Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach

Michelangelo Ceci, Annalisa Appice, and Donato Malerba

Dipartimento di Informatica, Università degli Studi
via Orabona, 4, 70126 Bari, Italy
{ceci,appice,malerba}@di.uniba.it

Abstract. In this paper we propose a novel spatial associative classifier method based on a multi-relational approach that takes spatial relations into account. Classification is driven by spatial association rules discovered at multiple granularity levels. Classification is probabilistic and is based on an extension of naïve Bayes classifiers to multi-relational data. The method is implemented in a Data Mining system tightly integrated with an object relational spatial database. It performs the classification at different granularity levels and takes advantage from domain specific knowledge in form of rules that support qualitative spatial reasoning. An application to real-world spatial data is reported. Results show that the use of different levels of granularity is beneficial.

1 Introduction

The rapidly expanding amount of spatial data gathered by collection tools, such as satellite systems or remote sensing systems have paved the way for advances in spatial data structures [12], spatial reasoning [8] and computational geometry [23] to serve multiple tasks including storage and sophisticated treatment of real-world geometry in a spatial database. A spatial database contains (spatial) objects that are characterized by a geometrical representation (e.g. point, line, and region in a 2D context) as well as several non-spatial attributes. The widespread use of spatial databases in real-world applications (e.g geo-marketing or environmental analysis) is leading to an increasing interest in Spatial Data Mining, i.e. in mining interesting and useful but implicit knowledge. Classification of spatial objects is a fundamental task in Spatial Data Mining, where training data consists of multiple target spatial objects (primary data), possibly spatially-related with other non-target spatial objects (secondary data). The goal is to learn the concept associated with each class on the basis of the interaction of two or more spatially-referenced objects or space-dependent attributes, according to a particular spacing or set of arrangements [15].

While a lot of research has been conducted, both in propositional and multi-relational setting, on mining classification models from data eventually stored in multiple tables of a relational database, only a few works deal with classification models to be discovered in spatial database. Indeed, mining spatial classification models presents two main sources of complexity, that is, the implicit definition of spatial relations and the granularity of the spatial objects. The former is due to the fact that the geometrical representation (e.g. point, line, and region in a 2D context) and the relative positioning of spatial objects with respect to some reference system, define implicitly spatial relations of different nature, such as directional and topological.

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 99–111, 2004.

Modeling these spatial relations is a key challenge in classification problems that arise in spatial domains [24]. Indeed, both the attribute values of the object to be classified and the attribute values of spatially related objects may be relevant for assigning an object to a class from a given set of classes. The second source of complexity refers to the fact that spatial objects can be described at multiple levels of granularity. For instance, UK census data can be geo-referenced with respect to the hierarchy of areal objects:

$$\text{ED} \rightarrow \text{Ward} \rightarrow \text{District} \rightarrow \text{County},$$

based on the inside relationship between locations. Therefore, some kind of *taxonomic knowledge* of task-relevant geographic layers may also be taken into account to obtain descriptions at different granularity levels (*multiple-level classification*).

In this paper we propose a novel spatial classification method based on a multi-relational approach that takes spatial relations into account. Classification is probabilistic and is based on the extension of naive Bayes classifiers to multi-relational data. Classification rules are automatically generated by means of a spatial association rule discovery system characterized by the capability of generating association rules at multiple levels of granularity. In this way, the proposed method can deal with both sources of complexity presented above. The proposed method has been implemented in a Data Mining system tightly integrated with an object-relational spatial database. It can perform the classification at different levels of granularity and takes advantage from domain specific knowledge expressed in form of rules to support qualitative spatial reasoning. Finally, it handles categorical as well as numerical data through a contextual discretization method.

The paper is organized as follows. In the next section we discuss the background of this research and some related works. The mining of multi-level spatial association rules for classification purpose is presented in Section 3 while the multi-relational Naïve Bayes classification is described in Section 4. Section 5 describes the system architecture. Finally, an application is presented in Section 6 and some conclusions are drawn.

2 Background and Motivations

The problem of classifying spatial objects has been investigated by some researchers. Ester et al. [10] proposed a neighbourhood graph based extension of decision trees that considers both non-spatial attributes of the classified objects and relations with neighbouring objects. However, the proposed method does not take into account hierarchical relations defined on spatial objects as well as non-spatial attributes (e.g. number of residents) of neighbouring objects. In contrast, Kopersky [15] described an efficient method that classifies spatial objects by considering both spatial and hierarchical relations between spatial objects and takes into account non-spatial attributes for neighbouring objects. However this method suffers from severe limitations due to the restrictive representation formalism known as *single-table assumption* [26]. More specifically, it is assumed that data to be mined are represented in a single table of a relational database, such that each row (or tuple) represents an independent unit of the sample population and columns correspond to properties of units. This requires that

non-spatial properties of neighboring objects be represented in aggregated form causing a consequent loss of information and a change in the units of analysis.

