

Propositionalization Through Relational Association Rules Mining

Michelangelo Ceci Annalisa Appice Donato Malerba

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

Abstract. In this paper we propose a novel (multi-)relational classification framework based on propositionalization. Propositionalization makes use of discovered relational association rules and permits to significantly reduce feature space through a feature reduction algorithm. The method is implemented in a Data Mining system tightly integrated with a relational database. It performs the classification at different granularity levels and takes advantage from domain specific knowledge in form of rules that support qualitative reasoning. An application of classification in real-world geo-referenced census data analysis is reported.

1 Introduction

Classification is one of the fundamental tasks in data mining and in the traditional data mining classification setting [15] data are generated independently and with an identical unknown distribution P on some domain \mathbf{X} and are labeled according to an unknown function g . The domain of g is spanned by m independent (or predictor) random variables X_i (both numerical and categorical), that is $\mathbf{X} = X_1 \times X_2 \times \dots \times X_m$, while the range of g is a finite set $Y = \{C_1, C_2, \dots, C_L\}$, where each C_i is a distinct class. A data mining algorithm takes a training sample $S = \{(\mathbf{x}, y) \in \mathbf{X} \times Y \mid y = g(\mathbf{x})\}$ as input and returns a function f which is hopefully close to g on the domain \mathbf{X} . In practice, the goal is to learn the target concept associated with each class by finding regularities in examples of a class that characterize the class in question and discriminate it from the other classes. This formalization of the classification problem well faces with the *single-table assumption* [17] underlying most of classification methods proposed in the literature. Data are represented as a single table, where each row corresponds to an example and each column to a predictor variable or to the *target* variable Y . However, the growing importance of knowledge discovery and data mining in practical real world requires increasingly sophisticated solutions for classification problems when data consists of a large amount of records that may be stored in *several tables* of a relational database. In this context, single table assumption appears too restrictive since both the predictor variables and the target variable are represented as attributes of distinct tables (relations).

(Multi-) relational data mining [16] is the branch of data mining research that overcomes limitations imposed by single table assumption and investigates methods

able to extract knowledge from data stored in multiple relational tables. (Multi-) relational data mining methods are typically based on two alternative approaches: a structural approach and a propositional approach. While in the first case the whole hypothesis space is directly explored by the mining method, in the second case, known as *propositionalisation*, a transformation of relational learning problems into attribute-value representations amenable for conventional data mining methods is performed. In principle, methods implementing structural approaches are more powerful than methods implementing propositional approaches since information about how data were originally structured is not lost. Nevertheless, approaches to (multi-)relational data mining based on propositionalisation have gained significant interest in the last few years. This is mainly due to the fact that in many practical cases propositionalisation allows the reduction of the search space to a minimal subset including features obtained as transformation of the original (multi-)relational feature space. Another reason is given by the observation that the transformation of an original (multi-)relational problem into a single table format allows one to directly apply conventional propositional data mining methods, thus making a wider choice of robust and well-known algorithms available [9].

Several multi-relational data mining methods based on propositionalisation have been proposed in literature for classification tasks. Generally, they assume an individual-centered data transformation such that there is a one-to-one correspondence between each tuple in the original target table (table containing the target attribute) and each tuple in the single table obtained after the propositionalisation process. At now, multi-relational classification through propositionalization problems has been extensively investigated by resorting to the field of Inductive Logic Programming (ILP). For instance, LINUS [10] as well as its successor SINUS [9] and ST [8] are able to construct a set of boolean features defined in terms of conjunctions of literals starting from relational data expressed in some first-order formalism. Nevertheless, a different approach to propositionalisation is supported by systems that directly work with relational databases. Indeed, they generally construct a single central relation by simply summarizing and/or aggregating information found in other tables and performing join operations according to foreign key constraints [7]. In both cases, propositional features are constructed by considering the structure imposed on data by relationships (e.g. foreign key constrains) between tuples in the target table and tuples in related tables, while ignoring the structure eventually imposed by hierarchical relationships on tuples of the same table. This means that these methods are not able to derive a propositional description of the same individual exploiting different levels of granularity. This turns to be an interesting aspect in many applications that would benefit from concept hierarchies [6]. Another limitation in the existing approaches is the huge number of new features they produce. In particular, propositionalisation tends to produce large numbers of features, many of which are highly correlated or even logically redundant. A simple example of a redundant feature is one which is never (or always) satisfied: e.g., ‘a molecule having an atom which has a bond with itself’. While some forms of redundancy can be recognised at feature generation time, others can only come to light by examining the data [2].

