

## Spatial associative classification: propositional vs structural approach

Michelangelo Ceci · Annalisa Appice

© Springer Science + Business Media, LLC 2006

**Abstract** In Spatial Data Mining, spatial dimension adds a substantial complexity to the data mining task. First, spatial objects are characterized by a geometrical representation and relative positioning with respect to a reference system, which implicitly define both spatial relationships and properties. Second, spatial phenomena are characterized by autocorrelation, i.e., observations of spatially distributed random variables are not location-independent. Third, spatial objects can be considered at different levels of abstraction (or granularity). The recently proposed SPADA algorithm deals with all these sources of complexity, but it offers a solution for the task of spatial association rules discovery. In this paper the problem of mining spatial classifiers is faced by building an associative classification framework on SPADA. We consider two alternative solutions for associative classification: a *propositional* and a *structural* method. In the former, SPADA obtains a propositional representation of training data even in spatial domains which are inherently non-propositional, thus allowing the application of traditional data mining algorithms. In the latter, the Bayesian framework is extended following a multi-relational data mining approach in order to cope with spatial classification tasks. Both methods are evaluated and compared on two real-world spatial datasets and results provide several empirical insights on them.

**Keywords** Spatial classification · Associative classification · Naive Bayesian classification · Association rules discovery

---

M. Ceci (✉) · A. Appice  
Dipartimento di Informatica,  
Università degli Studi di Bari  
via Orabona, 4 - 70126 Bari, Italy  
e-mail: ceci@di.uniba.it

A. Appice  
e-mail: appice@di.uniba.it

## 1 Introduction

The number of applications using spatial or geographic data has been increasing over the last decades. Some examples are traffic and fleet management, environmental and ecological modeling, robotics, computer vision, and, more recently, computational biology and mobile computing applications. Spatial or geo-referenced data are collected in spatial databases and Geographical Information Systems (GIS) at a rate which requires the application of automated data analysis methods in order to extract implicit, previously unknown, and potentially useful information. Emerging data mining technology already provides several data analysis tools for a variety of tasks, both predictive and descriptive. However, the presence of a spatial dimension in the data adds substantial complexity to the data mining tasks.

Spatial objects, indeed, are characterized by a geometrical representation and position with respect to a reference system. The former implicitly defines a number of spatial attributes (e.g., “orientation”), while the latter defines spatial relations of a different nature, such as topological (“intersects”) and geometric (“above”), not explicitly encoded in a spatial database. Modeling these implicit spatial properties (attributes and relations) in order to associate them with a clear semantics and a set of efficient procedures for their computation is the first challenge met by data miners when facing a spatial data mining problem (Shekhar, Schrater, Vatsavai, Wu, & Chawla, 2002). Some well-known formalizations, such as the 9-intersection model for topological relationships (Egenhofer, 1991), are unsatisfactory in many applications, since the end-user of a data mining solution is often interested in human-interpretable properties and relations between spatial objects, independent of their geometrical representation in the database.

Another challenge is represented by spatial autocorrelation. In spatial domains, everything is generally related to everything else, but nearby things are more related than distant things. This means that the effect of an explanatory (independent) or response (dependent) variable at any site may not be limited to the specific site. For instance, the presence of communal establishments (schools, hospitals) in an enumeration district (ED) may affect the number of migrants both in that ED and the nearby EDs. Therefore, when properties of some units of analysis are investigated, attributes of spatially related units of analysis must be taken into account as well. More in general, attributes of *any* spatial object in the neighborhood of a unit of analysis may have a certain influence. This leads distinguishing between the *reference objects* of analysis and other *task-relevant spatial objects*, and to represent their spatial interactions. Several traditional data mining methods do not make this distinction, nor do they allow the representation of any kind of interaction (spatial or not).

Spatial objects can often be organized in hierarchies of classes. By descending or ascending through a hierarchy, it is possible to view the same spatial object at different levels of abstraction (or granularity). Confident patterns are more likely to be discovered at low granularity levels. On the other hand, large support is more likely to exist at higher granularity levels. Therefore, the data mining methods should be able to explore the search space at different granularity levels in order to find the most interesting patterns (e.g., the most supported and confident). In the case of granularity levels defined by a containment relationship, this corresponds to exploring both global and local aspects of the underlying phenomenon. Very few data mining techniques do automatically support this multiple-level analysis. In general, the user is forced to repeat independent experiments on different

representations, and the results obtained for high granularity levels are not used to control search at low granularity levels (or vice versa). Therefore, the discovery of patterns at different levels of granularity is another challenge for research in spatial data mining.

Spatial reasoning is the process by which information about objects in space and their relationships are gathered through measurement, observation or inference, and used to arrive at valid conclusions regarding the object relationships (Sharma, 1996). Several theories of spatial reasoning have been developed (Aiello, 2001) and several rules have been proposed to support sound inferential mechanisms. They offer a source of domain independent knowledge which should be taken into account when searching for interesting patterns. Despite this, little research has been reported in the literature on the usage of background knowledge derived from spatial reasoning for effective spatial data mining (Wang, Liu, Wang & Liu, 2006).

In a recent work (Appice, Ceci, Lanza, Lisi & Malerba, 2003), the Spatial Pattern Discovery Algorithm (SPADA) has been proposed to deal with all these challenges in the descriptive task of association rules discovery from spatial data. SPADA supports the multi-level analysis by mapping hierarchies of spatial objects into different levels of granularity and by extracting association rules for each granularity level. Pattern generation proceeds from more general to more specific granularity levels, so that it is possible to profitably exploit information collected at higher levels to prevent the generation and evaluation of weak patterns at lower levels. SPADA distinguishes between the reference objects (the units of analysis) and the task-relevant objects (the spatial objects which are relevant for the task in hand but are not necessarily the main subject of the generalization). When units of analysis are considered both as reference objects and task-relevant objects, then autocorrelation effects can also be investigated. A flexible language bias has been designed for SPADA, so that constraints on patterns and association rules can be easily defined. They help to prevent the generation and/or the evaluation of uninteresting association rules. Finally, the generation of patterns also takes into account some background knowledge expressed in the form of Prolog clauses. In this way it is possible to simulate inferential mechanisms defined within a spatial reasoning theory. Moreover, by specifying both a background knowledge and some suitable pattern constraints, it is possible to change the representation language used for spatial patterns, making it more abstract (human-comprehensible) and less tied to the physical representation of spatial objects in the database.

Although SPADA offers a sufficiently complete solution to the challenges posed by the spatial domain, it is specially designed for a descriptive task. In this work, we exploit potentialities of SPADA as a spatial data mining algorithm to deal with predictive tasks and, more precisely, the classification task. In classification, the goal is to learn the concept associated with each class by finding regularities which characterize the class in question and discriminate it from the other classes. Examples of spatial applications where classification is highly demanded are the recognition of geographic objects for topographic map interpretation (Kramer, 1999; Malerba, Esposito, Lanza, Lisi & Appice, 2003) and the classification of deprived areas for urban planning activities (Fitzpatrick, 2001).

In order to exploit the potentialities of SPADA for classification purposes, we resort to recent studies on associative classification (Li, Han, & Pei, 2001; Liu, Hsu, & Ma, 1998; Yin & Han, 2003) which have investigated the opportunity of building classifiers by careful selection of high quality association rules discovered from

training data sets. Following the main inspiration in these studies, we use SPADA in an associative classification framework. In particular, SPADA is used:

1. to determine attributes of an attribute-value representation of reference objects;
2. to extract rules to be used in a Naïve Bayesian classifier which operates on *structured* representations.

In the first case, SPADA permits to obtain a *propositional* representation of training data even in spatial domains which are inherently non-propositional, thus making possible the application of traditional data mining algorithms (Krogel, Rawles, Zelezny, Flach, Lavrač, & Wrobel, 2003) to spatial data mining problems. In the second case, the Bayesian framework is extended following the multi-relational data mining approach (Džeroski & Lavrač, 2001) in order to cope with spatial classification tasks. In this paper, both the propositional and the structural approaches are investigated and tested on two geo-referenced census datasets. The problems considered are the classification of deprivation areas in Greater Manchester and the classification of the cost of renting apartments in Munich. Although experimental results cannot be used in practice because of the census data available is obsolete and there is a lack of experts who can help to interpret discovered patterns, they provide at least some empirical insights on the different solutions proposed for spatial associative classification.

The paper is organized as follows. In the next section, the background of this research and some related works are discussed. The general framework of multi-level spatial associative classification is presented in Section 3, while the extraction of multi-level spatial association rules used by the spatial associative classifier is briefly presented in Section 4. Two spatial associative classification methods based on a propositional and a structural approach respectively are described in Section 5. Section 6 reports the applications of both methods on the Greater Manchester and Munich census data together with a systematic analysis of results. Finally, in Section 7, the lessons learned are reported, conclusions are drawn and future work is proposed.