In [20], the authors proposed to exploit the expressive power of predicate logic to represent both spatial relations and background knowledge, such as spatial hierarchies. In addition the logical notions of generality order and of downward refinement operator on the space of patterns may be profitably used to define both the search space and the search strategy. For this purpose, the ILP system ATRE [21] has been integrated in the data mining server of a prototypical Geographical Information System (GIS), named INGENS, which allows, among other things, to mine classification rules for geographical objects stored in an object-oriented database. Training is based on a set of examples and counterexamples of geographic concepts of interest to the user (e.g., ravine or steep slopes). The first-order logic representation of the training examples is automatically extracted from maps, although it is still controlled by the user who can select a suitable level of abstraction and/or aggregation of data by means of a data mining query language [19].

Similarly, the discovery of spatial association rules, that is spatial and a-spatial relationships among spatial objects, has been investigated both in propositional and multi-relational setting. A *spatial association rule* is a rule of the form "$P \rightarrow Q$ (s, c)" such that both P (body) and Q (head) are sets of literals, some of which refer to spatial properties, and $P \cap Q = \emptyset$. $P \cup Q$ is named *pattern*. The support s estimates the probability $p(P \cup Q)$, while the confidence c estimates the probability $p(Q|P)$.

Koperski and Han [14] implemented the module Geo-associator of the spatial data mining system GeoMiner that mines rules from data represented in a single relation (table) of a relational database. In contrast, in [16], the authors proposed an ILP approach to spatial association rules discovery. The algorithm SPADA (Spatial Pattern Discovery Algorithm), reported in their work, allows the extraction of multi-level spatial association rules, that is, association rules involving spatial objects at different granularity levels. SPADA has been implemented as a module of the system ARES (Association Rules Extractor from Spatial data) [2], which also supports users in the complex processes of extracting spatial objects from the spatial database, specifying the background knowledge on the application domain and defining a search bias.

Despite the fact that spatial association rule mining is a descriptive task, while classification of spatial objects is a predictive task, recent studies in Data Mining and Machine Learning have investigated the opportunity of combining association rules discovery and classification, by taking advantage of employing association rules for classification purpose [6, 3]. This approach is named associative classification [17] and several advantages are reported in the literature for this approach. First, differently from most of classifiers as decision trees, association rules consider the simultaneous correspondence of values of different attributes, hence allowing to achieve better accuracy [3]. Second, it makes association rule mining techniques applicable to classification tasks. Third, the user can decide to mine both association rules and a classification model in the same data mining process [17]. Fourth, the associative classification approach helps to solve *understandability* problems [4, 25] that may occur with some classification methods. Indeed, many rules produced by standard classification systems are difficult to understand because these systems often use only domain independent biases and heuristics, which may not fulfil user's expectation. With the associative classification approach, the problem of finding understandable

rules is reduced to a post-processing task [17]; filtering based on user-defined rule template may help in extracting understandable rules.

Although associative classification methods present several interesting aspects, they also suffer from some limitations. First, most of methods reported in the literature work under the *single-table assumption*, which is a strong limitation in those application domains characterized by a spatial dimension. Second, they have a categorical output which convey no information on the potential uncertainty in classification. Small changes in the attribute values of an object being classified may result in sudden and inappropriate changes to the assigned class. Missing or imprecise information may prevent a new object from being classified at all. In alternative, to overcome these deficiencies, we propose to use a probabilistic classifier that returns, in addition to the result of the classification, the confidence of the classification. This is an important aspect because of the increasing attention on the ROC curve analysis [11] that defines an evaluation measure to take into account the confidence of the classification. Third, reported methods require additional heuristics to identify the most effective rule at classifying a new object. Alternatively, in the proposed approach, the evaluation of the class is based on the computation of probabilities taking into account all the rules.

3 Multi-level Spatial Association Rules

In [2] the problem of mining spatial association rules has been formalized as follows:

Given a spatial database (SDB), a set S of *reference objects* tagged with a class label $c_j \in \{C_1, C_2, ..., C_L\}$, some sets R_k, $1 \leq k \leq m$, of *task-relevant objects*, a background knowledge BK including some *spatial hierarchies* H_k on objects in R_k, M *granularity levels* in the descriptions (1 is the highest while M is the lowest), a set of *granularity assignments* ψ_k which associate each object in H_k with a granularity level, a couple of thresholds $minsup[l]$ and $minconf[l]$ for each granularity level, a language bias LB that constrains the search space;

Find strong multi-level spatial association rules, that is, association rules involving spatial objects at different granularity levels.

