[27] A. Bhalerao and R. G. Wilson, "Multiresolution image segmentation combining region and boundary information," Tech. Rep. RR154, Dept. of Comput. Sci., Univ. of Warwick, Coventry, UK, 1989.

Classification in Noisy Environments Using a Distance Measure Between Structural Symbolic Descriptions

Floriana Esposito, Donato Malerba, and Giovanni Semeraro

Abstract—**A definition of distance measure between structural descriptions, which is based on a probabilistic interpretation of the matching predicate, is proposed. It aims at coping with the problem of classification when noise causes both local and structural deformations. The distance measure is defined according to a top-down evaluation scheme: distance between disjunctions of conjuncts, conjunctions, and literals. At the lowest level, the similarity between a feature value in the pattern model (G) and the corresponding value in the observation (Ex) is defined as the probability of observing a greater distortion. The classification problem is approached by means of a multilayered framework in which the cases of single perfect match, no perfect match, and multiple perfect match are treated differently. Another possible application of the distance measure is in the field of concept acquisition. A plausible solution for the problem of completing the attribute and structure spaces, based on the probabilistic approach, is also given. Finally, both a comparison with other related works and an application in the domain of layout-based document recognition are illustrated.**

Index Terms—**Distance measure, flexible matching, incomplete descriptions, learning structural descriptions from examples, learning systems, pattern classification in noisy environments.**

I. Introduction

Learning tasks in real-world domains are often affected by some degree of uncertainty due to either the presence of noise or to the imprecision of measuring instruments or simply to the variability of the phenomena themselves.

Sometimes, when the origin of uncertainty is only noise, it is possible to select noiseless observations for the training phase, but this does not assure that future events will be noiseless. Therefore, the classification phase still remains affected by uncertainty. In such a situation, a simple classifier based on the match/unmatch criterion [1], [2] is useless, and it becomes necessary to rely on a less rigid definition of similarity between observations, namely, a measure taking on values in a continuous range and not in a boolean set. Such a measure of similarity is strictly connected to the concept of distance because if the two objects are more distant, the less they can be considered similar [3]–[7].

A classical distance measure is the Euclidean metric, defined on feature vectors whose components are interval or ratio level measurements. Although, such a metric is attractive for its computational simplicity, it forces us to represent objects by numerical feature vectors only, which is a serious limitation for real-world problems [8]. Even if some extensions are possible in order to include other kinds of measurements, such as nominal or ordinal-level measurements [9], there still remains the difficulty in representing the *structure* of an object that may be vital in many real applications. To describe an object, it may be decomposed by successive refinements until atomic parts, called *primitives*, are defined. Once these subparts and their mutual relationships are identified, the structure is obtained. Although feature vectors are adequate in representing the attributes of each primitive, they are unsuitable for describing relations among a variable number of subparts. Consequently, distances based on such a representation are only able to cope with *local deformations*, that is, differences at the primitive attribute level between the pattern recognized (ideal pattern) and the new object to be classified [10].

However, *structural deformations* may also arise when the number of subparts of the ideal pattern and the new events are different or the relationships between component parts change substantially. This problem is particularly felt in the field of pattern recognition when image segmentation is performed [11]. In fact, earlier research that dealt with structural deformations [3], [4], [10] used the terminology and representation of structured objects, through attributed relational graphs, which is typical of the field of pattern recognition. Later works [12] have pointed out the generality of the problem and the utility of the concept of distance also in the area of machine learning.

For instance, some possible applications of a more general definition of distance may be the following:

1) Classifying examples that do not exactly have all the regularities appearing in the corresponding recognition rule
2) evaluating the validity of the hypotheses generated during or at the end of an inductive inference process
3) clustering a given collection of objects (observations, situations, etc.) in a hierarchical structure of meaningful subcategories
4) selecting typical training examples in order to optimize the learning process
5) resolving ambiguity in multiple group classification with emphasis on what an example *is* versus what an example *contains* [13].

In this paper, the definition of a distance measure between structural descriptions, expressed as disjunctive *well-formed formulas* (wff's) in VL_{21} representation language, is proposed. First, some definitions and basic notions about the problem of pattern matching are introduced. Then, we proceed to formalize, by a top-down evaluation scheme, a measure of fitness suitable for structural deformations: the measure of fitness between formulas represented in disjunctive normal form in Section III, between conjunctions (products of selectors) in Section IV, and between selectors in Section V. In Section VI, we raise some questions about the application of the distance measure to both the conceptual clustering and the process of learning from examples. The problem of completing the attribute and structure spaces is addressed in Section VII. Section VIII is devoted to the illustration of an application of the distance to digitized document recognition. Finally, a comparison with related works both in the area of machine learning and pattern recognition is sketched out.

II. The Problem of Matching Two Formulas

Because the definition of the distance measure is closely related to the problem of matching two formulas and searching for the most general unifier, some notions concerning the matching process and its potential optimization will be introduced. Initially, however, we must briefly present the VL_{21} representation language in which formulas are expressed.

A. The VL_{21} Language

The VL_{21} system is a multivalued version of a first-order predicate logic (FOPL) whose basic component is the *selector* or relational

Manuscript received March 15, 1990; revised March 31, 1991.

The authors are with the Dipartimento di Informatica, Università di Bari, Bari, Italy.

IEEE Log Number 9102092.

statement, which is written as

$$[L = R]$$

where

- L, known as the *referee*, is a functional symbol with its arguments.
- R, known as the *reference*, is a disjunction of values of the referee's domain.

Function symbols of referees are called *descriptors*, and they are n-adic typed functions ($n \geq 1$) mapping onto one of three different kinds of domains: *nominal, linear*, and *tree structured*.

Selectors can be combined by applying different operators, some of which are AND, OR, and decision operator (::>) in order to represent facts and recognition rules. (For a more complete definition of VL_{21}, refer to [14]).

A conjunctive VL_{21} wff can be represented as a graph with labeled nodes and directed labeled edges [15], [16]. The labels on the nodes can be either a selector containing an n-ary descriptor without its argument list or a quantified variable. The edges are optionally labeled with integers $1, 2, \ldots$ that refer to the position of the argument at the head of the edge. The label is omitted if the argument position is irrelevant. For instance, the VL_{21} formula

$$[\text{on\_top}(x1, x2)][\text{shape}(x1) = \text{square}][\text{shape}(x2) = \text{rectangle}]$$

has the following graph representation:

$$\begin{array}{lcl} \exists x1 & \longrightarrow & [\text{shape} = \text{square}] \\ & {\scriptstyle 1}\searrow & \\ & & [\text{on\_top} = \text{true}] \\ & {\scriptstyle 2}\nearrow & \\ \exists x2 & \longrightarrow & [\text{shape} = \text{rectangle}] \end{array}$$

A conjunctive VL_{21} wff is *connected* if its graph structure representation is completely connected, i.e., there exists a traversal of the entire graph from any starting node [15].

Such a constraint on the connection of formulas actually represents a restriction on the variables appearing as arguments of the VL_{21} descriptors but has the advantage of speeding the matching process up. Moreover, it is always possible to turn a VL_{21} formula into an equivalent connected formula; for example, the formula

$$[\text{on\_top}(x1, x2)][\text{shape}(x1) = \text{square}][\text{shape}(x3) = \text{rectangle}]$$

may be redefined as

$$[\text{part\_of}(Ex, x1)][\text{part\_of}(Ex, x2)][\text{part\_of}(Ex, x3)]$$
$$[\text{on\_top}(x1, x2)][\text{shape}(x1) = \text{square}][\text{shape}(x3) = \text{rectangle}]$$

which is a connected one.

B. The Problem of Pattern Matching

Generally, if $S1$ and $S2$ are two selectors, the problem of matching $S1$ against $S2$ can be formally stated as follows: to find a substitution σ for the variables in $S1$ such that $\sigma(S1) = S2$. This last condition is generally weakened in classification problem solving, and it is required that $S2 \Rightarrow \sigma(S1)$, where $\Rightarrow$ is the logic implication. Thus, in order to match two selectors, the following conditions are to be satisfied:

1) They must have the same descriptor.
2) A consistent binding must exist between the variables appearing as arguments for the descriptors.
3) The reference of $S1$ should be "more general" than that of $S2$.

This classical matching paradigm is also known as syntactic matching [17]. The output is a boolean value since a pattern $S1$ either does or does not match against $S2$.

The problem of matching two VL_{21} wff's, Ex and G, is more complex. Once again, it is said that Ex matches against G if there exists a substitution σ for the variables in G such that $Ex \Rightarrow \sigma(G)$. In the field of structural pattern recognition, where objects or patterns are usually represented by labeled graphs, such a problem is known as subgraph isomorphism [18]. Unfortunately, it can be shown that such a problem is NP complete [19]; thus, we can either try to find pattern matching algorithms that, on average, perform quickly or try to find approximate algorithms that produce acceptable answers in an acceptable amount of time.

A way to optimize the matching process, thus reducing its complexity of some order of magnitude, is based on the decomposition of the generalized description G as follows:

$$G = G1 \wedge G2$$

where $G1 = Sel_1 \wedge Sel_2 \wedge \ldots \wedge Sel_i \wedge \ldots \wedge Sel_n$ is a conjunction of selectors such that the referee of Sel_i contains the maximum nonnull number of variables not appearing in the referees of $Sel_1, Sel_2, \ldots, Sel_{i-1}$;$G2$ is the conjunction of the remaining selectors of G.

Such a sort rearranges the selectors of G with the aim of minimizing the number of selectors in $G1$ to optimize the backtracking process. Afterwards, we will exploit such an optimization criterion in order to define some heuristic that produces an acceptable answer in an acceptable amount of time.

III. A Distance Measure Based on Flexible Matching

The central idea of this correspondence is that a distance measure between two formulas can be defined if a probability distribution on the results of a matching process is given. Below, we first extend the definition of pattern matching and then give a definition of distance measure according to our notation.

