defined.

The first stage of the Double Clustering Framework performs a multi-dimensional clustering on the dataset D, providing a collection of multi-dimensional prototypes

$$\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_p \in X \tag{8.3}$$

being $\mathbf{c}_i = \left(c_i^{(1)}, c_i^{(2)}, \ldots, c_i^{(n)}\right),\ i = 1, 2, ..., p.$

The multi-dimensional prototypes are then projected onto each dimension, resulting in n sets

$$C^{(j)} = \left\{c_i^{(j)} \in [m_j, M_j] : i = 1, 2, \ldots, p\right\} \tag{8.4}$$

for $j = 1, 2, \ldots, n$. The points of each $C^{(j)}$ are subject to one-dimensional clustering in the second step, yielding to one-dimensional prototypes

$$P^{(j)} = \left\{p_1^{(j)}, p_2^{(j)}, \ldots, p_{K_j}^{(j)}\right\} \tag{8.5}$$

being K_j the number of clusters in the j-th dimension, which must be low in accordance to the "justifiable number of elements" interpretability constraint.

The last stage of the Double Clustering Framework involves the fuzzification of the information granules. This is achieved by first fuzzifying the one-dimensional granules defined by the prototypes in each $P^{(j)}$ and then by aggregating one-dimensional fuzzy sets to form multi-dimensional fuzzy information granules.

For each dimension $j = 1, 2, \ldots, n$, the K_j extracted clusters are transformed into as many interpretable fuzzy sets. Different types of membership functions can be used to characterize fuzzy sets; here, Gaussian fuzzy sets are considered, with the following membership functions:

$$\mu_{A_k^{(j)}}(x) = \exp\left(-\frac{\left(x - \omega_k^{(j)}\right)^2}{2\left(\sigma_k^{(j)}\right)^2}\right) \tag{8.6}$$

for $k = 1, 2, \ldots, K_j$. To complete the mathematical definition of fuzzy sets, the values of the centers $\omega_k^{(j)}$ and the widths $\sigma_k^{(j)}$ must be derived.

The definition of such values must take into account the information provided by the clustering stages and, at the same time, must consider the required interpretability constraints. To meet both requirements, the following "cut points" are defined:

$$t_k^{(j)} = \begin{cases} 2m_j - t_1^{(j)} & \text{for} \quad k = 0 \\ \left(p_k^{(j)} + p_{k+1}^{(j)}\right)/2 & \text{for} \quad 0 < k < K_j \\ 2M_j - t_{K_j-1}^{(j)} & \text{for} \quad k = K_j \end{cases} \tag{8.7}$$

Note that there are $K_j + 1$ cut point for each dimension. Cut points can be used to define centers and widths of the Gaussian membership functions so that the interpretability constraints can be met. Their values can be derived by the following relations:

$$\omega_k^{(j)} = \frac{t_{k-1}^{(j)} + t_k^{(j)}}{2} \tag{8.8}$$

and

$$\sigma_k^{(j)} = \frac{t_k^{(j)} - t_{k-1}^{(j)}}{2\sqrt{-2\ln\varepsilon}} \tag{8.9}$$

where ε is the maximum allowed overlap (possibility) between two adjacent fuzzy sets[2].

It is easy to show that for each dimension, the fuzzy sets defined in (8.6) meet all required interpretability constraints, so qualitative linguistic labels can be attached within the newly formed frame of cognition. More specifically:

- Normality, convexity and continuity are verified by the Gaussian shape of fuzzy sets;
- Proper ordering is verified by defining the order relation of fuzzy sets reflecting the order of the prototypes.
- Justifiable number of elements is verified by an appropriate choice of K_j;

[2]On the basis of theoretical results derived in Chapter 3, the possibility measure is used to quantify distinguishability

- Distinguishability is verified by construction of the fuzzy sets, which have possibility measure not exceeding the allowed overlap ε;
- Completeness is verified by construction of the fuzzy sets, which guarantee ε-coverage;
- Leftmost/Rightmost fuzzy sets by construction of the first and last fuzzy sets, whose prototypes coincide with the limits of Universe of Discourse.

Once the fuzzification process is completed, for each dimension a Frame of Cognition can be defined, by attaching proper linguistic labels to the derived fuzzy sets. Because of the verification of the interpretability constraints, the association of linguistic labels is a straightforward process.

Let

$$\mathbf{F}^{(j)} = \left\langle U^{(j)}, \mathbf{F}^{(j)}, \preceq^{(j)}, \mathcal{L}^{(j)}, v^{(j)} \right\rangle \tag{8.10}$$

be the Frames of Cognition defined after the granulation process. Multi-dimensional fuzzy information granules are formed by combining one-dimensional fuzzy sets, one for each frame. A naïve approach, however, would lead to an exponentially high number of information granules, which violates the compactness interpretability constraint. Indeed, the total number of information granules is

$$\prod_{j=1}^{n} K_j \sim \Omega\left(2^n\right) \tag{8.11}$$

Furthermore, the so derived information granule would not convey any useful information concerning the distribution of data and, hence, a description of their underlying relationships. To avoid combinatorial explosion of multi-dimensional fuzzy information granules, only those granules that are most representative of the multi-dimensional prototypes $\mathbf{c}_i$ are considered. The selection of such granules is accomplished within each dimension, by considering, for each $i = 1, 2, \ldots, p$, the fuzzy set in the j-th dimension with highest membership value on the j-th projection of the i-th prototype. The index of such fuzzy set is defined as:

$$\overline{k_i^{(j)}} = \underset{k=1,2,\ldots,K_j}{\arg\max}\, \mu_{A_k^{(j)}}\left(c_i^{(j)}\right) \tag{8.12}$$

Once representative fuzzy sets are chosen, multi-dimensional fuzzy information granules can be defined as usual. Specifically, the semantics of each

information granule is defined by the Cartesian product of the selected fuzzy sets while its syntax is defined as:

$$v^{(1)} \text{ IS } \mathcal{L}\left(A_{k_i^{(j)}}^{(1)}\right) \text{ AND } \ldots \text{ AND } v^{(n)} \text{ IS } \mathcal{L}\left(A_{k_i^{(n)}}^{(n)}\right) \tag{8.13}$$

for $i = 1, 2, \ldots, p$. It is possible that two ore more information granules coincide, hence the total number of derived granules is upper bounded by the number of multi-dimensional prototypes p.

When the granulation process is completed, a fuzzy rule-based model can be built on the basis of the derived fuzzy granules. This is aimed to verify how much the fuzzy granules identified from data are useful in providing good mapping properties or classification capabilities.

The Double Clustering Framework can be customized by choosing appropriate clustering algorithms, either for the multi-dimensional or the one-dimensional stage. The unique requirement for such algorithms is to produce a set of prototypes in conformity with the granulation process outlined in the previous Section.

The choice of specific clustering algorithms defines an implementation of the Double Clustering Framework, and depends on applicative issues. In the following, some implementations of the framework are described, and potential applicative contexts are considered.

8.2.1 Fuzzy Double Clustering

A first implementation of the Double Clustering Framework integrates a fuzzy clustering algorithm in the first stage for multi-dimensional clustering, and a hierarchical clustering scheme for the one-dimensional clustering. This type of implementation is particularly suited to enhance existing applications of fuzzy clustering in order to enable interpretable fuzzy information granulation.

The fuzzy clustering scheme for the first stage could be any algorithm that carries out a set of prototypes as a result of its execution process. Examples of such scheme are the Fuzzy C-Means (Bezdek, 1981) and all its variants, like the Gustafson-Kessel algorithm (Gustafson and Kessel, 1979), the Gath-Geva algorithm (Abonyi et al., 2002), the Conditional Fuzzy C-Means (Pedrycz, 1996a), etc. Also, Possibilistic C-Means can be applied (Krishnapuram and Keller, 1993), as well as other clustering schemes (Baraldi and Blonda, 1999). In the choice of a specific clustering algorithm, the maximum number of fuzzy clusters that can be extracted must be a controlled parameter; indeed an uncontrolled number of clusters may hamper the compactness requirement that is fundamental for the interpretability of the fuzzy model. For this

reason, fuzzy clustering algorithms that require a fixed (or maximum) number of clusters as working parameters are preferred for the implementation of the first stage of the Double Clustering Framework.

The second stage of one-dimensional clustering does not require a fuzzy clustering scheme, since fuzzification in each dimension is an integral part of the Double Clustering Framework that is accomplished in its third stage. Moreover, since the number of points to be clustered coincides with the number of multi-dimensional clusters, and since the latter is very low to guarantee compactness of the fuzzy model, simple one-dimensional clustering algorithms can be adopted.

Hierarchical clustering is a crisp, simple clustering algorithm that has the additional property of being quite efficient for one-dimensional numerical data, provided that such data are sorted. These advantageous features justify its adoption within the considered implementation of the Double Clustering Framework.

The application of hierarchical clustering within the Double Clustering Framework is as follows. For each dimension $j = 1, 2, \ldots, n$, the set of prototype projections $C^{(j)}$ defined in (8.4) is sorted to provide the following array:

$$\overrightarrow{C^{(j)}} = \left\langle c_{i_1}^{(j)}, c_{i_2}^{(j)}, \ldots, c_{i_p}^{(j)} \right\rangle \tag{8.14}$$

such that $c_{i_h}^{(j)} \leq c_{i_{h+1}}^{(j)}$.

The hierarchical clustering operates by successive partitions of $\overrightarrow{C^{(j)}}$, and terminates when a partition of defined cardinality K_j is reached. The first partition is trivial, with each element of the array forming a single cluster (hereafter the index $^{(j)}$ is dropped for ease of notation):

$$H^0 = \left\langle h_1^0, h_2^0, \ldots, h_p^0 \right\rangle \tag{8.15}$$

being

$$h_l^0 = \{c_{i_l}\}, \ l = 1, 2, \ldots, p \tag{8.16}$$

For the partition H^0 a trivial set of prototypes is also defined as:

$$P^0 = \left\langle p_1^0, p_2^0, \ldots, p_p^0 \right\rangle \tag{8.17}$$

where

$$p_l^0 = c_{i_l}, \ l = 1, 2, \ldots, p \tag{8.18}$$

The clustering process proceeds by aggregating clusters from previous partitions. The clusters to be merged are chosen on the basis of the minimal distance between prototypes. Since data are one-dimensional and sorted, the selection of the two closest clusters can be achieved within a single scan of the prototype array, thus resulting in a very efficient process.

The aggregation process is carried out as follows. For $k = 1, 2, \ldots, p - K_j$, the two nearest prototypes from the $(k-1)$-th partition are selected and their position is considered:

$$l_k^* = \underset{l=1,2,\ldots,p-k-1}{\arg\max} \left| p_l^{k-1} - p_{l+1}^{k-1} \right| \tag{8.19}$$

The position index l_k^* enables the definition of a new partition H^k that is identical to H^{k-1} with the exception of the clusters $h_{l_k^*}^{k-1}$ and $h_{l_k^*+1}^{k-1}$ that are merged together. The formal definition of H^k is given as follows:

$$H^k = \left\langle h_1^k, h_2^k, \ldots, h_{p-k}^k \right\rangle \tag{8.20}$$

where, for each $l = 1, 2, \ldots, p - k$:

$$h_l^k = \begin{cases} h_l^{k-1} & \text{if} \quad l < l_k^* \\ h_l^{k-1} \cup h_{l+1}^{k-1} & \text{if} \quad l = l_k^* \\ h_{l+1}^{k-1} & \text{if} \quad l > l_k^* \end{cases} \tag{8.21}$$

The new prototype array $P^k = \left\langle p_1^k, p_2^k, \ldots, p_p^k \right\rangle$ can be computed on the basis of the new partition H^k:

$$p_l^k = \frac{1}{\left| h_l^k \right|} \sum_{c \in h_l^k} c, \;\; l = 1, 2, \ldots, p - k \tag{8.22}$$

Prototype calculation can be simplified with the observation that the values of P^k change from P^{k-1} only in one element. For $k = p - K_j$ the clustering process is stopped and the final prototype set $P^{(j)}$ is defined as:

$$P^{(j)} = P^{p-K_j} \tag{8.23}$$

Each cycle of the clustering process hence takes $O\left(p - k\right)$ operations and the number of cycles is $p - K_j + 1$. By adding the time necessary to sort the prototypes before clustering – $O\left(p^2\right)$ in the worst case – the time complexity of the hierarchical clustering procedure for all dimensions is

$$O\left(np^2 - nK^2\right) \tag{8.24}$$

being

$$K = \max_{j=1,2,\ldots,n} K_j \tag{8.25}$$

Since the number of prototypes is usually much lower that the cardinality of the dataset D, it can be affirmed that the overhead of the Double Clustering Framework is negligible when compared to the time complexity of the multidimensional clustering procedure, which is usually at least linear in the cardinality of the dataset, N.

8.2.2 Crisp Double Clustering

Fuzzy Double Clustering is useful when an existing fuzzy clustering application must be wrapped so as to accommodate interpretability constraints. When new systems have to be developed, it is more convenient to use a vector quantization technique in substitution to the fuzzy clustering algorithm in the first stage of the Double Clustering Framework. This choice is motivated by the unnecessity of the partition matrix within the Double Clustering Framework, which is usually calculated – with a significant time consumption – by fuzzy clustering algorithms.

The Crisp Double Clustering is based on a general-purpose vector quantization algorithm that follows the Linde-Buzo-Gray (LBG) formulation (Linde and Gray, 1980). From the dataset D a number of p multi-dimensional prototypes $\mathbf{c}_1^0, \mathbf{c}_2^0, \ldots, \mathbf{c}_p^0 \in D$ is randomly selected. Based on such prototypes, the dataset D is partitioned according to:

$$D_l^0 = \left\{ \mathbf{x} \in D : \left\| \mathbf{x} - \mathbf{c}_l^0 \right\| = \min_{k=1,2,\ldots p} \left\| \mathbf{x} - \mathbf{c}_k^0 \right\| \right\} \tag{8.26}$$

where $\|\cdot\|$ is a suitable norm. The partition $D_1^0, D_2^0, \ldots, D_p^0$ is used to define a new set of prototypes

$$\mathbf{c}_l^1 = \frac{1}{|D_l^0|} \sum_{\mathbf{x} \in D_l^0} \mathbf{x}, \; l = 1, 2, \ldots, p \tag{8.27}$$

The new set of prototype is used to calculate a new partition as in (8.26) and then a new set of prototypes as in (8.27). The process is repeated iteratively so as to generate a sequence of prototype vectors $\mathbf{c}_l^k$, for $k = 0, 1, 2, \ldots$. The iterative process is stopped when a termination criterion is met, e.g. when all prototypes do not change significantly:

$$\text{STOP when } \forall l = 1, 2, \ldots, p : \left\| \mathbf{c}_l^k - \mathbf{c}_l^{k-1} \right\| < \epsilon \tag{8.28}$$

where ϵ is a user-defined threshold. An accurate implementation of the LBG algorithm can avoid the explicit representation of the partitions D_l^k, thus saving significant amounts of memory space[3].

8.2.3 DCClass

The Crisp Double Clustering implementation is general-purpose and can be applied to all problems of numerical data granulation. However, for classification problems, the information deriving from the class partitioning of data can be effectively exploited to improve the granulation process. More specifically, the class division of data can be used to automatically determine the granularity level for each attribute, without the need of specifying the number K_j of fuzzy sets onto each dimension.

The first stage of DCClass is similar to Crisp Double Clustering, though better results can be achieved if a classification-oriented vector quantization algorithm is used in substitution to the classical LBG scheme. An example of algorithm that performs vector quantization in classification problems is LVQ1 (Learning Vector Quantization, version 1) (Kohonen, 1986), though more stable techniques may be adopted, like Self Organizing Maps, etc. (Kohonen, 1990; Kohonen, 1997). As an example, the LVQ1 algorithm is portrayed in the following.

Let $D_{\mathcal{C}}$ be the dataset of numerical data enriched with class information:

$$D_{\mathcal{C}} = \{\langle \mathbf{x}_i, \psi_i \rangle : i = 1, 2, \ldots, N\} \subseteq X \times \mathcal{C} \tag{8.29}$$

where $\mathcal{C}$ is the set of all class labels

$$\mathcal{C} = \{\Psi_1, \Psi_2, \ldots, \Psi_m\} \tag{8.30}$$

The LVQ1 algorithm operates by firstly selecting at random a number of p prototypes from $D_{\mathcal{C}}$, possibly reflecting the class distribution of the dataset. This selection corresponds to the first array of prototypes:

$$\left\langle \mathbf{c}_1^0, \psi_{c1} \right\rangle, \left\langle \mathbf{c}_1^0, \psi_{c2} \right\rangle, \ldots, \left\langle \mathbf{c}_p^0, \psi_{cp} \right\rangle \in D_{\mathcal{C}} \tag{8.31}$$

The prototypes are then updated interactively, by selecting at random an example $\left\langle \mathbf{x}_{i_k}, \psi_{i_k} \right\rangle$ from the dataset $D_{\mathcal{C}}$ and following an update rule

[3] This could be achieved by defining two arrays of p elements, namely `COUNT[1..p]` and `SUM[1..p]`. Each element $\mathbf{x}$ of D is compared to all prototypes; if the l-th prototype is the closest, then `COUNT[l]:=COUNT[l]+1` and `SUM[l]:=SUM[l]+x`. The new prototypes are calculated as `c[l]:=SUM[l]/COUNT[l]`; in this way the algorithm requires only an amount of memory proportional to $3p$ to compute prototypes.

that involves only the prototype $\mathbf{c}_l^k$ that is closest to the selected point $\mathbf{x}_{i_k}$. The update rule takes into account the class information, so as to bring the prototype closer to the data point if the class labels coincide, and to keep the prototype away from the data point if the class labels are different. Stated formally:

$$\mathbf{c}_l^k = \begin{cases} \mathbf{c}_l^{k-1} + \alpha_k \left(\mathbf{x}_{i_k} - \mathbf{c}_l^{k-1} \right) & \text{if} \quad \psi_{cl} = \psi_{i_k} \\ \mathbf{c}_l^{k-1} - \alpha_k \left(\mathbf{x}_{i_k} - \mathbf{c}_l^{k-1} \right) & \text{if} \quad \psi_{cl} \neq \psi_{i_k} \end{cases} \tag{8.32}$$

The update iterations are repeated until a convergence criterion is met, like the one define in (8.28).

A key feature of classification-oriented vector quantization algorithms is the possibility to clearly assign a class label to each prototype. This class information is also inherited by the prototype projections, so that each multi-dimensional prototype can be represented as

$$\langle \mathbf{c}_l, \psi_{cl} \rangle = \left\langle \left(c_{l1}^{(1)}, c_{l2}^{(2)}, \ldots, c_{ln}^{(n)} \right), \psi_{cl} \right\rangle \tag{8.33}$$

and the set of projection on the j-th dimension C_j as:

$$C^{(j)} = \left\{ \left\langle c_l^{(j)}, \psi_{cl} \right\rangle \in [m_j, M_j] \times \mathcal{C} : l = 1, 2, \ldots, p \right\} \tag{8.34}$$

By sorting each $C^{(j)}$ on the values $c_l^{(j)}$, the following array can be derived:

$$\overrightarrow{C^{(j)}} = \left\langle \left\langle c_{l_1}^{(j)}, \psi_{cl_1} \right\rangle, \left\langle c_{l_2}^{(j)}, \psi_{cl_2} \right\rangle, \ldots, \left\langle c_{l_p}^{(j)}, \psi_{cl_p} \right\rangle \right\rangle \tag{8.35}$$

A partition of elements of $\overrightarrow{C^{(j)}}$ is automatically induced by the class labels ψ_{cl_k} by keeping in the same cluster all adjacent elements that share the same class label. Formally, the one-dimensional projections are clustered according to the partition induced by the following equivalence relation:

$$c_{l_k}^{(j)} \equiv c_{l_{k+1}}^{(j)} \iff \psi_{cl_k} = \psi_{cl_{k+1}} \tag{8.36}$$

for $k = 1, 2, \ldots, p-1$. The number of one-dimensional clusters is not defined a priori, but it is upper bounded by p. As a consequence, if p is small (to meet the compactness requirement), also the number of clusters in each dimension is low.

For each one-dimensional cluster, one-dimensional prototypes can be calculated according to (8.22), so that the fuzzy information granules can be defined according to the Double Clustering Framework.

The granulation process carried out by DCClass is graphically exemplified in figs. 8.2–8.4. In fig. 8.2, data are quantized through the LVQ1 algorithms,

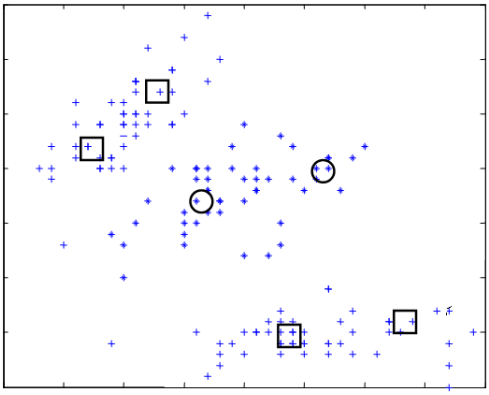

Figure 8.2: The first stage of DCClass. Data (dots and circles, according to the belonging class) are quantized into six prototypes, depicted as circles or squares depending on the associated class.

leading to a umber of prototypes that are distinguished by the associated class. In fig. 8.3, the prototypes are projected onto each dimension and clustered by aggregating projections belonging to the same class. Then, as illustrated in fig. 8.4,Gaussian fuzzy sets are defined and information granules are generated as required in the third stage of the Double Clustering Framework.

8.3 Illustrative examples

To illustrate the practical benefits of the Double Clustering Framework, a number of illustrative examples taken from real-world applicative contexts are portrayed. In such examples different implementations of the framework are used, so as to highlight their main differences.

8.3.1 Granulation example

The effectiveness of Fuzzy Double Clustering has been evaluated on the well-known Iris data set concerning classification of Iris flowers (Fisher, 1936). Three species of iris flowers (setosa, versicolor and virginica) are known. There are 150 samples of iris flowers, 50 of each class. A sample is a four-dimensional pattern vector representing four attributes of the iris flower concerning sepal length, sepal width, petal length, and petal width.

In the first experiment, to appreciate the ability of the proposed approach in deriving interpretable fuzzy granules, the whole dataset was considered to

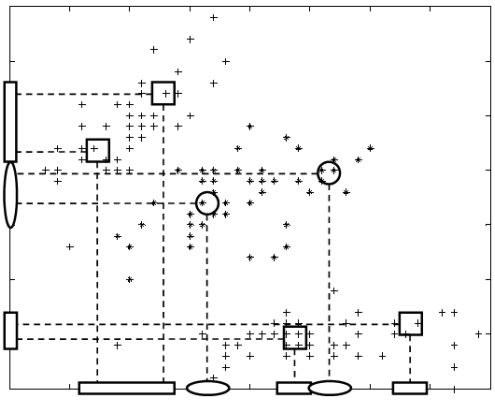

Figure 8.3: The second stage of DCClass. Multidimensional prototypes are projected onto each dimension and then are clustered by aggregating prototypes belonging to the same class.