The reference objects are the main subject of the description, that is, the observation units, while the task relevant objects are spatial objects that are relevant for the task in hand and are spatially related to the former. The sets R_k typically correspond to layers of the spatial database, while hierarchies H_k define *is-a* (i.e., taxonomical) relations of spatial objects in the same layer (e.g. river *is-a* water body). Objects of each hierarchy are mapped to one or more of the M user-defined description granularity levels in order to deal uniformly with several hierarchies at once. Both frequency of patterns and strength of rules depend on the granularity level l at which patterns/rules describe data. Therefore, a pattern P ($s\%$) at level l is *frequent* if $s \geq minsup[l]$ and all ancestors of P with respect to H_k are frequent at their corresponding levels. An association rule $Q \rightarrow R$ *(s%, c%)* at level l is *strong* if the pattern $Q \cup R$ ($s\%$) is frequent and $c \geq minconf[l]$.

The problem above is solved by the algorithm SPADA [16] that operates in three steps for each granularity level: i) pattern generation; ii) pattern evaluation; iii) rule

generation and evaluation. SPADA takes advantage of statistics computed at granularity level l when computing the supports of patterns at granularity level l+1.

In the system ARES (http://www.di.uniba.it/~malerba/software/ARES/index.htm) SPADA has been loosely coupled with a spatial database, since data stored in the SDB Oracle Spatial are pre-processed and then represented in a deductive database (DDB). For instance, spatial intersection between two objects X and Y is represented by the extensional predicate *crosses(X,Y)*. In this way, the expressive power of first-order logic in databases is exploited to specify both the background knowledge BK, such as spatial hierarchies and domain specific knowledge, and the language bias LB. Spatial hierarchies allow to face with one of the main issues of spatial data mining, that is, the representation and management of spatial objects at different levels of granularity, while the domain specific knowledge stored as a set of rules in the intensional part of the DDB supports qualitative spatial reasoning. On the other hand, the LB is relevant to allow the user to specify his/her bias for interesting solutions, and then to exploit this bias to improve both the efficiency of the mining process and the quality of the discovered rules. In SPADA, the language bias is expressed as a set of constraint specifications for either patterns or association rules. Pattern constrains allow to specify a literal or a set of literals that should occur one or more times in discovered patterns. During the *rule generation* phase, patterns that do not satisfy a pattern constraint are filtered out. Similarly, rule constraints are used do specify literals that should occur in the head or body of discovered rules.

In a more recent release of SPADA (3.1) a new rule constraint has been introduced in order to specify the maximum number of literal that should occur in the head of a rule. In this way users may define the head structure of a rule requiring the presence of exactly a specific literal and nothing more. In the case this literal describes the class label, multi-level spatial association rules discovered by ARES may be used for classification.

4 Naïve Bayes Classification

Once a set of rules has been extracted for each level, it is used in the construction of a naïve Bayesian classifier [5], which aims to classify any target object $o \in S$ by maximizing the *posterior probability* $P(C_i|o)$ that o is of class C_i, that is:

$$class(o) = arg\ max_i\ P(C_i|o)$$

By applying the Bayes theorem, $P(C_i|o)$ can be reformulated as follows:

$$P(C_i|o) = \frac{P(C_i)P(o|C_i)}{P(o)} \quad (1)$$

The term $P(o|C_i)$ is estimated by means of the *naïve Bayes assumption*:

$$P(o|C_i) = P(o_1, o_2, \ldots, o_m|C_i) = P(o_1|C_i) \times P(o_2|C_i) \times \ldots \times P(o_m|C_i)$$

where $o_1, o_2, \ldots, o_m$ represent the set of the properties, different from the class, used to describe the object. This assumption is clearly false if the predictor variables are statistically dependent. However, even in this case, the naïve Bayesian classifier can give good results [5].

In (1) the value $P(C_i)$ is the prior probability of the class C_i. Since $P(o)$ is independent of the class C_i, it does not affect $f(o)$, that is,

$$\text{class(o)} = \arg\max_i \; P(C_i)P(o|C_i) \tag{2}$$

However, this formulation of the problem holds in the *single-table assumption* data representation formalism, where an object represents an independent unit of the sample population described by means of a set of properties. In the multi-relational setting [7], the target object is related to other non-target objects. In order to take into account the relations of the target object, a modification of the problem formulation is necessary. For this purpose, a key role is played by the extracted association rules. In particular, the idea is to consider the set of rules to guide the computation of $P(o|C_i)$.