Definition 1: Let $\mathcal{S}$ denote the space of VL_{21} wff's, and let Match represent the canonical matching predicate defined on $\mathcal{S}$:

$$\text{Match} : \mathcal{S} \times \mathcal{S} \rightarrow \{false, true\}.$$

Then, it is known as *flexible matching* any function:

$$\text{Flex\_Match} : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$$

such that

$$\forall F1, F2 \in \mathcal{S}$$
$$\text{Flex\_Match}(F1, F2) = 1 \Leftrightarrow \text{Match}(F1, F2) = \text{true}$$
$$\text{Flex\_Match}(F1, F2) \in [0, 1) \Leftrightarrow \text{Match}(F1, F2) = \text{false}.$$

Generally, the value taken by a flexible matching function is a number indicating how well two descriptions match, i.e., the degree of similarity between two wff's in $\mathcal{S}$. Therefore, the definition of the flexible matching function should draw its inspiration from a theory that is able to quantify the degree of similarity between two objects. Because probability theory fulfills such requirements, we can assign to each pair of wff's in $\mathcal{S}$ the probability of precisely matching the two formulas, provided that a change is possibly made in the description $F2$. Consequently, we can define

$$\text{Flex\_Match}(F1, F2) = P(\text{Match}(F1, F2)). \quad (1)$$

Such a definition marks the transition from syntactic to probabilistic matching. Henceforth, in order to distinguish this particular function

from any flexible matching function, we will denote it by the term *measure of fitness* (MF).

Definition 2: Let $F1$ and $F2$ be two wff's; then, the *distance measure* Δ between $F1$ and $F2$, $\Delta(F1, F2)$, is

$$\Delta(F1, F2) = 1 - P(\text{Match}(F1, F2)) = 1 - MF(F1, F2). \quad (2)$$

In the following, attention will be devoted to the problem of concept recognition and classification; therefore, $F1$ and $F2$ will be denoted with G and Ex, respectively, where G stands for generalization and Ex for observation. Moreover, it is assumed that G is a disjunctive VL_{21} wff

$$G = \text{Or\_atom}_1 \vee \ldots \vee \text{Or\_atom}_n (n > 0)$$

and Ex is a conjunctive VL_{21} wff. The different form of G and Ex is justified because generally, in inductive learning, an observed event $Ex = Ex_1 \vee Ex_2 \vee \ldots \vee Ex_n$ can be considered to be the equivalent of a set of n different events Ex_i, where $i = 1, 2, \ldots n$. Under these assumptions, the distance measure may be rewritten as

$$\Delta(G, E_x) = 1 - MF(G, Ex) \quad (3)$$

where

$MF(G, Ex) = P(\text{Match}(G, Ex)) = P(\text{Match}(\text{Or\_atom}_1 \vee \ldots \vee \text{Or\_atom}_n), Ex) = = P(\text{Match}(\text{Or\_atom}_1, Ex) \vee \text{Match}(\text{Or\_atom}_2, Ex) \vee \ldots \vee \text{Match}(\text{Or\_atom}_n, Ex)) =$

and by replacing the symbol $\text{Match}(\text{Or\_atom}_i, Ex)$ with E_i, we have

$$= \sum_{i=1}^{n} P(E_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(E_i \wedge E_j) + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} P(E_i \wedge E_j \wedge E_k) + \ldots + (-1)^{n-1} \cdot P(E_1 \wedge E_2 \wedge \ldots E_n).$$

We assume here that all the elementary events E_i are mutually independent, i.e.

$$P(E_i | E_j) = P(E_i) \quad \forall i \neq j$$
$$P(E_i | E_j \wedge E_k) = P(E_i) \quad \forall i \neq j \neq k$$
$$\ldots$$

since the consideration of high-order joint probabilities is generally impractical. In such a case, it follows that

$$P(E_i \wedge E_j) = P(E_i)P(E_j) \quad \forall i \neq j$$
$$P(E_i \wedge E_j \wedge E_k) = P(E_i)P(E_j)P(E_k) \quad \forall i \neq j \neq k$$
$$\ldots$$

There could be two clear exceptions to this assumption. The first one arises when Or_atom_i is a specialization of another Or_atom_j within the same generalization rule G. In this case, we could have

$$P(\text{Match}(\text{Or\_atom}_i, Ex) | \sim \text{Match}(\text{Or\_atom}_j, Ex)) = 0.$$

However, we do not expect such a specialization to occur within the same generalization rule; otherwise, one of the or-atoms would be redundant.

The second exception arises when $E_i \wedge E_j = \varnothing$. In addition, in this case, we could have $P(E_i|E_j) = P(E_j|E_i) = 0$. However, mutual exclusion of the or-atoms of the same generalization is very unusual when dealing with structural descriptions. For instance, given the following generalization G

G : $[\text{ontop}(x1, x2) = \text{true}][\text{touch}(x1, x2) = \text{false}] \vee [\text{ontop}(x1, x2) = \text{false}][\text{touch}(x1, x2) = true]$

its or-atoms would seem disjunctive, but a possible example might be Ex : $[\text{ontop}(x1, x2) = \text{true}]$ $[\text{touch}(x1, x2) = \text{false}]$ $[\text{ontop}(x2, x3) = \text{true}][\text{ontop}(x3, x4) = \text{false}][\text{touch}(x3, x4) = \text{true}]$ which is covered by both or-atoms.

In other words, when no auxiliary information is provided, the independence of the events E_i seems a reasonable assumption, which simplifies the problem.

Finally, we have

$$\begin{aligned} MF(G, Ex) &= \sum_{i=1}^{n} P(E_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(E_i) \cdot P(E_j) \\ &+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} P(E_i) \cdot P(E_j) \cdot P(E_k) + \cdots \\ &+ (-1)^{n-1} \cdot \prod_{i=1}^{n} P(E_i) \\ &= \sum_{i=1}^{n} MF(\text{Or\_atom}_i, Ex) \\ &- \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} MF(\text{Or\_atom}_i, Ex) \cdot MF(\text{Or\_atom}_j, Ex) \\ &+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} MF(\text{Or\_atom}_i, Ex) \cdot \\ &MF(\text{Or\_atom}_j, Ex) \cdot MF(\text{Or\_atom}_k, Ex) + \cdots \\ &+ (-1)^{n-1} \cdot \prod_{i=1}^{n} MF(\text{Or\_atom}_i, Ex). \end{aligned} \quad (4)$$

In some situations, when the computation of formula (4) is computationally expensive, $MF(G, Ex)$ may be approximated by its lower limit, that is, the maximum value of $MF(\text{Or\_atom}_i, Ex), i \in \{1, 2, \ldots, n\}$.

At this stage, the problem has been reduced to the definition of a measure of fitness between conjunctive formulas, Or_atom_i and Ex, in terms of a matching function.

IV. A Measure of Fitness Between Conjunctive Formulas

Without loss of generality, let us suppose G is a conjunctive formula ($n = 1$) composed of $k \geq 1$ selectors $G = Sel_1 \wedge Sel_2 \wedge \ldots \wedge Sel_k$. Then, according to the definition of matching between two formulas, we have

$$\begin{aligned} \text{Match}(G, Ex) &= \text{Match}(Sel_1, Ex) \\ &\wedge \text{Match}(Sel_2, Ex) \wedge \ldots \wedge \text{Match}(Sel_k, Ex). \end{aligned} \quad (5)$$

It is easy to verify that the events $\text{Match}(Sel_i, Ex), i = 1, 2, \ldots, k$ are not independent because of the presence of the variables as arguments of a descriptor. The measurement of the fitness between two descriptions depends on two points: first, on the number of selectors that are unifiable and, second, on the possibility of determining a consistent binding for other selectors [15], [20].

Bearing in mind that G is decomposable into two parts $G1$ and $G2$ and supposing that the number of selectors in $G1$ is p, $0 < p \leq k$, then the measure of fitness

$$\begin{aligned} MF(G, Ex) &= P(\text{Match}(Sel_1, Ex) \wedge \\ &\text{Match}(Sel_2, Ex) \wedge \ldots \wedge \text{Match}(Sel_k, Ex)) \end{aligned}$$

can be rewritten as

$$\begin{aligned} MF(G, Ex) &= P(\text{Match}(G1, Ex) \wedge \\ &\text{Match}(Sel_{p+1}, Ex) \wedge \ldots \wedge \text{Match}(Sel_k, Ex)). \end{aligned}$$

Let us consider the event Match$(G1, Ex)$. It has been defined using the relationship Match, which is a partial order relation on the space of VL_{21} connected conjunctive descriptions. If we suppose that Match$(G1, Ex)$ is the event with probability equal to one, i.e., there exists at least one matching substitution σ_j between $G1$ and Ex, then the result is

$$Ex \Rightarrow \sigma_j(G1)$$

and

$$\begin{aligned} MF(G, Ex) &= P(\text{Match}(G1, Ex) \wedge \\ &\text{Match}(\text{Sel}_{p+1}, Ex) \wedge \ldots \wedge \text{Match}(\text{Sel}_k, Ex)) \\ &= \max_{\sigma_j} \prod_{i=p+1}^{k} P(\text{Match}_j(\text{Sel}_i, Ex)) \\ &= \max_{\sigma_j} \prod_{i=p+1}^{k} MF_j(\text{Sel}_i, Ex). \end{aligned} \tag{6}$$

Formula (6) must be interpreted as follows: While varying the considered operator σ_j, which is responsible for the consistent binding of the variables in $G1$, the measure of fitness between G and Ex is computed as the highest value given by the multiplication of the measures of fitness between each single selector of $G2$ and Ex.

Here, Match$_j$ denotes the relation Match when the substitutions fixed by the operator σ_j have been carried out, whereas MF_j denotes the measure of fitness between one selector and one formula when all the variables in G have been replaced according to the operator σ_j.

A proof of (6) through mathematical induction on k is given in Appendix A.

Until now, we have considered $P(\text{Match}(G1, Ex)) = 1$ as the fundamental hypothesis. If Match$(G1, Ex)$ is not satisfied, we can set $MF(G, Ex) = 0$ since G and Ex have no similarities, not even at a level of the subcomponents. Thus, (6) becomes

$$MF(G, Ex) = \begin{cases} \max_{\sigma_j} \prod_{i=p+1}^{k} MF_j(\text{Sel}_i, Ex) & \text{if there exists at least a substitution} \\ & \sigma_j \text{ such that } Ex \Rightarrow \sigma_j(G1) \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The constraint that $G1$ matches Ex by a canonical matching procedure could be interpreted as follows: There must be at least some correspondences between G and Ex, that is, $G1$ is a conjunction of *Must-relations* [21]. Unfortunately, the elimination of the above constraint causes the computational time to soar.