In this paper we propose a novel multi-relational propositionalisation-based classification framework that makes use of discovered multi-level association rules in

the propositionalisation step. Association rules are extracted by means of a (multi-) relational association rule discovery system embedded in the framework that is able to generate association rules at multiple levels of granularity according to some domain knowledge. This allows the classification at different levels of granularity by taking advantage from a qualitative reasoning on a domain specific knowledge expressed in form of rules. Discovered rules are subsequently used to create a relational table for each granularity level where each column represent a boolean feature. Moreover, a feature reduction algorithm has been integrated to remove redundant features and improve efficiency of classification step without affecting accuracy of classifier. Finally, four well-known data mining algorithms are applied to the resulting table for classification purposes.

The paper is organized as follows. In the next section we present the architecture of the proposed framework. In section 3 we briefly present the process of mining of multi-level relational association rules while the propositionalization approach is described in Section 4. Finally, an application to in real-world geo-referenced census data analysis is reported in Section 5 and some conclusions are drawn.

2 System Architecture

The framework we present in this paper has been implemented in the system MSRC (Multi-Step Relational Classifier).

The problem solved by MSRC can be formalized as follows:

Given:

- a training set represented by means of h relational tables $S=\{T_0, T_1, \dots, T_{h-1}\}$ of a relational database D ;
- a set of primary and foreign key constraints on tables in S ;
- a target relation (table) $T \in S$;
- a target discrete attribute y in T , different from the primary key of T ;
- a background knowledge BK including some *hierarchies* $H_k, k \in \{j \in \mathbb{N} \mid 0 \leq j \leq h-1 \wedge T_j \neq T\}$ (one for each table $T_k \in S - \{T\}$) on tuples in $S - \{T\}$. Hierarchies H_k define *is-a* (i.e., taxonomical) relationship of objects represented as tuples in the table T_k ;
- M *granularity levels* in the descriptions (1 is the highest while M is the lowest),
- a set of *granularity assignments* ψ_k which associate each object in H_k with a granularity level,
- a couple of thresholds $minsup[l]$ and $minconf[l]$ for each granularity level.

Find multi-level classifiers which predict the value of y for some individual represented as a tuple in T (with possibly UNKNOWN value for y) and related tuples in S according to foreign key constraints.

In Figure 1, the architecture of the multi-relational propositionalization-based classifier MSRC is presented.

The *MRSC Engine* manages the system by allowing user interaction and by coordinating the activity of all other components. It interfaces the system with the database module in order to store intermediate information.

The *Association rules discovery module* is in charge of the extraction of association rules from data. In particular, given a database schema, a background knowledge and

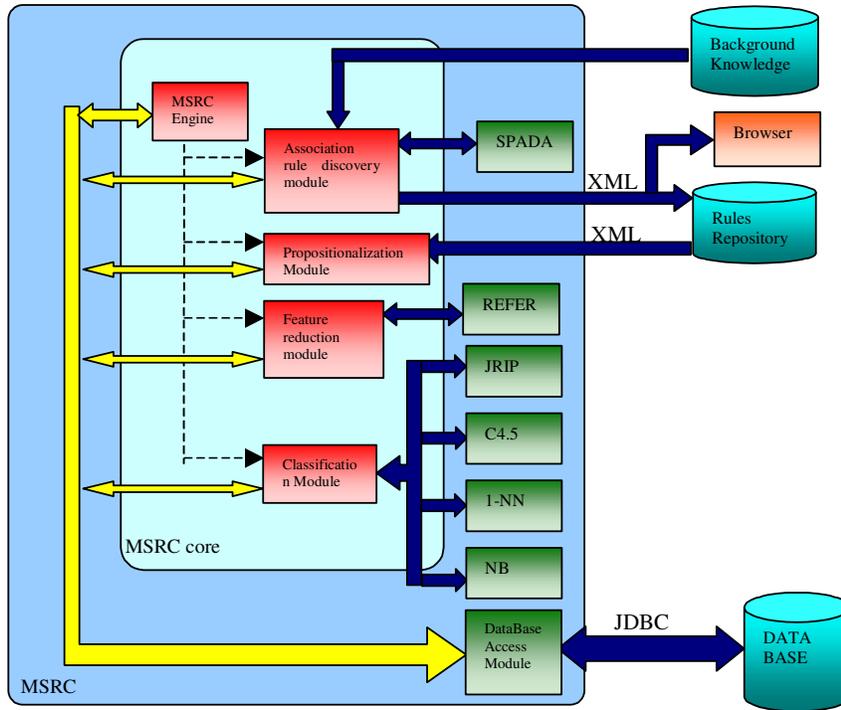


Figure 1. MSRC architecture.

a Language Bias, the system is able to extract association rules and store them in an appropriate repository for subsequent use. The system embeds, for this task, SPADA [1] that mines association rules at multiple levels of granularity. The expressive power of the language bias of SPADA allows us to filter out rules that are not useful for classification purposes (see section 3).