Given the object $o \in S$, we consider the subset of the extracted rules that can be used to classify o. More formally, we consider the subset R of rules whose body is satisfied by the object to be classified both in terms of the values of properties of involved spatial objects and in terms of the spatial relations between objects. For example, if S is the set of wards in a district, a ward w satisfies the rule:

mortality_rate(A, low) ← *wards_relatedTo_waters(A, B), waters_typewater(B, river), cars_per_person(A, high)*

when w is spatially related (intersects) to a river and is characterized by a high average number of cars per person.

We use R to estimate $P(o|C_i)$. In particular, we estimate $P(o|C_i)$ by means of the probabilities associated to both spatial relations (e.g. *wards_relatedTo_waters(A,B)*) and properties (e.g. *waters__typewater(B,RIVER)*, *cars_per_person(A,high)*) associated to each rule in R.

For instance, if $R = \{R_1, R_2\}$, where R_1 and R_2 are two association rules of class C_i extracted by SPADA:

$$R_1\colon \beta_{1,0} :- \beta_{1,1}, \beta_{1,2} \qquad R_2\colon \beta_{2,0} :- \beta_{2,1}, \beta_{2,2}$$

where $\beta_{1,1}$ and $\beta_{1,2}$ are spatial relations, $\beta_{1,2}$ and $\beta_{2,2}$ are properties and $\beta_{1,0} = \beta_{2,0}$ (class) then $P(\{R_1, R_2\}|C_i) = P(\beta_{1,0} \cap \beta_{1,1} \cap \beta_{2,1} \cap \beta_{1,2} \cap \beta_{2,2} \mid C_i) =$

$$P(\beta_{1,0} \cap \beta_{1,1} \cap \beta_{2,1} \mid C_i) \cdot P(\beta_{1,2} \cap \beta_{2,2} \mid \beta_{1,0} \cap \beta_{1,1} \cap \beta_{2,1} \cap C_i)$$

The first term takes into account the relations of the rules while the second term refers to the conditional probability of satisfying the property predicates in the rules given the relations. By means of the naïve Bayes assumption, the probabilities can be factorized as follows:

$$P(\beta_{1,0} \cap \beta_{1,1} \cap \beta_{2,1} \mid C_i) = P(\beta_{1,1} \mid C_i) \cdot P(\beta_{2,1} \mid C_i)$$
$$P(\beta_{1,2} \cap \beta_{2,2} \mid \beta_{1,0} \cap \beta_{1,1} \cap \beta_{2,1} \cap C_i) = P(\beta_{1,2} \mid \beta_{1,1} \cap \beta_{2,1} \cap C_i) \cdot P(\beta_{2,2} \mid \beta_{1,1} \cap \beta_{2,1} \cap C_i)$$

Since $\beta_{1,2}$ and $\beta_{2,2}$ do not depend from $\beta_{2,1}$ and $\beta_{1,1}$ respectively, then:

$$P(\beta_{1,2} \cap \beta_{2,2} \mid \beta_{1,0} \cap \beta_{1,1} \cap \beta_{2,1} \cap C_i) = P(\beta_{1,2} \mid \beta_{1,1} \cap C_i) \cdot P(\beta_{2,2} \mid \beta_{2,1} \cap C_i)$$

By generalizing to a set of rules we have:

$$P(C_i)P(o \mid C_i) = P(C_i) \prod_{k \in |R|} (P(relations_k \mid C_i) \prod_j P(property_{k,j} | relations_k, C_i)) \tag{3}$$

where the term $relations_k$ represents the event that the set of spatial relations expressed in the k-th rule is satisfied, while the term $property_{k,j}$ represents the event that the j-th property of the k-th rule is satisfied.

If $relations_k = \{ relation(Set_1, Set_2) \mid Set_1, Set_2 \in \{S\} \cup \{R_k, 1 \le k \le m\}, Set_1 \neq Set_2 \}$ is a set of binary relations between spatial objects (either task relevant or reference) involved in the k-th rule, the probability $P(relations_k | C_i)$ is computed by means of the naïve Bayes assumption:

$$P(relations_k | C_i) = \prod_{l \in |relations_k|} P(relation(Set_{l_1}, Set_{l_2}) \mid C_i)$$

where:

$$P(relation(Set_{l_1}, Set_{l_2}) \mid C_i) = P(relation(Set'_{l_1}, Set'_{l_2})) = \frac{|relation(Set'_{l_1}, Set'_{l_2})|}{|Set'_{l_1}| \cdot |Set'_{l_2}|} \quad (4)$$

where Set'_l is a subset of objects in Set_l that are related, by means of spatial relations, with objects in S of class C_i, while $|relation(Set'_{l_1}, Set'_{l_2})|$ is the number of relations between objects of Set'_{l_1} and objects of Set'_{l_2}.