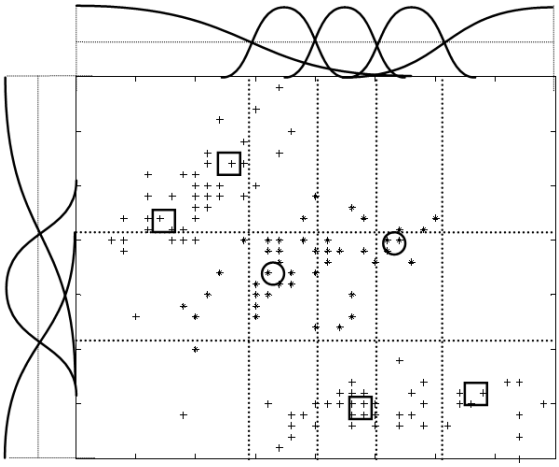

Figure 8.4: The third stage of DCClass. The clustered one-dimensional prototypes determine the necessary information to generate one-dimensional Gaussian fuzzy sets. The multi-dimensional prototypes enables the correct combination of such fuzzy sets to generate multidimensional information granules.

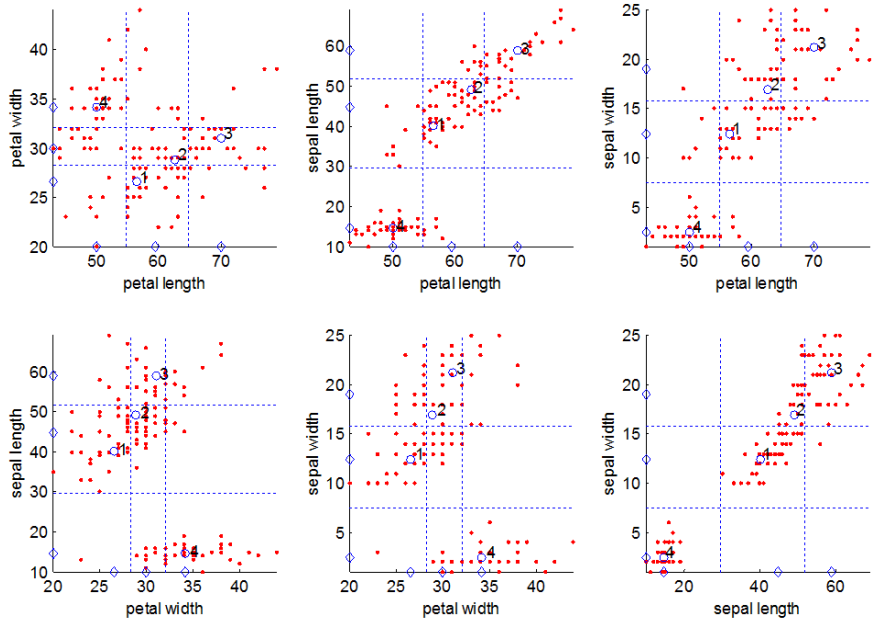

Figure 8.5: Result of the information granulation process plotted in two dimensions. The four cluster prototypes discovered by Fuzzy C-means (circles) are projected on each axis and further clustered to produce three prototypes (diamonds) on each dimension, resulting in three fuzzy sets per input variable. Dashed lines represent intersection points between adjacent fuzzy sets.

perform information granulation. In the first clustering step, the Fuzzy C-Means was applied to discover four granules (clusters) in the four-dimensional input space. In the second clustering step, hierarchical clustering was applied to the cluster prototypes projected along each dimension, providing three one-dimensional clusters per dimension, that were afterwards quantified into as many fuzzy sets. A two-dimensional plot of the results of the information granulation process is illustrated in fig. 8.5.

Following common sense, the three fuzzy sets on each dimension were assigned with the labels "Low", "Medium" and "High". On the basis of the fuzzy sets, four information granules have been derived and symbolically represented as:

1. PETAL LENGTH IS LOW AND PETAL WIDTH IS HIGH AND SEPAL LENGTH IS LOW AND SEPAL WIDTH IS LOW

2. PETAL LENGTH IS MEDIUM AND PETAL WIDTH IS LOW AND SEPAL

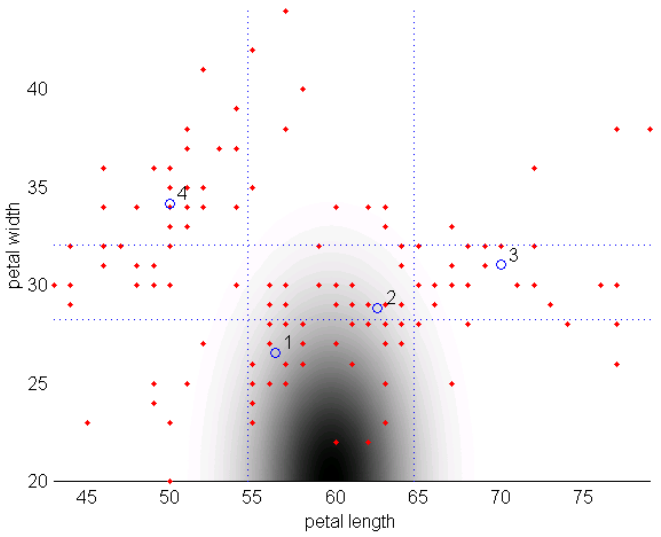

Figure 8.6: Distribution of membership degrees for the first information granule on the petal length – petal width plane. The darker the area, the higher the membership degree.

LENGTH IS MEDIUM AND SEPAL WIDTH IS MEDIUM

3. PETAL LENGTH IS MEDIUM AND PETAL WIDTH IS MEDIUM AND SEPAL LENGTH IS LOW AND SEPAL WIDTH IS LOW

4. PETAL LENGTH IS HIGH AND PETAL WIDTH IS MEDIUM AND SEPAL LENGTH IS HIGH AND SEPAL WIDTH IS HIGH

Due to the fuzzy semantics of the information granules, all input patterns have non zero membership degree to each proposition, even if they do not fall in regions of pattern space covered by the granules. This effect is depicted in figs. 8.6–8.9, where the influence of each information granule on the input data is shown in the subspace "petal length-petal width".

To verify how much the granules induced by the proposed approach are useful in providing good mapping properties, a further experiment was carried out using a 10-fold cross validation technique. Specifically, in each of the ten trials, the training set was used to perform information granulation and to build fuzzy rule-based model on the basis of the extracted granules (as described in Chapter 2), while the test set was used to check the classification ability of the constructed fuzzy classifiers. Such classifiers (denoted

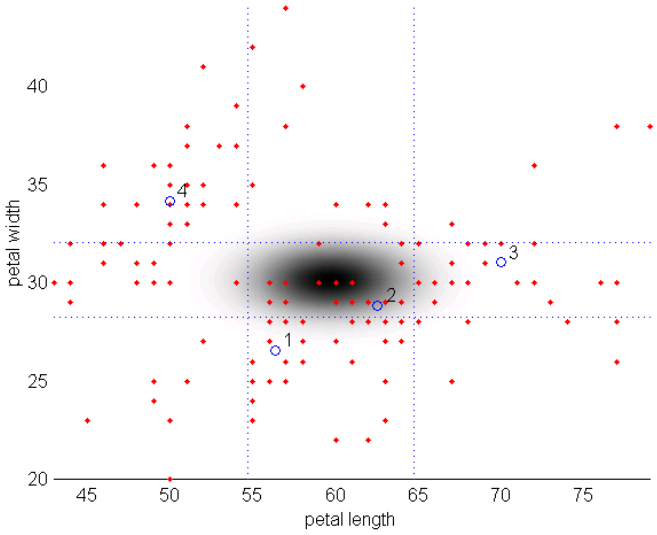

Figure 8.7: Distribution of membership degrees for the second information granule on the petal length – petal width plane. The darker the area, the higher the membership degree.

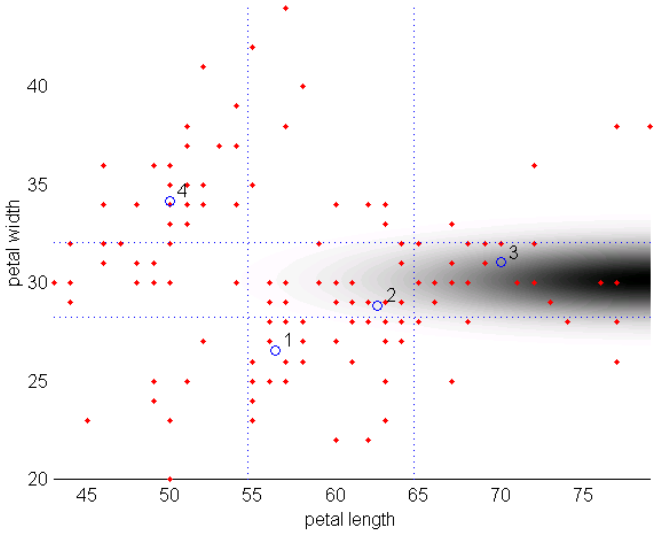

Figure 8.8: Distribution of membership degrees for the third information granule on the petal length – petal width plane. The darker the area, the higher the membership degree.

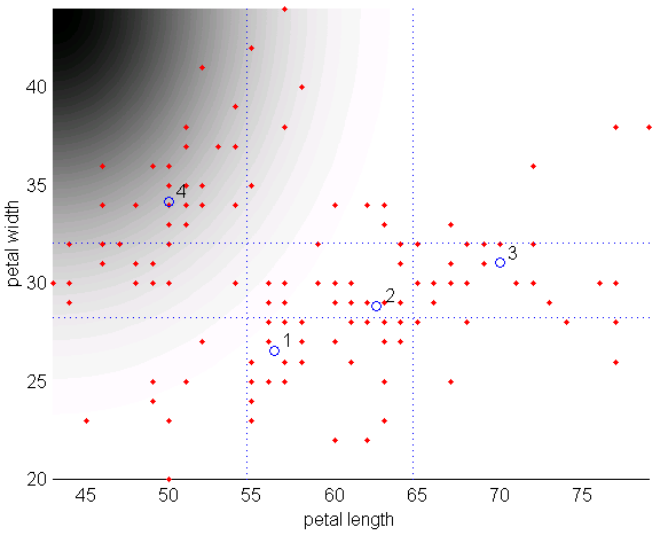

Figure 8.9: Distribution of membership degrees for the fourth information granule on the petal length – petal width plane. The darker the area, the higher the membership degree.

in the following by DC) were derived by considering 3 and 5 fuzzy sets per dimension (fspd) respectively.

Moreover, the DC classifiers were compared with fuzzy classifiers based on standard FCM, with different number of p clusters, where

$$p \in \{3, 4, 5, 6, 8, 10, 15, 20\}$$

. In the latter case, the classification mapping was defined in the following way. Given the partition matrix U, the class membership for each cluster prototype is given by $U \cdot K$ scaled columnwise by the cluster cardinality, where $K = [k_{tl}]$ such that k_{tl} is 1 if the t-th pattern belongs to the l-th class, 0 otherwise. After that, fuzzy granules were derived from prototypes by associating a Gaussian membership function with center in each prototype and circular amplitude found by trial-and-error (selection of the amplitude with minimum classification error).

The classification results of the fuzzy classifiers are summarized in table 8.1 and illustrated in fig. 8.10. As it can be seen, the quality of the DC fuzzy classifiers overcomes the FCM-based fuzzy classifiers both in terms of classification ability and compactness of the rule base. Indeed, as the number of clusters increases, the number of rules of DC classifiers keeps low with respect to that of FCM classifiers. In addition, fuzzy rules of DC classifiers

Table 8.1: Classification results of the fuzzy classifiers. Classification error is expressed in terms of misclassification rate (%)

	FCM			**DC 3 fspd**				**DC 5 fspd**			
	Test Error		Rul	Test Error		Rules		Test Error		Rules	
p	Mean	Std	es	Mean	Std	Min	Max	Mean	Std	Min	Max
3	20.0	10.8	3	14.0	7.6	3	3	-	-	-	-
4	18.0	7.3	4	14.7	6.5	4	4	-	-	-	-
5	16.7	8.6	5	12.7	7.0	5	5	19.3	11.3	5	5
6	16.0	8.0	6	10.7	8.0	5	6	14.0	7.6	6	6
8	13.3	18.1	8	6.7	8.0	6	8	7.1	8.5	8	8
10	5.3	7.8	10	7.3	4.7	6	8	5.3	5.0	9	10
15	5.3	4.0	15	4.0	3.3	8	10	4.7	3.1	12	15
20	3.3	4.5	20	5.3	5.8	7	9	5.3	6.5	15	19

are nicely interpretable, as it can be seen in the following, where the rule base of one of the DC fuzzy classifiers with 6 rules and 3 fuzzy sets per dimension is described, and from fig. 8.11 in which the fuzzy sets on each dimension are plotted.

1. IF PETAL LENGTH IS LOW AND PETAL WIDTH IS LOW AND SEPAL LENGTH IS MEDIUM AND SEPAL WIDTH IS MEDIUM THEN FLOWER IS SETOSA WITH DEGREE 0.02, VIRGINICA WITH DEGREE 0.86, VERSICOLOUR WITH DEGREE 0.11

2. IF PETAL LENGTH IS LOW AND PETAL WIDTH IS MEDIUM AND SEPAL LENGTH IS LOW AND SEPAL WIDTH IS LOW THEN FLOWER IS SETOSA WITH DEGREE 0.94, VIRGINICA WITH DEGREE 0.05, VERSICOLOUR WITH DEGREE 0.02

3. IF PETAL LENGTH IS LOW AND PETAL WIDTH IS HIGH AND SEPAL LENGTH IS LOW AND SEPAL WIDTH IS LOW THEN FLOWER IS SETOSA WITH DEGREE 0.94, VIRGINICA WITH DEGREE 0.05, VERSICOLOUR WITH DEGREE 0.02

4. IF PETAL LENGTH IS MEDIUM AND PETAL WIDTH IS HIGH AND SEPAL LENGTH IS LOW AND SEPAL WIDTH IS MEDIUM THEN FLOWER IS SETOSA WITH DEGREE 0.01, VIRGINICA WITH DEGREE 0.63, VERSICOLOUR WITH DEGREE 0.36

5. IF PETAL LENGTH IS MEDIUM AND PETAL WIDTH IS MEDIUM AND SEPAL LENGTH IS HIGH AND SEPAL WIDTH IS HIGH THEN FLOWER IS

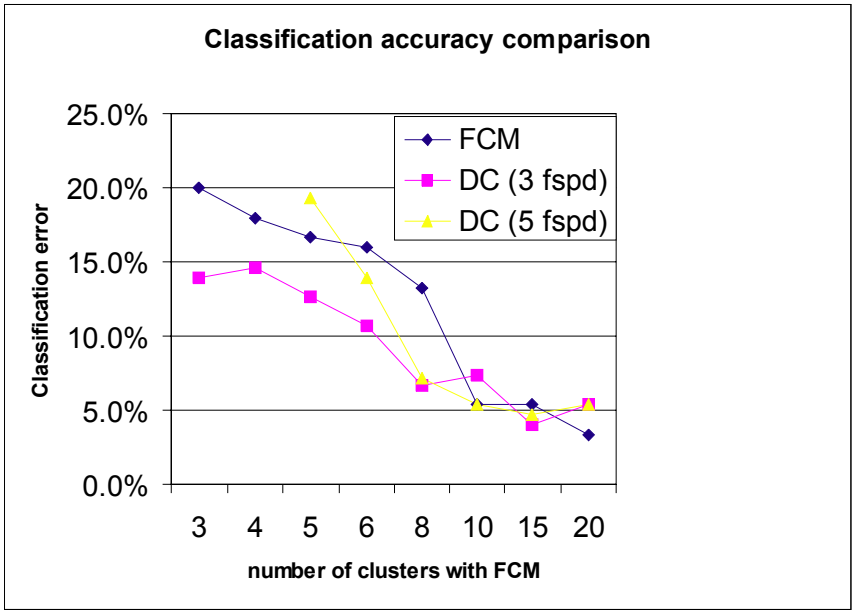

Figure 8.10: Comparison of the fuzzy classifiers in terms of classification error

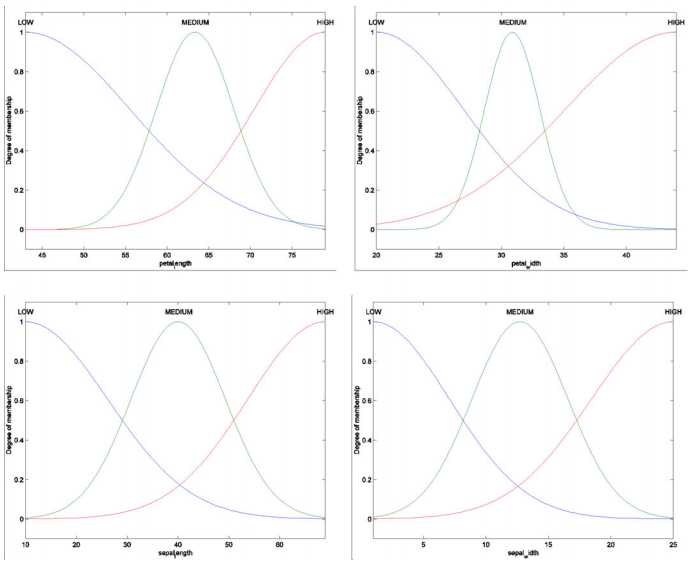

Figure 8.11: Fuzzy sets used to define the rulebase for Iris classification

SETOSA WITH DEGREE 0.01, VIRGINICA WITH DEGREE 0.17, VERSICOLOUR WITH DEGREE 0.82

6. IF PETAL LENGTH IS HIGH AND PETAL WIDTH IS MEDIUM AND SEPAL LENGTH IS HIGH AND SEPAL WIDTH IS HIGH THEN FLOWER IS SETOSA WITH DEGREE 0.01, VIRGINICA WITH DEGREE 0.11, VERSICOLOUR WITH DEGREE 0.87

8.3.2 Fuzzy Diagnosis

To assess the effectiveness of the Double Clustering Framework in solving real-world problems, the well-known Wisconsin Breast Cancer (WBC) classification problem has been considered. The original dataset consists of 699 examples of breast cancer diagnosis. Each example is formed by ten attributes plus the class diagnosis (benign/malignant). Full details about the dataset can be found in (Blake and Merx, 1998).

In order to effectively apply the proposed framework, the dataset has been cleaned by removing one attribute (corresponding to the patient IDs) and all the examples with at least one missing value. The total number of removed examples is 16, hence the cardinality of the dataset used in the fuzzy information granulation process is 683.

The dataset has been split according to the stratified 10-fold cross-validation scheme. For each fold, the Double Clustering Framework has been applied, and the resulting multidimensional granules have been used to define fuzzy rules. Such rules have been employed to infer breast cancer diagnosis for both the training set and the test set of the fold.

Two implementations of the Double Clustering Framework have been considered – namely the Crisp Double Clustering and DCClass – in order to observe their different behavior in solving the same classification problem. For the Crisp Double Clustering, the simulation has been repeated with different choices of both the number of multidimensional prototypes and the number of fuzzy set for each dimension. In particular, the number of prototypes ranges from 2 to 10, while the number of fuzzy sets per dimension varies from 2 to 4. In each trial, the coverage level ε is fixed to 0.5.

In table 8.2, an example of fuzzy rule is reported. As it can be seen, readability of the extracted rules is immediate. In figs. 8.12 and 8.13, the derived one-dimensional fuzzy sets for two input features are illustrated. Moreover, classification results are reported in tables 8.3 and 8.4 for the training set and the test set respectively, by varying number of multidimensional prototypes and fuzzy sets per input (equal for each dimension). Each value is the average for all the 10 folds used in the simulation.

Table 8.2: An example of rule induced by the proposed method

IF
ClumpThickness is HIGH AND
UniformityOfCellSize is MED-LOW AND
UniformityOfCellSize is MED-LOW AND
MarginalAdhesion is MED-HIGH AND
SingleEpithelialCellSize is MED-LOW AND
BareNuclei is HIGH AND
BlandChromatin is MED-LOW AND
NormalNucleoli is MED-HIGH AND
Mitoses is MED-LOW
THEN
cancer is BENIGN with degree 0.99171,
MALIGN with degree 0.0082883.

The obtained results on the breast-cancer problem are comparable with those obtained by other techniques proposed in literature (see, e.g. (Abonyi et al., 2002)), with the additional feature of that the derived rules are expressed in a nicely human-interpretable way.

The same classification problem has been tackled through the application of DCClass. In this case, however, the number of fuzzy sets per input is not specified as it is automatically defined within the granulation process.

By fixing the number of multidimensional prototypes to six, the application of DCClass provides ten rule sets with average classification error on the test sets of 3.975%, while the mean number of rules is 3.6. When compared with similar approaches, like NEFCLASS (Nauck and Kruse, 1999), the achieved results confirm that the proposed tool is a valid technique to extract accurate knowledge from data.

To assess the transparency of the resulting information granules, a rule set of five rules has been chosen, which provides a classification error of 1.471% on the test set. The one-dimensional fuzzy sets derived from the application of DCClass satisfy the interpretability constraints defined in Section 8.1, as also illustrated in figs. 8.14–8.16 for three input dimensions. For such fuzzy sets, the association of meaningful linguistic label is straightforward.