## 2 Related work

In order to clarify the background and the motivations for this work, related research on associative, propositional, and structural approaches to (spatial) classification are reported below.

### 2.1 Background for spatial associative classification

In the literature there are already several works on spatial classification. Ester, Kriegel and Sander (1997) have proposed an extension of decision trees based on neighborhood graphs to consider both thematic attributes of the classified objects and relations with neighboring objects. Hierarchical relations are not considered. Conversely, the spatial classification method proposed in Koperski (1999) deals with both spatial and hierarchical relations between spatial objects. Thematic attributes of neighboring objects are involved in the classification, but data is assumed to be stored

in a single table of a relational database. A structural approach to classification is reported in Malerba et al. (2003), where the authors exploit the expressive power of first-order logic to represent not only spatial relations, but also background knowledge. However, the learned rules are quite brittle, since conditions in the body must perfectly match to the object description in order to classify it, and hierarchies of spatial objects are not considered.

A seminal work on spatial association rules discovery is that of Koperski and Han (1995) for extraction of multi-level spatial association rules by a progressive deepening of the levels. More recently, Morimoto (2001) has exploited the concept of “frequent neighboring class sets” to partition the spatial dataset according to the *closeness* relation and discover patterns on each single partition. A common limitation of all these methods is that they operate on data stored in a single table, that is, the spatial database has to be “flattened” somehow before applying the data mining method. Moreover, discovered patterns are not relational, hence they cannot properly express spatial relations.

Relational frequent patterns are generated by WARMR (Dehaspe & Toivonen, 1999), which adapts Mannila and Toivonen’s levelwise method (Mannila & Toivonen, 1997) to a search space of conjunctive first-order formulas. Although *is-a* hierarchies (taxonomic relations) can be represented in WARMR, the system is not able to perform multi-level analysis because it lacks the mechanisms for taxonomic reasoning. Another difficulty in using WARMR for spatial data analysis is that it provides no support to extract properties of reference or task-relevant objects from spatial databases or GIS. In contrast, SPADA (Lisi & Malerba, 2004) is designed to extract multi-level association rules and is wrapped up in the spatial data mining system ARES, which provides SPADA with an interface to a commercial spatial DBMS.

In the last years, studies on association rules have resulted in methods for building classifiers through a careful selection of high quality association rules discovered in training data (Liu et al., 1998; Yin & Han, 2003). ‘Associative classification’ methods, as they are called, present several advantages. First, differently from most of tree-based classifiers, whose univariate tests cause the ‘horizon effect’, associative classifiers are based on association rules which consider the simultaneous correspondence of values of different attributes, hence promising to achieve better accuracy (Baralis & Garza, 2003). Second, associative classification 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 (Liu et al., 1998). Fourth, the associative classification approach helps to solve *understandability* problems (Pazzani, Mani, & Shankle, 1997) which 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 (Liu et al., 1998).

Although associative classification methods reported in the literature present several interesting aspects, they also suffer from some limitations. First, they do not deal with issues characteristic of the 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. To overcome these deficiencies, a probabilistic classifier can be used to return, in addition to the result of the classification, the confidence in the result itself. Third, reported methods require additional heuristics to identify the most effective rule at classifying an object. This suggests evaluating the class with computing probabilities according to all the rules.

## 2.2 Propositional and structural classification

Since spatial objects generally belong to different layers (e.g. roads, railways, waters) and are characterized by different properties, a natural way to model them is by relational tables of a relational data model. Hence, we can resort to (multi-) relational data mining (Džeroski & Lavrač, 2001) which investigates methods able to extract knowledge from data stored in multiple data tables. (Multi-)relational data mining methods are typically based on two alternative approaches: propositional and structural (or relational).

The former requires the transformation of relational data into a propositional (or feature-based or attribute-value) representation by building features which capture relational properties of data. This kind of transformation, named *propositionalization*, decouples feature construction from model construction (Kramer, Lavrač, & Flach, 2001) such that conventional propositional classification methods may be applied to transformed data, thus allowing a wider choice of robust and well-known algorithms (Krogel et al., 2003). Propositionalization learners have been proposed by resorting to either database-oriented techniques (Knobbe, Haas, & Siebes, 2001) or logic-oriented techniques (Kramer, 1999; Krogel, 2005; Lavrač & Džeroski, 1994). The former deal with data stored in a relational database by materializing joins according to foreign key associations and allowing for fast aggregations. The latter require data represented as Prolog facts and deal with complex background knowledge. This makes logic-oriented propositionalization able to provide expressive first-order models by constructing first-order logic features. For example, Srinivasan and King (1996) have proposed to enhance expert provided features with boolean features constructed from PROGOL clauses (Muggleton, 1995) learned from examples and background knowledge, while (Dehaspe & Toivonen, 2000) have constructed boolean features from frequent conjunctions of literals extracted from multiple relations.

Differently, the structural approach takes into account the structure of original data by providing functionalities to navigate relational structure in its original format and generate potentially new forms of evidence not readily available in a flattened single table representation. Hence, the whole hypothesis space is directly explored by the mining method. The ILP community has been investigating, for a number of years, both theoretical and practical aspects of inducing structural classifiers from data stored as Prolog facts. Most of proposed methods (Blockeel, 1998; Ceci, Appice & Malerba, 2003; Flach & Lachiche, 2004; Getoor, 2001; Leiva, 2002) have been obtained by upgrading standard statistical approaches, such as Bayesian networks, decision trees, naïve Bayesian, and Markov networks to the (multi-) relational setting. In any case, only a few of them (Ceci et al., 2003; Getoor, 2001; Leiva, 2002), exploit a tight-integration with a relational database where database schema is provided free of charge to guide the learning process.

Propositional and structural approaches to (multi-)relational data mining have been compared on several application domains (e.g., biology and chemistry) (Kramer, 1999). Studies emphasize that methods implementing structural approaches are, in principle, more accurate than their corresponding propositional approaches which lose information about how data were originally structured (Ceci et al., 2003). On the other hand, the propositionalisation reduces the search space to a minimal subset including features obtained as transformation of the original multi-relational feature space. Sometimes this transformation may speed up the learning process of several orders of magnitude, thus permitting a wider exploration of the search space. This explains why propositional approaches can produce interesting results despite the original loss of information (Kroegel, 2005).

### 3 Multi-level spatial associative classification

The multi-level spatial associative classification problem can be formalized as follows:

*Given* a spatial database (SDB), a set  $S$  of reference spatial objects tagged with a class label  $y \in \{c_1, \dots, c_L\}$ , some sets  $R_k$ ,  $1 \leq k \leq m$  of task-relevant spatial objects, a background knowledge  $BK$  including 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  $\psi_k$  which associate each object in  $H_k$  with a granularity level to deal with several hierarchies at once, a couple of thresholds  $\text{minsup}[l]$  and  $\text{minconf}[l]$  for each granularity level, a language bias  $LB$  which constrains the search space;

*Find*, for each granularity level, *i*) a set of strong spatial association rules capturing associations between the properties of reference objects and the corresponding class labels  $y$  and *ii*) a classifier to predict  $y$  for reference objects.

Spatial association rules are extracted in the form  $A \Rightarrow C (s, c)$ , where  $A$  (antecedent) and  $C$  (consequent) are sets of items with  $A \cap C = \emptyset$ . The support  $s$  estimates the probability  $p(A \cap C)$ , while the confidence  $c$  estimates the probability  $p(C|A)$ . The conjunction of literals in  $A \cup C$  is named spatial pattern since the relations and the objects involved in this pattern have a spatial nature. An example of spatial pattern is

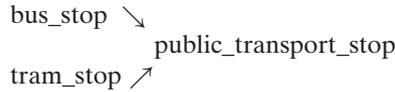
$$\text{is\_a}(X, \text{ed}) \wedge \text{no\_car\_owning\_households}\%(X, \text{high}) \wedge \text{contains}(X, Y) \wedge \\ \text{is\_a}(Y, \text{public\_transport\_stop}),$$

which expresses a spatial containment relation between a reference ED and some task-relevant spatial object classified as public transport stop, while the following:

$$\text{is\_a}(X, \text{ed}) \wedge \text{no\_car\_owning\_households}\%(X, \text{high}) \Rightarrow \\ \text{contains}(X, Y) \wedge \text{is\_a}(Y, \text{public\_transport\_stop}) (40\%, 60\%)$$

is a spatial association rule. It is noteworthy that items are first-order logic atoms, that is,  $n$ -ary predicates applied to  $n$  terms. In the previous examples, terms are either *variables* ( $X$  and  $Y$ ), or *constants* (*high* and *public\_transport\_stop*). In order to deal with discretized attributes, intervals (e.g., '[12..15]') are managed as constants as well.