To compute the probability $P(property_{k,j} | relations_k, C_i)$ in (3), we use the Laplace estimation:

$$P(property_{k,j} | relations_k, C_i) = \frac{|relations_k \wedge property_{k,j} \wedge C_i| + 1}{|relations_k \wedge C_i| + F} \quad (5)$$

where F is the number of possible admissible values of the property. Laplace's estimate is used in order to avoid null probabilities in equation (2). In practice, the value at the nominator is the number of target objects of class C_i that are related to other spatial objects by means of spatial relations expressed in $relations_k$ and for which $property_{k,j}$ is satisfied. The value of the denominator is the number of target objects of class C_i that are related to other spatial objects by means of spatial relations expressed in $relations_k$ plus F.

In order to avoid the problem that the same relation or the same property is considered more than once in the computation of probabilities in formula (3), the values computed in formula (4) and (5) are effectively determined and included in formula (3) only if the values have not been computed before.

5 A Spatial Associative Classification Framework

The integration of multi-level spatial association rules discovery with naïve Bayesian classification is realized in a spatial associative classification system based on a client-server model (see Fig. 1). Both the spatial association rule miner SPADA and the multi-relational naïve Bayes classifier are on the server side, so that several data mining tasks can be run concurrently by multiple users. SPADA fully exploits the flexibility of ILP to specify the background knowledge *BK* (i.e hierarchies and domain specific knowledge) as well as the language bias *LB* (i.e. search constraints). Hierarchies are expressed by a collection of ground atoms and represent spatial objects at different granularity level while domain specific knowledge is expressed as sets of definite clauses and support a spatial qualitative reasoning. Conversely, search

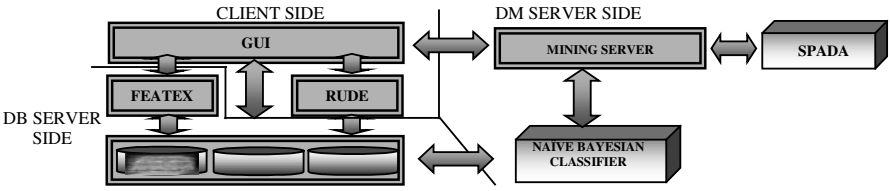

Fig. 1. Spatial associative classification system.

constraints are used to bias the search in order to fulfil user expectations. In this framework, constraints are also used to partially fix the structure of extracted rules in order to discover spatial association rules that contain only the class label in the head. For each granularity level, extracted rules concur in building the spatial classification model by exploiting a multi-relational naïve Bayesian classifier integrated with the SDB.

On the client side, the framework includes a Graphical User Interface (GUI), which provides users with facilities for controlling all parameters of the mining process.

SPADA, like many other association rule mining algorithms, cannot process numerical data properly, so it is necessary to perform a discretization of numerical features with a relatively large domain. For this purpose, the framework includes in the client side the module RUDE (relative unsupervised discretization algorithm) which discretizes a numerical attribute of a relational database in the context defined by other attributes [18].

The SDB (Oracle Spatial) can run on a third computation unit. Many spatial features (relations and attributes) can be extracted from spatial objects stored in the SDB. Feature extraction requires complex data transformation processes to make spatial relations explicit and representable as ground Prolog atoms. Therefore, a middle layer module, named FEATEX (Feature Extractor), is required to make possible a loose coupling between SPADA and the SDB by generating features of spatial objects (points, lines, or regions). The module is implemented as an Oracle package of procedures and functions, each of which computes a different feature [2]. Transformed data are also stored in SDB tables.

6 The Application: Mining North West England Census Data

In this section we present a real-world application concerning the mining of both spatial association rules and classification models for geo-referenced census data interpretation. We consider both census and digital map data provided in the context of the European project SPIN! (Spatial Mining for Data of Public Interest) [22]. They concern Greater Manchester, one of the five counties of North West England (NWE). Greater Manchester is divided into ten metropolitan districts, each of which is decomposed into censual sections or wards, for a total of two hundreds and fourteen wards. Spatial analysis is enabled by the availability of vectorized boundaries of the 1998 census wards as well as by other Ordnance Survey digital maps of NWE, where several interesting layers are found, namely road net, rail net, water net, urban area and green area (see Table 1).

Table 1. Geographic layers.