8.4 Final remarks

In this Chapter, a framework for fuzzy information granulation has been presented. Its key feature is the ability of generating information granules from

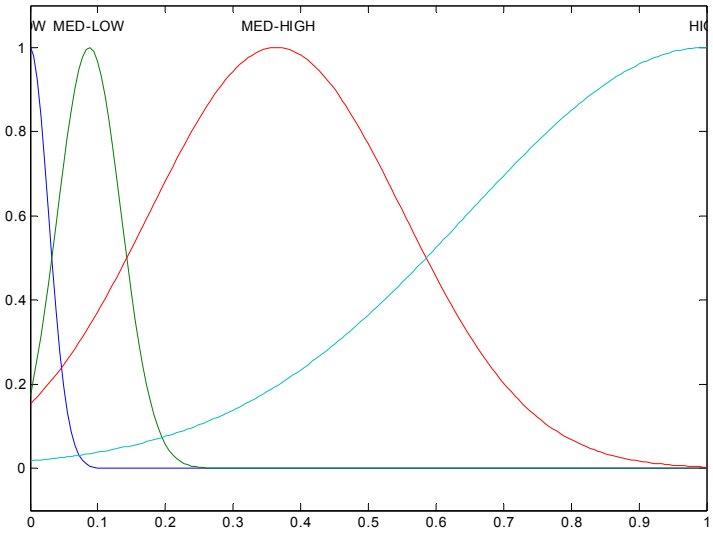

Figure 8.12: One-dimensional fuzzy sets derived for the "Bare Nuclei" feature

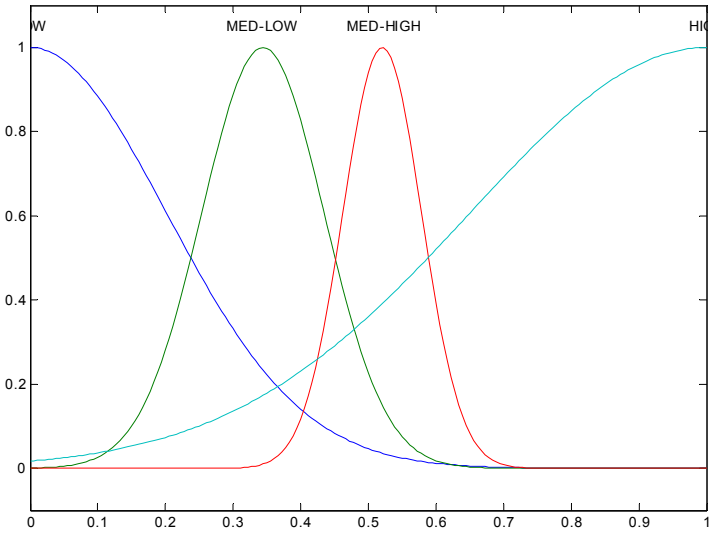

Figure 8.13: One-dimensional fuzzy sets derived for the "Uniformity of Cell Size" feature

Table 8.3: Crisp Double Clustering: Mean classification error for the training set

	Fuzzy sets per input		
p	2	3	4
2	3.80%	-	-
3	3.17%	4.71%	-
4	3.74%	4.21%	4.89%
5	3.93%	3.98%	3.84%
6	4.29%	3.37%	3.72%
7	6.10%	3.51%	3.58%
8	4.57%	3.33%	3.89%
9	5.19%	3.12%	3.93%
10	5.93%	3.28%	3.84%

Table 8.4: Crisp Double Clustering: Mean classification error for the test set

	Fuzzy sets per input		
p	2	3	4
2	3.97%	-	-
3	3.24%	4.86%	-
4	3.83%	4.56%	4.56%
5	3.53%	3.68%	3.68%
6	4.42%	3.83%	4.71%
7	6.48%	3.38%	3.68%
8	4.41%	3.53%	4.12%
9	4.42%	4.27%	4.12%
10	5.44%	3.39%	3.97%

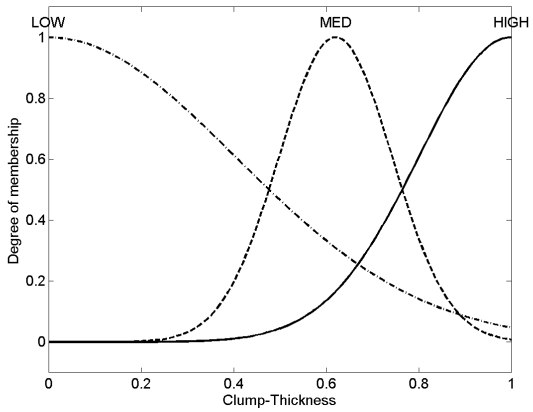

Figure 8.14: The Frame of Cognition defined for the attribute "Clump-Thickness"

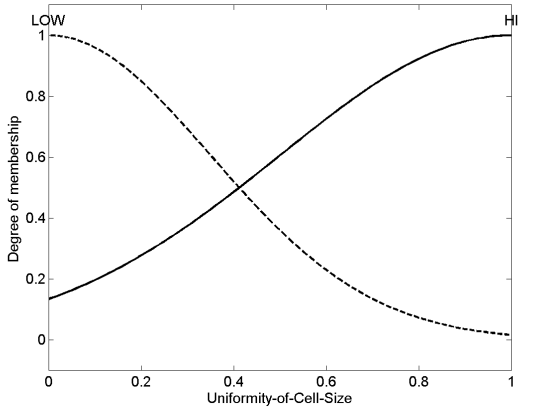

Figure 8.15: The Frame of Cognition defined for the attribute "Uniformity of cell size"

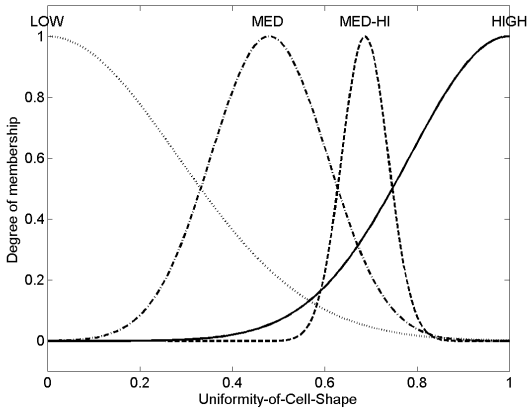

Figure 8.16: The Frame of Cognition defined for the attribute "Uniformity of cell shape"

data which can be represented in terms of qualitative properties on data. As supported by experimental examples, this feature turns very useful in all applicative contexts – like fuzzy diagnosis – where an intelligible description of a phenomenon (like symptoms/diagnosis relationships) is of great importance.

The framework is general enough to be extended with more sophisticated techniques that may support a greater number of interpretability constraints, which could lead to the generation of knowledge structures that are more comprehensible by human users.

Chapter 9

Refinement of Interpretable Information Granules through Neural Learning

9.1 Introduction

Fuzzy information granules can be used as building blocks for designing fuzzy models, often rule-based, which can be effectively employed to solve soft-computing problems such as predictions, classifications, system identification, etc. The added value of fuzzy models over "black-box" models (like neural networks) is the explicit representation of the knowledge base in terms of symbolic structures, with the additional capability of handling vaguely defined concepts by means of fuzzy sets. In this context, interpretable fuzzy information granulation plays an essential role in providing information granules that can be naturally associated to linguistic labels coming from the natural language. As a result of such association, the knowledge base embodied into a fuzzy model is easy to read and understand by human users.

From a functional point of view, fuzzy models are approximative representations of an underlying functional input/output relationship that has to be discovered. To achieve such objective, often a dataset is available, which provides a finite collection of input/output examples from which the model's knowledge has to be built. Information granulation techniques operate on such dataset to properly extract information granules that will be used to define the model's knowledge base. In this context, information granulation can be accomplished in two main ways:

1. The complete dataset of examples is granulated, involving both input and output information. At the end of the granulation process, the

knowledge base of the fuzzy model is complete and can be immediately used for approximating the underlying functional relationship. A typical example of fuzzy model defined by this kind of knowledge base is the Mamdani Fuzzy Inference System;

2. The examples are split in the input and output components, and the granulation process is executed only in the input part. This choice is typical in classification problems, where outputs are class labels that belong to a symbolic domain, which cannot be granulated. This form of granulation is also chosen when accuracy has a great impact in the assessment of the model quality.

When the second form of granulation is chosen, the resulting information granules are able to describe the relationships among input data, but they miss to correctly represent the underlying input/output relationship. When output information is available, the knowledge base of the fuzzy model has to be *refined* in order to better approximate the unknown functional relationship, by means of supervised learning schemes[1].

The problem of model refinement through supervised learning has been effectively solved by means of neuro-fuzzy networks. Neuro-fuzzy networks are translational architectures that map each element of the fuzzy model into an element of a suited neural network and vice-versa (Cloete and Zurada, 2000). This bijective mapping allows for a dual view of the same system: as a fuzzy model or as a neural network. In this way, learning techniques available for neural networks can be effectively applied to modify the parameters of the fuzzy model, whose knowledge base can thus be refined to accommodate the underlying functional mapping.

Neural learning schemes are very flexible and are able to adjust the network parameters so as to finely approximate the unknown input/output relationship described by the available dataset of examples. However, this strong flexibility may turn as a drawback in interpretable fuzzy modelling, since parameters of the network – which are reflected into parameters of the model's knowledge base and hence of the information granules – can easily change in a way that violates the interpretability constraints which guided the information granulation process. This drawback is particularly felt when information granules represent qualitative properties on data, since interpretability constraints on such kind of granules are more stringent.

[1]It should be noted that information granulation on the input space can be considered as an unsupervised learning scheme. The fuzzy model resulting from both unsupervised and supervised learning is then a hybrid model.

To protect interpretability of the model's knowledge base, several approaches have been proposed. Some of them apply a regularization term to the learning objective, so as to penalize those configurations of parameters that violate interpretability constraints (Jin et al., 1998a; Jin and Sendhoff, 2003; Lozowski and Zurada, 2000; Valente de Oliveira, 1999b). Regularized learning is effective for interpretability constraints that can be partially violated, and the violation degree must be represents as an explicit function of the network parameters. This condition cannot be easily applied to all interpretability constraints (e.g. leftmost/rightmost fuzzy sets, natural zero positioning, etc.), thus regularized learning could be ineffective in some cases. Furthermore, regularized learning procedures introduce new hyper-parameters whose automatic handling is difficult unless computationally complex mathematical techniques (Bengio, 2000; Craven and Wabba, 1979) – or time-consuming trial-and-error strategies – are adopted.

An alternative to regularized learning procedure is the adoption of genetic algorithms, which provide for an evolutive process of network configurations that combine themselves and mutate in order to survive over a selection process that promotes accurate and interpretable knowledge representations (Cordón et al., 2003; Marìn-Blàzquez and Shen, 2002; Riid et al., 2000; Herrera et al., 1995; Ishibuchi and Yamamoto, 2002; Ishibuchi et al., 1997). As an additional feature, the selection process can evaluate complex and multi-objective fitness functions that can represent interpretability constraints of different representations. The main disadvantage of genetic algorithms is their computational cost, which could be too high to be acceptable in many applicative contexts.

Other approaches have been proposed to cope with interpretable neural learning, including the one described in (Altug et al., 1999; Chow et al., 1999) that make use of a classical back-propagation learning scheme integrated with a procedure that projects the actual configuration of parameters computed by back-propagation into another configuration of parameters that is close to the actual but verifies a set of interpretability constraints. The projection operation is applied after each back-propagation iteration, resulting in high computational cost.

To deal with computational cost of interpretable learning schemes, Nauck and Kruse propose a family of fuzzy models, NEFCLASS, NEFCON and NEFPROX, which are translated in specific neural architectures that are trained by means of heuristic learning rules to preserve interpretability while simultaneously adapting the knowledge base to data (Nauck et al., 1997). This approach is interesting for its low computational cost, but the introduction of heuristic learning rules poses some serious questions concerning their convergence properties and optimality of the found solutions. Furthermore,

good interpretation of the learning results cannot always be guaranteed, especially for high-dimensional problems. Hence, in (Nauck and Kruse, 1999) the NEFCLASS algorithm is added with interactive strategies for pruning rules and variables so as to improve readability. This approach provides good results, but it results in a long interactive process that cannot extract automatically rules from data but requires the ability of the user to supervised and interpret the learning procedure in all its stages.

The problem of interpretable learning from data can be theoretically analyzed by observing that the space of the network parameters actually contains a small subset of configuration referring to an interpretable knowledge. The major difficulty is hence to characterize this subset in such a way to provide a specific neural architecture whose possible parameter configurations belong to this subspace only. Once this neural architecture is defined, classical learning schemes can be safely applied, since the adaption process will entirely take place within the subspace of interpretable configurations. As a result, the knowledge base mapped in the neural network will be interpretable before, during and after the learning process. The rest of the Chapter is devoted in describing this approach in a greater detail.

9.2 A neuro-fuzzy network to learn interpretable information granules

In this Section, a neuro-fuzzy architecture is proposed whose parameter space is restricted so that only interpretable fuzzy models are representable. First, such subspace (named "*subspace of interpretable configurations*") is mathematically characterized. On the basis of such characterization, a neuro-fuzzy architecture is defined. Finally, a learning scheme based on the well-known gradient-descent strategy is formalized and analyzed.

9.2.1 The subspace of interpretable configurations

To characterize the subspace of interpretable configurations, the following interpretability constraints are considered:

- Normality;
- Convexity;
- Unimodality;
- Continuity;

- Weak proper ordering;
- Coverage (more specifically, ε-coverage for $\varepsilon > 0$);
- Distinguishability;
- Leftmost/Rightmost fuzzy sets.

It should be noted that such constraints are the same required for qualitative information granulation. In this way, interpretable fuzzy information granules resulting from this form of granulation can be directly applied to design fuzzy models which can be implemented according to the proposed architecture. The chosen characterization for distinguishability is by means of the possibility function. For a simplified analysis, the maximum allowed possibility in a Frame of Cognition is set to ε, i.e. the same threshold for coverage. Furthermore, the value ε is assumed to be constant for each Frame of Cognition[2].

The aim of this Section is to characterize the subspace of interpretable configurations independently from the specific information granulation technique used to build the knowledge base. As a consequence, the simplest assumptions will be made on the knowledge base, which is characterized by a set of R rules of the form

$$\text{RULE } r \text{ :IF } v_1 \text{ IS } \mathcal{L}\left(A^{(1)}_{g(r,1)}\right) \text{ AND } \ldots \text{ AND } v_n \text{ IS } \mathcal{L}\left(A^{(n)}_{g(r,n)}\right)$$
$$\text{THEN } <\textit{Consequent}>$$

where $\mathcal{L}(A^{(i)}_{g(r,i)})$ is the label associated to fuzzy set $A^{(i)}_{g(r,i)}$ according to the Frame of Cognition $\mathbf{F}_i$ of the i-th input feature. The consequent of the rule is left unspecified since interpretability constraints do not apply to this part of the rule. For convenience, the index specification of each fuzzy set is represented as a function $g : \{1, 2, \ldots, R\} \times \{1, 2, \ldots, n\} \rightarrow \mathbb{N}$, so that $g(r, i) = k$ if in rule r the k-th fuzzy set of the Frame of Cognition i is used. The adoption of function g facilitates the representation of the rule base within the neural architecture[3].

For the i-th input dimension, $i = 1, 2, \ldots, n$, a Frame of Cognition consisting of K_i fuzzy sets is defined and properly labelled. The definition of such fuzzy sets, as well as the combinations of them occurring in the rules

[2] As will be clear forth, this is not a stringent requirement, which is set only for easining the discussion.

[3] Specifically, $g(i, j) = k$ means that the k-th node of the j-th input in the Membership Layer is connected to the i-th neuron of the Rule Layer

antecedents is determined by the information granulation algorithm, like any implementation of the Double Clustering Framework.

In most cases, fuzzy sets belonging to Frames of Cognition have a membership function with a fixed functional form, and one fuzzy set differ from any other by the values of a set of parameters. As an example, Gaussian fuzzy set differ for their centers and amplitudes. As a consequence, the array of fuzzy sets occurring in a frame of cognition, properly sorted by the related order relation of the frame, can be mapped bijectively into an array of parameters. For Gaussian fuzzy sets of center ω and width σ, a Frame of Cognition represented by fuzzy sets

$$\left\langle A_1^{(i)}, A_2^{(i)}, \ldots, A_{K_i}^{(i)} \right\rangle \tag{9.1}$$

is directly mapped into the vector that includes both centers and widths of each fuzzy set. The resulting parameter space can be defined as:

$$\Omega^{(i)} = \left\{ \begin{array}{c} \left(\omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_{K_i}^{(i)}, \sigma_1^{(i)}, \sigma_2^{(i)}, \ldots, \sigma_{K_i}^{(i)}\right) \in \mathbb{R}^{2K_i} : \\ \forall h : m_i \leq \omega_h^{(i)} < \omega_{h+1}^{(i)} \leq M_i \wedge \sigma_i^{(i)} > 0 \end{array} \right\} \tag{9.2}$$

where the order relation $\omega_h^{(i)} \leq \omega_{h+1}^{(i)}$ is given by the proper ordering of fuzzy sets and by unimodality of Gaussian membership functions. It is also assumed that each fuzzy set of the Frame of Cognition is defined on the Universe of Discourse $U_i = [m_i, M_i]$, as usual.

Due to the independence of the elements of the parameter vector, the dimensionality of each parameter space is $2K_i$. The entire parameter space of the fuzzy model is defined as

$$\Omega = \Omega^{(1)} \times \cdots \times \Omega^{(n)} \tag{9.3}$$

and has dimensionality

$$\dim \Omega = 2 \sum_{i=1}^{n} K_i \tag{9.4}$$

Let

$$\Omega^* = \Omega^{*(1)} \times \cdots \times \Omega^{*(n)} \tag{9.5}$$

be the subspace of interpretable configurations, i.e. the subspace of parameters that correspond to fuzzy sets satisfying the aforementioned interpretability constraints. If Gaussian fuzzy sets are used, then the properties of normality, unimodality, continuity and convexity are satisfied for any value

assumed by centers and amplitudes. As a consequence, the constraints on coverage and distinguishability – together with leftmost and rightmost constraints – contribute in restricting the set of admissible values for centers and amplitudes.

To characterize Ω^* by means of formal conditions that enable the evaluation of its dimensionality, the following lemma is particularly useful.

LEMMA 9.1 *For each pair of one-dimensional Gaussian membership functions (defined on the same Universe of Discourse) with centers and widths ω_1, ω_2 (such that $\omega_1 < \omega_2$) and σ_1, σ_2 respectively, there exists a unique point t between their centers, for which the membership values are identical.*
Proof. *The explicit representation of the Gaussian membership functions are:*

$$\mu_k(x) = \exp\left(-\frac{(x-\omega_k)^2}{2\sigma_k^2}\right), \ k = 1, 2 \tag{9.6}$$

The intersection points of the two functions are:

$$T = \{t \in \mathbb{R} : \mu_1(t) = \mu_2(t)\} \tag{9.7}$$

By solving the equation characterizing the set T, the following point are obtained:

$$T = \begin{cases} \left\{\frac{\sigma_1\omega_2 - \sigma_2\omega_1}{\sigma_1 - \sigma_2}, \frac{\sigma_1\omega_2 + \sigma_2\omega_1}{\sigma_1 + \sigma_2}\right\} & \text{if} \quad \sigma_1 \neq \sigma_2 \\ \left\{\frac{\sigma_2\omega_1 + \sigma_1\omega_2}{\sigma_1 + \sigma_2}\right\} & \text{if} \quad \sigma_1 = \sigma_2 \end{cases} \tag{9.8}$$

In both cases, the point:

$$t = \frac{\sigma_1\omega_2 + \sigma_2\omega_1}{\sigma_1 + \sigma_2} \tag{9.9}$$

is located between ω_1 and ω_2 since it is a convex combination of the two centers. If $\sigma_1 > \sigma_2$, then:

$$\omega_1 < \omega_2 \implies \frac{\sigma_1\omega_2 - \sigma_2\omega_1}{\sigma_1 - \sigma_2} > \omega_2 \tag{9.10}$$

Conversely, if $\sigma_1 < \sigma_2$, then

$$\omega_1 < \omega_2 \implies \frac{\sigma_1\omega_2 - \sigma_2\omega_1}{\sigma_1 - \sigma_2} < \omega_1 \tag{9.11}$$

In both cases, the point $\frac{\sigma_1\omega_2 - \sigma_2\omega_1}{\sigma_1 - \sigma_2}$ is outside the range $[\omega_1, \omega_2]$; hence the point t is unique. ■

If the fuzzy sets of a Frame of Cognition are constrained so as to satisfy ε-coverage and ε-possibility[4], then the following proposition can be also asserted:

[4]With the term ε-possibility, it is intended that the maximum allowed possibility in a Frame of Cognition is ε.

PROPOSITION 9.1 *Let* $\left\langle A_1^{(i)}, A_2^{(i)}, \ldots, A_{K_i}^{(i)} \right\rangle$ *be the family of ordered Gaussian fuzzy set defined in a Frame of Cognition* $\mathbf{F}_i$*. Suppose that the frame verifies* ε*-coverage and* ε*-possibility. Then, for any two adjacent fuzzy sets* $A_k^{(i)}, A_{k+1}^{(i)}$ *there exists a unique intersection point* t_k *such that* $\mu_{A_k^{(i)}}(t_k) = \varepsilon = \mu_{A_{k+1}^{(i)}}(t_k)$.

Proof. *Consider the intersection point* t_k *between two adjacent fuzzy sets* $A_k^{(i)}, A_{k+1}^{(i)}$ *located between the two centers* $\omega_k^{(i)}, \omega_{k+1}^{(i)}$*. By virtue of the previous Lemma, such point is unique. Three cases can be considered to analyze the value of* $\mu_{A_k^{(i)}}(t_k)$*. If, ad absurdum,* $\mu_{A_k^{(i)}}(t_k) > \varepsilon$*, then* $\Pi\left(A_k^{(i)}, A_{k+1}^{(i)}\right) > \varepsilon$*, but this is not possible since* ε*-possibility is verified by hypotheses. Suppose, still ad absurdum, that* $\mu_{A_k^{(i)}}(t_k) < \varepsilon$*; by* ε*-coverage there would be a third fuzzy set* $A_h^{(i)}$ *such that* $\mu_{A_h^{(i)}}(t_k) \geq \varepsilon$ *and* $h \notin \{k, k+1\}$*. Because of the proper ordering of the frame of cognition, the center* $\omega_h^{(i)}$ *is not located between* $\omega_k^{(i)}$ *and* $\omega_{k+1}^{(i)}$*. As a first case, consider* $\omega_h^{(i)} < \omega_k^{(i)}$*. In the interval* $\left[\omega_h^{(i)}, t_h\right[$ *the membership value of* $A_h^{(i)}$ *is always greater than* ε *and, in the unique intersection point* t'_{hk} *between* $\omega_h^{(i)}$ *and* $\omega_k^{(i)}$ *there would be* $\mu_{A_h^{(i)}}(t'_{hk}) = \mu_{A_k^{(i)}}(t'_{hk}) > \varepsilon$*, thus violating the* ε*-possibility constraint. Similarly, if* $\omega_h^{(i)} > \omega_{k+1}^{(i)}$ *there would exist a point* t'_{hk+1} *between* $\omega_h^{(i)}$ *and* $\omega_{k+1}^{(i)}$ *such that* $\mu_{A_h^{(i)}}\left(t'_{hk+1}\right) = \mu_{A_{k+1}^{(i)}}\left(t'_{hk+1}\right) > \varepsilon$*. This is absurd, hence the fuzzy set* $A_h^{(i)}$ *cannot exist and* $\mu_{A_k^{(i)}}(t_k)$ *cannot be less than* ε*. Since Gaussian membership functions are continuos and have range* $]0, 1]$*, then it must be* $\mu_{A_k^{(i)}}(t_k) = \varepsilon = \mu_{A_{k+1}^{(i)}}(t_k)$. ■

As an important corollary of the proposition, the following theorem can be stated:

THEOREM 9.1 *Let* $\left\langle A_1^{(i)}, A_2^{(i)}, \ldots, A_{K_i}^{(i)} \right\rangle$ *be the family of ordered Gaussian fuzzy set defined in a Frame of Cognition* $\mathbf{F}_i$*. Suppose that the frame verifies* ε*-coverage and* ε*-possibility. The, there exists a unique vector* $\mathbf{t}^{(i)} = \left(t_1^{(i)}, t_2^{(i)}, \ldots, t_{K_i-1}^{(i)}\right)$ *of intersection points between adjacent fuzzy sets of the frame. Such intersection points verify the following inequalities:*

$$m_i < t_1^{(i)} < t_2^{(i)} < \ldots < t_{K_i-1}^{(i)} < M_i \tag{9.12}$$

Proof. *The existence and uniqueness of* $\mathbf{t}^{(i)}$ *is a trivial consequence of Proposition 9.1. The inequalities are also a direct consequence of the weak proper ordering of the fuzzy sets in the Frame of Cognition.* ■

The last theorem is important because it reduces the vector of parameters of a Frame of Cognition into a vector of $K_i - 1$ intersection points between

adjacent fuzzy sets. By varying the parameters of the fuzzy sets in the frame, also the vector $\mathbf{t}^{(i)}$ varies in a space that can be defined as:

$$T^{(i)} = \left\{ \begin{array}{c} \left(t_1^{(i)}, t_2^{(i)}, \ldots, t_{K_i-1}^{(i)}\right) \in \mathbb{R}^{K_i-1} : \\ m_i < t_1^{(i)} < t_2^{(i)} < \ldots < t_{K_i-1}^{(i)} < M_i \end{array} \right\} \tag{9.13}$$

An important question concerns the possibility of expanding a vector of intersection points into a vector of parameters characterizing the membership functions of a Frame of Cognition. The following theorem affirmatively answers the question.