By taking into account hierarchies on task-relevant objects, we obtain descriptions at different granularity levels. For instance, if we assume the spatial hierarchy:



a finer-grained spatial association rule could be:

$$\begin{aligned} \text{is\_a}(X, \text{ed}) \wedge \text{no\_car\_owning\_households}\%(X, \text{high}) \Rightarrow \\ \text{contains}(X, Y) \wedge \text{is\_a}(Y, \text{bus\_stop})(33\%, 80\%), \end{aligned}$$

which provides better insight on the nature of the task-relevant object  $Y$ . In the problem formulation, spatial objects of each hierarchy  $H_k$  can be mapped to one or more of the  $M$  granularity levels to deal uniformly with several hierarchies at once. 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  $s \geq \text{minsup}[l]$  and all ancestors of  $P$  with respect to  $H_k$  are frequent at their corresponding levels. The support  $s$  estimates the probability  $p(P)$ . An association rule  $A \Rightarrow C$  ( $s\%$ ,  $c\%$ ) at level  $l$  is *strong* if the pattern  $A \cup C$  ( $s\%$ ) is frequent and  $c \geq \text{minconf}[l]$ .

In our proposal, we use SPADA (Lisi & Malerba, 2004) to extract strong multi-level spatial association rules. Since we are interested in strong rules with exactly one literal representing the class label in the consequent, some constraints are specified in SPADA language bias. Such strong rules are then used to build a multi-level spatial classifier according to either a propositional or a structural approach.

In the propositional case, we follow the same intuition of Dehaspe and Toivonen (1999) that is, patterns which succeed with a certain frequency are considered the basis for the construction of relevant features. Hence, for each granularity level, strong rules are used to create a relational table whose columns represent boolean features. This permits us to derive a propositional description of the same individual at different levels of granularity and overcome one limitation of state-of-art propositionalization learners, that is, propositional features are constructed by ignoring the structure possibly imposed by hierarchical relation on objects of the same type. Since propositionalization tends to produce a large number of features, many of which are highly correlated or even logically redundant, a feature reduction algorithm is applied to remove redundant features and improve the efficiency of the learning algorithm without generally affecting accuracy (Appice, Ceci, Rawles & Flach, 2004a). Finally, for each level of granularity, it is possible to apply traditional data mining algorithms which take as input the reduced relational table.

In the structural case, the set of rules extracted for each level is evaluated within a naïve Bayesian classification framework that, given a new reference object to be classified, estimates the posterior probability of the class on the basis of the object description. The classifier returns a posterior probability for each class, thus enhancing standard associative classification methods which return a categorical output and convey no information on the potential uncertainty in classification.

#### 4 Mining spatial association rules with SPADA

The basic idea of SPADA (Lisi & Malerba, 2004) is that a spatial database (SDB) boils down to a deductive relational database ( $D(S)$ ) once the spatial relations between reference objects and task-relevant objects have been extracted. For instance, spatial intersection between objects is represented in a derived relation  $intersects(X, Y)$ .  $D(S)$  is obtained by augmenting the user supplied  $BK$  with the data extracted from SDB and concerns both the reference ( $S$ ) and the task-relevant objects  $\{R_k\}$ . The ground facts<sup>1</sup> in  $D(S)$  can be grouped into distinct (not necessarily disjoint) subsets. Each group  $O[s]$  is uniquely identified by a reference object  $s \in S$  and corresponds to the description of a spatial unit of analysis. The uniqueness of the reference object associated to a spatial unit allows us to well define both the support and confidence of a spatial association rule. More precisely, the spatial association rule:  $A \Rightarrow C (s\%, c\%)$ , means that in  $s\%$  of spatial units of analysis, both conjunctions  $A$  and  $C$  hold and in  $c\%$  of spatial units of analysis where  $A$  is true, also  $C$  holds.

The task of spatial association rules discovery performed by SPADA is split into two sub-tasks: find frequent spatial patterns and generate highly-confident spatial association rules. The discovery of frequent patterns is performed according to the *levelwise* method described in Mannila and Toivonen (1997), that is, a breadth-first search in the lattice of patterns spanned by a generality order between patterns. In SPADA the generality order is based on  $\theta$  substitution. The pattern space is searched one level at a time, starting from the most general patterns and iterating between candidate generation and evaluation phases.

Once large patterns have been generated, it is possible to generate strong spatial association rules. For each pattern  $P$ , SPADA generates antecedents suitable for rules being derived from  $P$ . The consequent corresponding to an antecedent is simply obtained as the complement of atoms in  $P$  and not in the antecedent. Rule constraints are used to specify literals which should occur in the antecedent or consequent of discovered rules. In a more recent release of SPADA (3.1), new pattern (rule) constraints have been introduced in order to specify exactly both the minimum and maximum number of occurrences for a literal in a pattern (antecedent or consequent of a rule). An additional rule constraint has been introduced to eventually specify the maximum number of literals to be included in the consequent of a rule. In this way we are able to constrain the consequent of a rule requiring the presence of only the literal representing the class label and obtain patterns useful for classification purposes.

In the ARES<sup>2</sup> system, SPADA has been loosely coupled with the Oracle Spatial database in order to extract the logical description of the spatial units of analysis. Numerical data are discretized according to a context sensitive approach (Ludl & Widmer, 2000) which is more suitable for association rule mining.

<sup>1</sup>In this work we assume that ground facts concern either taxonomic “is\_a” relationships or binary spatial relationships  $\alpha(s, r)$  or object properties.

<sup>2</sup><http://www.di.uniba.it/~malerba/software/ARES/index.htm>

## 5 Propositional vs structural associative classification

Once strong spatial association rules with only the class label in the consequent are extracted for each granularity level, they are used to mine either propositional or structural spatial classifiers.

### 5.1 Propositional approach

A propositional classifier can be built by converting strong association rules to a set of boolean features and using the result as input of some attribute-value classification algorithm. Boolean features are constructed for each granularity level  $l$ , therefore, units of analysis are described by  $M$  distinct relational tables  $B_l$ , one for each level.

More precisely, let  $\mathfrak{R}_l = \{A_i \Rightarrow y(X, c_i)\}$  be the set of strong spatial association rules discovered at granularity level  $l$ , whose consequent contains only one literal, that is, the class label. Then the schema of the relational table  $B_l$  includes  $|\mathfrak{R}_l|+1$  attributes, that is,  $A_1, \dots, A_{|\mathfrak{R}_l|}, Y$ . Each row of  $B_l$  corresponds to a reference object  $s \in S$ . The column  $Y$  in  $B_l$  represents the class label  $y$  associated to  $s$ . If the antecedent of a rule  $(A_i \Rightarrow y(X, c_i)) \in \mathfrak{R}_l$  covers  $O[s]$ , that is, a substitution  $\theta$  exists such that  $A_i\theta \subseteq O[s]$ , then the  $i$ -th value of the row in  $B_l$  corresponding to  $s$  is set to *true*, *false* otherwise.

*Example 1* Let us consider the sets of literals:

is\_a( $s$ , ed),  $y(s$ , high\_deprivation), inhabitants( $s$ , [1, 000..3, 000]),  
 contains( $s$ ,  $s_1$ ), is\_a( $s_1$ , school), area( $s_1$ , [210..613]), students( $s_1$ , [50..200]),  
 contains( $s$ ,  $s_2$ ), is\_a( $s_2$ , shop), employed( $s_2$ , [3..5]),  
 contains( $s$ ,  $s_3$ ), is\_a( $s_3$ , shop), employed( $s_3$ , [1..2]), ...

which describes an enumeration district (ED)  $s$  and some spatially related task-relevant objects, namely  $s_1$ ,  $s_2$  and  $s_3$  belonging to the “School” and “Shop” layers of a spatial database. Spatial arrangement is here expressed by means of predicates “contains” and “area”. This relational description of  $s$  can be converted into  $\{\langle true, false, high\_deprivation \rangle\}$  attribute-value representation according to the class label (*high\_deprivation*) and according to the following spatial association rules:<sup>3</sup>

$R_1$  : is\_a( $X$ , ed)  $\wedge$  inhabitants( $X$ , [1, 000..3, 000])  $\wedge$  contains( $X$ ,  $Y$ )  
 $\wedge$  is\_a( $Y$ , shop)  $\wedge$  employed( $Y$ , [1..2])  $\Rightarrow y(X$ , high\_deprivation).  
 $R_2$  : is\_a( $Z$ , ed)  $\wedge$  inhabitants( $Z$ , [1, 000..3, 000])  $\wedge$  contains( $Z$ ,  $W$ )  
 $\wedge$  is\_a( $W$ , school)  $\wedge$  students( $W$ , [12..49])  $\Rightarrow y(Z$ , low\_deprivation).