The *Propositionalization Module* is in charge of the transformation of the tables in S in a single relational table. In the resulting table, columns represent the extracted rules, rows represent instances of the target table T and values are obtained by means of the propositionalization step (see section 4).

The *Feature Reduction Module* reduces the number of columns in order to remove redundant features that would uselessly increase the complexity of the learning task and negatively affect the classification effectiveness (see section 4).

Finally, the *Classification Module* is in charge of executing both the learning process and the classification process. The learning process takes as input the reduced relational table and returns a classification model. The classification process, takes in input the classification model and a tuple in T (with possibly UNKNOWN value for y) and returns the value of y . The classification model can be computed by means of four different learning algorithms, namely the JRIP[3], C4.5 [16], 1-NN and the Naive Bayesian classifier (NBC) [4].

In the following Sections, we first introduce the problem of discovering multi-level relational association rules and then we describe how these rules can be used to construct boolean features to be used in the classification step.

3 Relational Association rules Mining.

The discovery of relational association rules is a descriptive mining task that aims at detecting associations between *target objects* (represented as tuples of the target table T) and some *target-relevant objects* (represented as tuples in $S\text{-}\{T\}$). The former are the main subject of the description, while the latter are objects that are relevant for the task in hand and are related (by means of foreign key constraints) to the former.

In general, association rules are a class of regularities that can be expressed by the implication: $P \Rightarrow Q (s, c)$, where P (body) and Q (head) are a set of literals, called *items*, such that $P \cap Q = \emptyset$, the support s estimates the probability $p(P \cup Q)$, and the confidence c , estimates the probability $p(Q \mid P)$. The conjunction $P \wedge Q$ is called *pattern*. An example of relational association rule is:

$$is_a(X, molecule), contains(X, Y), is_a(Y, atom), charge_of(Y, [0.7... 0.88]) \Rightarrow contains(X, Z), is_a(Z, atom), charge_of(Z, [0.2 ... 0.3]), Z \neq Y (91\%, 100\%)$$

to be read as “If a molecule X contains an atom with charge between 0.7 and 0.88, then X contains another atom Z distinct from Y with charge between 0.2 and 0.3 with 91% support and 100% confidence.” By taking into account some kind of taxonomic knowledge on target-relevant objects (e.g. atom is specialized as atom_A and atom_B) it is possible to obtain descriptions at different granularity levels (*multiple-level association rules*). For instance, a finer-grained association rules can be the following:

$$is_a(X, molecule), contains(X, Y), is_a(Y, atom_A), charge_of(Y, [0.7... 0.88]) \Rightarrow contains(X, Z), is_a(Z, atom_B), charge_of(Z, [0.2 ... 0.3]), Z \neq Y (65\%, 85\%)$$

In this context, the process of discovery multi-level relational association rules correspondes to mine strong multi-level relational association rules, that is, relational association rules involving target-relevant objects at different granularity levels.

Hierarchies H_k define *is-a* (i.e., taxonomical) relationship of objects in the same table (e.g. atom-A *is-a* atom). Objects of each hierarchy are mapped to one or more of the M user-defined description granularity levels in order to deal uniformly with several hierarchies at once. Both frequency of patterns and strength of rules depend on the granularity level l at which patterns/rules describe data. Therefore, a pattern $P (s\%)$ at level l is *frequent* if $s \geq minsup[l]$ and all ancestors of P with respect to H_k are frequent at their corresponding levels. An association rule $Q \rightarrow R (s\%, c\%)$ at level l is *strong* if the pattern $Q \cup R (s\%)$ is frequent and $c \geq minconf[l]$.