THEOREM 9.2 *Let*

$$\mathbf{t}^{(i)} = \left(t_1^{(i)}, t_2^{(i)}, \ldots, t_{K_i-1}^{(i)}\right) \in T^{(i)} \tag{9.14}$$

be a set of intersection points in $[m_i, M_i]$. *Then there exists a unique vector*

$$\left(\omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_{K_i}^{(i)}, \sigma_1^{(i)}, \sigma_2^{(i)}, \ldots, \sigma_{K_i}^{(i)}\right) \in \Omega^{*(i)} \tag{9.15}$$

that corresponds to an ordered collection of Gaussian fuzzy sets

$$\left\langle A_1^{(i)}, A_2^{(i)}, \ldots, A_{K_i}^{(i)} \right\rangle \tag{9.16}$$

with ε*-coverage,* ε*-possibility and intersection point* $t_k^{(i)}$ *between two adjacent fuzzy sets* $A_k^{(i)}, A_{k+1}^{(i)}$, *for* $k = 1, 2, \ldots, K_i - 1$.
Proof. *Consider the following dummy points:*

$$t_0^{(i)} = 2m_i - t_1^{(i)} \tag{9.17}$$

and

$$t_{K_i}^{(i)} = 2M_i - t_{K_i-1}^{(i)} \tag{9.18}$$

and consider the following values:

$$\omega_k^{(i)} = \frac{t_{k-1}^{(i)} + t_k^{(i)}}{2} \tag{9.19}$$

and

$$\sigma_k^{(i)} = \frac{t_k^{(i)} - t_{k-1}^{(i)}}{2\sqrt{-2\ln\varepsilon}} \tag{9.20}$$

for $k = 1, 2, \ldots, K_i$. *Since* $\forall k : t_{k-1}^{(i)} < t_k^{(i)} < t_{k+1}^{(i)}$, *then* $\omega_k^{(i)} < \omega_{k+1}^{(i)}$ *and* $\sigma_k^{(i)} > 0$, *hence the vector* $\left(\omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_{K_i}^{(i)}, \sigma_1^{(i)}, \sigma_2^{(i)}, \ldots, \sigma_{K_i}^{(i)}\right)$ *belongs to* $\Omega^{(i)}$. *The unique intersection point between the Gaussian fuzzy sets* $A_k^{(i)}, A_{k+1}^{(i)}$ *induced by parameters* $\omega_k^{(i)}, \omega_{k+1}^{(i)}$ *and* $\sigma_k^{(i)}, \sigma_{k+1}^{(i)}$ *is*

$$t = \frac{\sigma_k^{(i)}\omega_{k+1}^{(i)} + \sigma_{k+1}^{(i)}\omega_k^{(i)}}{\sigma_k^{(i)} + \sigma_{k+1}^{(i)}} = \\ = \frac{\left(t_k^{(i)} - t_{k-1}^{(i)}\right)\left(t_k^{(i)} + t_{k+1}^{(i)}\right) + \left(t_{k+1}^{(i)} - t_k^{(i)}\right)\left(t_{k-1}^{(i)} + t_k^{(i)}\right)}{2\left(t_{k+1}^{(i)} + t_{k-1}^{(i)}\right)} = \\ = \frac{2t_k^{(i)}\left(t_{k+1}^{(i)} + t_{k-1}^{(i)}\right)}{2\left(t_{k+1}^{(i)} + t_{k-1}^{(i)}\right)} = t_k^{(i)} \quad (9.21)$$

The membership value at the intersection point is

$$\mu_{A_k^{(i)}}\left(t_k^{(i)}\right) = \exp\left(-\frac{\left(t_k^{(i)} - \omega_k^{(i)}\right)^2}{2\left(\sigma_k^{(i)}\right)^2}\right) = \exp\left(-\frac{\left(t_k^{(i)} - \frac{t_{k-1}^{(i)} + t_k^{(i)}}{2}\right)^2}{2\left(\frac{t_k^{(i)} - t_{k-1}^{(i)}}{2\sqrt{-2\ln\varepsilon}}\right)^2}\right) = \varepsilon \quad (9.22)$$

For $x \in \left[\omega_k^{(i)}, t_k^{(i)}\right]$ *then* $\mu_{A_k^{(i)}}(x) > \varepsilon$ *because* $A_k^{(i)}$ *is Gaussian. Similarly, for* $x \in \left[t_k^{(i)}, \omega_{k+1}^{(i)}\right]$ *results* $\mu_{A_{k+1}^{(i)}}(x) > \varepsilon$; *hence* ε*-coverage is verified by the fuzzy sets* $A_k^{(i)}$ *and* $A_{k+1}^{(i)}$ *in the interval* $\left[\omega_k^{(i)}, \omega_{k+1}^{(i)}\right]$. *Since:*

$$\bigcup_{k=1}^{K_i - 1}\left[\omega_k^{(i)}, \omega_{k+1}^{(i)}\right] = [m_i, M_i] \quad (9.23)$$

then ε*-coverage is verified in the entire domain of the* i*-th feature. To verify* ε*-possibility, consider two adjacent fuzzy sets* $A_k^{(i)}, A_{k+1}^{(i)}$; *for* $x < t_k^{(i)}$*:*

$$\mu_{A_{k+1}^{(i)}}(x) < \mu_{A_k^{(i)}}\left(t_k^{(i)}\right) = \varepsilon \implies \min\left\{\mu_{A_{k+1}^{(i)}}(x), \mu_{A_k^{(i)}}(x)\right\} < \varepsilon \quad (9.24)$$

Similarly, for $x > t_k^{(i)}$*:*

$$\mu_{A_k^{(i)}}(x) < \mu_{A_{k+1}^{(i)}}\left(t_k^{(i)}\right) = \varepsilon \implies \min\left\{\mu_{A_{k+1}^{(i)}}(x), \mu_{A_k^{(i)}}(x)\right\} < \varepsilon \quad (9.25)$$

As a consequence:

$$\Pi\left(A_{k+1}^{(i)}, A_k^{(i)}\right) = \sup_{x \in [m_i, M_i]} \min\left\{\mu_{A_{k+1}^{(i)}}(x), \mu_{A_k^{(i)}}(x)\right\} = \varepsilon \quad (9.26)$$

Thus ε-possibility is verified between $A_{k+1}^{(i)}$ and $A_k^{(i)}$. Consider a third fuzzy set $A_h^{(i)}$, and suppose $h > k+1$; if $\Pi\left(A_k^{(i)}, A_h^{(i)}\right) > \varepsilon$ then the unique intersection point t'_{hk} between $A_k^{(i)}$ and $A_h^{(i)}$ within $\left[\omega_k^{(i)}, \omega_h^{(i)}\right]$ would be greater than ε, as the possibility value coincides with the membership value at the intersection point. As a consequence, in the interval $\left[t'_{hk}, \omega_h^{(i)}\right]$ the membership value of $A_h^{(i)}$ is greater than ε. If $\omega_{h-1}^{(i)} \in \left[t'_{hk}, \omega_h^{(i)}\right]$ then the possibility $\Pi\left(A_{h-1}^{(i)}, A_h^{(i)}\right)$ would be greater than ε, but this is absurd because it has been proved that the possibility of two adjacent fuzzy sets is always equal to ε. If $\omega_{h-1}^{(i)} \in \left[\omega_k^{(i)}, t'_{hk}\right]$, then $\Pi\left(A_k^{(i)}, A_{h-1}^{(i)}\right) > \varepsilon$ since $\mu_{A_k^{(i)}}(x) > \varepsilon$ for $x \in \left[\omega_k^{(i)}, t'_{hk}\right]$. In this case, by performing a change of variable $h-1 \longmapsto h$ it is possible to repeat the same considerations as above, which will lead to an absurdum or to another change of variables. In all cases, it is proved that $\Pi\left(A_k^{(i)}, A_h^{(i)}\right) \leq \varepsilon$ for $h > k+1$. In a similar manner, it can be proved that $\Pi\left(A_k^{(i)}, A_h^{(i)}\right) \leq \varepsilon$ for $h < k$. ■

The theorems 9.1 and 9.2 state that there is a bijective correspondence between the spaces $\Omega^{*(i)}$ and $T^{(i)}$. Since each space $T^{(i)}$ has a geometric dimensionality $K_i - 1$, then the dimensionality of the space of interpretable configurations is:

$$\dim \Omega^* = \sum_{k=1}^{n} \dim \Omega^{(i)*} = \sum_{k=1}^{n} (K_i - 1) \tag{9.27}$$

Comparing (9.27) with (9.4), it is possible to affirm that:

$$\dim \Omega^* < \dim \Omega \tag{9.28}$$

that is, the number of free parameters necessary to generate interpretable configurations is smaller than the free parameters corresponding to the totality of centers and widths of all Gaussian fuzzy sets involved in the rule antecedents.

9.2.2 The Neuro-Fuzzy architecture

Proofs of theorems 9.1 and 9.2 provide a constructive procedure to transform elements of the space $T^{(i)}$ in elements of spaces $\Omega^{*(i)}$ of interpretable configurations of one-dimensional fuzzy sets. This procedure suggests a particular neuro-fuzzy architecture that is able to represent interpretable fuzzy rules defined by interpretable fuzzy information granules.

The network architecture is depicted in fig 9.1. It is a feed-forward, partially connected architecture defined by the following layers:

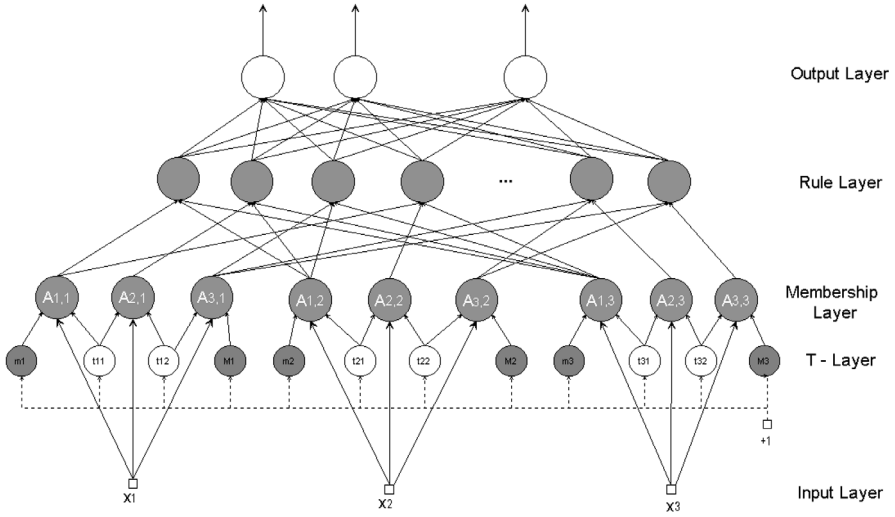

Figure 9.1: The neuro-fuzzy architecture for learning interpretable rules.

Input Layer It simply spreads input vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to the Membership Layer, described forth.

T-Layer The nodes of such layer are grouped into blocks, where each block corresponds to an input feature. Each block is made up of $K_i + 1$ linear Single Input Single Output nodes, ordered from the left to the right. The first node is defined by the following transfer function:

$$o^{\mathcal{T}}_{i,0} = u^{\mathcal{T}}_{i,0} \cdot m_i = m_i \tag{9.29}$$

being m_i the infimum of the i-th feature range and the input $u^{\mathcal{T}}_{i,0}$ is the input signal, which is kept constantly to $+1$. The transfer function of the last node of the group is similarly defined as

$$o^{\mathcal{T}}_{i,K_i} = u^{\mathcal{T}}_{i,K_i} \cdot M_i = M_i \tag{9.30}$$

being M_i the supremum of the i-th feature range and the input $u^{\mathcal{T}}_{i,K_i}$ is constantly $+1$. These two nodes are not adjustable, i.e. they have no free parameters. The remaining $K_i - 1$ nodes have a similar linear activation function, defined as

$$o^{\mathcal{T}}_{i,k} = u^{\mathcal{T}}_{i,k} \cdot t_{i,k} = t_{i,k}, \;\; k = 1, 2, \ldots, K_i - 1 \tag{9.31}$$

Each of such nodes has one free parameter, $t_{i,k}$, which is subject of adaptation during learning. The input signal $u_{i,k}^{\mathcal{T}}$ is always fed to $+1$. Each group of the T-Layer represents an element of the subspace $T^{(i)}$. The constraint $m_i < t_{i,1} < \cdots < t_{i,K_i} < M_i$ must be guaranteed by the initialization process as well as by the network learning process.

Membership Layer The nodes of this layer can are also grouped into blocks, with each block corresponding to an input feature. The nodes of the layer provide for the calculation of the membership degrees for each input feature, and are defined by three inputs and one output. The Membership Layer is connected both to the Input Layer and the T-Layer in a way that can be formalized by the definition of the input signals for each node:

$$u_{i,k}^{\mathcal{M},1} = o_{i,k-1}^{\mathcal{T}} \tag{9.32}$$

$$u_{i,k}^{\mathcal{M},2} = x_i \tag{9.33}$$

$$u_{i,k}^{\mathcal{M},3} = o_{i,k}^{\mathcal{T}} \tag{9.34}$$

For $k = 1, 2, \ldots, K_i$. The transfer functions represent Gaussian membership functions, with the centers and widths calculated by means of outputs coming from the T-Layer. The definition of the transfer functions are:

$$\begin{aligned} o_{i,1}^{\mathcal{M}} &= \exp\left(-\frac{\left(u_{i,k}^{\mathcal{M},2} - u_{i,k}^{\mathcal{M},1}\right)^2}{2\left(\sigma\left(2u_{i,k}^{\mathcal{M},1} - u_{i,k}^{\mathcal{M},3}, u_{i,k}^{\mathcal{M},3}\right)\right)^2}\right) \\ &= \exp\left(-\frac{(x_i - m_i)^2}{2\left(\sigma\left(2m_i - t_{i,1}, t_{i,1}\right)\right)^2}\right) = \mu_{A_1^{(i)}}(x_i) \end{aligned} \tag{9.35}$$

$$o_{i,k}^{\mathcal{M}} = \exp\left(-\frac{\left(u_{i,k}^{\mathcal{M},2} - \omega\left(u_{i,k}^{\mathcal{M},1}, u_{i,k}^{\mathcal{M},3}\right)\right)^2}{2\left(\sigma\left(u_{i,k}^{\mathcal{M},1}, u_{i,k}^{\mathcal{M},3}\right)\right)^2}\right) \tag{9.36}$$

$$= \exp\left(-\frac{\left(x_i - \omega\left(t_{i,k-1}, t_{i,k}\right)\right)^2}{2\left(\sigma\left(t_{i,k-1}, t_{i,k}\right)\right)^2}\right) = \mu_{A_k^{(i)}}(x_i), \tag{9.37}$$

$$k = 1, 2, \ldots, K_i \tag{9.38}$$

$$o_{i,K_i}^{M} = \exp\left(-\frac{\left(u_{i,K_i}^{\mathcal{M},2} - u_{i,K_i}^{\mathcal{M},3}\right)^2}{2\left(\sigma\left(u_{i,K_i}^{\mathcal{M},1}, 2u_{i,K_i}^{\mathcal{M},3} - u_{i,K_i}^{\mathcal{M},1},\right)\right)^2}\right) \tag{9.39}$$

$$= \exp\left(-\frac{(x_i - M_i)^2}{2\left(\sigma\left(t_{i,K_i}, 2M_i - t_{i,K_i}\right)\right)^2}\right) = \mu_{A_{K_i}^{(i)}}(x_i) \tag{9.40}$$

The distinguished definition of the extreme nodes vs. the middle nodes is necessary to satisfy the "leftmost/rightmost fuzzy sets" interpretability constraint. It is noteworthy that extreme nodes have no free parameters, since centers and widths are determined by the two functions ω and σ defined as:

$$\omega(t', t'') = \frac{t' + t''}{2} \tag{9.41}$$

and

$$\sigma(t', t'') = \frac{t'' - t'}{2\sqrt{-2\ln\varepsilon}} \tag{9.42}$$

In this way, interpretability constraints are always guaranteed due to the theoretical results shown in the previous Section.

Rule Layer This Layer combines fuzzy membership values coming from the Membership Layer to form the rule activation strengths. Each nodes has n inputs and one output. The i-th input signal corresponds to one output of the i-th block of the Membership Layer. The specific node of such layer that is connected to a node of the Rule layer is determined by the function g, which is in turn determined by the information granulation algorithm. The Rule Layer is composed of R nodes, one for each rule, and the r-th node of the layer is characterized by the following input signals:

$$u_r^{\mathcal{R},i} = o_{i,g(r,i)}^{\mathcal{M}},\ i = 1, 2, \ldots n \tag{9.43}$$

The transfer function computes the rule activation strength, which is generally determined by a t-norm composition of the membership values occurring in the rule antecedent. For the design of the neuro-fuzzy architecture, the product is chosen as a suitable t-norm because it is everywhere differentiable. As a consequence, the transfer function is defined as:

$$o_r^{\mathcal{R}} = \prod_{i=1}^{n} u_r^{\mathcal{R},i} = \prod_{i=1}^{n} \mu_{A_{g(r,i)}^{(i)}}(x_i) = \mu_r(\mathbf{x}) \tag{9.44}$$

Again, the nodes of the Rule Layer have not free parameters, as the output signal is fully determined by the input signals.

Output Layer The role of such layer is to complete the inference process of the fuzzy model implemented by the neural network. For simplicity, a 0-order Takagi-Sugeno fuzzy model is chosen, though more sophisticated models can be also considered. If the fuzzy model has m outputs, then there will be m nodes on the Output Layer. There are R input signals for each node, defined as:

$$u_j^{\mathcal{O},r} = o_r^{\mathcal{R}},\ r = 1, 2, \ldots, R \tag{9.45}$$

for $j = 1, 2, \ldots, m$. It can be noted that the definition of the input signal does not depend on the output index j, thus the Rule Layer and the Output Layer are fully connected. The transfer function of each node reflects the inference rule of the fuzzy model:

$$o_j^{\mathcal{O}} = \frac{\sum_{r=1}^{R} u_j^{\mathcal{O},r} \cdot v_{j,r}}{\sum_{r=1}^{R} u_j^{\mathcal{O},r}} = \frac{\sum_{r=1}^{R} u_j^{\mathcal{O},r} \cdot v_{j,r}}{\sum_{r=1}^{R} u_j^{\mathcal{O},r}} = \frac{\sum_{r=1}^{R} \mu_r(\mathbf{x}) \cdot v_{j,r}}{\sum_{r=1}^{R} \mu_r(\mathbf{x})} = y_j(\mathbf{x}) \tag{9.46}$$

Each node of the Output Layer has r free parameters $v_{j,r}$ which linearly contribute to the determination of the final output. However, such parameters are not involved in the definition of interpretable configurations of the rules.

9.2.3 The Neuro-Fuzzy learning scheme

The learning process for the proposed neural model can be accomplished in several ways, including hybrid ones that exploit the linear contributions of the free parameters in the Output Layer. The simplest learning algorithm is, however, the back-propagation algorithm, which follows the gradient descent rule for updating the free parameters. Apart from simplicity, Back-propagation is able to update also the parameters of the T-Layer, thus contributing to the refinement of the information granules codified in the network.

For the purpose of describing the learning algorithm, it is supposed that a finite training set D of input/output pairs is available:

$$D = \left\{ \left\langle \mathbf{x}^{(1)}, \mathbf{d}^{(1)} \right\rangle, \left\langle \mathbf{x}^{(2)}, \mathbf{d}^{(2)} \right\rangle, \ldots, \left\langle \mathbf{x}^{(N)}, \mathbf{d}^{(N)} \right\rangle \right\} \tag{9.47}$$

where $\mathbf{x}^{(p)} = \left(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}\right) \in X$ is an input pattern and $\mathbf{d}^{(p)} = \left(d_1^{(p)}, d_2^{(p)}, \ldots, d_m^{(p)}\right) \in \mathbb{R}^m$ is the corresponding desired output vector. The objective of learning is to modify the free parameters of the neural network so as to minimize an error function generally defined as

$$\mathbf{E} = \frac{1}{N} \sum_{p=1}^{N} E\left(S\left(\mathbf{x}^{(p)}\right), \mathbf{d}^{(p)}\right) \tag{9.48}$$

where $S : X \rightarrow \mathbb{R}^m$ is the functional mapping realized by the neural network and E is a punctual error function that is averaged over all training patterns to obtain the final error value $\mathbf{E}$.

The error function E can have different formulations, depending on the applicative context of the neuro-fuzzy model. For simplicity, it is temporarily left undefined. To derive the learning rule of the Back-propagation algorithm, the derivatives of the global error vale $\mathbf{E}$ with respect to any free parameter must be explicated. Let ξ be a generic free parameter (of any layer); the general schema for the derivatives is:

$$\frac{\partial \mathbf{E}}{\partial \xi} = \frac{1}{N} \sum_{p=1}^{N} \frac{\partial E}{\partial \xi}\left(S\left(\mathbf{x}^{(p)}\right), \mathbf{d}^{(p)}\right) \tag{9.49}$$

which corresponds to:

$$\frac{\partial \mathbf{E}}{\partial \xi} = \frac{1}{N} \sum_{p=1}^{N} \left[\frac{\partial E}{\partial S}\left(S, \mathbf{d}^{(p)}\right) \cdot \frac{\partial S}{\partial \xi}\left(\mathbf{x}^{(p)}\right)\right]_{S=S\left(\mathbf{x}^{(p)}\right)} \tag{9.50}$$

In general, S is multidimensional, hence the product within the summation is a scalar product, which can be explicated as:

$$\frac{\partial \mathbf{E}}{\partial \xi} = \frac{1}{N} \sum_{p=1}^{N} \left[\sum_{j=1}^{m} \frac{\partial E}{\partial y_j}\left(S, \mathbf{d}^{(p)}\right) \cdot \frac{\partial y_j}{\partial \xi}\left(\mathbf{x}^{(p)}\right)\right]_{S=(y_1, y_2, \ldots, y_m)\left(\mathbf{x}^{(p)}\right)} \tag{9.51}$$

This general formulation decouples the derivatives of the Error Function, which depend only on the desired output $\mathbf{d}^{(p)}$ and the actual output S, and the derivatives that depend on the structure of the neural network. The latter depend on the specific free parameter ξ. If $\xi \equiv v_{k,r}$, the derivative formulation is:

$$\frac{\partial y_j}{\partial v_{k,r}}\left(\mathbf{x}^{(p)}\right) = \frac{\partial}{\partial v_{k,r}} \frac{\sum_{r=1}^{R} \mu_r\left(\mathbf{x}^{(p)}\right) \cdot v_{k,r}}{\sum_{r=1}^{R} \mu_r\left(\mathbf{x}^{(p)}\right)} = \delta_{jk} \frac{\mu_r\left(\mathbf{x}^{(p)}\right)}{\sum_{r=1}^{R} \mu_r\left(\mathbf{x}^{(p)}\right)} \tag{9.52}$$

where δ_{jk} is the Kronecker Symbol, $\delta_{jk} = 1$ if $j = k$, otherwise $\delta_{jk} = 0$. The following definition:

$$\lambda_r^{(p)}\left(\mathbf{x}^{(p)}\right) = \frac{\mu_r\left(\mathbf{x}^{(p)}\right)}{\sum_{r=1}^{R}\mu_r\left(\mathbf{x}^{(p)}\right)} \tag{9.53}$$

will be useful in successive calculations.