Indeed, there is a substitution  $\theta = \{X \leftarrow s, Y \leftarrow s_3\}$  such that antecedent( $R_1$ ) $\theta \subseteq O[s]$ , while there is no substitution  $\theta$  such that antecedent( $R_2$ ) $\theta \subseteq O[s]$ .

The number of attributes in  $B_l$  depends on the cardinality of  $\mathfrak{R}_l$ . Since, the number of discovered association rules is usually high and many rules may be strongly

<sup>3</sup>Henceforth, we will assume that variables of association rules are standardized apart, that is, two distinct association rules do not share variables.

correlated, this may generate boolean attributes (or features) which are highly correlated or even logically redundant. More precisely, a feature  $p$  is identified as *redundant* with respect to another feature  $q$  for distinguishing *positive* from *negative* examples of a class  $c$  if  $p$  is *true* for at least the same positive examples as  $q$ , and *false* for at least the same negative examples as  $q$  (Modrzejewski, 1993). Association rule-based propositionalization can be profitably combined with redundant feature elimination to reduce the hypothesis space by excluding boolean features which are redundant for learning.

Redundancy elimination is important to improve the efficiency of the learning process without affecting the accuracy of the learned classifier. In this work it is based on the method REFER (Appice et al., 2004a) specially designed for multi-class problems. In Appice et al. (2004a), it is formally proved that REFER preserves the existence of a complete and consistent theory for each class label when eliminating redundant features and empirically proved that REFER significantly reduces the number of features when combined with propositionalization without reducing in accuracy.

Finally, each reduced relational table  $B_l$  obtained by removing redundant attributes from  $B_l$  is input to a propositional learning algorithm. In this work we consider four learning methods implemented in the WEKA package, namely, JRIP (Cohen, 1995), C4.5 (Quinlan, 1993), 1-NN (Mitchell, 1997) and naïve Bayesian classifier (NBC) (Domingos & Pazzani, 1997). JRIP is a rule learner similar to the commercial rule learner Ripper, C4.5 is a decision tree learner, 1-NN is an instance-based learner which assigns the class of the closest training example to a new one, and NBC is a probabilistic classifier based on the Bayes theorem for conditional probability.

## 5.2 Structural approach

In the structural approach, strong spatial association rules extracted at each granularity level are directly used to build a multi-level naïve Bayesian classifier which classifies any reference object  $s$  at each granularity level  $l$  by maximizing the *posterior probability*  $P(c_i|s)$  such that  $s$  is of class  $c_i$ , that is,  $\text{class}(s) = \arg \max_i P(c_i|s)$ . By applying the Bayes theorem,  $P(c_i|s)$  is reformulated as follows:

$$P(c_i|s) = \frac{P(c_i)P(s|c_i)}{P(s)}. \quad (1)$$

Since  $P(s)$  is independent of the class  $c_i$ , it does not affect  $\text{class}(s)$ , that is,  $\text{class}(s) = \arg \max_i P(c_i)P(s|c_i)$ . Under the conditional independence assumption (*naïve Bayes assumption*), the likelihood  $P(s|c_i)$  can be factorized:  $P(s|c_i) = P(s_1, \dots, s_m|c_i) = P(s_1|c_i) \times \dots \times P(s_m|c_i)$ , where  $s_1, \dots, s_m$  represent the set of attribute values, different from the class label, used to describe the object in hand. Surprisingly, naïve Bayesian classifiers have been proved accurate even when the conditional independence assumption is grossly violated. This is due to the fact that when the assumption is violated, estimates of posterior probabilities can be poor, but the correct class still has the highest estimate, leading to correct classification (Domingos & Pazzani, 1997).

The formulation reported above for naïve Bayesian classifiers is clearly limited to propositional representations. In the case of structural representations, some

extensions are necessary. The basic idea is that of using a set of relational patterns to describe an object to be classified, and then to define a suitable decomposition of the likelihood  $P(s|c_i)$  à la naive Bayesian classifier to simplify the probability estimation problem. For each granularity level  $l$ ,  $P(s|c_i)$  is computed on the basis of the subset  $\mathfrak{R}_l(s) \subseteq \mathfrak{R}_l$  of strong spatial association rules whose antecedent covers the reference object  $s$ :

$$P(s|c_i) = P \left( \bigwedge_{R_k \in \mathfrak{R}_l(s)} \text{antecedent}(R_k)|c_i \right). \tag{2}$$

The straightforward application of the naïve Bayes independence assumption to all literals in  $\bigwedge_{R_k \in \mathfrak{R}_l(s)} \text{antecedent}(R_k)$  is not correct, since it may lead to underestimate the probabilities for the case of classes for which several similar association rules are found. For instance, suppose that  $\mathfrak{R}_l(s) = \{R_1, R_2\}$ , such that:

$$R_1 : \text{is\_a}(X, \text{ed}) \wedge \text{inhabitants}(X, [1, 000..3, 000]) \wedge \text{contains}(X, Y) \wedge \text{is\_a}(Y, \text{shop}) \Rightarrow y(X, c_i).$$

$$R_2 : \text{is\_a}(Z, \text{ed}) \wedge \text{inhabitants}(Z, [1, 000..3, 000]) \wedge \text{contains}(Z, W) \wedge \text{is\_a}(W, \text{school}) \Rightarrow y(Z, c_i).$$

where the variable  $X$  denotes the reference object to be classified. The simple application of the naïve Bayes independence assumption would produce this factorization:

$$\begin{aligned} P(s|c_i) &= P(\text{antecedent}(R_1) \wedge \text{antecedent}(R_2)|c_i) \\ &= P(\text{is\_a}(X, \text{ed})|c_i) \times P(\text{inhabitants}(X, [1, 000..3, 000])|c_i) \\ &\quad \times P(\text{contains}(X, Y)|c_i) \times P(\text{is\_a}(Y, \text{shop})|c_i) \\ &\quad \times P(\text{is\_a}(Z, \text{ed})|c_i) \times P(\text{inhabitants}(Z, [1, 000..3, 000])|c_i) \\ &\quad \times P(\text{contains}(Z, W)|c_i) \times P(\text{is\_a}(W, \text{school})|c_i) \\ &= P(\text{is\_a}(X, \text{ed})|c_i)^2 \times P(\text{inhabitants}(X, [1, 000..3, 000])|c_i)^2 \\ &\quad \times P(\text{contains}(X, Y)|c_i)^2 \times P(\text{is\_a}(Y, \text{shop})|c_i) \times P(\text{is\_a}(Y, \text{school})|c_i) \end{aligned}$$

since the first three literals of the two antecedents can be unified according to the substitution  $\sigma = \{Z \leftarrow X, W \leftarrow Y\}$ . Therefore, there is a quadratic contribution of some probabilities, and if one of them is small  $P(s|c_i)$  will approach zero.

To prevent this problem we resort to the logical notion of factorization (Robinson, 1965) which is given for clauses (i.e., disjunctions of literals) but we adapt it to the notion of pattern.

**Definition 1** Let  $P$  be a pattern, which has a non-empty subset  $Q \subseteq P$  of unifiable literals with most general unifier (mgu)  $\theta$ . Then  $P\theta$  is called a *factor* of  $P$ .

A factor of a pattern  $P$  is obtained by applying a substitution  $\theta$  to  $P$  which unifies one or more literals in  $P$ , and then deleting all but one copy of these unified literals. In our context we are interested in particular factors, namely those that are obtained by substitutions  $\theta$  which satisfy three conditions:

1.  $\text{Domain}(\theta) = \bigcup_{R_k \in \mathfrak{R}_l(s)} \text{Vars}(\text{antecedent}(R_k))$  that is, the domain of  $\theta$  includes all variables occurring in the association rules  $R_k \in \mathfrak{R}_l(s)$ ;

2.  $\text{Domain}(\theta) \cap \text{Range}(\theta) = \emptyset$ , that is,  $\theta$  renames all variables occurring in the association rules  $R_k \in \mathfrak{R}_l(s)$  with new variable names;
3.  $\theta|_{\text{Vars}(R_k)}$  is injective, that is, the restriction of  $\theta$  on the variables occurring in  $R_k$  is injective.

In the previous example,  $\theta = \{X \leftarrow U, Y \leftarrow V, Z \leftarrow U, W \leftarrow V\}$  satisfies all these conditions, therefore, the factor of interest is:

$(\text{antecedent}(R_1) \wedge \text{antecedent}(R_2))\theta$ :

$$\text{is\_a}(U, \text{ed}) \wedge \text{inhabitants}(U, [1, 000..3, 000]) \wedge \text{contains}(U, V) \wedge \text{is\_a}(V, \text{shop}) \wedge \text{is\_a}(V, \text{school}).$$

For each pattern  $P$ , a factor always exists. In the trivial case, it coincides with  $P$  up to a redenomination of variables in  $P$ . A factor  $P\theta$  is minimal, when there are no other factors of  $P$  with less literals than  $P\theta$ . Again, in the previous example,  $(\text{antecedent}(R_1) \wedge \text{antecedent}(R_2))\theta$  is minimal.