In MSRC, rules are extracted by means of the algorithm SPADA [12] that operates in three steps for each granularity level: i) pattern generation; ii) pattern evaluation; iii) rule generation and evaluation. SPADA takes advantage of statistics computed at granularity level l when computing the supports of patterns at granularity level $l+1$. In particular, SPADA exploits the expressive power of first-order logic to specify both the background knowledge BK, such as hierarchies and domain specific knowledge, and the language bias LB. Hierarchies allow to represent and manage objects at different levels of granularity, while the domain specific knowledge stored as a set of rules supports qualitative reasoning. On the other hand, the LB is relevant to allow data miners to specify his/her bias for interesting solutions, and then to exploit this bias to improve both the efficiency of the mining process and the quality of the

discovered rules. In SPADA, the language bias is expressed as a set of constraint specifications for either patterns or association rules. Pattern constraints allow specifying a literal or a set of literals that should occur one or more times in discovered patterns. During the rule generation phase, patterns that do not satisfy a pattern constraint are filtered out. Similarly, rule constraints are used to specify literals that should occur in the head or body of discovered rules. In a more recent release of SPADA (3.1) [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 (head or body of a rule). An additional rule constraint has been introduced to eventually specify the maximum number of literals to be included in the head of a rule. In this way we are able to constraint the head structure of a rule requiring the presence of only the literal representing the class label and obtain patterns useful for classification purposes.

4 Transforming Association Rules in boolean features.

Once relational association rules having only the class label in the head have been extracted for each level, they are converted in a set of boolean features such that the result can be used as input for attribute-value classification algorithms. More precisely, for each granularity level l , boolean features construction is performed by transforming the original set of tables S in a single relational table B whose columns correspond to each body of association rules discovered at level l . Before describing how B is obtained, we introduce:

- R' : a set of relational association rules whose head contains a literal representing the class label;
- $I_i \in T$: a target object instance (tuple of T);
- B : the output table with $|R'|+1$ attributes. Tuples in B correspond to tuples in T .

The instance I_i can be logically represented as a set of ground facts describing both the target tuple in T and all target-relevant tuples that are related (e.g. foreign key associated) to the target table (see Figure 2). The body of a rule $R_j \in R'$ covers $I_i \in T$ if there exists a substitution θ , such that $body(R_j)\theta \subseteq I$.

In the case that the body of the rule R_j covers I_i , the j -th value of the tuple in B associated to the i -th instance of T is *true*, otherwise it is *false*. The $(|R'|+1)$ -th column in B represents the target attribute y .

It is noteworthy that the number of attributes of B depends on the number of discovered association rules in R' . Since, the number of discovered association rules is usually high and many rules are strongly correlated, this may lead to generate boolean features which are highly correlated or even logically redundant.

The definition of feature redundancy we adopt in this work properly follows from reducts in rough sets theory [14], in which boolean features are redundant if their removal does not change the set of example-pairs having the same value for each feature. Hence, a feature f can be identified as *redundant* with respect to another feature g for distinguishing *positive* from *negative* examples of a class c if f is true for at least the same positive examples as g , and false for at least the same negative examples as g . Coherently with this definition, we may combine the association rule based propositionalisation with redundant feature elimination (feature reduction) in

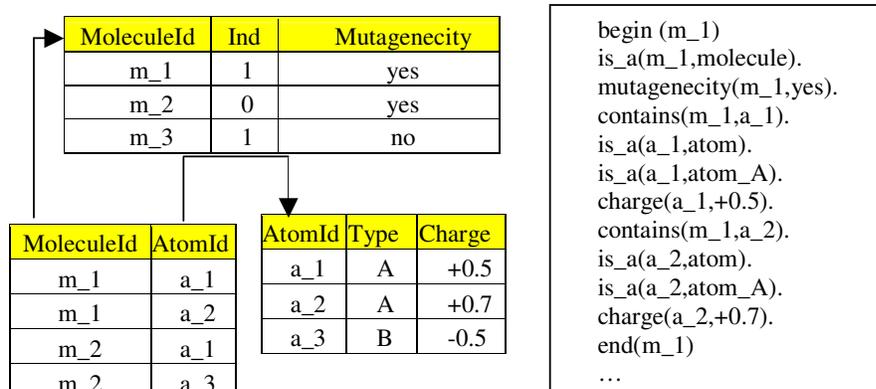


Figure 2. Set of ground atoms representing relational data concerning molecules (target objects) and atoms (target-relevant objects). In this case a two level hierarchy is defined on atom, that is, atom is specialized as atom_A and atom_B.

order to reduce the hypothesis space by excluding boolean features (or literals in ILP) which are redundant for learning. For this purpose, we integrate REFER [2], that is, a feature reduction method for multi-class problems that operates in an ILP setting.