V. A Measure of Fitness Between Selectors

$MF_j(Sel_i, Ex)$ is determined by considering the fitness between the selector $\sigma_j(Sel_i) = GSel_i$ (derived from the substitution of each variable in the referee of Sel_i with the corresponding variable fixed by σ_j) and only one selector of Ex, $ExSel_i$, which has the same referee as $GSel_i$. Consequently

$$\begin{aligned} MF_j(Sel_i, Ex) &= MF(GSel_i, ExSel_i) \\ &= MF(\text{reference}(GSel_i), \text{reference}(ExSel_i)). \end{aligned} \tag{8}$$

Let us suppose that reference$(GSel_i) = \{g_1, g_2, \ldots, g_y\}$ and that reference$(ExSel_i) = \{e_1, e_2, \ldots, e_w\}$. It should be pointed out that the reference of $GSel_i$ can contain more than one element $(y \geq 1)$ since G is the description of a concept, whereas the reference of $ExSel_i$ usually contains only one element $(w = 1)$ since Ex describes an observation. Indeed, a multiple-value reference for $ExSel_i$ denotes uncertainty in the measurement process, and it should be dealt in a different way. From condition 3) in Section II-B and the definition of flexible matching, we can state that $MF(GSel_i, ExSel_i)$ is equal to 1 if and only if the reference of $ExSel_i$ is more specific than that of $GSel_i$. The notion of specialization is intended as set inclusion if the descriptor is a nominal or linear one. This interpretation can be easily extended to tree-structured descriptors; each single element in the reference of $GSel_i$ is replaced by all the values representing the leaves of the subtree having just that element as its root. When the set inclusion does not hold, the definition of $MF(GSel_i, ExSel_i)$ takes into account the probability that the value in the reference of $ExSel_i$ becomes equal to one of the y values in the reference of $GSel_i$.

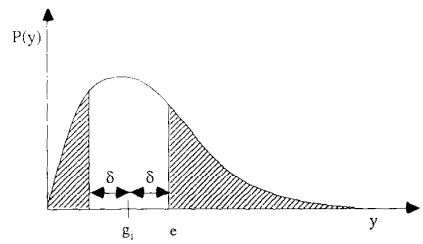


Fig. 1. In a domain whose pdf of its values is represented by $P(v)$, the probability $P(\text{EQUAL}(g_i, e))$ equals the shaded areas.

Let $EQUAL(x, y)$ denote the matching predicate defined on any two values x and y of the same domain. According to our criteria of considering the most probable match for the computation of the measure of fitness between G and Ex, we can write the following formula:

$$MF(GSel_i, ExSel_i) = \max_{i \in [1,y]} P(\text{EQUAL}(g_i, e)) \tag{9}$$

where e is the only element in the reference of $ExSel_i$ when there is a reasonable certainty about the value taken by the descriptor in $ExSel_i$.

To sum up, the definition of a measure of fitness between G and Ex has been reduced to the computation of the probability of the event EQUAL(g_i, e). It represents the probability that an observation e may be considered a distortion of g_i; thus, we define

$$P(EQUAL(g_i, e)) = P(\delta(g_i, X) \geq \delta(g_i, e)) \tag{10}$$

where

- X is a random variable assuming values in the domain D_f of the descriptor f in Sel_i
- δ is a distance defined on the domain itself.

In Fig. 1, a geometrical interpretation of this definition is shown.

When no information is available on the probability distribution of X, we can assume that each value from the domain of the descriptor in Sel_i has the same probability, that is, $1/C$, where C is the number of elements of D_f.

The definition of δ has to be specialized according to the type of VL_{21} descriptors. In particular, we propose for nominal descriptors

$$\delta(x, y) = \begin{cases} 0 \text{ if } x = y \\ 1 \text{ otherwise} \end{cases} \tag{11}$$

and for linear descriptors with a finite domain

$$\delta(x, y) = |ord(x) - ord(y)| \tag{12}$$

where $ord(x)$ denotes the ordinal number given to the value x of the domain D_f.

In fact, if the referee of Sel_i is a linear descriptor with a totally ordered domain $D_f = \{y_0, y_1, \ldots, y_{C-1}\}$, it is always possible that consecutive elements of D_f are assigned consecutive integers (starting from 0 or from any other); for instance, $ord(y_0) = 0, ord(y_1) = 1, \ldots, ord(y_{C-1}) = C - 1$.

It should be observed that other reasonable choices of δ are possible; nevertheless, the value of $P(\text{EQUAL}(g_i, e))$ does not change since we compute the probability over the distance and not merely the geometrical distance. This key point also allows us to disregard problems with scaling when the similarity is computed over the whole set of features.

The specialization of δ has repercussions on the definition of $P(\text{EQUAL}(g_i, e))$, which is adapted to nominal, linear, and tree-structured descriptors.

For nominal descriptors, it is defined by

$$P(\text{EQUAL}(g_i, e)) = \begin{cases} 1 & \text{if } g_i = e \\ (C-1)/C & \text{otherwise.} \end{cases} \tag{13}$$

For linear descriptors it becomes as shown in (14), at the bottom of the page, where

$$\text{step}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

(derivations of (13) and (14) are given in Appendix B).

For tree-structured descriptors, each element in the reference of Sel_i is replaced by the values representing the leaves of the subtrees that have that element as their root. Then, (13) or (14) are adopted, depending on whether the generalization hierarchy for the descriptor is unordered or ordered, respectively. The only change to be made both in (13) and in (14) consists of replacing C with the cardinality of LEAVES(D_f), where LEAVES(D_f) represents the set composed of all the leaves of the tree representing the domain of the tree-structured descriptor in Sel_i.

The distance measure may be easily extended taking into account other kinds of descriptors or different probability distributions of the values. An example might be the descriptor *number_of_calls* (whose domain is the set of nonnegative integers) counting the number of calls occurring in a unit time and for which a Poisson distribution appears more appropriate.

Using the definitions (8), (9), (13), and (14), (7) can be rewritten as

$$MF(G, Ex) = \begin{cases} \max\limits_{\sigma_j} \prod\limits_{i=1}^{k} MF_j(\text{Sel}_i, Ex) & \text{if there exists at least a substitution } \sigma_j \\ & \text{such that } Ex \Rightarrow \sigma_j(G1) \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Another possible extension could take into account the weights w_i associated with all the descriptors. They should range in the interval [0,1] and might represent users preferences or descriptor relevances. Then, the previous formula becomes

$$MF(G, Ex) = \begin{cases} \max\limits_{\sigma_j} \prod\limits_{i=1}^{k} w_i \cdot MF_j(\text{Sel}_i, Ex) & \text{if there exists at least a substitution } \sigma_j \\ & \text{such that } Ex \Rightarrow \sigma_j(G1) \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

which is an extension of the definition of the measure of fitness in the case where weights of descriptors are available.

VI. Further Remarks on the Applicability of the Measure of Fitness

In the previous sections, we have defined a measure of fitness based essentially on the probability that a single selector of an example Ex matches with a selector of a recognition rule G. The measure of fitness $MF(G, Ex)$ actually computes the probability that a new event Ex may come from the class described by G : $P(Ex|G)$ (within-class probability). In fact, the $MF(GSel_i, ExSel_i)$ defined for single selectors may be interpreted as the probability that a random variable X defined on the domain of the descriptor in $ExSel_i$ takes a value x_i farther than the reference of $ExSel_i$ from the reference of $GSel_i$, given that $GSel_i$ is the centroid. Therefore, since the definition of $MF(G, Ex)$ is essentially based on the measure of fitness for single selectors, we can state that it computes the probability that any observation of the concept described by G would be as far from the centroid G as the case Ex being considered. If this value is small, it signifies the possibility that Ex does not belong to the concept G, even though it is the "closest."

As pointed out by Shapiro and Haralick [22], the definition of a distance measure between structural descriptions allows exploitation of the Bayesian decision framework. Thus, it is possible to compute

$$P(G_i|Ex) = \frac{P(Ex|G_i)P(G_i)}{P(Ex)} = \frac{P(Ex|G_i)P(G_i)}{\sum\limits_j P(Ex|G_j)P(G_j)}$$

which is the *a posteriori* probability used in Bayesian classification. When there is little or no knowledge available on the population distribution, empirical Bayes methods are applicable, making some choices for the *a priori* probabilities $P(Gi)$ [23]. Another subjective choice may be a threshold t for the maximum value of $P(Ex|G_i) = MF(G_i, Ex), i = 1, 2, ..., K$ (where K is the number of concepts recognized), below which, an observation Ex is rejected because its similarity with any concept is too low.

Classification based on a distance measure is more expensive than a canonical match/unmatch procedure; therefore, a multilayered framework is preferable. At first, a canonical matching is applied in order to classify an observation Ex. There are three possible outcomes:

a) *Single Match*: No further processing is required; the observation is assigned to the class represented by the matching recognition rule G.
b) *No Match*: The definition of measure of fitness is exploited in order to have a more flexible matching.
c) *Multiple Match*: In this case, the decision is taken in favor of the matching rule G, which minimizes the sum of weights associated with the descriptors of the selectors belonging to $Ex - G$, where $Ex - G$ indicates the part of Ex obtained

$$P(\text{EQUAL}(g_i, e)) = \begin{cases} [1 + \text{ord}(e) + (C - 2\text{ord}(g_i) + \text{ord}(e)) \cdot \text{step}(C - 1 - 2\text{ord}(g_i) + \text{ord}(e))]/C & \text{if } g_i > e \\ 1 & \text{if } g_i = e \\ [C - \text{ord}(e) + (2\text{ord}(g_i) - \text{ord}(e) + 1) \cdot \text{step}(2\text{ord}(g_i) - \text{ord}(e))]/C & \text{if } g_i < e \end{cases} \tag{14}$$

ignoring those selectors that are unifiable with the matching rule G.

Besides classification, the distance measure can also play an important role in the selection of hypotheses generated by an inductive generalization process. For instance, in those methodologies based on event covering, such as STAR [14], it is possible to define for each class described by G_i a measure of separation from other classes as follows:

$$\alpha_i = 1/n_i \sum_{j=1}^{n_i} MF(G_i, Ex_{ij}) - 1/(K-1) \cdot \sum_{l=1}^{K} 1/n_l \sum_{j=1}^{n_l} MF(G_i, Ex_{lj}) \quad (18)$$

where n_i is the number of examples for class i, Ex_{ij} is the jth example from class i, and K is the number of classes.