As stated previously, a straightforward application of the naïve Bayes independence assumption may result in totally unreliable probability estimates because of the presence of redundant literals. For this reason, we impose that  $P(s|c_i) = P(F|c_i)$  for any minimal factor  $F$  of  $\bigwedge_{R_k \in \mathfrak{R}_l(s)} \text{antecedent}(R_k)$ .

By separating the contribution of the conjunctions of literals corresponding to spatial relations ( $\text{relations}(F)$ ) from the contribution of the conjunction of literals corresponding to attributes ( $\text{attributes}(F)$ ) we have:

$$P(s|c_i) = P(\text{relations}(F)|c_i) \times P(\text{attributes}(F)|\text{relations}(F) \wedge c_i) \tag{3}$$

Under the naïve Bayes independence assumption,  $P(\text{relations}(F)|c_i)$  can be factorized as follows:

$$P(\text{relations}(F)|c_i) = \prod_{\text{rel}_j(A, B) \in \text{relations}(F)} P(\text{rel}_j(A, B)|c_i), \tag{4}$$

The estimation of  $P(\text{rel}_j(A, B)|c_i)$  is based on the substitution  $\theta$  used to compute the factor  $F$ . More precisely, for each rule  $R_k \in \mathfrak{R}_l(s)$  such that  $\text{rel}_j(A, B)\theta|_{R_k}^{-1} \in \text{antecedent}(R_k)$ , we compute the set  $\Sigma_k^i$  of substitutions  $\sigma$  such that  $\text{relations}(R_k)\sigma$  covers some reference object  $s \in S$  of class  $c_i$ . Then the cardinality of the set  $\bigcup_{R_k \in \mathfrak{R}_l(s)} \{\text{rel}_j(A, B)\theta|_{R_k}^{-1}\sigma | \sigma \in \Sigma_k^i\}$  represents the absolute frequency of  $\text{rel}_j(A, B)$  for the class  $c_i$ . The relative frequency is obtained by dividing this absolute frequency by the maximum cardinality of  $\text{rel}_j(A, B)$  for the class  $c_i$ , that is:

$$\hat{P}(\text{rel}_j(A, B)|c_i) = \frac{|\bigcup_{R_k \in \mathfrak{R}_l(s)} \{\text{rel}_j(A, B)\theta|_{R_k}^{-1}\sigma | \sigma \in \Sigma_k^i\}|}{\text{maxcard}(\text{rel}_j(A, B)|c_i)} \tag{5}$$

where  $\text{maxcard}(\text{rel}_j(A, B)|c_i) = |\bigcup_{R_k \in \mathfrak{R}_l(s)} \{\text{is\_a}(A, \text{type\_of}(A))\theta|_{R_k}^{-1}\sigma | \sigma \in \Sigma_k^i\}| \times |\bigcup_{R_k \in \mathfrak{R}_l(s)} \{\text{is\_a}(B, \text{type\_of}(B))\theta|_{R_k}^{-1}\sigma | \sigma \in \Sigma_k^i\}|$  and the functor  $\text{type\_of}(A)$  returns the name of the set of objects to which  $A$  belongs. It is noteworthy that  $\hat{P}(\text{rel}_j(A, B)|c_i)$  is never zero, since we assume that association rules returned by SPADA have a non-null support.

The naïve Bayes conditional independence can also be assumed for the computation of  $P(\text{attributes}(F)|\text{relations}(F) \wedge c_i)$ , in which case

$$\begin{aligned}
 P(\text{attributes}(F)|\text{relations}(F) \wedge c_i) &= \\
 &= \prod_{\text{attr}_j(A,v) \in \text{attributes}(F)} P(\text{attr}_j(A,v)|\text{relations}(F) \wedge c_i). \quad (6)
 \end{aligned}$$

Once again, the estimation of  $P(\text{attr}_j(A,v)|\text{relations}(F) \wedge c_i)$  is based on the substitution  $\theta$  used to compute the factor  $F$ . More precisely, for each rule  $R_k \in \mathfrak{R}_l(s)$  such that  $\text{attr}_j(A,v)\theta_{|R_k}^{-1} \in \text{antecedent}(R_k)$ , we compute the set  $\Sigma_k^i$  of substitutions  $\sigma$  such that  $(\text{relations}(R_k) \wedge \text{attr}_j(A,v))\sigma$  covers some reference object  $s \in S$  of class  $c_i$ . Then the cardinality of the set  $\bigcup_{R_k \in \mathfrak{R}_l(s)} \{\text{attr}_j(A,v)\theta_{|R_k}^{-1}\sigma | \sigma \in \Sigma_k^i\}$  represents the absolute frequency of  $\text{attr}_j(A,v)$ , given  $\text{relations}(F)$  and the class  $c_i$ . The relative frequency is obtained by dividing this absolute frequency by the maximum cardinality of  $\text{attr}_j(A,v)$  given  $\text{relations}(F)$  and the class  $c_i$ , that is:

$$\hat{P}(\text{attr}_j(A,v)|\text{relations}(F) \wedge c_i) = \frac{|\bigcup_{R_k \in \mathfrak{R}_l(s)} \{\text{attr}_j(A,v)\theta_{|R_k}^{-1}\sigma | \sigma \in \Sigma_k^i\}|}{|\bigcup_{R_k \in \mathfrak{R}_l(s)} \{\text{is\_a}(A, \text{type\_of}(A))\theta_{|R_k}^{-1}\sigma | \sigma \in \Sigma_k^i\}|} \quad (7)$$

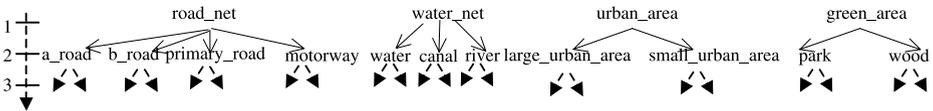
Also in this case  $\hat{P}(\text{attr}_j(A,v)|\text{relations}(F) \wedge c_i)$  cannot be null.

## 6 Experiments

We have evaluated both propositional and structural spatial associative classifiers in two real-world problems concerning North West England and Munich Census data.

### 6.1 Mining North West England census data

For this study we have considered both census and digital map data provided in the European project SPIN! (<http://www.ais.fraunhofer.de/KD/SPIN/project.html>). These data 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 (wards), for a total of 214 wards. Census data are available at the ward level and provide socio-economic statistics (e.g. mortality rate—the percentage of deaths with respect to the number of inhabitants) and measures describing the deprivation of each ward according to information provided by Census combined into single index scores. We consider Jarman Underprivileged Area Score which is designed to measure the need for primary care, the indices developed by Townsend and Carstairs which are used in health-related analyses, and the Department of the Environment’s (DoE) index which is used in targeting urban regeneration funds. The higher the index value the more deprived a ward is. Deprivation indices values and mortality rate are all numeric, but association rules discovery deals with discrete values corresponding to intervals determined by the context sensitive discretization. Jarman index, Townsend index, DoE index and Mortality rate have been discretized in (*low, high*), while Carstairs index has been discretized in (*low, medium, high*).



**Fig. 1** Spatial hierarchies defined for road net, water net, urban area and green area

By considering Greater Manchester wards as reference objects, we focus our attention on mining a classification model to predict discrete value of DoE index by exploiting not only socio-economic census factors but also geographical factors represented in some linked topographic maps. Spatial analysis is enabled by the availability of vectorized boundaries of the 1998 census wards as well as by other Ordinance Survey digital maps of NWE, where several interesting layers such as road net (1,687 lines), rail net (805 lines), water net (716 lines), urban area (115 lines) and green area (9 lines) are found. These additional layers are task-relevant objects of our analysis. Both ward-referenced census data and map data are stored in Oracle Spatial 9i database.