Layer name		*Geometry*
Road net	A-road; B-road; Motorway; Primary road	Line
Rail net	Railway	Line
Urban area	Large urban area; Small urban area	Line
Green area	Wood; Park:	Line
Water net	Water; River; Canal	Line
Greater Manchester Ward	Ward	Region

Census data are available at ward level. They provide socio-economic statistics (e.g. mortality rate, that is, the percentage of deaths with respect to the number of inhabitants) as well as some measures describing the deprivation level. Indeed, the material deprivation of an area may be estimated according to information provided by Census combined into single index scores [1]. Over the years different indices have been developed for different applications: the Jarman Underprivileged Area Score was designed to measure the need for primary care, the indices developed by Townsend and Carstairs have been used in health-related analyses, while the Department of the Environment's Index (DoE) has been used in targeting urban regeneration funds. Thereby, we have considered the values of Jarman index, Townsend index, Carstairs index and DoE index. The higher the index value the more deprived a ward is. Both index values as well as mortality rate are all numeric and have been discretized by means of RUDE. More precisely, Jarman index, Townsend index, DoE index and Mortality rate have been automatically discretized in (*low, high*), while Carstairs index has been discretized in (*low, medium, high*).

For this application, we have considered Greater Manchester wards as reference (target) objects. In particular, three different experimental settings have been analysed by varying the target property among mortality rate, Jarman index and DoE index. We have chosen Jarman and DoE indices because they are defined on the basis of different social factors. For each setting, we have focused our attention on investigating dependencies between the target property and socio-economic factors represented in census data as well as geographical factors represented in linked topographic maps. These dependencies are detected in form of spatial association rules having only the target property in the head. Rules in this form may be employed for spatial subgroup mining, that is, discovery of interesting groups of spatial objects with respect to a certain property of interest [13] as well as for classification purpose.

For this analysis, we have formulated queries involving the FEATEX *relate* function to compute topological relationships between reference objects and task relevant objects. For instance, a relationship extracted by FEATEX is *crosses(ward_135, urbareaL_151)*, where *ward_#* denotes a specific Greater Manchester ward, while *urbanareaL#* refers to a large urban area crossing the interested ward. The topological relationship *crosses* is computed according to the 9-intersection model [9]. The number of computed relationships is 784,107.

To support a spatial qualitative reasoning, a domain specific knowledge (BK) has been expressed in form of a set of rules. Some of these rules are:

crossed_by_urbanarea(X,Y) :- connects(X,Y), is_a(Y, urban_area). …
crossed_by_urbanarea(X,Y) :- inside(X,Y), is_a(Y, urban_area).

Here the use of the predicate *is_a* hides the fact that a hierarchy has been defined for spatial objects which belong to the urban area layer. In detail, five different hierarchies have been defined to describe the following layers: road net, rail net, water net, urban area and green area (see Fig. 2). The hierarchies have depth three and are straightforwardly mapped into three granularity levels. They are also part of the BK.

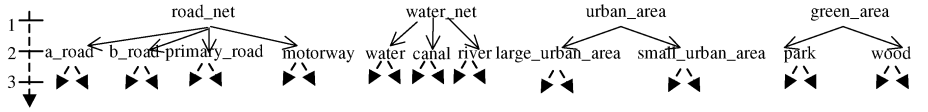

Fig. 2. Spatial hierarchies defined for road net, water net, urban area and green area.

Finally, we have specified a language bias (LB) both to constrain the search space and to filter out uninteresting spatial association rules. In particular, we have ruled out all spatial relations (e.g. crosses, inside, and so on) directly extracted by FEATEX and asked for rules containing topological predicates defined by means of BK. Moreover, by combining the rule filters *head_constraint([mortality_rate(_),1,1)* and *rule_head_length(1,1)* we have asked for rules containing only *mortality rate* in the head. Similar considerations apply to the classification tasks concerning the Jarman and the DoE indices. In addition, we have specified the maximum number *K* of refinement steps (i.e. number of literals in the body of rules).

For each setting, a ten-fold cross validation has been performed and results are evaluated. For instance, by analyzing spatial association rules extracted with parameters *minsup* = 0.1, *minconf* = 0.6 we discover the following rule:

mortality_rate(A, high) ← *is_a(A, ward), crossed_by_urbanarea(A, B),*
is_a(B, urban_area), townsendidx_rate(A, high) (40.72%, 72.47%)

which states that a high mortality rate is observed in a ward *A* that includes an urban area *B* and has a high value of Townsend index. The support (40.72%) and the high confidence (72.47%) confirm a meaningful association between a geographical factor, such as living in deprived urban areas, and a social factor, such as the mortality rate. It is noteworthy that SPADA generates the following rule:

mortality_rate(A, high) ← *is_a(A,ward), crossed_by_urbanarea(A,B),*
is_a(B, urban_area) (56.7%, 60.77%)

which has a greater support and a lower confidence. These two association rules show together an unexpected association between Townsend index and urban areas. Apparently, this means that this deprivation index is unsuitable for rural areas.