It is easy to verify that α_i ranges in $[-1,1]$. The value 1 is assumed when each example from class i matches with G_i and when the matching condition on $G_i 1$ for all the counterexamples does not hold. At the other extreme, the value -1 is assumed when all the examples from class i have the maximum distance from G_i and when all the counterexamples match exactly with G_i.

Since α_i takes into account both the completeness and consistency conditions and since the measure of fitness computed according to (17) includes the weight of the descriptors, it would be interesting to introduce α_i as a basic criterion in the lexicographic evaluation function (LEF). The LEF is used both in the trimming of a partial star and in the selection of the best consistent hypothesis generated for a class i. It is a global preference criterion of the generalization rules, expressed as a list of elementary preference criteria, such as consistency, completeness, or cost.

Moreover, a new definition of consistency can be provided when a threshold t for the measure of fitness $MF(G, Ex)$ is fixed by the user; it is the *t-weak consistency*, defined as $MF(G_i, Ex) < t$ for all the counterexamples Ex. By analogy, it is possible to define a *t-weak completeness*, defined as $MF(G_i, Ex) > t$ for all the positive examples. Some advantages arising from weaker definitions of consistency and completeness are both a higher noise immunity in the process of inductive generalization and a reduction of complexity of the rule base [5], [6]. An optimal value of t might be one that leads to the best tradeoff between the accuracy and the complexity of a rule base.

VII. The Problem of Completing the Attribute and Structure Spaces

Up to now, we dealt with the problem of information containing errors due to noise, and we deliberately disregarded an allied problem: incomplete information. The description of an object may be incomplete owing to a variety of reasons:

1) The value of a descriptor is unknown because it was not possible to "measure" it (*unknown* value).
2) It does not make sense to set a value for the descriptor (*meaningless* value).
3) It does not matter at all ("*don't care*" value).

Henceforth, we will denote unknown values with "?," meaningless values with "NA" (Not Applicable), and "don't care" values with "*" (i.e., all possible values of the domain).

Generally speaking, although "*" is allowed to appear in each VL_{21} wff, its usage in an example contradicts the concept that such an example is an observation of a class. Indeed an "example" with a "*" would really be a generalization in itself. On the contrary, meaningless values are not desirable for recognition rules because a description of a class based on meaningful properties is usually sought. The latter reflections can also be applied to the unknown values.

Of course, the measure of fitness is strongly influenced by the above taxonomy. For instance, let us suppose the following:

Domain(color) = {white, grey, blue, black}
G_1Sel_i = [color($s1$)=white]
Ex_1Sel_i = [color($x1$)=*]
Ex_2Sel_i = [color($x1$)=?]
Ex_3Sel_i = [color($x1$)=NA]
Ex_4Sel_i = [color($x1$)=grey].

Then

a) $MF(G_1Sel_i, Ex_1Sel_i)$ does not make sense
b) $MF(Ex_1Sel_i, G_1Sel_i) = 1$ because "*" ≡{white, grey, blue, black} ⊃{white}
c) $MF(G_1Sel_i, Ex_3Sel_i) = 0$ because the attribute color must be meaningful for the class
d) $MF(G_1Sel_i, Ex_4Sel_i) = 3/4$ according to formula (13)

but what about the value of $MF(G_1Sel_i, Ex_2Sel_i)$?

This is the case in which the attribute color is not specified for $x1$ when $\sigma = \{s1 \leftarrow x1\}$ is a matching substitution for G and Ex.

In order to answer this question, let us suppose

$$D = \text{Domain}(\text{referee}(ExSel_i)) = \{y_0, y_1, \ldots, y_{C-1}\}$$

then

$$\begin{aligned} P(\text{EQUAL}(g,?)) &= P(\delta(g,X) \geq \delta(g,Y) \cap Y \in D) \\ &= \sum_{i=0}^{C-1} P(\delta(g,X) \geq \delta(g,Y) | Y = y_i) P(Y = y_i) \\ &= \sum_{i=0}^{C-1} P(\delta(g,X) \geq \delta(g,y_i)) P(y_i). \end{aligned} \quad (19)$$

As $P(\text{EQUAL}(g, y_i))$ specializes according to the type of descriptor, we can conclude that for nominal descriptors

$$\begin{aligned} P(\text{EQUAL}(g,?)) &= 1/C \sum_{i=0}^{C-1} P(\delta(g,X) \geq \delta(g,y_i)) \\ &= [(C-1)(C-1)/C + 1]/C = (C^2 - C + 1)/C^2 \end{aligned} \quad (20)$$

according to (13). For linear descriptors with a finite domain

$$\begin{aligned} P(\text{EQUAL}(g,?)) = \\ 1/C \sum_{i=0}^{C-1} \{ &\text{step}(g-e) \cdot \text{step}(e-g) + (1 - \text{step}(e-g)) \\ &\cdot [(C - 2g + e) \cdot \text{step}(C - 1 - 2g + i) + e + 1]/C \\ &+ (1 - \text{step}(g-e)) \cdot [(2g - e + 1) \\ &\cdot \text{step}(2g - e) + C - e]/C \} \end{aligned} \quad (21)$$

where, for simplicity, we have denoted $ord(g)$ and $ord(i)$ with g and e, respectively.

Some interesting properties may be easily derived from this.

Property 1: For each nominal domain D whose values have the same probability, $P(\text{EQUAL}(g,?))$ does not depend on g, and for each $y \in D - \{g\}$:

$$P(\text{EQUAL}(g,y)) < P(\text{EQUAL}(g,?)) < P(\text{EQUAL}(g,g)).$$

Property 2: For each finite linear domain D whose values have the same probability, there are two values y_s and y_t, with $s < t$, so that for each $y \in \{y_s, y_{s+1}, y_{s+2}, \ldots, y_t\}$:

$$P(\text{EQUAL}(g,?)) \leq P(\text{EQUAL}(g,y)).$$

The first property suggests that

$$MF(G_1Sel_i, Ex_4Sel_i) \leq MF(G_1Sel_i, Ex_2Sel_i)$$

which is intuitively true. In fact, if we observed that color($x1$)=grey, the fit with color($s1$)=white would take into account only the uncertainty due to noise, whereas the unknown value would take into account both the probability for the value being a distortion of white and the probability of observing just white.

For a linear domain $D = \{0, 1, 2, \ldots, 9\}$, we have

$$P(\text{EQUAL}(7, ?)) = 0.57 \quad P(\text{EQUAL}(7, 5)) = 0.7$$
$$P(\text{EQUAL}(7, 9)) = 0.7 \quad P(\text{EQUAL}(7, 4)) = 0.5$$

that is, if the observed value ranges between 5 and 9, its fit to 7 is higher than the unknown value.

Other solutions to the problem of unknown attribute values have been proposed by Quinlan [5], Hand [8] and Chan *et al.* [24]. Nevertheless, their works cannot be exploited for structural descriptions because they are strictly connected to feature vector representation for which there is only one possible unification between G and Ex. Going up to a higher level, first-order predicate logic, it is no longer possible to estimate the pdf for each descriptor, and it becomes necessary for the user to define it. When he cannot provide such information, the pdf is assumed to be uniform.

Previous considerations are valid when an attribute related to a subpart of Ex is omitted in the description. What happens for relations? First of all, we should observe that completing the attribute space is generally less expensive than completing the structure space. For instance, if O is an object made up of n homogeneous primitive parts for which m attributes are definable, then a complete description of O should contain $n \cdot m$ attributional selectors. It is also possible to define a set of relations on each n-tuple of subparts, for instance, r binary relations that are valid for each couple of primitive parts. Thus, the complete description of O should include $n \cdot (n-1) \cdot r$ structural selectors or even $n^2 r$ if we considered relations between a subpart and itself as well. Such an example clearly illustrates the complexity and undesirability of descriptions with a complete structure space.

Consequently, one prefers to represent only those relations that really hold, for instance, [ontop($x1, x2$) = true], rather than [ontop($x1, x3$) = false]. Therefore, if $GSel_i$ = [ontop($s1, s2$)] and cannot be unified with a selector from Ex, we would suppose the existence of a matching selector $ExSel_i$ = [ontop($x3, x5$) = false], provided that unifying $G1$ with Ex, the following correspondence is fixed:

$$\sigma = \{s1 \leftarrow x3, s2 \leftarrow x5\}.$$

Here, the problem of distinguishing an unholding relation from an unknown one arises. A solution might be just the opposite of that adopted for attributes: unknown relations must be specified (for instance, [ontop($x3, x5$) =?]), and the computation of the measure of fitness (in this case MF([ontop($s1, s2$)], [ontop($x3, x5$) =?])) should be done by using the same formulas defined for the unknown attribute values.

Nevertheless, relations cannot be expressed by predicates only, as is shown by the following relation:

$$[\text{alignment}(block1, block2) = \text{both\_rows}]$$

and therefore, it should be specified to be a default value in the domains (for instance, no_alignment), which should be considered every time a relation is not specified.

Finally, another point might concern the application of inference rules to flexible matching. Again, let us consider the following descriptions:

$$G : [\text{ontop}(s1, s2)][\text{size}(s2) = \text{large}]$$
$$Ex : [\text{ontop}(x1, x2)][\text{size}(x2) = \text{small}].$$

Then

$$MF(G, Ex) = MF([\text{size}(s2) = \text{large}], [\text{size}(x2) = ?]).$$

However, if we hypothesized the validity of the following inference rule

$$[\text{ontop}(t1, t2)] \Rightarrow [\text{size}(t1) \leq \text{size}(t2)]$$

i.e., the upper part is never greater than the lower part, an important question arises: How would this influence the computation of a distance measure? Future research may address this open problem.

VIII. Application to Office Automation

The measure of fitness has been implemented and tested on digitized office document classification [25]. By a document, we mean a related collection of printed objects (characters, columns, paragraphs, titles, figures, etc.) on a paper or microform, for example, technical journals or reports. Here, only single-page documents will be considered. Provided there is a set of documents with common page layout features, an optically scanned document can be classified in the early phase of its processing flow by using a defined set of relevant and invariant layout characteristics: the *page layout signature*. As a human is generally able to classify any document in a specific environment by a perceptive point of view, recognizing the structure of a form or reading only the content of particular parts of the document, it is also possible to classify digitized documents without using optical character recognition or syntactic descriptions of the document given by the user. In fact, a printed page is treated by dealing only with automatically detected and constructed characteristics of the document, namely, the geometrical characteristics of the blocks (height, width, spacing, and alignment), and the document structure, whose description is created in a symbolic notation.