Central to REFER is the notion of a *neighborhood* that is a subset of examples belonging to the same class such that the number of different features between examples is relatively small. Each neighbourhood is uniquely identified by two examples. The first example is where the neighbourhood construction starts and the second one is the termination point. Let e_s be a random starting example for the construction of a neighbourhood, the corresponding termination point is the closest example in e_t tagged with a different class label, referred to here as the *point* (or *example*) of *class change*. The neighbourhood $E(e_s, e_t)$ contains the set of training examples $e_{s_1}, e_{s_2}, \dots, e_{s_k}$ such that:

$$class(e_s) = class(e_{s_1}) = \dots = class(e_{s_k}), \text{ distance}(e_s, e_{s_h}) \leq \text{distance}(e_s, e_t) \quad \forall h=1, \dots, k$$

where the distance between two examples is computed as the *Hamming distance*, that is, the number of features whose values differ between the two examples. The neighbourhood construction proceeds in $E \setminus E(e_s, e_t)$ by considering the last example of class change as the current starting point and the process is repeated until the entire set of training examples is partitioned in neighbourhoods.

Redundant features are then eliminated by applying a revised version of REDUCE [11] method to each pair of neighbourhoods of different class. In particular, let E_l and E_m be a pair of neighbourhoods in E of different classes ($C_l \neq C_m$), the goal is to detect which features $f \in B$ describing examples in $E_l \cup E_m$ are non-redundant for discriminating between the classes C_l and C_m . We thereby eliminate features estimated redundant according to properties and dependencies among the features. Formally, a feature $f \in F$ covers a feature $g \in F$ with respect to $E_l \cup E_m$ if $T(g) \subseteq T(f)$ and $F(g) \subseteq F(f)$, where $T(f)$ ($T(g)$) is the set of all examples $e_i \in E_l$ such that f (g) has the value *true* for e_i ; and $F(f)$ ($F(g)$) is the set of all example $e_j \in E_m$ such that f (g) has the value *false* for e_j . The intuition is that a feature f is better than another feature g for distinguishing C_l from C_m if f is true for at least the same C_l examples as g , and false for at least the same C_m examples as g . The implicit assumption is that class C_l is the

positive class we are trying to describe. This suggests the notion of useless features, those for which $T(f) = \emptyset$ or $F(g) = \emptyset$. Such features can be immediately removed from the set of features F regardless of the properties of other features. Furthermore, a feature $g \in F$ is considered as a redundant feature to be removed if there exists another feature $f \in F$ ($f \neq g$) such that f covers g .

Since, it is proved that REFER preserves the existence of a complete and consistent theory for each class label when eliminating redundant features [2], this method appears well-suited to reduce the set of features describing data at each granularity level, thus improving efficiency of classification step without comprising the existence of some classification model discriminating between examples tagged with different class label.

5 Experimental Results.

In this section we present a real-world application MSRC in geo-referenced census data interpretation. For this study we consider both census and digital map data provided in the context of the European project SPIN! (Spatial Mining for Data of Public Interest). 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 census sections or wards, for a total of two hundreds and fourteen wards. Census data are available at ward level and provide socio-economic statistics (e.g. mortality rate, that is, the percentage of deaths with respect to the number of inhabitants) as well as some measures describing the deprivation level of each ward according to information provided by Census combined into single index scores. We consider Jarman Underprivileged Area Score that is designed to measure the need for primary care, the indices developed by Townsend and Carstairs that is used in health-related analyses, and the Department of the Environment's Index (DoE) that is used in targeting urban regeneration funds. The higher the index value the more deprived a ward is. Both deprivation indices values as well as mortality rate are all numeric, but discrete values are here obtained by applying the Relative Unsupervised Discretization RUDE algorithm [13] that discretizes each continuous variable in the context defined by remaining continuous variables. More precisely, Jarman index, Townsend index, DoE index and Mortality rate have been automatically discretized in (*low, high*), while Carstairs index has been discretized in (*low, medium, high*).