At a granularity level 2, SPADA specializes the task relevant object B by generating the following rule which preserves both support and confidence:

mortality_rate(A, high) ← *is_a(A, ward), crossed_by_urbanarea(A, B),*
is_a(B, urban_areaL), townsendidx_rate(A,high) (40.72%, 72.47%)

This rule clarifies that the urban area B is large.

The average predictive accuracy of mined multi-level spatial classification model is evaluated by varying *minsup*, *minconf* and *K* for each setting,. Results are reported in Table 2, 3 and 4. In the first setting, results show that, predictive accuracy of the Bayesian classifier is slightly better than the accuracy (0.567) of the trivial classifier that returns the most probable class. We explain this result with the inherent complex-

Table 2. Mortality Rate average accuracy.

MORTALITY Avg. Accuracy		*K=4*	*K=5*	*K=6*	*K=7*
minsup=0.1	*Level=1*	0.5932	0.5915	0.5932	0.628
minconf=0.6	*Level=2*	0.5932	0.596	0.5932	0.628
minsup=0.2	*Level=1*	0.5932	0.602	0.5932	0.623
minconf=0.65	*Level=2*	0.5932	0.602	0.5932	0.623

Table 3. Jarman average accuracy.

JARMAN Avg. Accuracy		*K=4*	*K=5*	*K=6*	*K=7*
minsup=0.1	*Level=1*	0.8176	0.8176	0.8176	0.8176
minconf=0.6	*Level=2*	0.8176	0.8176	0.8176	0.8176
minsup=0.2	*Level=1*	0.528	0.528	0.528	0.528
minconf=0.8	*Level=2*	0.528	0.528	0.6272	0.6705

Table 4. DoE average accuracy.

DoE Avg. Accuracy		*K=4*	*K=5*	*K=6*	*K=7*
minsup=0.1,	*Level=1*	0.912	0.912	0.912	0.912
minconf=0.6	*Level=2*	0.912	0.912	0.912	0.912
minsup=0.2,	*Level=1*	0.875	0.875	0.875	0.821
minconf=0.8	*Level=2*	0.875	0.9028	0.883	0.874

ity of the task. Different conclusions can be drawn from both Jarman and DoE results, where the Bayesian classifiers significantly improve the trivial classifiers (acc. 0.542 and 0.625, respectively). Another consideration is that the average predictive accuracies of classification models discovered at higher granularity levels (i.e. level=2) are always better or equal to the corresponding accuracies at lowest levels. This means that the classification model takes advantage of the use of the hierarchies defined on spatial objects. Furthermore, results show that by decreasing the number of extracted rules (higher support and confidence) we have lower accuracy. This means that there are several rules that strongly influence classification results and often such rules are not characterized by high values of support and confidence. Finally, we observe that, generally, the higher the number of refinement steps, the better the model.

7 Conclusions

In this paper we have presented a spatial associative classifier that combines spatial association rule discovery with naïve Bayes classification. Domain specific knowledge may be defined as a set of rules that makes possible the qualitative spatial reasoning. In addition, hierarchies on spatial objects are expressed by a collection of ground atoms and are exploited to mine classification models at different granularity levels. Search constraints are used to bias the spatial association rules discovery in order to fulfil user expectations. In particular, constraints are also used to partially fix the structure of extracted rules in order to discover spatial association rules that contain only the class label in the head. Finally, for each granularity level, extracted rules concur in building the spatial classification model by exploiting a multi-relational naïve Bayesian classifier integrated with the SDB.

Experiments on real-world spatial data show that the use of different levels of granularity generally increases the accuracy of the mined classification model. As future work, we intend to frame the work within the context of hierarchical Bayesian classifiers, in order to exploit the multi-level nature of extracted association rules.

Acknowledgments

We would like to thank Jim Petch, Keith Cole and Mohammed Islam (University of Manchester) for expert collection, collation, editing and delivery of the several data sets made available through Manchester Computing in the context of the IST European project SPIN! (Spatial Mining for Data of Public Interest).