The derivatives of one output with respect to the center of a Gaussian fuzzy set is

$$\frac{\partial y_j}{\partial \omega_h^{(i)}}\left(\mathbf{x}^{(p)}\right) = \sum_{s=1}^{R}\frac{\partial y_j}{\partial \mu_s}\left(\mathbf{x}^{(p)}\right)\cdot\frac{\partial \mu_s}{\partial \omega_h^{(i)}}\left(\mathbf{x}^{(p)}\right) \tag{9.54}$$

where:

$$\frac{\partial y_j}{\partial \mu_s}\left(\mathbf{x}^{(p)}\right) = \frac{\partial}{\partial \mu_s}\frac{\sum_{r=1}^{R}\mu_r\left(\mathbf{x}^{(p)}\right)\cdot v_{j,r}}{\sum_{r=1}^{R}\mu_r\left(\mathbf{x}^{(p)}\right)} = \frac{v_{j,s} - y_j\left(\mathbf{x}^{(p)}\right)}{\sum_{r=1}^{R}\mu_r\left(\mathbf{x}^{(p)}\right)} \tag{9.55}$$

and:

$$\frac{\partial \mu_s}{\partial \omega_h^{(i)}}\left(\mathbf{x}^{(p)}\right) = \delta_{g(s,i),h}\cdot\mu_s\left(\mathbf{x}^{(p)}\right)\cdot\frac{x_i^{(p)} - \omega_h^{(i)}}{\left(\sigma_h^{(i)}\right)^2} \tag{9.56}$$

By combining (9.55) and (9.56) in (9.54), it results:

$$\frac{\partial y_j}{\partial \omega_h^{(i)}}\left(\mathbf{x}^{(p)}\right) = \frac{x_i^{(p)} - \omega_h^{(i)}}{\left(\sigma_h^{(i)}\right)^2}\sum_{\substack{s=1\\ g(s,i)=h}}^{R}\left(v_{j,s} - y_j\left(\mathbf{x}^{(p)}\right)\right)\cdot\lambda_s^{(p)}\left(\mathbf{x}^{(p)}\right) \tag{9.57}$$

Similarly, the derivatives of one output with respect to the width of a Gaussian fuzzy set can be derived:

$$\frac{\partial y_j}{\partial \sigma_h^{(i)}}\left(\mathbf{x}^{(p)}\right) = \frac{\left(x_i^{(p)} - \omega_h^{(i)}\right)^2}{\left(\sigma_h^{(i)}\right)^3}\sum_{\substack{s=1\\ g(s,i)=h}}^{R}\left(v_{j,s} - y_j\left(\mathbf{x}^{(p)}\right)\right)\cdot\lambda_s^{(p)}\left(\mathbf{x}^{(p)}\right) \tag{9.58}$$

To complete the calculation, the derivatives with respect to the cut points must be computed. From (9.19) and (9.20) it can be observed that the h-th center and width of i-th input feature depend on the $(h-1)$-th and h-th cut point. As a consequence, the derivative of y_j with respect to $t_{i,h}$ could be written as:

$$\frac{\partial y_j}{\partial t_{i,h}} = \frac{\partial y_j}{\partial \omega_h^{(i)}}\cdot\frac{\partial \omega_h^{(i)}}{\partial t_{i,h}} + \frac{\partial y_j}{\partial \omega_{h+1}^{(i)}}\cdot\frac{\partial \omega_{h+1}^{(i)}}{\partial t_{i,h}} + \frac{\partial y_j}{\partial \sigma_h^{(i)}}\cdot\frac{\partial \sigma_h^{(i)}}{\partial t_{i,h}} + \frac{\partial y_j}{\partial \sigma_{h+1}^{(i)}}\cdot\frac{\partial \sigma_{h+1}^{(i)}}{\partial t_{i,h}} \tag{9.59}$$

for $h = 1, 2, \ldots, K_i - 1$ and $i = 1, 2, \ldots, n$. Special cases must be considered for $k = 1$ and $k = K_i - 1$, since the relative cut points are associated to leftmost and rightmost fuzzy sets. For $h = 2, 3, \ldots, K_i - 2$ derivatives can be explicated as:

$$\frac{\partial \omega_h^{(i)}}{\partial t_{i,h}} = \frac{1}{2} = \frac{\partial \omega_{h+1}^{(i)}}{\partial t_{i,h}} \tag{9.60}$$

and:

$$\frac{\partial \sigma_h^{(i)}}{\partial t_{i,h}} = -\frac{1}{2\sqrt{-2\ln\varepsilon}} = -\frac{\partial \sigma_{h+1}^{(i)}}{\partial t_{i,h}} \tag{9.61}$$

Leftmost and rightmost derivatives are calculated as,

$$\omega_1^{(i)} = \omega\left(2m_i - t_{i,1}, t_{i,1}\right) = m_i \implies \frac{\partial \omega_1^{(i)}}{\partial t_{i,1}} = 0 \tag{9.62}$$

$$\omega_{K_i}^{(i)} = \omega\left(t_{K_i-1}, 2M_i - t_{K_i-1}\right) = M_i \implies \frac{\partial \omega_{K_i}^{(i)}}{\partial t_{i,K_i-1}} = 0 \tag{9.63}$$

and

$$\begin{aligned} &\sigma_1^{(i)} = \sigma\left(2m_i - t_{i,1}, t_{i,1}\right) = \frac{2t_{i,1} - 2m_i}{2\sqrt{-2\ln\varepsilon}} \implies \frac{\partial \sigma_1^{(i)}}{\partial t_{i,1}} = \frac{1}{\sqrt{-2\ln\varepsilon}} \\ &\sigma_{K_i}^{(i)} = \sigma\left(t_{K_i-1}, 2M_i - t_{K_i-1}\right) = \frac{2M_i - 2t_{K_i-1}}{2\sqrt{-2\ln\varepsilon}} \\ &\qquad\qquad \implies \frac{\partial \sigma_{K_i}^{(i)}}{\partial t_{i,K_i-1}} = -\frac{1}{\sqrt{-2\ln\varepsilon}} \end{aligned} \tag{9.64}$$

To complete the computation of derivatives, the error function must now be taken into account. More specifically, the derivatives $\partial E / \partial y_j$ must be computed. If the neuro-fuzzy model is used for function approximation, the most appropriate error function is Mean Squared Error (MSE), defined as

$$E_{MSE}\left(\mathbf{y}, \mathbf{d}\right) = \sum_{j=1}^{m} \left(y_j - d_j\right)^2 \tag{9.65}$$

In such a case, the required derivatives are easily computed:

$$\frac{\partial E}{\partial y_j} = \frac{\partial E_{MSE}}{\partial y_j} = 2\left(y_j - d_j\right) \tag{9.66}$$

If the neuro-fuzzy model is used for classification tasks, other error functions can be used. As an example, the Approximate Differentiable Empirical Risk Functional (ADERF) proposed in (Castellano et al., 2004) could be adequate. The ADERF functional is defined as:

$$E_{ADERF}\left(\mathbf{y},\mathbf{d}\right)=\frac{1}{2}\sum_{j=1}^{m}\left(\left(\frac{y_j}{\left\|\mathbf{y}\right\|_w}\right)^u-d_j\right)^2 \tag{9.67}$$

where $\left\|\mathbf{y}\right\|_w$ is the Minkowski norm of order w. The derivative of such error function has the following form:

$$\begin{aligned}\frac{\partial E}{\partial y_j}&=\frac{\partial E_{ADERF}}{\partial y_j}=\\ &u\left(\frac{y_j^u}{\left\|\mathbf{y}\right\|_w^u}-d_j\right)y_j^{u-1}\frac{1-\left\|\mathbf{y}\right\|_w^{-w}y_j^w}{\left\|\mathbf{y}\right\|_w^w}+\\ &-u\frac{y_j^{w-1}}{\left\|\mathbf{y}\right\|_w^{u+w}}\sum_{\substack{k=1\\k\neq j}}^{m}\left(\frac{y_k^u}{\left\|\mathbf{y}\right\|_w^u}-d_k\right)y_k^u\end{aligned} \tag{9.68}$$

With all derivatives explicated, the gradient descent learning rule can computed. At any discrete instant τ and for any network parameter ξ, the learning rule follows the law:

$$\xi\left[\tau+1\right]=\xi\left[\tau\right]-\eta\left[\tau\right]\frac{\partial\mathbf{E}}{\partial\xi}\left[\tau\right] \tag{9.69}$$

where $\eta\left[\tau\right]$ is the learning rate. Variations of (9.69) can also be considered, e.g. to provide a lightweight learning procedure by taking into account only one pattern at time. Furthermore, derivatives formulations can be rearranged so as to follow the Back-propagation update formulas.

When the learning procedure is applied to train the neuro- fuzzy network, an important question arises on the order of the elements of each $T^{(i)}$, as defined in (9.13). The respect of such constraint, which is fundamental to guarantee understandability of fuzzy rules, depends on the value of the learning rate. Indeed, if the learning rate is too high, the update formulas of parameters $t_{i,h}$ could exchange the relative position of a couple $t_{i,h}$ and $t_{i,h+1}$. If the exchange is not avoided, the updated configuration of network parameters would not belong to $T^{(i)}$ anymore, thus resulting in non-interpretable configurations. On the other hand, if the learning rate is too small, convergence could be too slow to be acceptable. An important question arises on the quantification of the upper bound allowed for the learning rate to avoid dangerous exchanges of parameter positions.

In order to provide for such a quantification, for each discrete instant τ two consecutive cut points $t_{i,h}$ and $t_{i,h+1}$ on a generic dimension i are considered. The following quantity is taken into account:

$$\Delta_{i,h}[\tau] = t_{i,h+1}[\tau] - t_{i,h}[\tau] \tag{9.70}$$

By hypothesis, $\Delta_{i,h}[0] > 0$ since the the initial configuration of parameters is assumed to be in $T^{(i)}$. The sequence $\Delta_{i,h}[0], \Delta_{i,h}[1], \ldots, \Delta_{i,h}[\tau], \ldots$ is derived by the following relation:

$$\begin{aligned}
\Delta_{i,h}[\tau+1] &= t_{i,h+1}[\tau+1] - t_{i,h}[\tau+1] = && (9.71)\\
&= t_{i,h+1}[\tau] - \eta[\tau]\frac{\partial \mathbf{E}}{\partial t_{i,h+1}}[\tau] - t_{i,h}[\tau] + \eta[\tau]\frac{\partial \mathbf{E}}{\partial t_{i,h}}[\tau] && (9.72)\\
&= \Delta_{i,h}[\tau] + \eta[\tau]\left(\frac{\partial \mathbf{E}}{\partial t_{i,h}}[\tau] - \frac{\partial \mathbf{E}}{\partial t_{i,h+1}}[\tau]\right) && (9.73)
\end{aligned}$$

It is noteworthy observing that:

$$\Delta_{i,h}[\tau+1] < \Delta_{i,h}[\tau] \iff \frac{\partial \mathbf{E}}{\partial t_{i,h}}[\tau] < \frac{\partial \mathbf{E}}{\partial t_{i,h+1}}[\tau] \tag{9.74}$$

By assuming that $\Delta_{i,h}[\tau] > 0$ (i.e. non exchange between $t_{i,h}$ and $t_{i,h+1}$ at time τ), exchange at time $\tau+1$ is avoided if $\Delta_{i,h}[\tau+1] \geq \Delta_{i,h}[\tau]$. However, if $\Delta_{i,h}[\tau+1] < \Delta_{i,h}[\tau]$ then exchange can be still avoided if:

$$\underbrace{\Delta_{i,h}[\tau]}_{>0} + \underbrace{\eta[\tau]}_{>0}\underbrace{\left(\frac{\partial \mathbf{E}}{\partial t_{i,h}}[\tau] - \frac{\partial \mathbf{E}}{\partial t_{i,h+1}}[\tau]\right)}_{<0} > 0 \tag{9.75}$$

that is:

$$0 < \eta[\tau] < \frac{\Delta_{i,h}[\tau]}{\frac{\partial \mathbf{E}}{\partial t_{i,h+1}}[\tau] - \frac{\partial \mathbf{E}}{\partial t_{i,h}}[\tau]} \tag{9.76}$$

Since the learning rate does not depend on the specific cut points considered, nor on the specific input dimension, the upper bound for the learning rate is:

$$\eta[\tau] < \min_{h,i} \frac{\Delta_{i,h}[\tau]}{\frac{\partial \mathbf{E}}{\partial t_{i,h+1}}[\tau] - \frac{\partial \mathbf{E}}{\partial t_{i,h}}[\tau]} \tag{9.77}$$

Since $\Delta_{i,h}[\tau]$ by hypotheses, the upper bound of the learning rate is always greater than zero. As a consequence, relation (9.77) proves the existence of a non-empty range of values admissible for learning rates. The exact calculation of the upper bound is unnecessary, since a simple trial-and-error procedure can be adopted: if a learning rate causes the exchange of two cut points, then it is reduced (e.g. halved) and the update procedure is repeated.

9.3 Illustrative examples

To illustrate the effectiveness of the proposed neuro-fuzzy model, two simulation stages has been carried out. The first experimental stage is concerned with the design of a neuro-fuzzy model that learns interpretable rules to solve a benchmark problem of non-linear system identification. The second experimental setting involves the solution of a diagnosis problem based on real-world medical data. Such simulation is especially aimed at comparing the trained model both in terms of interpretability and accuracy with respect to existing approaches.

9.3.1 Non-linear system identification

The goal of this first simulation is to show how the proposed approach can extract a fuzzy rule base from data and how this rule base turns out to be interpretable and accurate as well. A very simple example concerning the identification of a nonlinear system has been considered. The results were compared with those obtained by an ANFIS network (Jang, 1993) implementing a 0-order TSK fuzzy model. The ANFIS code was taken from the Matlab® Fuzzy Toolbox.

The system to be identified is a static nonlinear system with two inputs and a single output (Narendra and Parthasarathy, 1990). The input/output relation of such system is described by:

$$y = \left(1 + x_1^{-2} + x_2^{-1.5}\right)^2, \ 1 \leq x_1, x_2 \leq 5 \tag{9.78}$$

A three-dimensional I/O graph of this nonlinear system is depicted in fig. 9.2.

The training set was obtained by computing the function (9.78) on 50 pairs (x_1, x_2) randomly taken in $[1, 5]^2$. For a fair comparison with the ANFIS model, no specific information granulation techniques has been applied on the input data. To generate the initial configuration of the knowledge base, the input Universe of Discourse has been first normalized in the interval $[-1, 1]$. Then it has been partitioned according to the classical grid partition technique by setting five fuzzy sets per input (see fig. 9.3). As a consequence, a total of 25 rules have been generated.

The same structure and configuration of parameters has been set both for the ANFIS network and the proposed neuro-fuzzy model. Furthermore, both networks have been trained for 5000 epochs with a learning rate fixes to 0.01 in each epoch. The standard MSE has been used as cost function during learning.

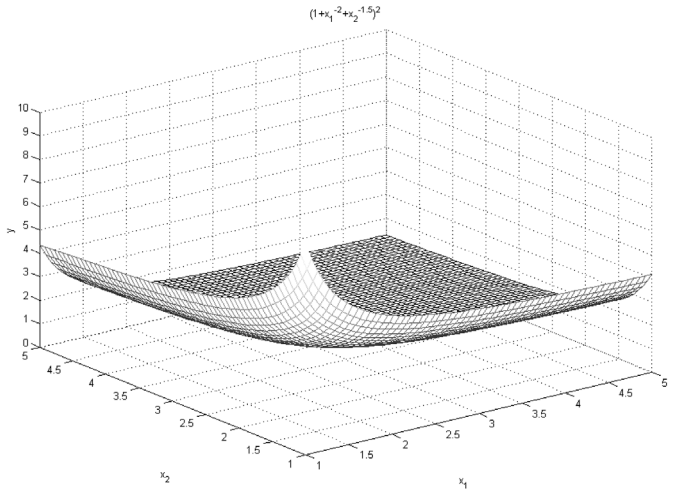

Figure 9.2: Output surface of the nonlinear system

Figure 9.4 compares the frames of cognition obtained after the learning process for the two models. It can be seen that the fuzzy sets generated by the proposed neuro-fuzzy model approach are still interpretable, while those obtained by ANFIS violate several interpretability constraints. More specifically, the fuzzy sets belonging to the frames of cognition after ANFIS training do not verify ε-coverage, distinguishability, leftmost/rightmost fuzzy sets, proper ordering and even normality[5]. In this case, the fuzzy sets do not match the metaphorical meaning of the attached linguistic labels; hence the knowledge base of the fuzzy model is not interpretable.

In addition, as it can be seen from fig. 9.5, the output surface provided by the proposed neuro-fuzzy model approximates quite well the desired input/output mapping, while the approximation provided by the ANFIS network is rather poor (see fig. 9.6). Moreover, as shown in figs. 9.7 and 9.8, the trend of the MSE in the case of the proposed learning algorithm is smoother in comparison to the ANFIS learning algorithm, providing a final MSE of 0.0053 which is lower than the final MSE (0.0301) achieved in the ANFIS case.

The neuro-fuzzy model defined by the proposed approach overcomes also other fuzzy models in terms of accuracy. For example, the Sugeno-Yasukawa model (Sugeno and Yasukawa, 1993) and the fuzzy model in (Huang and Chu,

[5]subnormal fuzzy sets appear because the prototype values of the Gaussian fuzzy sets lay outside the Universe of Discourse.

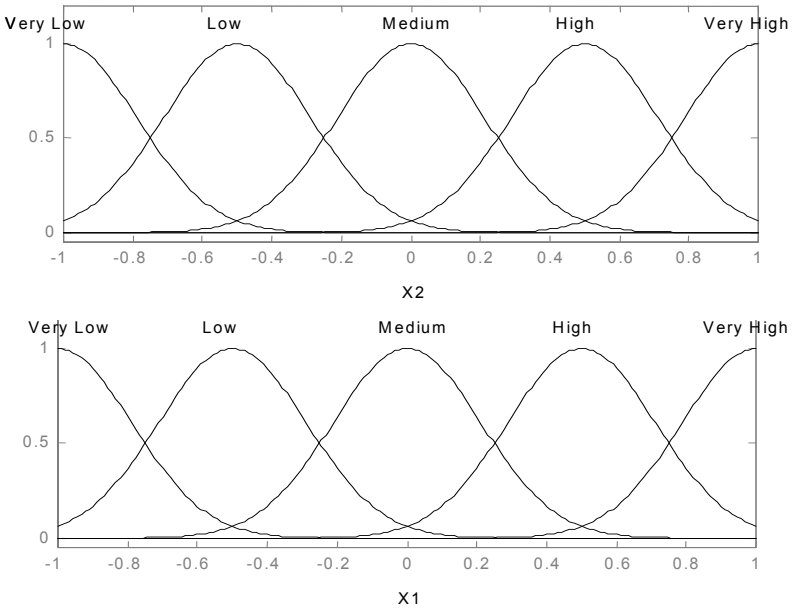

Figure 9.3: Initial frames of cognition of the two input variables. 0.5-coverage and 0.5-possibility have been set to generate the frames.

1999) provide, for the same data set, a MSE of 0.079 and 0.022, respectively. Comparison in terms of interpretability was not possible since no semantic issue is addressed by such fuzzy modeling methods.

In conclusion, through this example, we have illustrated how the proposed approach is able to extract a knowledge base with interpretable fuzzy sets and with a good approximation ability.

9.3.2 Fuzzy Medical Diagnosis

To assess the effectiveness of the proposed approach, a more realistic example, with higher dimensionality, has been considered to provide an idea of the network behavior in practice. The example problem concerns breast cancer diagnosis. To train the neuro-fuzzy model, the Wisconsin Breast Cancer (WBC) data set has been used, which is provided by W.H. Wolberg from the University of Wisconsin Hospitals, Madison (Wolberg and Mangasarian, 1990). The data set contains 699 cases belonging to one of two classes (benign: 458 cases, or malignant: 241 cases). Each case is represented by an ID number and nine attributes (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei,

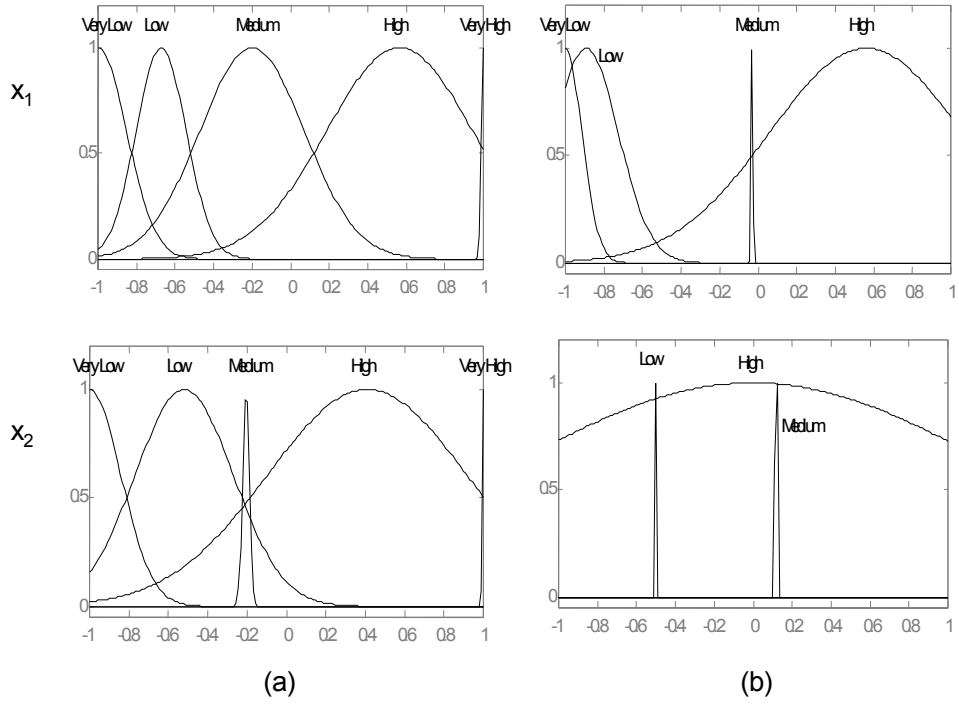

Figure 9.4: The frames of Cognition resulting after training: (a) within the proposed neuro-fuzzy model; (b) within the ANFIS model

bland chromatin, normal nucleoli, mitoses). All attribute values are integers from the domain 1..10. There are 16 cases with missing values. Since the proposed neuro-fuzzy model cannot yet deal with missing values, only the complete 683 cases has been considered: 444 in class 'benign' and 239 in class 'malignant'.