By considering Greater Manchester wards as target objects, we focus our attention on mining a classification model to predict discrete value of DoE index by exploiting not only socio-economic factors represented in census data but also geographical factors represented in some linked topographic maps. Spatial analysis is then enabled by the availability of vectorized boundaries of the 1998 census wards as well as by other Ordnance Survey digital maps of NWE, where several interesting layers (road net, rail net, water net, urban area and green area) forming target-relevant objects are found (see Table 1). Both ward-referenced census data and map data are stored in an Object-Relational spatial database, i.e., Oracle Spatial 9i database, as a set of spatial tables, one for each layer. Each spatial table includes a geometry attribute that allows storing the geometrical representation (e.g. point, line, and region in a 2D context) and the

positioning of a spatial object with respect to some reference system in a single geometry field within a spatial table.

Table 1. Geographic layers.

| | <i>Layer name</i> | <i>Geometry</i> | <i>Number of objects</i> |
|------------|--|-----------------|--------------------------|
| Road net | A-road; B-road; Motorway; Primary road | Line | 2798 |
| Rail net | Railway | Line | 1054 |
| Urban area | Large urban area; Small urban area | Line | 381 |
| Green area | Wood; Park: | Line | 13 |
| Water net | Water; River; Canal | Line | 1067 |
| Ward | Ward | Region | 241 |

It is noteworthy that the spatial nature of both target objects and target-relevant objects poses two main degree of complexity that is the granularity of the spatial objects and the implicit definition of spatial relations. The former is due to the fact that non-target objects can be described at multiple levels of granularity. In this case, five different hierarchies can be defined to describe target-relevant layers (see Fig. 3). The hierarchies have depth three and are straightforwardly mapped into three granularity levels. The second source of complexity refers to the fact that both geometrical representation and relative positioning define implicitly spatial relations of different nature (e.g. topological) which can be explicitly modeled as many relational tables as the number of objects type and spatial relations. Modeling these spatial relations has a key role in this classification problem, since both the attribute values of the object to be classified and the attribute values of spatially related objects may be relevant for assigning an object to a class from the given set of classes.

We adopt here some geometrical and topological algorithm based on the 9-intersection model [5] to extract topological relationships between target objects and target-relevant objects. For instance the relationship *crosses(ward_135, urbareaL_151)* denotes that *ward_135*, that is a specific Greater Manchester ward, is crossed by *urbanareaL_151* that is a large urban area. The number of computed relationships is 784,107.

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

ward_urbanarea(X,Y) :- is_a(X,ward), connects(X,Y), is_a(Y,urban_area). ...
ward_urbanarea(X,Y) :- is_a(X,ward), inside(X,Y), is_a(Y,urban_area).

The use of the predicate *is_a* hides the fact that a hierarchy is defined for spatial objects which belong to the urban area layer. According to spatial relations introduced with BK, we easily obtain a relational representation of classification problem in question (see Figure 4).

Performances of MSRC are evaluated by means of a 10-fold cross validation.

The first step consists in the extraction of association rules at different levels of granularity according to hierarchies defined on target-relevant objects. Rules contain useful information about relational patterns frequently occurring in data. For instance, by analyzing spatial association rules extracted with parameters *minsup* = 0.1, *minconf* =

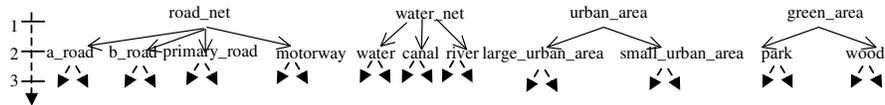


Figure 3. Spatial hierarchies defined for road net, water net, urban area and green area.

0.6 we discover the following rule:

$$doe(A, low) \leftarrow is_a(A, ward), ward_urbanarea(A, B), is_a(B, urban_area), jarman(A, low) \quad (52.6\%, 100\%)$$

which states that a low DoE index value is observed in a ward A that includes an urban area B and has a low value of Jarman index. The support (52.6%) and the high confidence (100%) confirm a meaningful association between a geographical factor, such as living in urban areas where primary care are well satisfied, and low level of derivation when considering targeting urban regeneration funds. In the feature construction step, this rule defines the boolean feature “ $is_a(A, ward) \wedge ward_urbanarea(A, B) \wedge is_a(B, urban_area) \wedge jarman(A, low)$ ” that is true for each Greater Manchester ward with a low value of Jarman index such that there is at least an urban area that is topologically related (e.g. connects or inside) to the ward in question. At a granularity level 2, SPADA specializes the target-relevant object B by generating the following rule that preserves both support and confidence:

$$doe(A, low) \leftarrow is_a(A, ward), ward_urbanarea(A, B), is_a(B, large_urban_area), jarman(A, low) \quad (52.6\%, 100\%)$$

This rule clarifies that the urban area B is large.