In order to produce the classification rules, some significant examples of document classes of interest in a specific environment are used as training samples to determine the layout similarities within each class. A set of 75 single-page documents has been considered, namely, printed letters and magazine indexes, belonging to eight different classes (the last one is a reject class, representing "the rest of the world"). As training examples, 40 instances were selected (five for each class), leaving the remaining 35 documents for the testing process. All the sample documents are real letters received or sent by some firms or copies of the indexes of international magazines, therefore, several forms of noise actually affected them.

Once a document has been digitized, its page layout is produced by segmenting it through a run length smoothing algorithm (RLSA) and by grouping together some segments (or blocks), thus satisfying some predetermined requirements such as closeness, equal width, same type, etc. An example of document page layout is shown in Fig. 2. The numerical output of the layout analysis is automatically translated into VL_{21} descriptions, whose standard descriptors are listed in Table I. The third processing step concerns the classification of the document by matching its description with the generalizations of the document classes produced by the system INDUBI, which was inspired by Michalski's INDUCE [14]. The recognition rules for each class are listed in Table II. Of course, as noisy documents are handled as well, it is not possible to use a strict matching procedure for classifying the test documents; therefore, the proposed distance measure defined for VL_{21} descriptions was introduced for a flexible matching. When a strict matching does not allow for the singling

TABLE I
PAGE LAYOUT DESCRIPTORS

Function symbol	Domain
WIDTH(Block)	linear domain: [very_small, small, medium_small, medium, medium_large, large, very_large]
HEIGHT(Block)	linear domain: [smallest, very_very_small, very_small, small, medium_small, medium, medium_large, large, very_large, very_very_large, greatest]
TYPE(Block)	nominal domain: [text, hor_line, picture, ver_line, graphic, mixture]
CONTAIN_IN_POS(Doc,Block)	denotes the relative position of the region Block within Doc: [east, north_east, north, north_west, west, south_west, south, south_east, centre]
ONTOP(Block1,Block2)	**true** if Block1 is above Block2.
TORIGHT(Block1,Block2)	**true** if Block1 is on the right of Block2.
ALIGN(Block1,Block2)	mutual alignment between Block1 and Block2: [no_align, starting_col, ending_col, middle_col, starting_row, ending_row, middle_row]

out of a membership class, one makes use of the flexible matching in order to compute the measure of fitness for each concept. If all the measures of fitness are not greater than a fixed threshold (0.85), the document is rejected; otherwise, it is classified into the document class with the highest *a posteriori* probability $P(G_i|Ex)$.

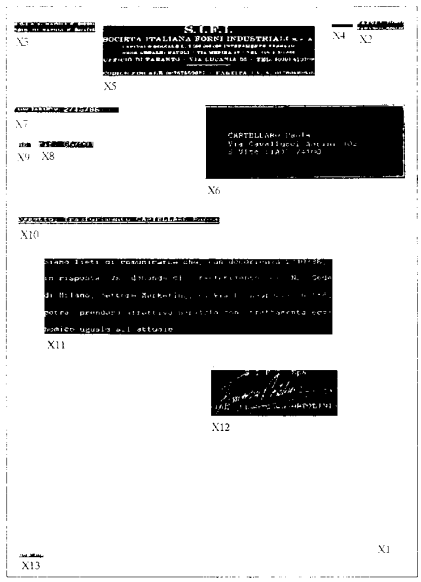
Table III shows the correct classification for the testing sample together with the results of the application of both the strict matching procedure and the multilayered framework. Moreover, the values of the highest measures of fitness are reported. From this table, it emerges that the canonical matching procedure classifies only 23 documents correctly, rejects nine documents, and presents a double classification in three cases. By adopting the multilayered framework for classification, seven no-match cases are recovered through the computation of the distance measure.

In Fig. 2, the page layout of a nonstrictly-matching document is displayed, as well as its VL_{21} symbolic description. This document is correctly classified when its measure of fitness with each generated class description is computed.[1] In the following, we report an example of computation of the distance measure between the description (Ex) of such a document and the description (G) of the correct class, i.e., SIFI letters. At first, we should observe that the "$G1$ part" of that rule is given by the first three selectors:

$$G1 = [\text{TORIGHT}(S1, S2)][\text{ONTOP}(S3, S4)] \\ [\text{ALIGN}(S5, S6) = \text{starting\_col.}]$$

There are 17 possible substitutions σ_j such that $Ex \Rightarrow \sigma_j(G1)$. Thus, the value of $MF(G, Ex)$ is given by the maximum of the MF_j values computed according to the different substitutions. In Table IV, we have reported all of the 17 substitutions and the corresponding values of the measure of fitness. The highest of these values is obtained in correspondence with the substitution $\sigma_{12} = \{S1 \leftarrow X4, S2 \leftarrow X2, S3 \leftarrow X5, S4 \leftarrow X6, S5 \leftarrow X3, S6 \leftarrow X7\}$.

[1] Here, we are assuming the same *a priori* probability in each class. Thus, the maximum $P(G_i|Ex)$ corresponds to the maximum value of $MF(G_i, Ex)$.



WIDTH(X2) = medium_small
WIDTH(X3) = medium
WIDTH(X4) = small
WIDTH(X5) = large
WIDTH(X6) = large
WIDTH(X7) = medium_large
WIDTH(X8) = medium_small
WIDTH(X9) = small
WIDTH(X10) = large
WIDTH(X11) = very_large
WIDTH(X12) = medium_large
WIDTH(X13) = medium_small

HEIGHT(X2) = very_very_small
HEIGHT(X3) = very_very_small
HEIGHT(X4) = smallest
HEIGHT(X5) = medium_small
HEIGHT(X6) = medium
HEIGHT(X7) = very_very_small
HEIGHT(X8) = very_very_small
HEIGHT(X9) = very_very_small
HEIGHT(X10) = very_very_small
HEIGHT(X11) = medium
HEIGHT(X12) = medium_small
HEIGHT(X13) = smallest

TYPE(X2) = text
TYPE(X3) = text
TYPE(X4) = text
TYPE(X5) = mixture
TYPE(X6) = mixture
TYPE(X7) = text
TYPE(X8) = text
TYPE(X9) = text
TYPE(X10) = mixture
TYPE(X11) = text
TYPE(X12) = graphic
TYPE(X13) = text

CONTAIN_IN_POS(X1,X2) = north_east
CONTAIN_IN_POS(X1,X3) = north_west
CONTAIN_IN_POS(X1,X4) = north_east
CONTAIN_IN_POS(X1,X5) = north
CONTAIN_IN_POS(X1,X6) = north_east
CONTAIN_IN_POS(X1,X7) = north_west
CONTAIN_IN_POS(X1,X8) = north_west
CONTAIN_IN_POS(X1,X9) = north_west
CONTAIN_IN_POS(X1,X10) = west
CONTAIN_IN_POS(X1,X11) = centre
CONTAIN_IN_POS(X1,X12) = centre
CONTAIN_IN_POS(X1,X13) = south_west

ONTOP(X5,X6) = true
ONTOP(X7,X8) = true
ONTOP(X7,X9) = true

TORIGHT(X3,X5) = true
TORIGHT(X9,X8) = true
TORIGHT(X5,X4) = true
TORIGHT(X5,X2) = true
TORIGHT(X4,X2) = true

ALIGN(X2,X6) = ending_col
ALIGN(X3,X5) = starting_row
ALIGN(X3,X2) = ending_row
ALIGN(X3,X7) = starting_col
ALIGN(X3,X8) = ending_col
ALIGN(X7,X6) = starting_row
ALIGN(X9,X8) = starting_row
ALIGN(X9,X8) = ending_row
ALIGN(X7,X8) = middle_col
ALIGN(X8,X6) = middle_row
ALIGN(X9,X10) = starting_col
ALIGN(X5,X4) = starting_row
ALIGN(X10,X13) = starting_col
ALIGN(X4,X2) = starting_row
ALIGN(X4,X2) = ending_row

Fig. 2. Example of document page layout and its VL_{21} description.

Indeed, in this case, we have

$$\begin{aligned} MF_{12}(G, Ex) = {} & MF_{12}([\text{TORIGHT}(X5, X4)], \\ & [\text{TORIGHT}(X5, X4)]) \cdot \\ & \cdot MF_{12}([\text{TORIGHT}(X3, X5)], \\ & [\text{TORIGHT}(X3, X5)]) \cdot \\ & \cdot MF_{12}([\text{TORIGHT}(X5, X2)], \\ & [\text{TORIGHT}(X5, X2)]) \cdot \\ & \cdot MF_{12}([\text{ALIGN}(X3, X4) = \text{starting\_row}], \\ & [\text{ALIGN}(X3, X4) = \text{no\_align}]) \cdot \\ & \cdot MF_{12}([\text{WIDTH}(X4) = \text{small}], \\ & [\text{WIDTH}(X4) = \text{small}]) \\ = {} & 1 \cdot 1 \cdot 1 \cdot 6/7 \cdot 1 = 6/7 \approx 0.857. \end{aligned}$$

The reason why there is no perfect match is the lack of alignment in the top row between blocks $X4$ and $X5$ of the document. Therefore, in accordance with what was stated about the problem of completing the attribute and structure spaces, we have implicitly assumed that the following relation holds: [ALIGN$(X3, X4)$ = no_align].