To cope with the high-dimensionality of this data set, the number of input variables was reduced by applying a feature selection algorithm developed in (Castellano and Fanelli, 2000). After the feature selection process, the most significant attributes found are clump thickness, uniformity of cell shape and bare nuclei, while the less significant attributes are uniformity of cell size and single epithelial cell size, both related to cell size, as also stated in (Duch et al., 2001). Hence the three selected features have been used, onto which simple Frames of Cognition have been defined, consisting of only two Gaussian fuzzy sets per feature. By this, a total of 8 rules have been generated with null consequents, which was used to establish the structure and initial parameters of the proposed neuro-fuzzy network.

The data set was split randomly in a training set of 342 cases and a

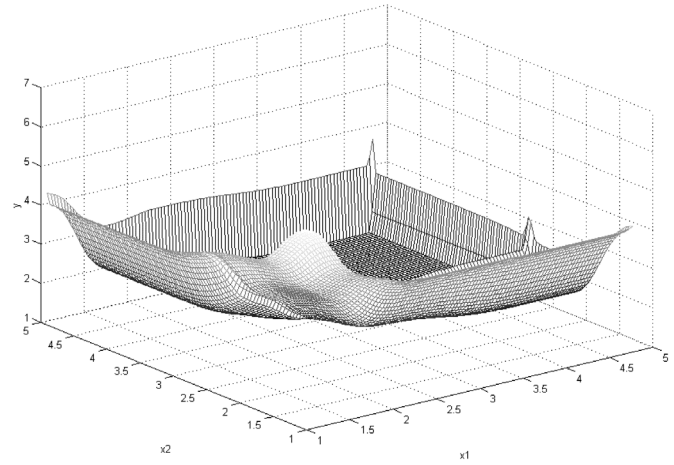

Figure 9.5: The input/output mapping realized by the proposed neuro-fuzzy model

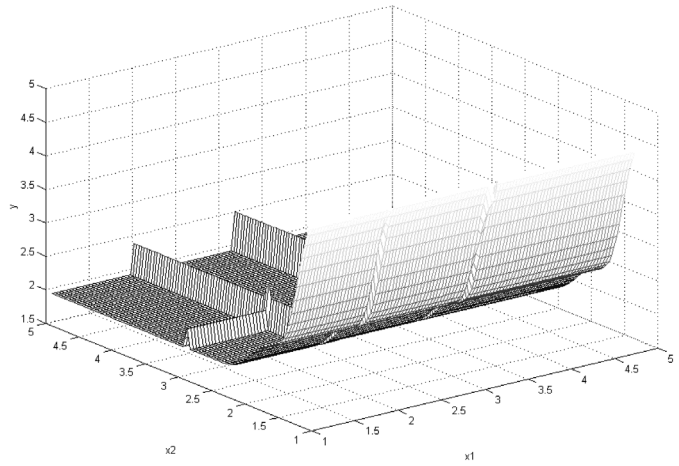

Figure 9.6: The input/output relationship realized by the ANFIS network

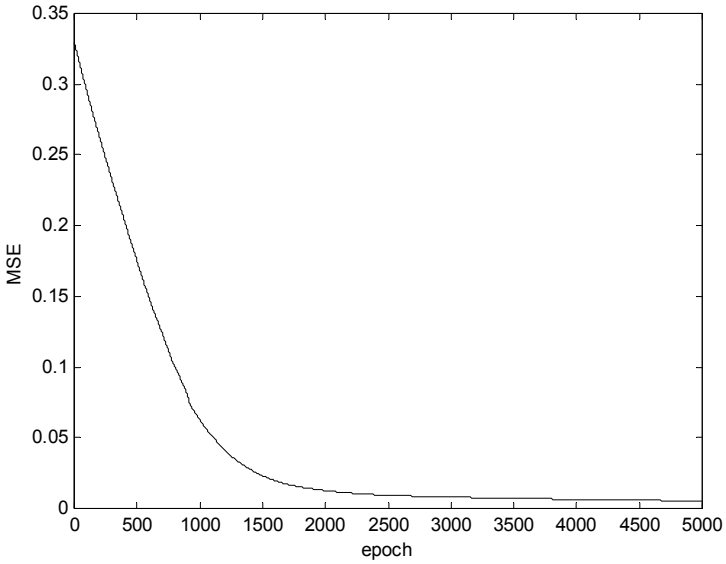

Figure 9.7: The trend of MSE during training epochs for the proposed neuro-fuzzy model

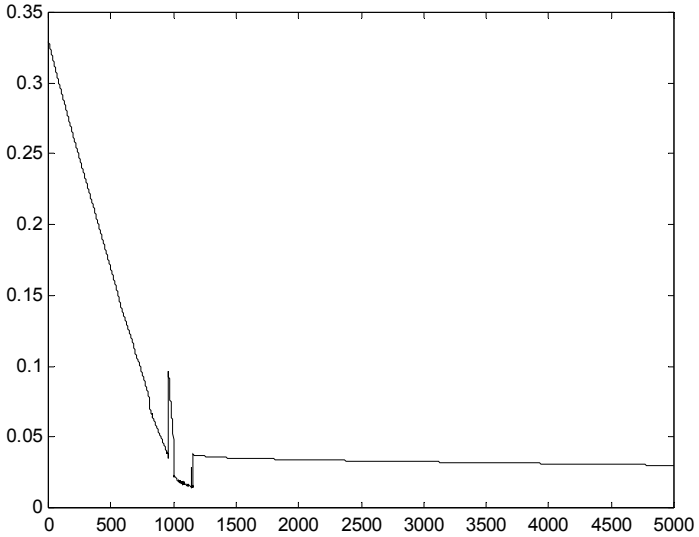

Figure 9.8: The trend of MSE during training epochs for the ANFIS model

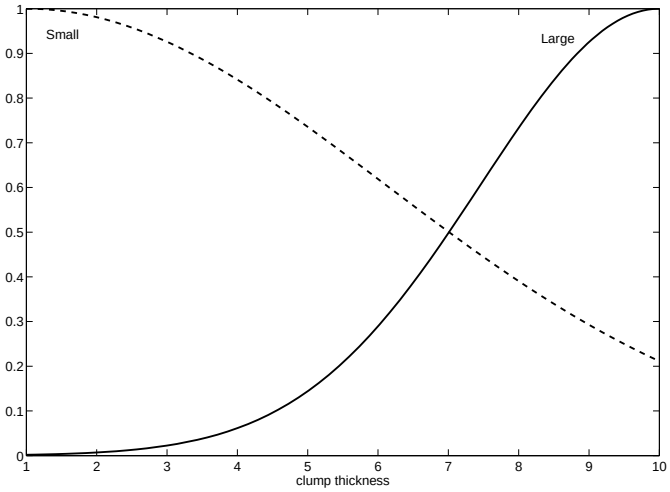

Figure 9.9: The refined frame of cognition for the feature "clump thickness"

test set of 341 cases, so that each set contains roughly the same number of patterns for each class. After 100 epochs of the proposed learning algorithm, a classification rate of 95.32% on the training set (16 errors), 96.48% on the test set (12 errors), and 95.90% (28 errors) on the whole dataset has been attained. For each variable, the two fuzzy sets (which have been labeled 'SMALL' and 'LARGE') refined after learning are represented in figs. 9.9-9.11. The fuzzy sets and the frames of cognition verify all the required interpretability constraints.

Since in application areas like medicine interpretability is of great concern, a small number of rules is highly desirable. As a consequence, the refined knowledge base has been pruned according to the technique described in (Castellano and Fanelli, 1996), which automatically reduces the number of rules while completely preserving the accuracy and the interpretability of the knowledge base. After rule simplification, there are only four fuzzy rules in the rule base with unchanged fuzzy sets and accuracy with respect to the 8-rule base. In the following, the discovered diagnostic rules are listed.

RULE 1: IF (CLUMP THICKNESS IS SMALL) AND (UNIFORMITY OF CELL SHAPE IS SMALL) AND (BARE NUCLEI IS SMALL) THEN PATIENT BELONGS TO CLASS BENIGN WITH DEGREE 0.67 AND TO CLASS MALIGNANT WITH DEGREE 0.11

RULE 2: IF (CLUMP THICKNESS IS SMALL) AND (UNIFORMITY OF CELL SHAPE IS SMALL) AND (BARE NUCLEI IS LARGE)

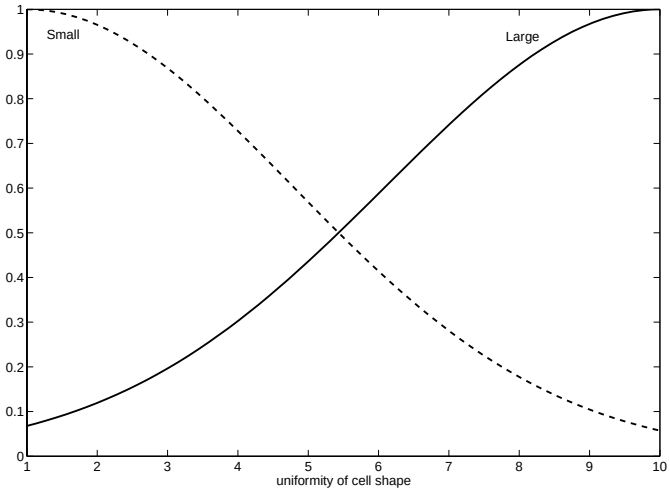

Figure 9.10: The refined frame of cognition for the feature "uniformity of cell shape"

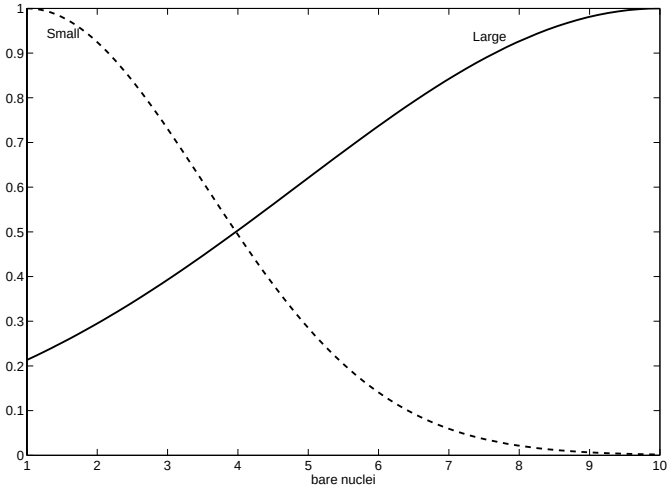

Figure 9.11: The refined frame of cognition for the feature "bare nuclei"

THEN Patient belongs to class benign with degree 0.022 and to class malignant with degree 0.47

Rule 3: IF (clump thickness IS small) AND (uniformity of cell shape IS large) AND (bare nuclei IS small) THEN Patient belongs to class benign with degree 0.00 and to class malignant with degree 0.54

Rule 4: IF (clump thickness IS large) AND (uniformity of cell shape IS small) AND (bare nuclei IS small) THEN Patient belongs to class benign with degree 0.081 and to class malignant with degree 0.32

To evaluate the effectiveness of the resulting model in terms of accuracy, it has been compared with the NEFCLASS neuro-fuzzy system, also applied to this dataset in (Nauck and Kruse, 1999), under the same experimental setting (i.e. removing 16 cases with missing values and partitioning the data set into 342 samples for training and 341 for testing). Using 4 rules and "best-per-class" rule learning (which can be regarded as a kind of pruning strategy), NEFCLASS achieves 8 errors on the training set (97.66% correct), 18 errors on the test set (94.72% correct) and 26 errors (96.2% correct) on the whole set. Despite the slightly better accuracy of NEFCLASS on the whole dataset, it should be noted that in case of the proposed neuro-fuzzy model, higher accuracy on the test set (generalization ability) is achieved with even a very small number of input variables with respect to the nine features used by NEFCLASS, thus resulting in a more simple and interpretable rule base. Furthermore, it should be noted that results of the proposed model come from the application of automatic procedures, both for learning and simplification, that do not require human intervention unlike the NEFCLASS system.

In addition, to obtain a more feasible estimate of the classification error, a 10-fold cross validation has been carried out. The data set was randomly split into ten equally sized parts without changing the class frequency. Each part was used as test set for the knowledge base discovered from the remaining data. The estimate of the classification accuracy was then computed as average of classification errors of all 10 neuro-fuzzy models on their test set. The mean error on the 10 test sets was 96.08% and the average number of rules was 4.2 (ranging from 3 to 7 rules). A comparison with other neural, fuzzy and traditional classifiers developed for the same dataset is summarized in table 9.1. It can be seen that the estimated classification of the proposed neuro-fuzzy model is comparable with most of the considered models. Indeed, most of the modeling methods reported in the table pursue only accuracy as ultimate goal and take no care about the interpretability of the knowledge representation.

Table 9.1: Accuracy comparison of the proposed neuro-fuzzy model w.r.t. other approaches for the WBC dataset

Method	**Accuracy**	**Reference**
IncNet	97.1%	(1)
k-NN	97.0±0.12	(2)
Fisher LDA	96.8%	(3)
MLP with Backprop	96.7%	(3)
LVQ	96.6%	(3)
Bayes (pairwise dep.)	96.6%	(3)
Naïve Bayes	96.4%	(3)
DB-CART	96.2%	(4)
LDA	96.0%	(3)
LFC,ASI,ASR	94.4%–95.6%	(3)
CART	93.5%	(4)
Quadratic DA	34.5%	(3)
FSM, 12 rules	96.5%	(2)
SSV, 3 rules	96.3%±0.2%	(2)
NEFCLASS-X, 2 rules, 5-6 feat.	95.06%	(5)
Proposed model	*96.08%*	

(1): (Jankowski and Kadirkamanathan, 1997)
(2): (Duch et al., 2001)
(3): (Ster and Dobnikar, 1996)
(4): (Shand and Breiman, 1996)
(5): (Nauck and Kruse, 1999)

9.4 Final remarks

In this Chapter, a neuro-fuzzy architecture and its learning algorithm for refinement of fuzzy models with interpretability protection. Interpretability of fuzzy models is preserved during learning by allowing the free parameters to vary in a parameter subspace containing only configurations satisfying a set of formal properties.

A simple benchmark and a real-world example have been considered to illustrate the key features of the proposed model and its related learning algorithm. The given examples highlight that the proposed neural architecture can be effective for model refinement, since the refined fuzzy models have an accuracy comparable to other state-of-art models. Also, simulation results confirm the essential feature of the proposed neuro-fuzzy architecture, that is the ability to produce final rule bases that are also interpretable.

Further extensions of the proposed model may concern the fulfillment of additional interpretability requirements. As an example, the "Uniform Granulation" interpretability constraint may be taken into account during neural learning by checking the distance between consecutive cut points. To embody such constraint, a threshold (or a penalization term) can be included in the learning process so that such distances do not become too dissimilar.

Part IV

Conclusions and Future Research

The purpose of computing is insight,
not numbers.
(R. Hamming)

One of the fundamental motivations for the development of Granular Computing is to bring computer-based information processing closer to human perception and reasoning. In this way it is hoped that complex problems could be tackled and solved by computer systems following the well-known principle postulated by Lotfi Zadeh:

> [...] to exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost.

To cope with complex problems, the knowledge-oriented approach appears the most successful, since it associates a general-purpose inference mechanism with a problem-specific knowledge base that represents the problem in a formal –hence machine treatable – way. The key feature of models based on the Granular Computing paradigm is the organization of the knowledge base into chunks of information, the so-called information granules. This organization has several advantages, like hierarchical decomposition, information hiding and user-centered modelling. Furthermore, the introduction of Fuzzy Logic into Granular Computing – which gives rise to the Theory of Fuzzy Information Granulation – enables the processing of vague knowledge that highly pervades real-world problems.

As a valuable consequence of modelling with the Theory of Fuzzy Information Granulation, the knowledge base embodied in the model could be expressed in linguistic terms that are immediately intelligible by human users. However, interpretability of the knowledge base is not granted *ipso facto*: fuzzy information granules must define a semantics that matches the metaphorical meaning conveyed by the associated linguistic representation.

Interpretability within the Theory of Fuzzy Information Granulation raises a number of issues that only recently have been taken into consideration. The aim of this thesis has been to provide a number of contributions in this direction, both of theoretical and algorithmic type.

The first issue taken under consideration in the thesis concerns the characterization of interpretability in a mathematical fashion. A deep study of existing literature revealed that the constraint-based approach is the most effective for such a characterization, but there is not a general agreement of which constraints mathematically define the blurry notion of interpretability in an exhaustive way. The first contribution of the thesis has hence been a

comprehensive study of interpretability constraints, which have been homogeneously formalized and described from several perspectives, also including psychological, epistemological and computational viewpoints.

This survey of interpretability constraints opens new directions for future research, which would include a deeper study of them under more robust psychological, cognitive and philosophical underpinnings. Such deeper insights would help to discover new interpretability constraints or by generalizing existing ones towards a more definitive characterization of the interpretability notion. Furthermore, a different organization of the constraints could be identified, which would isolate a kernel of general-purpose – hence mandatory – constraints from application specific ones that could be selected according to practical necessities.

The study of interpretability constraints has led to a close examination of some of them, resulting in new theoretical contributions reported in the thesis. Specifically, the distinguishability constraint has been studied in depth, with special focus on its numerical quantification. In this perspective, the two most common measures of distinguishability – namely, possibility and similarity – have been compared and related at the computational level. From this comparison, some theoretical results have been derived, which unveil an relationship of the two measures under weak conditions. As a consequence of such study, the possibility measure – which is more efficient than similarity – is fully justified for measuring distinguishability between fuzzy sets.

Another interpretability constraint that has been studied in depth concerns the error-free reconstruction of fuzzy interfaces. Optimal interfaces (i.e. interfaces that enable error-free reconstruction) have a great impact in the interpretability of the inference process carried out by a fuzzy model. As a consequence, the study of interface optimality, though theoretical in nature, has a relevant importance in model design. The study of interface optimality has been carried out by taking into account the type of interface. For input interfaces, two classes of fuzzy sets have been defined, namely the strictly and the loosely bi-monotonic fuzzy sets. It has been proved that input interfaces made up of bi-monotonic fuzzy sets are optimal under mild conditions, usually met in interpretable fuzzy modelling. This results generalize those known in literature, which are confined to narrower classes of fuzzy sets.

For output interfaces, optimality is hard to meet except for trivial cases. Sub-optimality of output interfaces is directly observable by the spurious nonlinearities that afflict the mappings realized by most fuzzy models. To study this phenomenon quantitatively, a new measure, called optimality degree, has been proposed. The optimality degree can be viewed as a fuzzy version of the classical optimality condition, and helps to assess the quality of different output interfaces in terms of information reconstruction. The optimality degree

could be viewed also as a starting point for newly devised output interfaces that are specifically aimed at maximizing this measure. Such new output interfaces would drastically alleviate the problem of spurious nonlinearities, thus improving the interpretability of the inference process. The definition and the implementation of such new interface is under investigation.

Fuzzy information granules are often acquired from a set of available observations. The process of fuzzy information granulation, which builds information granules from data, must be tailored to verify interpretability constraints, otherwise the resulting knowledge cannot be expressed with linguistic terms. This calls for brand new algorithms, or extension of existing ones, for information granulation. Different algorithms can be devised on the basis of the desired representation of the information granules, e.g. as a quantitative or as a qualitative description of data.

A number of granulation algorithms have been proposed throughout this work. For quantitative information granules, an algorithm for Gaussian information granulation has been proposed, which can be used to wrap existing clustering algorithms (like Fuzzy C-Means) in order to derive interpretable information granules that can expressed as fuzzy numbers or fuzzy vectors. The granulation algorithm operates by solving a constrained quadratic optimization problem based on the results of a fuzzy clustering algorithm. It is shown that the solution of such optimization problem leads to compact information granules that can be effectively used to build accurate yet interpretable fuzzy models.

For interval-valued information granules, a different approach has been proposed. It extends the well-known Fuzzy C-Means clustering algorithm by processing Minkowski distances of any order on data. With the Minkowski order approaching infinity (in practice, for any order greater than about ten), this extended version of fuzzy clustering is able to discover boxlike-shaped clusters. The analysis of interference between clusters helps in choosing appropriate settings for the final definition of hyper-boxes, which could be decomposed as Cartesian product of intervals, either of crisp or of fuzzy type.

The added value deriving from the adoption of intervals instead of scalar values consists in the representation of an entire range of possibilities instead of a single quantity. This key feature has suggested the proposal of extending the inference process of a fuzzy model to derive a prediction interval instead of a single numerical value. Prediction intervals, which have a sound meaning within the Theory of Probability, embody the knowledge about the uncertainty of the prediction carried out by a fuzzy model. In this way, the user can perceive the uncertainty of the model in providing the predicted value. In this sense, the interpretability of the model is significantly improved. For such reason, prediction intervals have been adopted to

extend a class of fuzzy models, and the added value of such approach has been observed on a complex real-world problem.

In some applicative contexts, a qualitative representation of information granules is preferable. To generate interpretable information granules that can be described in terms of linguistic adjectives, like "small", "cold", etc., an algorithmic framework has been proposed in this work. The framework, called "Double Clustering" is able to acquire from data information granules that verify a number of interpretability constraints so that they can be labelled by linguistic adjectives of immediate semantics. The framework can be implemented in several ways, according to applicative requirements. Furthermore, it is open to novel implementations that can further enhance the capabilities of the framework. It is under investigation an implementation of the framework that makes use of a sophisticated informed search strategy, which could automatize the process of information granulation by selecting the best granularity level in describing data so as to minimize the description length of the final representation.

Fuzzy models design by means of information granulation processes could be further refined if the internal knowledge is not completely defined by the derived information granules. When this occurs, a refinement process is necessary to complete the definition of the knowledge base. Neural learning is an effective approach for this task, but without any control, the interpretability of the generated information granules can be seriously hampered. To cope with such a problem, a new neuro-fuzzy architecture has been proposed, together with its learning scheme, which enable the refinement of the model knowledge through neural learning without violating the interpretability constraints for the tuned information granules. As a further enhancement, the proposed neuro-fuzzy network can be extended so as to satisfy a greater number of interpretability constraints, thus providing an even better interpretability of the resulting knowledge base.