By varying granularity level as well as the value of $minsup$, $minconf$ and the number of refinement steps K (pattern length) in association rule discovery, we obtain several experimental settings. For each setting, the set of discovered rules is transformed in a set of boolean features that is the conjunctions of literals derived from the body of each rule, and redundant features are removed to reduce feature space without affecting the existence of a complete and consistent theory for each class label. Finally, we apply the four classification learners embedded in MSRC and compare the percentage of correct classification averaged over 10-folds of cross-validation. Results on average reduction percentage and average accuracy are reported in both Table 2. Results confirms that high number of association rules typically lead to redundant features. Percentage of feature reduction increases when the refinement step increases. This is due to the high number of similar rules (thus producing redundant features). Another consideration is that the average predictive accuracies of classification models discovered at higher granularity levels (i.e. level=2) are sometimes better than the corresponding accuracies

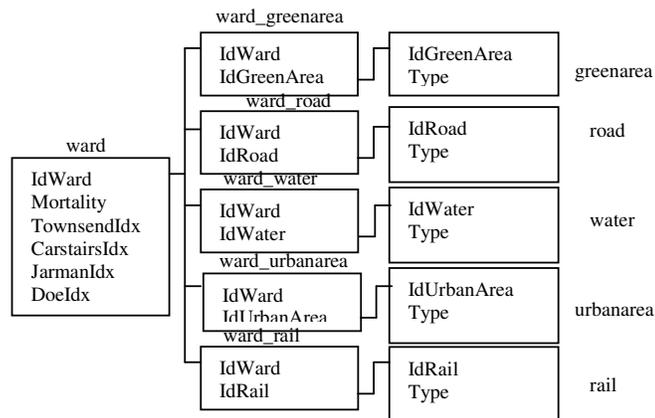


Figure 4. Relational data describing ward-referenced census data and map data

at lowest levels. This means that the classification model takes advantage of the use of the hierarchies defined on spatial objects. In this case, results at different abstraction

Table 2. DoE Index average accuracy

| min sup | min conf | Granularity level | K | Avg. number of original features | Avg. perc. of feature reduction | Avg. Accuracy | | | |
|---------|----------|-------------------|---|----------------------------------|---------------------------------|---------------|--------------|--------------|--------------|
| | | | | | | NBC | 1-NN | C4.5 | JRIP |
| 0.2 | 0.8 | 1 | 5 | 27.5 | 37% | 81.64 | 80.73 | 83.01 | 83.01 |
| 0.2 | 0.8 | | 6 | 221.1 | 44% | 80.28 | 80.73 | 83.01 | 83.01 |
| 0.2 | 0.8 | | 7 | 982.3 | 73% | 75.56 | 80.28 | 83.01 | 83.01 |
| 0.2 | 0.8 | 2 | 5 | 492.8 | 29% | 82.55 | 82.55 | 81.64 | 81.93 |
| 0.2 | 0.8 | | 6 | 278.3 | 11% | 82.10 | 82.55 | 82.10 | 83.01 |
| 0.2 | 0.8 | | 7 | 1860.1 | 66% | 80.28 | 82.55 | 81.19 | 81.02 |
| 0.1 | 0.6 | 1 | 5 | 233.2 | 48% | 83.01 | 80.73 | 83.01 | 83.01 |
| 0.1 | 0.6 | | 6 | 374 | 57% | 80.28 | 80.73 | 83.01 | 83.01 |
| 0.1 | 0.6 | | 7 | 904.2 | 74% | 80.73 | 80.73 | 83.01 | 83.01 |
| 0.1 | 0.6 | 2 | 5 | 442.2 | 28% | 83.01 | 81.30 | 82.10 | 81.02 |
| 0.1 | 0.6 | | 6 | 635.8 | 27% | 82.55 | 79.65 | 80.28 | 81.64 |
| 0.1 | 0.6 | | 7 | 3051.4 | 68% | 80.73 | 79.31 | 81.19 | 81.02 |

levels provide insights on what are the non-target objects that affect the classification. For instance, when $K=7$, $minsup = 0.1$ and $minconf = 0.6$, naïve-bayes classification is strongly improved when considering the size of the urban area, the type of road, and so on. Moreover, results show that by decreasing the number of extracted rules (higher support and confidence) we have lower accuracy. This means that there are boolean features that strongly influence classification results and often these features correspond to rules which are not characterized by high values of support and confidence. Finally, we observe that, in this specific task, the higher the number of refinement steps (boolean features involving more literals) not necessarily means the better the model. This is mainly due to the fact that rules become very specific.