Topological relationships between reference objects and task-relevant objects have been extracted by means of the feature extraction module of ARES. Their semantics is based on the 9-intersection model (Egenhofer, 1991). In these experiments, the number of computed ‘non disjoint’ spatial relationships is 5,313 (13 wards-green areas, 2,798 wards-roads, 1,067 wards-waters, 381 wards-urban areas and 1,054 wards-rails). To support a spatial qualitative reasoning, a domain specific knowledge has been expressed in the *BK* as a set of Prolog rules. Some of them 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).
```

To mine multi-level association rules, we enrich *BK* with five hierarchies on the task-relevant objects (see Fig. 1). They have a depth three and are mapped into three granularity levels. A detailed description of this data setting is reported in Appice et al. (2003), Appice, Lanza, Malerba and Turi (2004b).

In all experiments we specify a language bias (*LB*) to constrain the search space and to filter out uninteresting spatial association rules, namely those including spatial relationships directly extracted by ARES (e.g. connects, inside, and so on), while we keep rules containing topological predicates defined by means of *BK* rules (i.e., *crossed\_by\_urban\_area*, *crossed\_by\_road\_net*, *crossed\_by\_rail\_net*, *crossed\_by\_green\_area* and *crossed\_by\_water\_net*).

Spatial association rules are then discovered at different levels of granularity according to hierarchies defined on task-relevant objects. These rules contain useful information about spatial patterns frequently occurring in data. For instance, by setting parameters *minsup* = 0.1 and *minconf* = 0.6 the following rule is discovered:

```
is_a(A, ward), ward_urban_area(A, B), is_a(B, urban_area),
jarman(A, low) => doe(A, low)(52.6%, 100%),
```

which states that a low DoE index value is observed in a ward *A* which includes an urban area *B* and has a low value of Jarman index. The support (52.6%) and the high confidence (100%) confirm an association between living in urban areas where primary care are well satisfied, and low level of deprivation according to DoE index.

**Table 1** Feature reduction factor (propositional approach)

minsup	minconf	<i>l</i>	<i>K</i>	Average no. of original features	Average percent (%) of feature reduction
0.2	0.8		5	27.5	37
0.2	0.8	1	6	221.1	44
0.2	0.8		7	982.3	73
0.2	0.8		5	492.8	29
0.2	0.8	2	6	278.3	11
0.2	0.8		7	1,860.1	66
0.1	0.6		5	233.2	48
0.1	0.6	1	6	374.0	57
0.1	0.6		7	904.2	74
0.1	0.6		5	442.2	28
0.1	0.6	2	6	635.8	27
0.1	0.6		7	3,051.4	68

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

$$\text{is\_a}(A, \text{ward}), \text{ward\_urban\_area}(A, B), \text{is\_a}(B, \text{large\_urban\_area}), \text{jarman}(A, \text{low}) \Rightarrow \text{doe}(A, \text{low})(52.6\%, 100\%),$$

This rule clarifies that the urban area *B* is large. By varying granularity level, minsup, minconf and number of refinement steps *K* (pattern length) in association rule discovery, we obtain several experimental settings. For each setting, classifiers performances have been evaluated by means of a 10-fold cross validation.

In the propositional approach, the set of discovered rules is transformed in a set of boolean features and redundant features are removed to reduce feature space without affecting the existence of a complete and consistent theory for each class label. The results on the average feature reduction percentage are reported in Table 1.

**Table 2** DoE index average accuracy

minsup	minconf	<i>l</i>	<i>K</i>	Average Accuracy				
				Propositional approach				Structural approach
				NBC	1-NN	C4.5	JRIP	
0.2	0.8		5	81.64	80.73	83.01	83.01	<b>87.5</b>
0.2	0.8	1	6	80.28	80.73	83.01	83.01	<b>87.5</b>
0.2	0.8		7	75.56	80.28	<b>83.01</b>	<b>83.01</b>	82.1
0.2	0.8		5	82.55	82.55	81.64	81.93	<b>90.3</b>
0.2	0.8	2	6	82.10	82.55	82.10	83.01	<b>88.3</b>
0.2	0.8		7	80.28	82.55	81.19	81.02	<b>87.4</b>
0.1	0.6		5	83.01	80.73	83.01	83.01	<b>91.2</b>
0.1	0.6	1	6	80.28	80.73	83.01	83.01	<b>91.2</b>
0.1	0.6		7	80.73	80.73	83.01	83.01	<b>91.2</b>
0.1	0.6		5	83.01	81.30	82.10	81.02	<b>91.2</b>
0.1	0.6	2	6	82.55	79.65	80.28	81.64	<b>91.2</b>
0.1	0.6		7	80.73	79.31	81.19	81.02	<b>91.2</b>

They confirm that high number of association rules typically lead to redundant features. Percentage of feature reduction increases when the number of refinement steps increases. This is due to the high number of similar rules which produce redundancy.

The average accuracy and running time of classifiers are reported in Table 2 and Table 3, respectively. Classification is evaluated by varying minsup, minconf and  $K$  for each setting. Several considerations can be drawn.

Firstly, both the propositional and the structural associative classifiers significantly improve the accuracy of the trivial classifier which returns the most probable class (acc. 0.625). Secondly, the average accuracies of classification models discovered at lower granularity levels (i.e.,  $l = 2$ ) are sometimes (always, in the structural approach) better than the corresponding accuracies at highest levels. This means that the classification model may take advantage of the use of the hierarchies defined on spatial objects. In this case, results at different abstraction levels provide insights on what are the task-relevant objects which affect the classification. For instance, when  $K = 7$ , minsup = 0.2 and minconf = 0.8, both propositional and structural naïve Bayesian associative classifiers are strongly improved when considering the size of the urban area, the type of road, and so on. Thirdly, results show that by decreasing the number of extracted rules (higher support and confidence) we have lower accuracy. This means that there are association rules with low values of support and confidence which strongly influence classification results. Fourthly, we observe that, in this specific task, the higher the number of refinement steps (rules involving more literals) not necessarily means the better the model. This is mainly due to the fact that rules become very specific. Fifthly, it is noteworthy that by considering different levels of granularity it is possible to find the best trade-off between bias and variance of the classifier (German, Bienenstock, & Doursat, 1992) where bias denotes the lack of fitting the data, while variance is the probability of assigning a class which is different from the Bayes optimal one. In particular, at higher levels of granularity we have higher bias and lower variance. On the contrary, at lower levels of granularity we have lower bias and higher variance. Finally, results clearly show that the structural associative classifier outperforms the propositional one. On

**Table 3** DoE Index average running times (s). Results are obtained on a Intel Centrino (1.3 GHz) PC running WinXP

min sup	min conf	$l$	$K$	Propositional-ization	Feature reduction	Classification				Structural approach
						Propositional approach				
						NBC	1-NN	C4.5	JRIP	
0.2	0.8		5	12.7	0.017	0.010	0.011	0.03	0.046	20.1
0.2	0.8	1	6	10.10	0.016	0.011	0.010	0.029	0.041	28.3
0.2	0.8		7	40.7	0.030	0.010	0.027	0.035	0.047	78.0
0.2	0.8		5	25.1	0.025	0.010	0.018	0.049	0.066	56.3
0.2	0.8	2	6	24.0	0.018	0.010	0.014	0.043	0.064	56.8
0.2	0.8		7	102.5	0.041	0.013	0.030	0.064	0.075	343.8
0.1	0.6		5	11.3	0.016	0.010	0.015	0.029	0.041	21.5
0.1	0.6	1	6	16.2	0.022	0.017	0.011	0.033	0.046	35.3
0.1	0.6		7	40.6	0.028	0.011	0.013	0.032	0.045	105.6
0.1	0.6		5	21.6	0.023	0.011	0.024	0.028	0.071	81.3
0.1	0.6	2	6	27.9	0.029	0.013	0.020	0.048	0.128	120.4
0.1	0.6		7	160.7	0.064	0.024	0.046	0.087	0.120	757.2

the other hand, our propositional approach outperforms the structural one when considering efficiency. This depends on the reduction of the search space in the mining process.

## 6.2 Mining Munich census data

In this study we analyze the monthly rent per square meter of flats in Munich. We consider data collected in 1998 by Infratest Sozialforschung to develop the Munich rental guide in 1999. The dataset contains 2,180 flats geo-referenced within Munich subquarters, where Munich metropolitan area is divided into three areal zones, each of which is decomposed into 64 districts, for a total of 446 subquarters. Both vectorized boundaries of subquarters, districts and zones and map of public train stops (56 subway stops, 15 rapid train stops and 1 railway station) within Munich are stored in an Oracle 9i spatial database ([http://www.di.uniba.it/%7Ececci/micFiles/munich\\_db.tar.gz](http://www.di.uniba.it/%7Ececci/micFiles/munich_db.tar.gz)). A taxonomic relation allows us to describe the metropolitan areas hierarchy: *subquarters*  $\rightarrow$  *districts*  $\rightarrow$  *zone* according to a containment relation. Similarly, public train stops are organized according to a hierarchy representing stops at different abstraction levels.

Flats are described by the “monthly rent per square meter” (in German marks), the “floor space in square meters” and the “year of construction”. In addition, aggregated information on flats is available at subquarter level, that is, the percentage of flats having an extension within the intervals [19.0..69.0], [69.0..150.0] and [150.0..196.0] and the percentage of flats whose year of construction is within the years intervals [1800..1850], [1850..1958] and [1958..1992]. Numeric thematic attributes and spatial properties (area) are discretized by ARES. In our analysis, flats represent reference objects, while Munich metropolitan areas and public train stops play the role of task-relevant objects. The phenomenon under observation is the “monthly rent per square meter” which, after discretization, can be either low = [2.0..14.0] or high = [14.0..35.0]. The spatial arrangement of data is explicitly expressed by extracting both the *close\_to* relation between Munich metropolitan areas (autocorrelation) and the *inside* relation between public train stops and metropolitan areas.

As in NWE Census data, propositional and structural associative classifiers are evaluated by means of a ten-fold cross validation. Classification models are built at multiple granularity levels by exploiting hierarchies on both Munich metropolitan areas and public train stops. Several experimental settings are obtained by varying the granularity level ( $1 \leq l \leq 3$ ) and number of refinement steps  $K$  ( $3 \leq K \leq 7$ ) and setting  $\text{minconf}[l] = 0.2$ ,  $1 \leq l \leq 3$  and  $\text{minsup}[1] = 0.17$ ,  $\text{minsup}[2] = 0.15$ ,  $\text{minsup}[3] = 0.015$ . An example of a spatial association rule discovered at granularity level 1 with  $K = 7$  is:

$$\begin{aligned} & \text{in\_apartment}(A), \text{inside}(A, B), \text{is\_a}(B, \text{subquarter\_in\_munich}), \text{close\_to}(B, C), \\ & \text{is\_a}(C, \text{subquarter\_in\_munich}), \text{extension\_245\_255}(C, [0.0..0.166]) \\ & \Rightarrow \text{rent}(A, \text{low})(59.71\%, 60.36\%). \end{aligned}$$

Such rule states that a low rent value is observed for a flat  $A$  which is located inside a subquarter  $B$  of the Munich area.  $B$  is close to a subquarter  $C$  having a percentage of flats with extension between 245 and 255 m<sup>2</sup> lower than 16%. The support (59.71%) and the relatively high confidence (60.36%) confirm that flat rent

**Table 4** Flat rent average accuracy

<i>l</i>	<i>K</i>	Average Accuracy				
		Propositional approach				Structural approach
		NBC	1-NN	C4.5	JRIP	
1	3	64.57	64.57	64.57	64.57	<b>67.66</b>
	4	64.57	64.57	64.57	64.57	<b>71.74</b>
	5	62.50	64.20	64.38	63.83	<b>70.06</b>
	6	61.63	64.84	64.75	63.92	<b>69.60</b>
	7	59.29	65.62	66.45	64.89	<b>69.75</b>
2	3	64.57	64.57	64.57	64.57	<b>67.66</b>
	4	64.57	64.38	64.57	64.57	<b>71.74</b>
	5	55.48	60.80	60.16	60.66	<b>62.51</b>
	6	49.38	<b>60.76</b>	60.07	59.84	59.30
	7	49.42	<b>62.27</b>	62.18	61.31	59.35
3	3	62.84	62.84	62.84	62.84	<b>67.66</b>
	4	53.78	52.98	54.01	54.12	<b>71.18</b>
	5	50.00	52.17	51.94	52.06	<b>64.29</b>
	6	49.88	52.63	52.86	51.60	<b>64.29</b>
	7	–	–	–	–	–

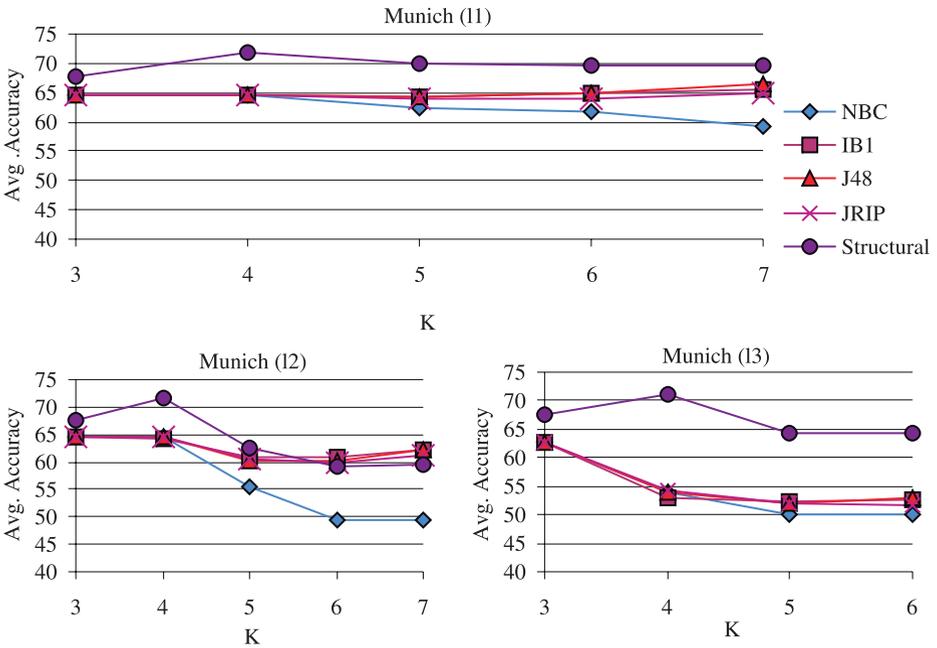
can be affected by the neighborhood. At higher granularity level ( $l = 2$ ), SPADA specializes the task-relevant objects  $B$  and  $C$  by generating the rule:

in\_apartment( $A$ ), inside( $A, B$ ), is\_a( $B$ , subquarter\_in\_middle\_zone),  
 close\_to( $B, C$ ), is\_a( $C$ , subquarter\_in\_middle\_zone),  
 extension\_245\_255( $C$ , [0.0..0.166])  $\Rightarrow$  rent( $A$ , low)(16.31%, 39.5%).

Spatial association rules discovered for this task also includes some characterization of subquarters within the Munich area according to the presence of some public train stops. For example, at granularity level 1, SPADA has discovered the spatial association rule:

in\_apartment( $A$ ), inside( $A, B$ ), is\_a( $B$ , subquarter\_in\_munich), inside( $C, B$ ),  
 is\_a( $C$ , public\_train\_stop), year\_1800\_1850( $B$ , [0.0..0.2])  
 $\Rightarrow$  rent( $A$ , [2.0..14.0])(19.63%, 57.8%).