All the contributions presented in this work finally emerge as a unified framework for interpretable information granulation and model design. In this direction, future developments will focus on the implementation of an extensible software tool that embodies all the proposed algorithms with possibility of further extension to include newly devised algorithms. The tool will allow the application of the proposed techniques for solving complex problems coming from the real world. This could promote further investigations on interpretability issues within the Theory of Fuzzy Information Granulation, both from theoretical and algorithmic standpoints.

List of Figures

List of Tables

Bibliography

Abonyi, J., Babuška, R., and Szeifert, F. (2002). Modified gath-geva fuzzy clustering for identification of takagi-sugeno fuzzy models. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 32(5):612–621.

Abonyi, J., Roubos, H., Babuška, R., and Szeifert, F. (2003). Interpretable semi-mechanistic fuzzy models by clustering, OLS and FIS model reduction. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 221–248. Springer-Verlag, Heidelberg.

Agarwal, L. and Taniru, M. (1992). A petri-net based approach for verifying the integrity of production systems. *International Journal of man-machine studies*, 36:447–468.

Altug, S., Chow, M.-Y., and Trussel, H. J. (1999). Heuristic constraints enforcement for training of and rule extraction from a Fuzzy/Neural architecture – part II: Implementation and application. *IEEE Transactions on Fuzzy Systems*, 7(2):151–159.

Angelov, P. (2004). An approach for fuzzy rule-base adaptation using on-line clustering. *International Journal of Approximate Reasoning*, 35:275–289.

Auephanwiriyakul, S. and Keller, J. (2002). Analysis and efficient implementation of a linguistic fuzzy c-means. *IEEE Transactions on Fuzzy Systems*, 10(5):563–582.

Babuška, R. (1998). *Fuzzy Modeling and Control*. Kluwer, Norwell, MA.

Babuška, R. (1999). Data-driven fuzzy modeling: Transparency and complexity issues. In *Proceedings 2nd European Symposium on Intelligent Techniques ESIT'99*, Crete, Greece. ERUDIT.

Baraldi, A. and Blonda, P. (1999). A survey on fuzzy clustering algorithms for pattern recognition I/II. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 29(6):778–801.

Baranyi, P., Yam, Y., Tikk, D., and Patton, R. (2003). Trade-off between approximation accuracy and complexity: TS controller design via HOSVD based complexity minimization. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 249–277. Springer-Verlag, Heidelberg.

Bargiela, A. (2001). Interval and ellispoidal uncertainty in water system state estimation. In Pedrycz, W., editor, *Granular Computing: An Introduction*, pages 23–57. Physica-Verlag.

Bargiela, A. and Pedrycz, W. (2003a). *Granular Computing: An Introduction*. Kluwer Academic Publishers, Boston, Dordrecht, London.

Bargiela, A. and Pedrycz, W. (2003b). A model of granular data: A design problem with the tchebyschev fuzzy c-means. *Soft Computing*, (online first).

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ.

Bellman, R. and Giertz, M. (1973). On the analytic formalism of the theory of fuzzy sets. *Information sciences*, 5:149–156.

Bellman, R., Kalaba, L., and Zadeh, L. (1966). Abstraction and pattern classification. *Journal of Mathematical Analysis and Applications*, 13:1–7.

Bengio, Y. (2000). Gradient-based optimization of hyper-parameters. *Neural Computation*, 12(8).

Berenji, H. and Khedkar, P. (1993). Clustering in product space for fuzzy inference. In *Proceedings of the Second IEEE International Conference on Fuzzy Systems*, pages 1402–1407, San Francisco. IEEE.

Bettini, C. and Montanari, A., editors (2000). *Spatial and Temporal Granularity: Papers from the AAAI Workshop*. The AAAI Press, Menlo Park, CA.

Bettini, C. and Montanari, A. (2002). Research issues and trends in spatial and temporal granularities. *Annals of Mathematics and Artificial Intelligence*, 36:1–4.

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.

Bezdek, J. (1992). On the relationship between neural networks, pattern recognition and intelligence. *Journal of Approximate Reasoning*, 6:85–107.

Bikdash, M. (1999). A highly interpretable form of sugeno inference systems. *IEEE Transactions on Fuzzy Systems*, 7(6):686–696.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Blake, C. and Merx, C. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/ mlearn/MLRepository.html.

Bonarini, A. (1997). Anytime learning and adaptation of hierarchical fuzzy logic behaviors. *Adaptive Behavior Journal, Special Issue on Complete Agent Learning in Complex Environments*, 5(3-4):281–315.

Butz, C. and Lingras, P. (2004). Granular jointree probability propagation. In *Proceedings of the 2004 Conference of the North American Fuzzy Information Processing Society (NAFIPS'04)*, pages 69–72, Banff, Alberta.

Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors (2003). *Interpretability Issues in Fuzzy Modeling*. Springer, Germany.

Castellano, G., Castiello, C., Fanelli, A., and Giovannini, M. (2003a). A neuro-fuzzy framework for predicting ash propeties in combustion processes. *Neural, Parallel and Scientific Computation*, 11:69–82.

Castellano, G., Castiello, C., Fanelli, A., and Mencar, C. (2003b). Discovering prediction rules by a neuro-fuzzy modeling framework. In Palade, V., Howlett, R., and Jain, L., editors, *Knowledge-Based Intelligent Information and Engineering Systems (KES 2003)*, volume I, pages 1242–1248. Springer.

Castellano, G. and Fanelli, A. (1996). Simplifying a neuro-fuzzy model. *Neural Processing Letters*, 4(6):75–81.

Castellano, G. and Fanelli, A. (2000). Variable selection using neural network models. *Neurocomputing*, 31(14):1–13.

Castellano, G., Fanelli, A., and Mencar, C. (2001). Automatic fuzzy encoding of complex objects. In *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, pages 407–410.

Castellano, G., Fanelli, A., and Mencar, C. (2004). An empirical risk functional to improve learning in a neuro-fuzzy classifier. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 34(1):725–731.

Castellano, G., Fanelli, A. M., and Mencar, C. (2002). A neuro-fuzzy network to generate human understandable knowledge from data. *Cognitive Systems Research, Special Issue on Computational Cognitive Modeling*, 3(2):125–144.

Chepoi, V. and Dumitrescu, D. (1999). Fuzzy clustering with structural constraints. *Fuzzy Sets and Systems*, 105:91–97.

Chiang, J.-H., Yue, S., and Yin, Z.-X. (2004). A new fuzzy cover approach to clustering. *IEEE Transactions on Fuzzy Systems*, 12(2):199–208.

Chow, M.-Y., Altug, S., and Trussel, H. J. (1999). Heuristic constraints enforcement for training of and knowledge extraction froma Fuzzy/Neural architecture–part i: Foundation. *IEEE Transactions on Fuzzy Systems*, 7(2):143–150.

Cios, K., Pedrycz, W., and Swiniarski, R. (1998). *Data Mining. Methods for Knowledge Discovery.* Kluwer.

Cloete, I. and Zurada, J., editors (2000). *Knowledge-Based NeuroComputing.* MIT Press.

Combs, W. E. and Andrews, J. E. (1998). Combinatorial rule explosion eliminated by a fuzzy rule configuration. *IEEE Transactions on Fuzzy Systems*, 6(1):1–11.

Cordón, O., Del Jesus, M., Herrera, F., Magdalena, L., and Villar, P. (2003). A multiobjective genetic learning process for joint feature selection and granularity and context learning in fuzzy rule-based classification systems. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 79–99. Springer-Verlag, Heidelberg.

Cordón, O. and Herrera, F. (2000). A proposal for improving the accuracy of linguistic modeling. *IEEE Transactions on Fuzzy Systems*, 8(3):335–344.

Corsini, P., Lazzerini, B., and Marcelloni, F. (2004). A fuzzy relational clustering algorithm based on a dissimilarity measure extracted from data. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 34(1):775–781.

Craven, P. and Wabba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley.

Cross, V. (1993). An analysis of fuzzy set aggregators and compatibility measures. Master's thesis, Dept. of Electronic Engineering, University of Delft.

Czogala, E. and Leski, J. (2000). *Fuzzy and Neuro-Fuzzy Intelligent Systems*. Physica-Verlag, Heidelberg, Berlin.

de Oliveira, J. (1993). On optimal fuzzy systems I/O interfaces. In *Proceedings of the 2nd IEEE International Conference on Fuzzy Systems*, pages 851–865, San Francisco, CA.

de Oliveira, J. (1995). A set-theoretical defuzzification method. *Fuzzy Sets and Systems*, 76:63–71.

de Oliveira, J. (1996). Sampling, fuzzy discretization, and signal reconstruction. *Fuzzy Sets and Systems*, 79:151–161.

Dempster, A. (1966). New methods for reasoning toward posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374.

Dick, S. and Kandel, A. (1999). Comment on "combinatorial rule explosion eliminated by a fuzzy rule configuration" (with author's reply). *IEEE Transactions on Fuzzy Systems*, 7(4):475–478.

Dreyfus, H. and Dreyfus, S. (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. MIT Press, Cambridge, MA.

Dubois, D. and Prade, H. (1980). *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, New York.

Dubois, D. and Prade, H. (1988). *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York.

Dubois, D. and Prade, H. (1992). Evidence, knowledge and belief functions. *International Journal of Approximate Reasoning*, 6:295–319.

Dubois, D. and Prade, H. (1996). What are fuzzy rules and how to use them. *Fuzzy Sets and Systems*, 84:169–185.

Dubois, D. and Prade, H. (1997). The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, 90:141–150.

Duch, W., Adamczak, R., and Grabczewski, K. (2001). A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks*, 12(2):277–306.

Dybowski, R. and Roberts, S. (2001). Confidence intervals and prediction intervals for feed-forward neural networks. In Dybowski, R. and Gant, V., editors, *Clinical Applications of Artificial Neural Networks*, pages 298–326. Cambridge University Press, Cambridge, UK.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Espinosa, J. and Vandewalle, J. (2000). Constructing fuzzy models with linguistic integrity from numerical data – AFRELI algorithm. *IEEE Transactions on Fuzzy Systems*, 8(5):591–600.

Espinosa, J. and Vandewalle, J. (2003). Extracting linguistic fuzzy models from numerical data – AFRELI algorithm. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 100–124. Springer-Verlag, Heidelberg.

Euzenat, J. (2001). Granularity in relational formalisms - with application to time and space representation. *Computational Intelligence*, 17:703–737.

Fagin, R. and Halpern, J. (1989). Uncertainty, belief and probability. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI-89)*, pages 1161–1167.

Farinwata, S., Filev, S., and Langari, R. (2000). *Fuzzy Control: Synthesis and Analysis*. Wiley.

Ferri, E., Kandel, A., and Langholz, G. (1999). Fuzzy negation. In Zadeh, L. and Kacprzyk, J., editors, *Computing with Words in Information/Intelligent Systems 1: Foundations*, pages 297–325. Physica-Verlag, Heidelberg, New York.

Fischer, M., Nelles, O., and Fink, A. (1998). Adaptive fuzzy model-based control. *Journal a*, 39(3):22–28.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7:179–188.

Fodor, J. A. (1975). *The Language of Thought*. Thomas Y. Crowell Co., New York.

Fonseca, C. and Fleming, P. (1995). An overview of evolitionary algorithms in multiobjective optimization. *Evolutionary computation*, 3(1):1–16.

Friedman, J. (1995). An overview of prediction learning and function approximation. In Cherkassky, V., Friedman, J., and Wechsler, H., editors, *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*. Springer Verlag, New York.

Gao, Q., Li, M., and Vitanyi, P. (2000). Applying MDL to learning best model granularity. *Artificial Intelligence*, 121(1-2):1–29.

Gaweda, A. E. and Zurada, J. M. (2001). Data driven design of fuzzy system with relational input partition. In *Proceedings of International IEEE Conference on Fuzzy Systems*, Melbourne, Australia.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the Bias/Variance dilemma. *Neural Computation*, 4:1–58.

Giunchigalia, F. and Walsh, T. (1992). A theory of abstraction. *Artificial Intelligence*, 56:323–390.

Glorennec, P. (1993). Adaptive fuzzy control. In Lowen, R. and Roubens, M., editors, *Fuzzy Logic: State of the Art*, pages 541–551. Kluwer Academic Press.

Goguen, J. (1967). L-fuzzy sets. *Journal of Mathematical Analysis and Applications*, 18:145–174.

Gómez-Skarmeta, A., Jiménez, F., and Ibánez, J. (1998). Pareto-optimality in fuzzy modelling. In *Proc. Of EUFIT'98*, pages 694–700, Aachen, Germany.

Gottwald, S. (1979). Set theory for fuzzy sets of higher level. *Fuzzy Sets and Systems*, 2:125–151.

Groenen, P. and Jajuga, K. (2001). Fuzzy clustering with squared minkowski distances. *Fuzzy Sets and Systems*, 120(2):227–237.

Groenen, P., Mathar, R., and Heiser, W. (1995). The majorization approach to multidimensional scaling for minkowski distances. *Journal of Classification*, 12:3–19.

Guillaume, S. (2001). Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, 9(3):426–443.

Guillaume, S. and Charnomordic, B. (2003). A new method for inducing a set of interpretable fuzzy partitions and fuzzy inference systems from data. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 148–175. Springer-Verlag, Heidelberg.

Guillaume, S. and Charnomordic, B. (2004). Generating an interpretable family of fuzzy partitions from data. *IEEE Transactions on Fuzzy Systems*, 12(3):324–335.

Gustafson, E. and Kessel, W. (1979). Fuzzy clustering with a covariance matrix. In *Proceedings of the IEEE Conference on Decision Control*, pages 761–766, San Diego, US.

Hájek, P. (1998). *Metamathematics of Fuzzy Logic*. Springer.

Han, J., Cai, Y., and Cercone, N. (1993). Data-driven discovery of quantitative rules in data bases. *IEEE Transactions on Knowledge and Data Engineering*, 5:29–40.

Hansen, E. (1975). A generalized interval arithmetic. *Lecture Notes in Computer Science*, 29:7–18.

Haykin, S. (1999). *Neural Networks. A Comprehensive Foundation (2nd Ed.)*. Prentice-Hall, New Jersey, NJ.

Hermann, C. S. (1997). Symbolic reasoning about numerical data: A hybrid approach. *Applied Intelligence*, 7:339–354.

Herrera, F., Lozano, M., and Verdegay, J. (1995). Generating fuzzy rules from examples using genetic algorithms. In Bouchon-Meunier, B., Yager, R., and Zadeh, L., editors, *Fuzzy Logic and Soft Computing*, pages 11–20. World Scientific.

Herrera, F., Lozano, M., and Verdegay, J. (1998). A learning process for fuzzy control rules using genetic algorithms. *Fuzzy Sets and Systems*, 100:143–158.

Heskes, T. (1997). Practical confidence and prediction intervals. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, pages 176–182. The MIT Press.

Hobbs, J. (1985). Granularity. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 432–435.

Höppner, F. and Klawonn, F. (2000). Obtaining interpretable fuzzy models from fuzzy clustering and fuzzy regression. In *Proc. Of the 4th Int. Conf. On Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES)*, pages 162–165, Brighton, UK.

Hornsby, K. (2001). Temporal zooming. *Transactions on Geographical Information Systems*, 5:255–272.

Huang, Y. and Chu, H. (1999). Simplifying fuzzy modeling by both gray relational analysis and data transformation methods. *Fuzzy Sets and Systems*, 104:183–197.

Ichihashi, H., Shirai, T. Nagasaka, K., and Miyoshi, T. (1996). Neuro-fuzzy ID3: A method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning. *Fuzzy Sets and Systems*, 81:157–167.

IEE, editor (1988). *IEE Colloquium Qualitative Modelling in Diagnosis and Control*, Edinburgh, UK.

IJCAI, editor (1995). *IJCAI'95 (11th International Joint Conference on AI) Workshop on Comprehensibility in Machine Learning*, Montreal, Canada.

Inuiguchi, M., Hirano, S., and Tsumoto, S., editors (2003). *Rough Set Theory and Granular Computing*. Springer, Berlin.

Ishibuchi, H., Murata, T., and Türksen, I. (1997). Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems. *Fuzzy Sets and Systems*, 89:135–150.

Ishibuchi, H., Nozaki, K., and Tanaka, H. (1994). Empirical study on learning in fuzzy systems by rice taste analysis. *Fuzzy Sets and Systems*, 64:129–144.

Ishibuchi, H., Nozaki, K., Yamamoto, N., and Tanaka, H. (1995). Selecting fuzzy if-then rules for classification problems using genetic algorithms. *IEEE Transactions on Fuzzy Systems*, 3:260–270.

Ishibuchi, H. and Yamamoto, T. (2002). Fuzzy rule selection by data mining criteria and genetic algorithms. In *Proc. Of Genetic and Evolutionary Computation Conference (GECCO 2002)*, pages 399–406, New York.

Jain, A., Duin, R., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1).

Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

Jamei, M., Mahfouf, M., and Linkens, D. A. (2001). Elicitation and fine tuning of mamdani-type fuzzy rules using symbiotic evolution. In *Proceedings of European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems (EUNITE 2001)*, Tenerife, Spain.

Jang, J. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics*, 23(3):665–684.

Jang, J. (1996). Input selection for ANFIS learning. In *Proceedings of IEEE ICFS*, New Orleans.

Jang, J.-S. (1992). *Neuro-Fuzzy Modeling: Architectures, Analyses, and Applications*. PhD thesis, University of California, Berkeley.

Jang, J.-S. and Sun, C.-T. (1995). Neuro-fuzzy modeling and control. *Proceedings of the IEEE*.

Janikow, C. Z. (1998). Fuzzy decision trees: Issues and methods. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 28(1):1–14.

Jankowski, N. and Kadirkamanathan, V. (1997). Statistical control of RBF-like networks for classification. In *Proceedings of the 7th International Conference on Artificial Neural Networks (ICANN'97)*, pages 385–390, Lausanne, Switzerland.

Jaulin, L., Kieffer, M., Didrit, O., and Walter, E. (2001). *Applied Interval Analysis*. Springer, London.

Jiménez, F., Gómez-Skarmeta, A., Roubos, H., and Babuška, R. (2001). A multi-objective evolutionary algorithm for fuzzy modeling. In *Proc. Of NAFIPS'01*, pages 1222–1228, New York.

Jin, Y. (2000). Fuzzy modeling of high-dimensional systems: Complexity reduction and interpretability improvement. *IEEE Transactions on Fuzzy Systems*, 8(2):212–221.

Jin, Y. and Sendhoff, B. (2003). Extracting interpretable fuzzy rules from RBF networks. *Neural Processing Letters*, 17:149–164.

Jin, Y., Von Seelen, W., and Sendhoff, B. (1998a). An approach to rule-based knowledge extraction. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 1188–1193, Anchorage, AK. IEEE.

Jin, Y., Von Seelen, W., and Sendhoff, B. (1998b). An approach to rule-based knowledge extraction. In *Proceedings IEEE International Conference on Fuzzy Systems*, pages 1188–1193, Anchourage, AK.

Jin, Y., Von Seelen, W., and Sendhoff, B. (1999). On generating FC3 fuzzy rule systems from data using evolution strategies. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 29(6):829–845.

Johansen, T. A., Shorten, R., and Murray-Smith, R. (2000). On the interpretation and identification of dynamic takagi-sugeno fuzzy models. *IEEE Transactions on Fuzzy Systems*, 8(3):297–313.

Johansson, U., Niklasson, L., and Köning, R. (2004). Accuracy vs. comprehensibility in data mining models. In *Proceedings of the Seventh International Conference on Information Fusion*, pages 295–300, Stockholm, Sweden.

Kandel, A. (1982). *Fuzzy Techniques in Pattern Recognition*. John Wiley and Sons, New York.

Karplus, W. (1996). cited in A. Bargiela, "Cycle of lectures on: Granular Computing, An Introduction", 1-3 july 2003, Department of Informatics, University of Bari, Italy.

Kaufmann, A. and Gupta, M. (1985). *Introduction to Fuzzy Arithmetic*. Reinhold, New York.

Kearfott, R. and Kreinovich, V., editors (1996). *Applications of Interval Computations*. Kluwer, Dordrecht.

Klawonn, K. and Keller, A. (1997). Fuzzy clustering and fuzzy rules. In *Proceedings of the 7th International Fuzzy Systems Association World Congress (IFSA'97)*, page 193.198, Prague.

Klir, G. (2001). Foundations of fuzzy set theory and fuzzy logic: A historical review. *Journal of General Systems*, 30(2):91–132.

Klir, G. and Folger, T. (1988). *Fuzzy Sets, Uncertainty and Information*. Prentice Hall, Englewood Cliffs.

Klir, G. and Ramer, A. (1990). Uncertainty in the dempster-shafer theory: A critical re-examination. *International Journal of General Systems*, 18:155–166.

Knight, K. (1990). Connectionist ideas and algorithms. *Communications of the ACM*, 33(11):59–74.

Knoblock, C. (1993). *Generating Abstraction Hierarchies: An Automated Approach to Reducing Search in Planning*. Kluwer Academic Publishers, Boston.

Kohonen, T. (1986). Learning vector quantization for pattern recognition, TKK-f-a601. Technical report, Helsinki University of Technology, Finland.

Kohonen, T. (1990). Improved versions of learning vector quantization. In *IEEE International Joint Conference on Neural Networks*, volume I, pages 545–550, San Diego, CA.

Kohonen, T. (1997). *Self-Organizing Maps*. Spinger-Verlag, Berlin, 2nd edition.

Kollios, G., Gunopulos, D., Koudas, N., and Berchtold, S. (2003). Efficient biased sampling for approximate clustering and outlier detection in large data sets. *IEEE Transactions on Knowledge and Data Engneering*, 15(5):1170–1187.

Kong, S.-G. and Kosko, B. (1992). Adaptive fuzzy systems for backing up a truck-and-trailer. *IEEE Transactions on Neural Networks*, 3:211–223.

Kosko, B. (1986). Fuzzy cognitive maps. *Internation Journal of Man-Machine Studies*, 24:65–75.

Kosko, B. (1992). *Neural Networks and Fuzzy Systems*. Prentice-Hall, Englewood Cliffs.

Kosko, B. (1995). Optimal fuzzy rules cover extrema. *International Journal of Intelligent Systems*, 10(2):249–255.

Kowalczyk, R. (1998). On linguistic approximation of subnormal fuzzy sets. In *Proceedings of the 1998 Conference of the North American Fuzzy Informatio Processing Society (NAFIPS'98)*, pages 329–333. IEEE.

Krishnapuram, R. and Keller, J. (1993). A possibilistc approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110.

Kruse, R., Gebhardt, J., and Klawonn, F. (1994). *Foundations of Fuzzy Systems*. Wiley, Chichester.

Kuipers, B. (1994). *Qualitative Reasoning: Modelinf and Simulation with Incomplete Knowledge*. MIT Press, Cambridge, MA.

Kuncheva, L. (2000). *Fuzzy Classifier Design*. Physica-Verlag, Hedelberg.