TABLE II
SYMBOLIC DESCRIPTIONS INDUCED FOR EACH DOCUMENT CLASS

Class name	No	Class description
Olivetti letter	1	[ALIGN(S1,S2)=starting_col][ONTOP(S3,S4)] [TORIGHT(S2,S3)][WIDTH(S1)=medium] [HEIGHT(S1)=very_small,small,medium_small,medium,medium_large]
SIET letter	2	[TORIGHT(S1,S2)][ONTOP(S3,S4)][ALIGN(S5,S6)=starting_col] [TORIGHT(S3,S1)][TORIGHT(S5,S3)][TORIGHT(S3,S2)] [ALIGN(S5,S1)=starting_row][WIDTH(S1)=small]
SIMS letter	3	[ALIGN(S1,S2)=starting_row][ONTOP(S1,S3)] [WIDTH(S2)=medium,medium_large] [HEIGHT(S2)=small,medium_small]
SOGEA letter	4	[TORIGHT(S1,S2)][CONTAIN_IN_POS(S3,S1)=north_west] [CONTAIN_IN_POS(S3,S2)=north][TYPE(S1)=mixture] [WIDTH(S1)=medium_large][WIDTH(S2)=large] [TYPE(S2)=mixture][HEIGHT(S2)=small]
IEEE Transactions on Pattern Analysis and Machine Intelligence index	5	[TORIGHT(S1,S2)][ONTOP(S3,S1)][ONTOP(S2,S4)] [ONTOP(S3,S2)][ONTOP(S1,S4)][ALIGN(S3,S1)=starting_col] [ALIGN(S1,S4)=starting_col][ALIGN(S1,S2)=middle_row] [HEIGHT(S2)=very_small]
IEEE Spectrum index	6	[ALIGN(S1,S2)=starting_row][ONTOP(S3,S4)] [ALIGN(S1,S2)=ending_row][ALIGN.S4,S1)=starting_col] [WIDTH(S1)=very_small][TYPE(S2)=text][WIDTH(S2)=medium]
IEEE Transactions on Computers index	7	[ALIGN(S1,S2)=starting_col][WIDTH(S1)=large] [HEIGHT(S2)=medium_small,medium,medium_large,large]

TABLE III
RESULTS OF THE CLASSIFICATION OF TESTING DOCUMENTS

Ex. No.	Correct Class	Predicted Class (Canonical Matching)	Predicted Class (Flexible Matching)	Highest MF value
1	1	1,7	1	1.0
2	1	-	-	0.5
3	1	1	1	1.0
4	1	1	1	1.0
5	1	1	1	1.0
6	2	-	2	0.857
7	2	2	2	1.0
8	2	-	2	0.857
9	2	2	2	1.0
10	2	2,7	2	1.0
11	2	-	2	0.857
12	3	3	3	1.0
13	3	3	3	1.0
14	3	3	3	1.0
15	3	3	3	1.0
16	3	3	3	1.0
17	4	4	4	1.0
18	4	4	4	1.0
19	4	4	4	1.0
20	4	4	4	1.0
21	4	4	4	1.0
22	5	5	5	1.0
23	5	-	7	0.909
24	5	5	5	1.0
25	5	-	-	0.779
26	5	5	5	1.0
27	5	5,7	5	1.0
28	6	6	6	1.0
29	6	6	6	1.0
30	6	6	6	1.0
31	7	-	7	0.909
32	7	7	7	1.0
33	7	-	7	0.909
34	7	7	7	1.0
35	7	-	7	0.909

The lowest MF value in Table IV is obtained with the substitutions $\sigma_6 - \sigma_{11}$ since only the selectors in $G1$ match perfectly. For instance, the measure of fitness computed according to σ_6 is given by

$$\begin{aligned} MF_6(G, Ex) = {} & MF_6([\text{TORIGHT}(X7, X5) = \text{true}], \\ & [\text{TORIGHT}(X7, X5) = \text{false}]) \cdot \\ & \cdot MF_6([\text{TORIGHT}(X9, X7) = \text{true}], \\ & [\text{TORIGHT}(X9, X7) = \text{false}]) \cdot \\ & \cdot MF_6([\text{TORIGHT}(X7, X4) = \text{true}], \\ & [\text{TORIGHT}(X7, X4) = \text{false}]) \cdot \\ & \cdot MF_6([\text{ALIGN}(X9, X5) = \text{starting\_row}], \\ & [\text{ALIGN}(X9, X5) = \text{no\_align}]) \cdot \\ & \cdot MF_6([\text{WIDTH}(X5) = \text{small}], \\ & [\text{WIDTH}(X5) = \text{large}]) \\ = {} & 1/2 \cdot 1/2 \cdot 1/2 \cdot 6/7 \cdot 2/7 \\ = {} & 3/98 \approx 0.031. \end{aligned}$$

It should be observed that the constraint on a perfect match for the only part $G1$ allowed us to cope effectively with the problem of structural deformations since the elimination of this constraint would require the computation of the measure of fitness according to more than 8 million possible substitutions.[2]

As has already been pointed out, in our experimentation, the adoption of a flexible matching allows recovery of most of the no-match cases. Nevertheless, in two cases, the value of the measure of fitness is quite low, namely, less than the fixed threshold 0.85. Thus, the documents are still rejected. Moreover, in one case, we have a misclassification because the distance of the document from the description of a wrong class is lower than the distance concerning the correct class. This disappointing result may be ascribed to the fact that the indexes of both IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON COMPUTERS have a similar layout; thus, a simple symbolic approach to inductive generalization and classification does not suffice. When more training samples are available, it is possible to integrate parametric (numerical) and conceptual (symbolic) learning methods in order to produce better classification rules to improve the classification rate [25]. In fact, the parametrical method works on the numerical output of the segmentation process rather than on the symbolic page layout description. Moreover, its result is mapped into a metadescriptor that is subsequently considered by the conceptual learning method while generating the document classification rules.

The very simplicity of the rule generated for the class of IEEE TRANSACTIONS ON COMPUTERS is also the main cause of some cases of multiple match. Once again, the use of the distance measure can help to eliminate ambiguity in these situations; this is done by evaluating the differences between the nonmatching parts of the document description Ex and class descriptions G_i, as explained in Section VI.

Concerning the system performance, we observed the following average time: about 6 s for generating the symbolic description of a document, about 3 s for matching and classifying a document, and less than 1 min for the complete processing of a single document (including the scanning process). The time needed for training the system was about 15 min, and of course, it is strongly dependent on the number of training examples and their complexity. At present, the learning step in INDUBI is accomplished according to the classical STAR algorithm, even though the introduction of the distance measure in the selection of hypotheses generated during the generalization process is at a project stage. The whole system has been implemented in C on a SUN3/280 under UNIX, even if the final classification expert module runs alone on an OLIVETTI PC M280 under MS-DOS.

IX. COMPARISON WITH PREVIOUS WORKS

A variety of distance measures have been proposed in previous works, both from the field of pattern recognition [3], [4] and from the area of machine learning [6], [9], [12]. Basically, they can be grouped into two categories according to the representation language:

[2] If p and q, $p \leq q$, are the number of variables in G and Ex, respectively, the number of possible substitutions σ_j is given by the permutation of p elements taken from a set of q elements $P(q, p)$. In this particular example, $p = 6, q = 13$, and $P(13, 6) = 8648640$

TABLE IV
SUBSTITUTIONS FOR THE FLEXIBLE MATCHING BETWEEN THE DESCRIPTION OF THE DOCUMENT IN FIG. 2 AND THE DESCRIPTION OF THE CLASS "SIFI letters"

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}	e_{16}	e_{17}
S1	X3	X3	X3	X9	X9	X5	X5	X5	X5	X5	X5	X4	X4	X4	X4	X4	X4
S2	X5	X5	X5	X8	X8	X4	X4	X4	X2	X2	X2	X2	X2	X2	X2	X2	X2
S3	X7	X7	X7	X5	X5	X7	X7	X7	X7	X7	X7	X5	X5	X5	X7	X7	X7
S4	X8	X8	X9	X6	X6	X8	X8	X9	X8	X8	X9	X6	X6	X6	X8	X8	X9
S5	X9	X10	X10	X3	X10	X9	X10	X10	X9	X10	X10	X3	X9	X10	X9	X10	X10
S6	X10	X13	X13	X7	X13	X10	X13	X13	X10	X13	X13	X7	X10	X13	X10	X13	X13
MF	.061	.061	.061	.107	.107	.031	.031	.031	.031	.031	.031	.857	.429	.429	.107	.107	.107

- Distance measures defined on feature vectors or languages from propositional calculus, which deal with only local deformations
- distance measures defined on graphs or languages from predicate calculus, which are concerned both with local and structural deformations.

In this section, we shall discuss some distances from the first category, restricting ourselves only to those defined on VL_1 (the ancestor of the VL_{21}), and some other distances from the second category, which are similar to the problem they are intended to solve.

A. Distance Measures for the VL_1 Language

The VL_1 language is an extension of the propositional calculus proposed by Michalski for representing knowledge in a family of inductive inference systems (AQ) based on the STAR methodology [26], [27]. The first distance measures for VL_1 propositions are presented in [9] and are concerned both with the problem of training example selection (ESEL) and the classification task (AQ11).

The distance measure implemented in ESEL is defined on only two levels:

1) Distance between the values of a descriptor (or feature) specialized according to the type of the descriptor
2) distance between events (feature vectors), which is defined as a weighted sum of distances between values of corresponding features of two events.

The most relevant points for this distance measure are as follows:

1) The distance for tree-structured descriptors is defined in terms of the number of branches on the shortest path linking the two values in the tree structure divided by the maximum number of branches on the shortest path linking any two nodes of the tree.
2) It is purely geometrical, and no probabilistic interpretation is given; indeed, the distance of two events is the sum of distances along each direction determined by features and not the product as in (16).

In the same paper, a measure of fitness used by AQ11, called the *degree of consonance*, is proposed for classification purposes. The degree of match is developed according to a wider evaluation scheme: degree of consonance of a selector, product of selectors (term), disjunctive VL_1 formula, and a set of formulas.

Even if less rigid definitions are sketched, the degree of consonance between a selector S and an event e implemented in AQ11 is

$$DC(S,e) = \begin{cases} 1 & \text{if the value of the appropriate descriptor in } e \text{ satisfies the selector } S \\ 0 & \text{if it does not satisfy } S \\ * & \text{if the value is unknown} \end{cases}$$

and the degree of consonance of a term is computed as the ratio of the number of selectors satisfied in the term to the total number of selectors in the term.

These definitions are inadequate for noise-affected data as they do not consider the type of descriptor and their weights in a match. On the other hand, the degree of consonance of a term suggests a sort of standardization that might prove useful when recognition rules with a quite different number of selectors are used for classifying new events. Indeed, in such a case, we may multiply the MF values by the proportion of matching selectors of the classification rule in order to reward the rule with the highest number of matching selectors.

Furthermore, the idea of extending the definition of the degree of consonance to a set of formulas appears to be interesting for those inductive systems generating more than one plausible hypothesis for each concept; in that case, taking the average of the degrees of consonance of the formulas in each set of hypotheses may improve the reliability of classification.