References

1. Andrienko, G., Andrienko, N.: Exploration of heterogeneous spatial data using interactive geo-visualization tools: study of deprivation indices in North-West England. North-West England Report. IST European project SPIN!(Spatial Mining for Data of Public Interest).
2. Appice, A., Ceci, M., Lanza, A., Lisi, F.A. Malerba, D.: Discovery of Spatial Association Rules in Georeferenced Census Data: A Relational Mining Approach, Intelligent Data Analysis. Special issue of Mining Official Data. 7(6) (2003).
3. Baralis, E., Garza, P.; Majority Classification by Means of Association Rules., Knowledge Discovery in Databases PKDD'03, LNCS 2838, Springer-Verlag (2003), 35-46.
4. Clark, P., Matwin, S.: Using qualitative models to guide induction learning. Proceedings of International Conference of Machine Learning, Morgan Kaufmann, (1993), 49-56.
5. Domingos, P. Pazzani, M.: On the optimality of the simple Bayesian classifier under zeo-ones loss. Machine Learning, 29 (2-3), (1997), 103-130.
6. Dong, G., Zhang, X., Wong, L. Li, J.: Classification by aggregating emerging patterns. Proceedings of DS'99 (LNCS 1721), Japan, (1999).
7. Dzeroski, S., Lavrac, N. (eds.): Relational Data Mining. Springer-Verlag, Berlin, (2001).
8. Egenhofer, M.J.: Reasoning about Binary Topological Relations. Proceedings of the Second Symposium on Large Spatial Databases, Zurich, Switzerland, (1991) 143-160.
9. Egenhofer, M.J., Franzosa, R.: Point-Set Topological Spatial Relations, International Journal of Geographical Information Systems, 5(2), (1991), 61-174.
10. Ester, M. Kriegel, H.P., Sander J.: Spatial Data Mining: A Database Approach. Proceedings International Symposium on Large Databases, Berlin, (1997), 47-66.
11. Fürnkranz, J., Flach, P.A.: An analysis of rule evaluation metrics. Proceedings of International Conference of Machine Learning, Morgan Kaufmann, (2003).
12. Güting, R.H.: An introduction to spatial database systems. VLDB Journal, 4(3) (1994)
13. Klösgen, W., May, M.: Spatial Subgroup Mining. Proceedings of European Syposium of Principles of Knowledge Discovery in Database PKDD'02. Springer-Verlag, (2002).
14. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases. Advances in Spatial Databases. LNCS 951, Springer-Verlag, (1995) 47-66.
15. Koperski, K.: Progressive Refinement Approach to Spatial Data Mining,Ph.D. thesis, Computing Science, Simon Fraser University, (1999).
16. Lisi, F.A., Malerba, D.: Inducing Multi-Level Association Rules from Multiple Relations. Machine Learning, (2004), to appear.
17. Liu, B., Hsu, W., Ma, Y.: Integratine classification and association rule mining. Proceedings of Knowledge Discovery in Databases KDD'98, New York, (1998).
18. Ludl, M.C., Widmer, G.: Relative Unsupervised Discretization for Association Rule Mining. PKDD'00, LNCS 1910, Springer-Verlag, (2000), 148-158.

19. Malerba D., A. Appice, & N. Vacca. SDMOQL: An OQL-based Data Mining Query Language for Map Interpretation Tasks. Proc. of the Workshop on Database Technologies for Data Mining (DTDM'02), Prague, Czech Republic, March 25-27, 2002
20. Malerba, D., Esposito, F., Lanza, A., Lisi, F.A., Appice, A.: Empowering a GIS with Inductive Learning Capabilities: The Case of INGENS. Journal of Computers, Environment and Urban Systems, Elsevier Science, 27 (2003). 265-281.
21. Malerba, D.: Learning Recursive Theories in the Normal ILP Setting, Fundamenta Informaticae, 57, 1, (2003), 39-77.
22. May, M.: Spatial Knowledge Discovery: The SPIN! System. In: Fullerton, K. (ed.): Proceedings of the 6th EC-GIS Workshop, Lyon, JRC, Ispra, (2000).
23. Preparata, F., Shamos, M.: Computational Geometry: An Introduction. Springer-Verlag, New York (1985).
24. Shekhar, S., Schrater, P.R., Vatsavai, R. R., Wu, W., Chawla, S.: Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. IEEE Transaction on Multimedia, 4(2) (2002) 174-188.
25. Pazzani, M., Mani, S., Shankle, W.R.: Beyond concise and colorful: learning intelligible rules. In Proceedings of Knowledge Discovery in Databases KDD'97, (1997).
26. Wrobel, S.: Inductive logic programming for knowledge discovery in databases. In: Džeroski, S., N. Lavra (eds.): Relational Data Mining, Springer: Berlin, (2001) 74-101.