Concerning classification results, the average accuracy is reported in Table 4. No classifier has been built when  $l = 7$  and  $K = 3$ , since SPADA was not able to discover any strong spatial association rule. Looking at the results which someways were unexpected, we observe that when the number of refinement steps is low, the accuracy remains constant varying the granularity level. When the number of refinement steps is higher, the accuracy even decreases by increasing the granularity level. A possible explanation for this phenomenon can be found in the “curse of dimensionality” which requires the number of samples per feature to increase exponentially with the number of features to maintain a given level of accuracy (Bellman, 1961). In fact, in this dataset, autocorrelation increases the number of features considered in the learning of a classifier by keeping constant the number of examples. Finally, as noted in NWE Census dataset, results clearly show that the structural associative classifier outperforms the propositional one. This confirms that structural approaches are more powerful than propositional ones since information about how data were originally structured is not lost (Fig. 2).



**Fig. 2** Average accuracy of propositional and structural classifiers at different granularity levels

### 7 Conclusions

In this work, we have compared propositional and structural approaches to spatial classifications in multi-relational data mining. Both approaches are investigated in the context of the associative classification framework which combines spatial association rules discovery and classification by taking advantage of employing association rules for classification purposes. In particular, association rules which contain only the class label in the consequent are discovered at multiple levels of granularity. In the propositional approach, they are transformed into a set of boolean features to be used in the classical attribute-value classification. Differently, in the structural approach, this set of rules contributes to estimate the posterior probability. Results show that the use of different levels of granularity permits us to find the best trade-off between bias and variance in spatial classification. Moreover, they confirm that spatial data mining applications benefit by preserving information about how data were originally structured in the spatial domain. This justifies the improved accuracy of the structural approach with respect to the propositional one. On the contrary, the propositional approach outperforms the structural one in terms of efficiency due to the reduction of the search space.

For future work, we intend to frame the proposed classification methods within the context of hierarchical Bayesian classifiers, in order to exploit multi-level nature of extracted association rules. Associative classification can be further extended in order to deal with temporal data dimension in addition to the spatial one. Finally, the associative classification proposed in this work uses simple statistics such as support and confidence to filter out uninteresting patterns and rules. Such measures do not

guarantee that discovered patterns are relevant to discriminate among separate classes. An alternative is posed by emerging patterns (Dong & Li, 1999), that is, patterns capturing multi-attribute contrasts between separate data classes and using such patterns either in the propositionalization step or structural classification.

**Acknowledgements** This work is partial fulfillment of the research objective of ATENEO-2005 project “Gestione dell’informazione non strutturata: modelli, metodi e architetture”. We are grateful to Paoline Buts for her help in re-reading the first draft of the paper.

## References

- Aiello, M. (2001). *Spatial reasoning: theory and practice*. PhD thesis, University of Amsterdam, Holland.
- Appice, A., Ceci, M., Lanza, A., Lisi, F. A., & Malerba, D. (2003). Discovery of spatial association rules in georeferenced census data: A relational mining approach. *Intelligent Data Analysis*, 7(6), 541–566.
- Appice, A., Ceci, M., Rawles, S., & Flach, P. A. (2004a). Redundant feature elimination for multi-class problems. In Greiner, R. & Schuurmans, D. (Eds.), *Proceedings of the 21st international conference on machine learning*, (pp. 33–40). New York: ACM.
- Appice, A., Lanza, A., Malerba, D., & Turi, A. (2004b). Mining spatial association rules from census data with ARES. In May, M. & Malerba, D. (Eds.), *Notes of the KdNet workshop symposium knowledge-based services for public sector: mining official data*.
- Baralis, E., & Garza, P. (2003). Majority classification by means of association rules. In Lavrac, N., Gamberger, D., Todorovski, L., & Blockeel, H. (Eds.), *Proceedings of the 7th European conference on principles and practice of knowledge discovery in databases*, volume 2838 of *LNAI*, (pp. 35–46). Berlin Heidelberg New York: Springer.
- Bellman, R. E. (1961). *Adaptive control processes*. New Jersey: Princeton University Press.
- Blockeel, H. (1998). *Top-down induction of first order logical decision trees*. PhD thesis, Department of Computer Science, Katholieke Universiteit, Leuven, Belgium.
- Ceci, M., Appice, A., & Malerba, D. (2003). Mr-SBC: A multi-relational naive bayes classifier. In Lavrac, N., Gamberger, D., Todorovski, L. & Blockeel, H. (Eds.), *Proceedings of the 7th European conference on principles and practice of knowledge discovery in databases*, volume 2838 of *LNAI*, (pp. 95–106). Berlin Heidelberg New York: Springer.
- Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the 12th international conference on machine Learning*, (pp. 115–124).
- Dehaspe, L. & Toivonen, H. (1999). Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery*, 3(1), 7–36.
- Dehaspe, L. & Toivonen, H. (2000). *Relational data mining*, Discovery of relational association rules (chapter), (pp. 189–208). Berlin Heidelberg New York: Springer.
- Domingos, P. & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under Zero-Ones loss. *Machine Learning*, 28(2–3), 103–130.
- Dong, G. & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *Knowledge discovery and data mining*, (pp. 43–52). New York: ACM.
- Džeroski, S. & Lavrač, N. (2001). *Relational data mining*. Berlin Heidelberg New York: Springer.
- Egenhofer, M. J. (1991). Reasoning about binary topological relations. In *Proceedings of the 2nd symposium on large spatial databases*, (pp. 143–160). Zurich, Switzerland.
- Ester, M., Kriegel, H., & Sander, J. (1997) Spatial data mining: A database approach. In *Proceedings international symposium on large databases*, (pp. 47–66).
- Fitzpatrick, J. (2001). *Geographic variations in health*. London: The Stationery Office.
- Flach, P. & Lachiche, N. (2004). Naive bayesian classification of structured data. *Machine Learning*, 57(3), 233–269.

- German, S., Bienenenstock, E., & Doursat, R. (1992) Neural networks and the bias/variance dilemma. 4, 1–58.
- Getoor, L. (2001). Multi-relational data mining using probabilistic relational models: research summary. In Knobbe, A. & Van der Wallen, D. M. G. (Eds.), *Proceedings of the 1st workshop in multi-relational data mining*, Freiburg, Germany.
- Knobbe, J., Haas, M., & Siebes, A. (2001) Propositionalisation and aggregates. In Raedt, L. D. and Siebes, A. (Eds.) *Proceedings of PKDD 2001*, volume 2168 of *LNAI*, (pp. 277–288). Berlin Heidelberg New York: Springer.
- Koperski, K. (1999) *Progressive refinement approach to spatial data mining*. PhD thesis, Computing Science, Simon Fraser University, British Columbia, Canada.
- Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. In *Proceedings of the 4th international symposium on large spatial databases: Advances in spatial databases*, LNCS, volume 951, 47–66. Berlin Heidelberg New York: Springer.
- Kramer, S. (1999). *Relational learning vs. propositionalization: Investigations in inductive logic programming and propositional machine learning*. PhD thesis, Vienna University of Technology, Vienna, Austria.
- Kramer, S., Lavrač, N., & Flach, P. (2001). *Relational data mining*, Propositionalization approaches to relational data mining, LNAI. (pp. 262–291). Berlin Heidelberg New York: Springer.
- Krogl, M. A. (2005). *On propositionalization for knowledge discovery in relational databases*. PhD thesis, Fakultät für Informatik, Germany.
- Krogl, M., Rawles, S., Zelezny, F., Flach, P., Lavrač, N., & Wrobel, S. (2003). Comparative evaluation of approaches to propositionalization. In Horvath, V. and Yamamoto, A. (Eds.), *Proceedings of international conference on inductive logic programming*, volume 2835 of *LNAI*, (pp. 197–214). Berlin Heidelberg New York: Springer.
- Lavrač, N., & Džeroski, S. (1994). *Inductive logic programming: Techniques and applications*. Chichester, UK: Ellis Horwood.
- Leiva, H. A. (2002). *MRDTL: A multi-relational decision tree learning algorithm*. Master's thesis, University of Iowa, USA.
- Li, W., Han, J., & Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple class-association rules. In *ICDM*, (pp. 369–376) San Jose, California.
- Lisi, F. A., & Malerba, D. (2004). Inducing multi-level association rules from multiple relations. *Machine Learning*, 55, 175–210.
- Liu, B., Hsu, W., & Ma, Y. (1998). Integrative classification and association rule mining. In *Proceedings of AAAI Conference of knowledge discovery in databases*.
- Ludl, M. C., & Widmer, G. (2000). Relative unsupervised discretization for association rule mining. In Zighed, D. A., Komorowski, H. J. & Zytkow, J. M. (Eds.), *Proceedings of the 4th European conference on principles of data mining and knowledge discovery*, volume 1910 of *LNCS*, (pp. 148–158). Berlin Heidelberg New York: Springer.
- Malerba, D., Esposito, F., Lanza, A., Lisi, F. A., & Appice, A. (2003). Empowering a gis with inductive learning capabilities: The case of ingens. *Journal of Computers, Environment and Urban Systems*, Elsevier Science, 27, 265–281.
- Mannila, H. & Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3), 241–258.
- Mitchell, T. (1997). *Machine learning*. New York: McGraw Hill.
- Modrzejewski, M. (1993). Feature selection using roughsets theory. In *Proceedings of the European conference on machine learning*, (pp. 213–226). Berlin Heidelberg New York: Springer.
- Morimoto, Y. (2001). Mining frequent neighboring class sets in spatial databases. In *Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 353–358).
- Muggleton, S. (1995). Inverse entailment and progol. *New Generation Computing, Special issue on Inductive Logic Programming*, 13(3-4), 245–286.
- Pazzani, M., Mani, S., & Shankle, W. (1997). Beyond concise and colorful: Learning intelligible rules. In *Proceedings of the 4th international conference on knowledge discovery and data mining*, (pp. 235–238). Menlo Park, California: AAAI.
- Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Mateo, California: Morgan Kaufmann.
- Robinson, J. A. (1965). A machine oriented logic based on the resolution principle. *Journal of the ACM*, 12, 23–41.
- Sharma, J. (1996). *Integrated spatial reasoning in geographic information systems: Combining topology and direction*. PhD thesis, University of Maine, Bangor, Maine.

- Shekhar, S., Schrater, P. R., Vatsavai, R., Wu, W. & Chawla, S. (2002). Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia*, 4(2), 174–188.
- Srinivasan, A. & King, R. (1996). Feature construction with inductive logic programming: A study of quantitative predictions of biological activity aided by structural attributes. In Muggleton, S. (Ed.), *Proceedings of the 6th international workshop on inductive logic programming*, (pp. 352–367). Stockholm University, Royal Institute of Technology, Stockholm, Sweden.
- Wang, S. S., Liu, D. Y., Wang, X. Y. & Liu, J. (2006). Spatial reasoning based spatial data mining for precision agriculture. In *Proceedings of APWeb workshops 2006*, (pp. 506–510).
- Yin, X. & Han, J. (2003). CPAR: Classification based on predictive association rules. In *SIAM International conference on data mining*.