In [6], the distance measure for VL_1 formulas evolved toward a probabilistic approach. In fact, the measure of fitness of a *complex* (product of selectors) and an event e is defined as

$$MF(Cpx_j, e) = \prod_k MF(Sel_k, e) \frac{\text{weight}(Cpx_j)}{\#\text{examples}}$$

where the ratio weight(Cpx_j)/#examples is the estimate of the *a priori* probability of the complex (i.e., the relative frequency of positive training examples covered by the complex). Nevertheless, the problem has not been dealt with in any depth because the MF of a selector is not specialized in the type of domain and is not supported by an equally probabilistic approach. For instance, if

$$Sel_k = [\text{color=green}]$$

while the corresponding selector Sel of e is

$$Sel = [\text{color=white}]$$

then, according to their paradigm, the following results:

$$MF(sel_k, e) = 1/\text{DomainSize}$$

which is a decreasing function of DomainSize. This goes against the intuitive notion that in a wider domain, it is easier to observe a distortion. For instance, it is easier to assign a wrong value to the variable color when one has to choose from a set of 100 different equally probable colors instead of a set reduced to only two colors. Such a shortcoming has repercussions on possible extensions of the measure of fitness when the classifier knows the pdf of the values of the domain.

B. Distance Measures for Structural Representations

In [12], the authors focus on the principle that a generalization of two examples as well as the process of obtaining this generalization give indications of the conceptual distance between the examples. Indeed, very different examples generalize to an expression that is very far from each of them, whereas identical examples generalize to themselves. However, a set of examples may be characterized by several generalizations, each suggesting a certain conceptual distance; in that case, the minimum of these is taken as the estimation of the real conceptual distance.

The estimation of the conceptual distance is obtained by transforming the examples until they acquire approximately the same form, then by generalizing them, and thus retaining only the common features. The operations made in order to obtain such generalizations are considered as indicators of similarities and dissimilarities between those examples.

Some substantial differences between our distance measure and Kodratoff's conceptual distance are in the representation language. In fact, even though structural descriptions are allowed and a mapping into the VL_{21} language is possible, there is no tie on the connection of formulas, on the distinction of variables, and on the kind of quantifiers, and this further complicates the computation of distance. Another basic difference with our proposal is the bias of the conceptual distance toward the conceptual clustering and its difficult application to classification or concept recognition.

In [3], a distance measure between nonhierarchical attributed relational graphs, characterized by a descriptive graph grammar (DGG), is presented. A DGG is used in the pattern segmentation process for finding the nodes and the structure of the graph but not to compute the distance measure.

Their distance measure is based on the computation of the minimum number of modifications required to transform an input graph (pattern to classify) into a reference graph (concept describing a class). The modifications considered are the following: node insertion, node deletion, branch insertion, branch deletion, node label substitution, and branch label substitution. The distance measure, which was especially designed for classification purposes, is computed according to a combination of costs and weights associated with each modification leading to the graph isomorphism.

Although such a distance measure seems very different from ours, and this is especially due to the representation language, it is possible to find some analogies. For instance, a node in a relational graph may correspond to the set of attribute selectors of a single VL_{21} variable, and the cost function for that node may be compared with the product of the MF values for those selectors. Indeed, the main difference is that Sanfeliu and Fu consider a nonnull cost of node deletion since the model pattern is described in a complete space. This is due to the fact that Fu's work is in the area of classical pattern recognition, where reference patterns are described extensively (for instance, they are canonical handwritten characters) versus rule-guided pattern recognition [14], where reference patterns are generalizations, that is to say, incomplete descriptions outlining only discriminant and/or characteristic features of a concept.

In [4], *attributed graphs* are adopted to represent knowledge. Such a representation allows us to describe primitives of an object and the binary relations between them. Wong and You also generalize the definition of attributed graphs into a *random graph* by associating a pdf with both nodes and arcs. A random graph gives a probabilistic description of a concept when uncertainty (due to noise) exists in a structural pattern, whereas a simple attributed graph is used for (un)classified observations.

In the same paper, the authors define a distance measure between two random graphs, an attributed graph and a random graph, or two attributed graphs. The distance measure is based on the computation of the minimum increase of Shannon's entropy before and after the synthesis of an ensemble of attributed graphs into the probability distribution of a random graph.

The applications of the proposed distance are as follows:

1) *Unsupervised Learning*: The ensemble of attributed graphs contains more than one class of patterns, and the synthesis process produces probability distributions corresponding to these various classes.
2) *Supervised Learning*: The graphs to be synthesized are designated as belonging to the same pattern class.
3) *Classification*: The attributed graph of the unclassified pattern is synthesized with each random graph representing a class, and then, it is assigned to the class with the minimum distance.

The matching process used by our distance measure corresponds to the search of a monomorphism T labeling the vertices of a graph in Wong's distance. Therefore, the number of nodes in a random graph representing a concept may be greater than the number of vertices in the graph representing an observation, or equivalently and according to our terminology, the number of variables in G (or $G1$) is greater than the number of variables in Ex.

Unlike Fu's distance measure, the cost of node and branch insertion is null, and thus, the distance between a graph of order n and its m extension ($m > n$) is null. Consequently, if a part of the object cannot be completely defined, it can only be totally nonexistent and not partially unknown. Once again, this comes from the type of problem solved by random graphs: the problem of handwritten English letter recognition.

X. Conclusions

In this paper, we endeavored to provide a paradigm for classifying structured patterns when various sources of noise affect them or there exists a substantial amount of variability among the patterns themselves. The key idea consisted of the computation of a distance measure or, conversely, a measure of fitness, between symbolic descriptions expressed in the VL_{21} language, which is an extension of the first-order predicate calculus. When an observation is matched against the prototypical description of a concept, a flexible matching instead of a strict matching is applied. The result is a real number in [0,1] representing the probability that the observation matches against the description of the concept and not a simple true/false value assumed by the canonical matching predicate. The computation of the distance measure was defined according to a top-down evaluation scheme (formula in disjunctive normal form, conjunction of selectors, selector) and is essentially based on the concept that if there is a low probability of observing a greater distortion from the ideal pattern, there is a correspondingly low distance between the observation and the concept.

The classification process is conducted according to a multilayered framework:

1) *Single Match*: No further processing is required.
2) *No Match*: The distance measure is used.
3) *Multiple Match*: Ambiguity is solved by minimizing the sum of weights of the selectors belonging to $Ex - G$.

The multilayered approach comprises a two-tiered concept representation [28] in which the meaning of an approximate (imprecise) concept is distributed between its *base representation* and its *inferential interpretation*. The former describes the typical context-independent properties of the concept, whereas the latter determines whether a given instance satisfies some inferential extensions of the base representation. The process of interpretation involves probabilistic inference based on a distance measure (flexible matching), contextual information, background knowledge, and analogical reasoning.

An open question we raised in this paper concerns the exploitation of background knowledge, which is expressed in the form of inference rules, in the computation of the distance measure.

When dealing with real-world data, the problem of missing values is particularly felt. We proposed a general scheme distinguishing various kinds of missing values, different patterns (concepts or observations), and different descriptors (attributes or relations). Some formulas, which are coherent with the probabilistic approach followed in the definition of the distance measure, have been suggested.

Finally, we illustrated the application of the distance measure in an experimental system for office document automatic classification. It proved effective in reducing the negative effects of noise without overly increasing the recognition time too much. In addition, we believe that further studies on the matching problem may overcome

some limitations of the current implementation and improve the computational efficiency.

APPENDIX A

In the proof, we shall omit the use of the maximum function since it is possible to consider the substitution σ_j as providing the highest value of fitness between G and Ex. As a consequence, we will use the relation Match and not Match$_j$: If $k = p + 1$, where p is the number of selectors in $G1$, we have $MF(G, Ex) = P(\text{Match}(G1, Ex) \wedge \text{Match}(Sel_{p+1}, Ex)) =$ and, according to the multiplication theorem of probability, it results in $= P(\text{Match}(G1, Ex)) \cdot P(\text{Match}(Sel_{p+1}, Ex) | \text{Match}(G1, Ex)) =$ which, if the hypothesis $P(\text{Match}(G1, Ex)) = 1$ is satisfied, becomes $= P(\text{Match}(Sel_{p+1}, Ex)) = MF(Sel_{p+1}, Ex)$.

Let us suppose (6) is true for $k - 1$, and now, let us prove it for k:

$$\begin{aligned} MF(G, Ex) = & \\ & = P(\text{Match}(G1, Ex) \wedge Match(Sel_{p+1}, Ex) \wedge \\ & \ldots \wedge \text{Match}(Sel_{k-1}, Ex) \wedge \text{Match}(Sel_k, Ex)) = \\ & = P(\text{Match}(G1, Ex) \wedge \text{Match}(Sel_{p+1}, Ex) \wedge \\ & \ldots \wedge \text{Match}(Sel_{k-1}, Ex)) \cdot \\ & \cdot P(\text{Match}(Sel_k, Ex) | \text{Match}(G1, Ex) \\ & \wedge \text{Match}(Sel_{p+1}, Ex) \wedge \ldots \\ & \wedge \text{Match}(Sel_{k-1}, Ex)) = . \end{aligned}$$

Since (6) is considered true for $k - 1$ and first keeping account of the fact that

$$\begin{gathered} \text{Match}(G1, Ex) \wedge \text{Match}(Sel_{p+1}, Ex) \wedge \\ \ldots \wedge \text{Match}(Sel_{k-1}, Ex) = \text{Match}(Sel_{p+1}, Ex) \wedge \\ \ldots \wedge \text{Match}(Sel_{k-1}, Ex) \end{gathered}$$

(because $\text{Match}(G1, Ex)$ is the event with probability equal to one), and second, that $\text{Match}(Sel_k, Ex)$ is independent from

$$\text{Match}(Sel_{p+1}, Ex) \wedge \ldots \wedge \text{Match}(Sel_{k-1}, Ex).$$

We finally have

$$\begin{aligned} & = \sum_{i=p+1}^{k-1} MF(Sel_i, Ex) \cdot P(\text{Match}(Sel_k, Ex)) \\ & = \sum_{i=p+1}^{k} MF(Sel_i, Ex). \end{aligned}$$ q.e.d.

APPENDIX B

Let us recall the definition (10) given above

$$P(\text{EQUAL}(g_i, e)) = P(\delta(g_i, X) \geq \delta(g_i, e)). \tag{1B}$$

Henceforth, in order to simplify our notation, we will use g instead of g_i. As has already been stated, (1B) specializes according to both the type of domain to which g and e belong and the probability distribution of the domain values.