Lazzerini, B. and Marcelloni, F. (2000). Some considerations on input and output partitions to produce meaningful conclusions in fuzzy inference. *Fuzzy Sets and Systems*, 113:221–235.

Lee, C. (1990). Fuzzy logic in control systems: Fuzzy logic controller – parts i and II. *IEEE Transactions on Systems, Man and Cybernetics*, 20:404–435.

Lee, S. and Lee, E. (1975). Fuzzy neural networks. *Mathematical Bioscience*, 23:151–177.

Liao, T., Celmins, R., and Hammel, I. (2003). A fuzzy c-means varian for the generation of fuzzy term sets. *Fuzzy Sets and Systems*, 135:241–257.

Lin, T. (1998). Granular computing on binary relations i: Data mining and neighborhood systems, II: Rough sets representations and belief functions. In Polkowski, L. and Skowron, A., editors, *Rough Sets in Knowledge Discovery*, pages 107–140. Physica-Verlag, Heidelberg.

Lin, T. (2001). Granular computing. In *Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, volume 2639 of *Lecture Notes in Artificial Intelligence*, pages 16–24.

Lin, T. and Cercone, N., editors (1997). *Rough Sets and Data Mining*. Kluwer Academic Publishers, Boston.

Lin, T., Yao, Y., and Zadeh, L., editors (2002). *Rough Sets, Granular Computing and Data Mining*. Physica-Verlag, Heidelberg.

Linde, A. and Gray, R. (1980). An algorithm for vector quantization design. *IEEE Transactions on Communications*, COM-28:84–95.

Linkens, D. and Chen, M. (1999). Input selection and partition validation for fuzzy modelling using neural networks. *Fuzzy Sets and Systems*, 107(3):299–308.

Lotfi, A., Andersen, H., and Chung Tsoi, A. (1996). Interpretation preservation of adaptive fuzzy inference systems. *International Journal of Approximate Reasoning*, 15:379–394.

Lozowski, A. and Zurada, J. (2000). Extraction of linguistic rules from data via neural networks and fuzzy approximation. In Cloete, J. and Zurada, J., editors, *Knowledge-Based Neurocomputing*. The MIT Press, Massachussets, MA.

Maciej, W. (2000). An axiomatic approach to scalar cardinalities of fuzzy sets. *Fuzzy Sets and Systems*, 110(2):175–179.

MacKay, D. (1992). A practical bayesan framework for back-propagation networks. *Neural Computation*, 4(3):448–472.

Malerba, D., Esposito, F., Ceci, M., and Appice, A. (2004). Top-down induction of model trees with regression and splitting nodes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):612–625.

Mani, I. (1998). A theory of granularity and its applications to problems of polysemy and underspecification of meaning. In *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning*, pages 245–255.

Marìn-Blàzquez, J. G. and Shen, Q. (2002). From approximative to descriptive fuzzy classifiers. *IEEE Transactions on Fuzzy Systems*, 10(4):484–497.

Matheron, G. (1975). *Random Sets and Integral Geometry*. Wiley.

McCalla, G., Greer, J., Barrie, J., and Pospisil, P. (1992). Granularity hierarchies. *Computer and Mathematics with Applications*, 23:363–375.

Meesad, P. and Yen, G. G. (2002). Quantitative measures of the accuracy, comprehensibility, and completeness of a fuzzy expert system. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ '02)*, pages 284–289, Honolulu, Hawaii.

Mendel, J. and John, R. (2002). Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, 10(2):117–127.

Mendel, J. M. and Liang, Q. (1999). Comments on "combinatorial rule explosion eliminated by a fuzzy rule configuration" (with authors' reply). *IEEE Transactions on Fuzzy Systems*, 7(3):369–372.

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Heidelberg.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97. url: http://www.well.com/user/smalin/miller.html (electronic reproduction by Stephen Malinowski).

Mitaim, S. and Kosko, B. (2001). The shape of fuzzy sets in adaptive function approximation. *IEEE Transactions on Fuzzy Systems*, 9(4):637–656.

Moore, R. (1962). *Interval Arithmetic and Automatic Error Analysis in Digital Computing*. PhD thesis, Department of Mathematics, Stanford University.

Morse, A. (1965). *A Theory of Sets*. Academic Press, San Diego, CA.

Mundici, D. (2002). IF-THEN-ELSE and rule extraction from two sets of rules. In Apolloni, B. and Kurfess, F., editors, *From Synapses to Rules. Discovering Symbolic Rules from Neural Processed Data*, pages 87–108. Kluwer Academic / Plenum Publishers, New York.

Narendra, K. and Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks*, 1:4–27.

Nauck, D., Klawonn, F., and Kruse, R. (1997). *Foundations of Neuro-Fuzzy Systems*. John Wiley and Sons, New York.

Nauck, D. and Kruse, R. (1994). Choosing appropriate neuro-fuzzy models. In *Proc. Of EUFIT'94*, pages 552–557, Aachen, Germany.

Nauck, D. and Kruse, R. (1998). A neuro-fuzzy approach to obtain interpretable fuzzy systems for function approximation. In *Proc. Of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'98)*, pages 1106–1111, Anchorage (AK).

Nauck, D. and Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. *Artificial Intelligence in Medicine*, 16:149–169.

Nauck, D., Nauck, U., and Kruse, R. (1996). Generating classification rules with the neuro-fuzzy system NEFCLASS. In *Proceedings of the Biennial Conference of the North American Fuzzy Information Processing Society (NAFIPS'96)*, pages 466–470.

Neal, R. (1996). Bayesian learning for neural networks. *Lecture Notes in Statistics Series*, 118.

Neter, J., Wasserman, W., and Kutner, M. (1985). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Eperimental Designs*. Irwin, Homewood, IL.

Nguyen, H. (1978). On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542.

Nguyen, S. and Skowron, A. (2001). Granular computing: A rough set approach. *Computational Intelligence*, 17(3):514–544.

Nix, D. and Weigen, A. (1994). Learning local error bars for nonlinear regression. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, pages 489–496. The MIT Press.

Norman, D. (1981). *Perspective on Cognitive Science*. Ablex Publishing Co., Norwood, NJ.

Novák, V., Pefilieva, I., and Mockor, J. (1999). *Mathematical Principles of Fuzzy Logic*. Kluwer, Boston, Dordrecht.

Ostasiewicz, W. (1999). Towards fuzzy logic. In Zadeh, L. and Kacprzyk, J., editors, *Computing with Words in Information/Intelligent Systems 1: Foundations*, pages 259–296. Physica-Verlag, Heidelberg.

Paiva, R. and Dourado, A. (2001). Merging and constrained learning fo interpretability in neuro-fuzzy systems. In *Proc. Of the to First International Workshop on Hybrid Methods for Adaptive Systems*, Tenerife, Spain.

Pal, S. and Skowron, A., editors (1999). *Rough Fuzzy Hybridization: A New Trend in Decision Making*. Springer-Verlag, Singapore.

Pal, S. K. (2004). Soft data mining, computational theory of perceptions, and a rough set approach. *Information Sciences*, 163:5–12.

Papadopoulos, G., Edwards, P., and Murray, A. (2001). Confidence estimation for neural networks: A praxtical comparison. *IEEE Transactions on Neural Networks*, 12(6).

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, 11:341–356.

Pawlak, Z. (1999). *Rough Sets. Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Dordrecht.

Pedrycz, W. (1993). *Fuzzy Control and Fuzzy Systems*. RSP Press, New York.

Pedrycz, W. (1994). Why triangular membership functions? *Fuzzy Sets and Systems*, 64(1):21–30.

Pedrycz, W. (1995). *Fuzzy Sets Engineering*. CRC Press.

Pedrycz, W. (1996a). Conditional fuzzy c-means. *Pattern Recognition Letters*, 17:625–631.

Pedrycz, W. (1996b). Interfaces of fuzzy models: A study in fuzzy information processing. *Intelligent Systems*, 90:231–280.

Pedrycz, W. (1997). *Computational Intelligence: An Introduction*. CRC Press.

Pedrycz, W., editor (2001). *Granular Computing: An Emerging Paradigm*. Physica-Verlag, Heidelberg, New York.

Pedrycz, W. (2002). Collaborative fuzzy clustering. *Pattern Recognition Letters*, 23(14):1675–1686.

Pedrycz, W. and Bargiela, A. (2002). Granular clustering: A granular signature of data. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 32(2):212–224.

Pedrycz, W. and de Oliveira, J. (1996). Optimization of fuzzy models. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 26(4):627–636.

Pedrycz, W. and Gomide, F. (1998). *An Introduction to Fuzzy Sets. Analysis and Design*. MIT Press, Cambridge (MA).

Pedrycz, W. and Smith, M. (1999). Granular correlation analysis in data mining. In *Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS99)*, pages 715–719, New York.

Pedrycz, W., Smith, M., and Bargiela, A. (2000). Granular clustering: A granular signature of data. In *Proceedings of the 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS2000)*, pages 105–109, Atlanta.

Pedrycz, W. and Vukovi, G. (1999). Quantification of fuzzy mappings: A relevance of rule-based architectures. In *Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS99)*, pages 105–109, New York.

Peña-Reyes, C.-A. and Sipper, M. (2003). Fuzzy CoCo: Balancing accuracy and interpretability of fuzzy models by means of coevolution. In Casillas, J., Cordon, O., Herrera, F., and Magdalena, L., editors, *Accuracy Improvements in Linguistic Fuzzy Modeling*, Studies in Fuzziness and Soft Computing, pages 119–146. Springer-Verlag.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford University Press, UK.

Polkowski, L. and Skowron, A. (1998). Towards adaptive calculus of granules. In *Proceedings of the 1998 IEEE International Conference on Fuzzy Systems (FUZZ'IEEE98)*, pages 111–116.

Ramot, D., Milo, R., Friedman, M., and Kandel, A. (2002). Complex fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 10(2):171–186.

Rapaport, W. (2003). Cognitive science. In Ralston, A., Reilly, E., and Hemmendinger, D., editors, *Encyclopedia of Computer Science, 4th Ed.*, pages 227–233. Wiley, England.

Raskin, J. (1994). Intuitive equals familiar. *Communications of the ACM*, 37(9):17–18.

Rieger, B. (2003). Understanding is meaning constitution. perception-based processing of natural language texts in procedural models of SCIP systems. In Wang, P., editor, *Proceedings of the 7th Joint Conference on Information Science (JCIS-03)*, pages 13–18, Research Triangle Park, Duke UP.

Riid, A., Jartsev, P., and Rüstern, E. (2000). Genetic algorithms in transparent fuzzy modeling. In *Proc. 7th Biennal Baltic Electronic Conference*, pages 91–94, Tallinn.

Riid, A. and Rüstern, E. (2000). Transparent fuzzy systems and modelling with transparency protection. In *Proc. Of IFAC Symposium on Artificial Intelligence in Real Time Control*, pages 229–234, US.

Riid, A. and Rüstern, E. (2003). Transparent fuzzy systems in modelling and control. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modelling*, pages 452–476. Springer, Berlin Heidelberg.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Rojas, I., Pomares, H., Ortega, J., and Prieto, A. (2000). Sel-organized fuzzy system generation from training examples. *IEEE Transactions on Fuzzy Systems*, 8:23–26.

Ross, T. (1997). *Fuzzy Logic with Engineering Applications*. McGraw Hill.

Roubos, H. and Setnes, M. (2001). Compact and transparent fuzzy models and classifiers through iterative complexity reduction. *IEEE Transactions on Fuzzy Systems*, 9(4):516–524.

Ruspini, E. (1969). A new approach to clustering. *Information and Control*, 15:22–32.

Ruspini, E., Lowrance, J., and Strat, T. (1992). Understanding evidential reasoning. *International Journal of Approximate Reasoning*, 6:401–424.

Russell, B. (1923). Vagueness. *The Australian Journal of Psychology and Philosophy*, 1:88–92.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Saaty, T. (1980). *The Analytic Hierarchy Processes*. McGraw Hill, New York.

Saitta, L. and Zucker, J.-D. (1998). Semantic abstraction for concept representation and learning. In *Proceedings of the Symposium on Abstraction, Reformulation and Approximation*, pages 103–120.

Scarpelli, H. and Gomide, F. (1994). Discovering potential inconsistencies in fuzzy knowledge bases using high level nets. *Fuzzy Sets and Systems*, 64:175–193.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–457.

Setnes, M. (1995). Fuzzy rule-base simplification using similarity measures. Master's thesis, Dept. of Electronic Engineering, University of Delft.

Setnes, M. (2003). Simplification and reduction of fuzzy rules. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 278–302. Springer-Verlag, Heidelberg.

Setnes, M., Babuška, R., Kaymak, U., and Van Nauta Lemke, H. R. (1998a). Similarity measures in fuzzy rule base simplification. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 28(3):376–386.

Setnes, M., Babuška, R., and Verbruggen, H. B. (1998b). Rule-based modeling: Precision and transparency. *IEEE Transactions on Systems, Man and Cybernetics, part C*, 28(1):165–169.

Setnes, M. and Roubos, H. (1999). Transparent fuzzy modeling using fuzzy clustering and GA's. In IEEE, editor, *18th Conference of the North American Fuzzy Information Processing Society, NAFIPS,*, pages 198–202, New York.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shafer, G. (1987). Probability judgement in artificial intelligence and expert systems. *Statistical Science*, 2:3–44.

Shand, N. and Breiman, M. (1996). Distribution based trees are more accurate. In *Proceedings of the International Conference on Neural Information Processing*, volume 1, pages 133–138, Hong Kong.

Shen, Q. and Chouchoulas, A. (2001). Selection of features in transparent fuzzy modelling. In *Proc. Of 2001 IEEE International Fuzzy Systems Conference*, pages 51–54. IEEE.

Shneiderman, B. and Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley.

Skowron, A. (2001). Toward intelligent systems: Calculi of information granules. *Bulletin of International Rough Set Society*, 5:193–236.

Skowron, A. and Stepaniuk, J. (1998). Information granules and approximation spaces. In *Proceedings of 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 354–361.

Skowron, A. and Stepaniuk, J. (2001). Information granules: Towards foundations of granular computing. *International Journal of Intelligent Systems*, 16(1):57–85.

Stamou, G. and Tzafestas, S. (1999). Fuzzy relation equations and fuzzy inference systems: An inside approach. *IEEE Transactions on Systems, Man and Cybernetics, part B*, 29(6):694–702.

Stell, J. and Worboys, M. (1998). Stratified map spaces: A formal basis for multiresolution spatial databases. In *Proceedings of the 8th International Symposium on Spatial Data Handling*, pages 180–189.

Ster, B. and Dobnikar, A. (1996). Neural networks in medical diagnosis: Comparison with other methods. In Bulsari, A., editor, *Proceedings of the European Conference on Artificial Neural Networks (EANN'96)*, pages 427–430.

Stoyan, D., Kendall, D., and Mecke, J. (1995). *Stochastic Geometry and Its Applications*. Wiley.

Sudkamp, T., Knapp, A., and Knapp, J. (2003). Effect of rule representation in rule base reduction. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 303–324. Springer-Verlag, Heidelberg.

Sugeno, M. and Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modelling. *IEEE Transactions on Fuzzy Systems*, 1:7–31.

Sun, C.-T. (1994). Rule-base structure identification in an adaptive-network-based fuzzy inference system. *IEEE Transactions on Fuzzy Systems*, 3:64–73.

Sunaga, T. (1958). Theory of interval algebra and its applications to numerical analysis. In *Gaukutsu Bunken Fukeyu-Kai*, Tokyo.

Takagi, T. and Sugeno, M. (1993). Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Fuzzy Systems*, 1:7–31.

Tarrazo, M. (2004). Schopenhauer's prolegomenon to fuzziness. *Fuzzy Optimization and Decision Making*, 3(3):227–254.

Tikk, D. and Baranyi, P. (2003a). Constrained optimization of fuzzy decision trees. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 125–147. Springer-Verlag, Heidelberg.

Tikk, D. and Baranyi, P. (2003b). Exact trade-off between approximation accuracy and interpretability: Solving the saturation problem for certain FRBs. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 587–604. Springer-Verlag, Heidelberg.

Tikk, D., Gedeon, T., and Wong, K. (2003). A feature ranking algorithm for fuzzy modelling problems. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 176–192. Springer-Verlag, Heidelberg.

Toth, H. (1997). Fuzziness: From epistemic considerations to terminological clarification. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5:481–503.

Toth, H. (1999). Towards fixing some 'fuzzy' catchwords: A terminological primer. In Zadeh, L. and Kacprzyk, J., editors, *Computing with Words in Information/Intelligent Systems 1. Foundations*, pages 154–181. Physica-Verlag, Heidelberg New York.

Trillas, E., de Soto, A., and Cubillo, S. (2000). A glance at implication and t-conditional functions. In Novák, V. and Perfilieva, I., editors, *Discovering the World with Fuzzy Logic*, pages 126–149. Physica-Verlag, Heidelberg New York.

Ungar, L., De Veaux, R., and Rosengarten, E. (1996). Estimating prediction intervals for artificial neural networks. In *Proc. Of Ninth Yale Workshop on Adaptive and Learning Systems*, US.

Valente de Oliveira, J. (1998). On the optimization of fuzzy systems using bio-inspired strategies. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 1229–1234, Anchorage, AK. IEEE.

Valente de Oliveira, J. (1999a). Semantic constraints for membership function optimization. *IEEE Transactions on Systems, Man and Cybernetics, part A*, 29(1):128–138.

Valente de Oliveira, J. (1999b). Towards neuro-linguistic modeling: Constraints for optimization of membership functions. *Fuzzy Sets and Systems*, 106:357–380.

Vanhoucke, V. and Silipo, R. (2003). Interpretability in multidimensional classification. In Casillas, J., Cordón, O., Herrera, F., and Magdalena, L., editors, *Interpretability Issues in Fuzzy Modeling*, pages 193–217. Springer-Verlag, Heidelberg.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks (special issue on VC Learning Theory and Its Applications)*, pages 988–1000.

Vuorimaa, P. (1994). Fuzzy self-organizing map. *Fuzzy Sets and Systems*, 66:223–231.

Wang, L. (1992). Fuzzy systems are universal approximators. In *Proceedings of the First IEEE Conference on Fuzzy Systems*, pages 1163–1170.

Wang, S., Chung, K., Shen, H., and Zhu, R. (2004). Note on the relationship between probabilistic and fuzzy clustering. *Soft Computing*, 8:523–526.

Warmus, M. (1956). Calculus of approximations. *Bulletin de l'Academie Polonaise des Sciences*, 4(5):253–259.

Wolberg, W. and Mangasarian, O. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of National Academy of Sciences*, 87:9193–9196.

Xie, X. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):841–846.

Yager, R. (1999). Approximate reasoning as a basis for computing with words. In Zadeh, L. and Kacprzyk, J., editors, *Computing with Words in Information/Intelligent Systems 1: Foundations*, pages 50–77. Physica-Verlag, Heidelberg.

Yager, R. and Filev, D. (1998). Operations for granular computing: Mixing words with numbers. In *Proceedings of the 1998 IEEE International Conference on Fuzzy Systems*, pages 123–128.

Yager, R. and Larsen, H. (1991). On discovering potential incosistencies in validating uncertain knowledge bases by reflecting on the input. *IEEE Transactions on Systems, Man and Cybernetics*, 21(4):790–801.

Yao, J., Dash, M., Tan, S., and Liu, H. (2000). Entropy-based fuzzy clustering and fuzzy modelling. *Fuzzy Sets and Systems*, 113:381–388.

Yao, Y. (2000a). Granular computing: Basic issues and possible solutions. In *Proceedings of the 5th Joint Coference on Information Sciences*, pages 186–189.

Yao, Y. (2000b). Granular computing: Basic issues and possible solutions. In *Proceedings of the 5th Joint Conference on Information Sciences*, pages 186–189.

Yao, Y. (2004). A partition model of granular computing. *Lecture Notes in Computer Sciences, Transactions on Rough Sets*, 1:232–253.

Yao, Y. and Zhong, N. (2002). Granular computing using information tables. In Lin, T., Yao, Y., and Zadeh, L., editors, *Data Mining, Rough Sets and Granular Computing*, pages 102–124. Physica-Verlag, Heidelberg.

Yen, J., Wang, L., and Gillespie, C. W. (1998). Improving the interpretability of TSK fuzzy models by combining global learning and local learning. *IEEE Transactions on Fuzzy Systems*, 6(4):530–537.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zadeh, L. (1975a). The concept of linguistic variable and its application to approximate reasoning - part 1. *Information Sciences*, 8:199–249.

Zadeh, L. (1975b). The concept of linguistic variable and its application to approximate reasoning - part 2. *Information Sciences*, 9:43–80.

Zadeh, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28.

Zadeh, L. (1979). Fuzzy sets and information granularity. In Gupta, M., Ragade, R., and Yager, R., editors, *Advances in Fuzzy Set Theory and Applications*, pages 3–18. North Holland, North Holland.

Zadeh, L. (1994). Soft computing and fuzzy logic. *IEEE Software*, 11(6):48–56.

Zadeh, L. (1996a). Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4:103–111.

Zadeh, L. (1996b). Fuzzy logic and the calculi of fuzzy rules and fuzzy graphs. *International Journal of Multiple-Valued Logic*, 1:1–39.

Zadeh, L. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 90:111–117.

Zadeh, L. (1998). Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. *Soft Computing*, 2:23–25.

Zadeh, L. (2000). Toward a logic of perceptions based on fuzzy logic. In Novák, V. and Perfilieva, I., editors, *Discovering the World with Fuzzy Logic*, pages 4–28. Physica-Verlag, Heidelberg.

Zadeh, L. (2002). Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. *Journal of Statistical Planning and Inference*, 105:233–264.

Zadeh, L. (2003). Computing with words and its application to information processing, decision and control. In *Proceedings of the 2003 IEEE Conference on Information Reuse and Integration*. The IEEE. keynote speech available at http://parks.slu.edu/IRI2003/.

Zadeh, L. (2004). Precisiated natural language (PNL). *AI Magazine*, 25(3):74–91.

Zadeh, L. and Kacprzyk, J. (1999). *Computing with Words in Information/Intelligent Systems*, volume 1,2. Physica-Verlag, Heidelberg.

Zahid, N., Abouelala, O., Limouri, M., and Essaid, A. (2001). Fuzzy clustering based on k-nearest-neighbours rule. *Fuzzy Sets and Systems*, 120:239–247.

Zhang, B. and Zhang, L. (1992). *Theory and Applications of Problem Solving*. North-Holland, Amsterdam.

Zhang, J.-S. and Leung, Y.-W. (2004). Improved possibilistc c-means clustering algorithms. *IEEE Transactions on Fuzzy Systems*, 12(2):209–217.

Zhang, L. and B., Z. (2003). The quotient space theory of problem solving. *Lecture Notes in Artificial Intelligence, Proceedings of the International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, 2639:11–15.

Zhong, N., Skowron, A., and Ohsuga, S., editors (1999). *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*. Springer-Verlag, Berlin.

Zimmermann, H. (2001). *Fuzzy Set Theory and Its Applications*. Kluwer Academic Publishers, Boston, 4th edition.

Zwick, R., Carlstein, E., and Budescu, D. (1987). Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1:221–242.