By assuming that the probability distribution is uniform and remembering the definition of δ for nominal domains, we have

$$\begin{aligned} & P(\text{EQUAL}(g, e)) = \\ & P(\delta(g, X) \geq \delta(g, e)) = \\ & \qquad \begin{cases} P(\delta(g, X) \geq 0) & \text{if } e = g \\ P(\delta(g, X) \geq 1) = (C - 1)/C & \text{if } e \neq g \end{cases} \end{aligned} \tag{2B}$$

where C is the number of elements of the domain.

For ordinal domains, (1B) becomes

$$\begin{aligned} & P(\text{EQUAL}(g, e)) = \\ & \quad P(|ord(g) - ord(X)| \geq |ord(g) - ord(e)|) = \end{aligned}$$

which can be rewritten in a simpler form by denoting $ord(g)$, $ord(e)$, and $ord(X)$ with g, e, and X, respectively:

$$= P(|g - X| \geq |g - e|). \tag{3B}$$

First Case: $g = e$

$$P(\text{EQUAL}(g, e)) = P(|g - X| \geq 0) = 1. \tag{4B}$$

Second Case: $g > e$

$$\begin{aligned} & P(|g - X| \geq |g - e|) \\ & = P(g - X < e - g \vee g - X = g - e \vee \\ & g - X > g - e \vee g - X = e - g) = \\ & = P(g - X < e - g) + P(g - X = g - e) \\ & + P(g - X = e - g) + P(g - X > g - e) = \\ & - P(X > 2g - e) + P(X = e) \\ & + P(X = 2g - e) + P(X < e) = \\ & = P(X \geq 2g - e) + P(X \leq e) = \\ & = [(C - 2g + e) \cdot \text{step}(C - 1 - 2g + e) + e + 1]/C. \end{aligned} \tag{5B}$$

Third Case: $g < e$

$$\begin{aligned} & P(|g - X| \geq |g - e|) \vee = P(g - X < g - e \vee g - X \\ & = e - g \vee g - X > e - g \vee g - X = g - e) = \\ & = P(g - X < g - e) + P(g - X = e - g) \\ & + P(g - X > e - g) + P(g - X = g - e) = \\ & = P(X > e) + P(X = 2g - e) \\ & + P(X < 2g - e) + P(X = e) = \\ & = P(X \leq 2g - e) + P(X \geq e) = \\ & = [(2g - e + 1) \cdot \text{step}(2g - e) + C - e]/C \end{aligned} \tag{6B}$$

where step(x) is the following function:

$$\text{step}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise.} \end{cases}$$

Finally, resubstistuting $ord(g)$ and $ord(e)$ to g and e, respectively, we have

$$\begin{aligned} & P(\text{EQUAL}(g_i, e)) = \\ & \begin{cases} [1 + \text{ord}(e) + (C - 2\text{ord}(g_i) + \text{ord}(e)) & \\ \quad \cdot \text{step}(C - 1 - 2\text{ord}(g_i) + \text{ord}(e))]/C & \text{if } g_i > e \\ 1 & \text{if } g_i = e \\ [C - \text{ord}(e) + (2\text{ord}(g_i) - \text{ord}(e) + 1) & \\ \quad \cdot \text{step}(2\text{ord}(g_i) - \text{ord}(e))]/C & \text{if } g_i < e. \end{cases} \end{aligned} \tag{7B}$$

A geometrical interpretation of (7B) may clarify the formula itself. $P(\text{EQUAL}(g, e))$ is the sum of probabilities of domain values not falling into a neighborhood of g with radius equal to $\delta(e, g)$ (Fig. 1). If g is the lowest (highest) value of the domain, then $P(\text{EQUAL}(g, e))$ constantly decreases along only one direction, whereas if g is the middle value, then $P(\text{EQUAL}(g, e))$ decreases along two directions as $\delta(g, e)$ increases. This explains why the event e represented in Fig. 3 is more probable as a distortion of g_1 rather than $g2$, even if it is at the same distance from the two centroids.

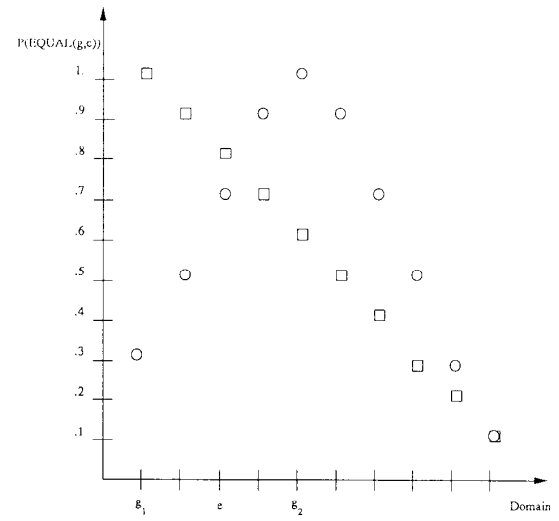


Fig. 3. Even if e is at the same geometrical distance from g_1 and g_2, a distortion of g_1 is more likely than a distortion of g_2.

ACKNOWLEDGMENT

The authors would like to thank E. Annese for his helpful discussions concerning the research presented in this paper as well as for making available the facilities of the OLIVETTI Systems and Networks Laboratory of Tecnopolis (Bari). They are also grateful to L. Saitta for her constructive criticism and stimulating remarks.

REFERENCES

[1] B. Hayes-Roth, "Implications of human pattern processing for the design of artificial knowledge systems," in *Pattern-Directed Inference Systems* (D.A. Waterman and F. Hayes-Roth, Eds.). Orlando: Academic, 1978, pp. 333–346.

[2] D. Ballard and C. Brown, *Computer Vision*. Englewood Cliffs, NJ: Prentice-Hall, 1982 .

[3] A. Sanfeliu and K. S. Fu, "A distance measure between attributed relational graphs for pattern recognition," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-13, pp. 353–362, 1983.

[4] A. K. C. Wong and M. You, "Entropy and distance of random graphs with application to structural pattern recognition," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-7, pp. 599–609, 1985.

[5] J. R. Quinlan, "The effect of noise on concept learning," in *Machine Learning: An Artificial Intelligence Approach, Vol. II* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds.). Los Altos: Morgan Kaufmann, 1986.

[6] R. S. Michalski, I. Mozetic, J. Hong, and N. Lavrac, "The AO15 inductive learning system: An overview and experiments," Intell. Syst. Group, Dept. of Comput. Sci., Univ. of Illinois, Urbana, IL, 1986.

[7] F. Bergadano, A. Giordana, and L. Saitta, "Automated concept acquisition in noisy environments," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 10, pp. 555–578, 1988.

[8] D. J. Hand, *Discrimination and Classification*: London, Wiley, 1981.

[9] R. S. Michalski and J. B. Larson, "Selection of the most representative training examples and incremental generation of VL_1 hypotheses: The underlying methodology and the description of programs ESEL and AQ11," Tech. Rep. UIUCDCS-R-78-867, Dept. of Comput. Sci., Univ. of Illinois, Urbana, IL, 1978.

[10] W. H. Tsai and K. S. Fu, "Subgraph error-correcting isomorphisms for syntactic pattern recognition," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-13, pp. 48–62, 1983.

[11] M. A. Eshera and K. S. Fu, "A graph distance measure for image analysis," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-14, pp. 398–408, 1984.

[12] Y. Kodratoff and G. Tecuci, "Learning based on conceptual distance," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 10, pp. 897–909, 1988.

[13] R. E. Stepp, "Machine learning from structured objects," in *Proc. Fourth Int. Workshop Machine Learning*. Irvine, CA: Morgan Kaufman, 1987.

[14] R. S. Michalski, "Pattern recognition as rule-guided inductive inference," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-2, pp. 349–361, 1980.

[15] R. E. Stepp, "Conjunctive conceptual clustering: A methodology and experimetation," Doctoral dissertation, Rep. UIUCDCS-R-84-1189, Dept. of Comput. Sci., Univ. of Illinois, Urbana, IL, 1984.

[16] J. B. Larson, "Inductive inference in the variable valued predicate logic system VL_{21}: Methodology and computer implementation," Doctoral dissertation, Dept. of Comput. Sci., Univ. of Illinois, Urbana, IL, 1977.

[17] D. G. Bobrow, "Dimensions of representation," in *Representation and Understanding* (D. G. Bobrow and A. Collins, Eds.). New York: Academic, 1975, pp. 1–34.

[18] M. G. Thomason, "Structural methods in pattern analysis," in *Pattern Recognition Theory and Applications* (P.A. Devijver and J. Kittler, Eds.). Berlin: Springer-Verlag, 1987.

[19] M. R. Garey and D. S. Johnson, *Computers and Intractability*. San Francisco, CA: W.H. Freeman, 1979.

[20] R. S. Michalski and R. E. Stepp, "Revealing conceptual structure in data by inductive inference," in *Machine Intelligence*. (J. E. Hayes, D. Michie, Y. -H. Pao, Eds.). London: Ellis Horwood, 1982.

[21] P. H. Winston, *Artificial Intelligence (2nd ed.)*. Reading, MA: Addison-Wesley, 1984, pp. 391–414.

[22] L. G. Shapiro and R.M. Haralick, "A metric for comparing relational descriptions," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-2, pp. 90–94, 1985.

[23] V. Barnett, *Comparative Statistical Inference*. London: Wiley, 1973.

[24] L. S. Chan and O. J. Dunn, "The treatment of missing values in discriminant analysis—I. The sampling experiment," *J. Amer. Stat. Assoc.*, vol. 67, pp. 473–477, 1972.

[25] F. Esposito, D. Malerba, G. Semeraro, E. Annese, and G. Scafuro, "An experimental page layout recognition system for office document automatic classification: An integrated approach for inductive generalization," in *Proc. 10th IEEE Int. Conf. Patt. Recogn.* (Atlantic City, NJ), 1990, pp. 557–562.

[26] R. S. Michalski, "On the quasi-minimal solution of the general covering problem," in *Proc. V Int. Symp. Inform. Processing (FCIP 69)* (Bled, Yugoslavia), 1969, vol. A3 (Switching Circuits).

[27] ——, "VL_1: variable-valued logic system," in *Proc. 1974 Int. Symp. Multiple-Valued Logic* (Morgantown, WV), 1974.

[28] ——, "How to learn imprecise concepts: A method for employing a two-tiered knowledge representation in learning," in *Proc. Fourth Int. Workshop Machine Learning*. Irvine, CA: Morgan Kaufman, 1987.