6 Conclusions

In this paper we present MSRC (Multi-Step Relational Classifier) that exploits relational association rules mining for inducing (multi-)relational classifiers. Classification is based on a propositionalization step where features are constructed by considering the structure imposed on data by relational association rules discovered at multiple levels of granularity according to some domain knowledge. This allows MSRC to support multi-level classification that is highly demanded in many real-world applications that would benefit from concept hierarchies. Since the number of features constructed starting from relational association rules is usually high for each level of granularity, MSRC integrates a feature reduction method that reduces feature space by removing redundant features without compromising classification effectiveness.

An application of the MSRC framework to a classification task in geo-referenced census data interpretation is reported and results shows that multi-level classification provides insights on what target-relevant objects really affect the classification. It is also noteworthy that, in our study, support and confidence do not seem to be valid

criterion to filter rules, since rules characterized by low values of support and confidence positively affect classification results. As future work we intend to extend experimental evaluation by comparing MSRC with a (multi-)relational structural approach to multi-level classification.

Acknowledgments

We thank Simon Rawles and Peter Flach for working together in defining and developing the REFER algorithm. This work has been supported by the annual Scientific Research Project "Metodi di apprendimento automatico e di data mining per sistemi di conoscenza basati sulla semantica" Year 2003 funded by the University of Bari.

References

- [1] Appice A., Berardi M., Ceci M., Lapi M., Malerba D., Turi A.: Mining interesting spatial association rules: two case studies. SEBD 2004: 86-97
- [2] Appice, A., Ceci, M. Rawles, S., Flach, P.. Redundant Feature Elimination for Multi-Class Problems. In . C. E. Brodley (Ed.): In the Proceedings of 21st International Conference on Machine Learning, ICML 2004. ACM 2004, 33-40, (2004).
- [3] Cohen, W. W.. Fast Effective Rule Induction. *Proc. of the 12th International Conference on Machine Learning*, (pp. 115-123), Morgan Kaufmann (1995)
- [4] Domingos, P. & Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss, 29, 103-130.]. (1997)
- [5] Egenhofer, M.J.: Reasoning about Binary Topological Relations. Proceedings of the Second Symposium on Large Spatial Databases, Zurich, Switzerland, (1991) 143-160.
- [6] Han J., Fu Y.: Discovery of multiple-level association rules from large databases., In 21st International Conference on Very Large Data Bases, VLDB'95, 1995, pp. 420-431
- [7] Knobbe A.J., Haas M. & Siebes A: Propositionalisation and aggregates. In 5th European Conf. on Principles of Data Mining and Knowledge Discovery, Springer-Verlag, 2001.
- [8] Kramer S., Pfahringer B., and Helma C.. In D. Page, editor, Proceedings of the 8th International Conference on Inductive Logic Programming, volume 1446 of Lecture Notes in Artificial Intelligence, pages 80--94. Springer-Verlag, 1998.
- [9] Krogel, M., Rawles, S., Zelezny, F., Flach, P., Lavrac, N., Wrobel S.: Comparative evaluation of approaches to propositionalization. *Proceedings of 13th International Conference on Inductive Logic Programming, ILP'03*. Springer-Verlag, 197-214, (2003).
- [10] Lavrac N. and Dzeroski S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, 1994.
- [11] Lavrač, N., Gamberger, D., & Jovanoski V.. A study of relevance for learning in deductive databases. The Journal of Logic Programming, 16, 215-249. (1999)
- [12] Lisi, F.A., Malerba, D.: Inducing Multi-Level Association Rules from Multiple Relations. Machine Learning, 55, 175-210, (2004).
- [13] Ludl, M.C., Widmer, G.: Relative Unsupervised Discretization for Association Rule Mining. PKDD'00, LNCS 1910, Springer-Verlag, (2000), 148-158.
- [14] Modrzejewski, M. (1993). Proceedings of the European Conference on Machine Learning. Vienna, Austria, Springer-Verlag.
- [15] Mitchell, T. Machine Learning. McGraw Hill, 1997.
- [16] Quinlan, J.. C4.5: Programs for machine learning. Morgan Kaufmann (1993)
- [17] Wrobel, S. Inductive logic programming for knowledge discovery in databases. In: Dzeroski, S., N. Lavrač(eds.): Relational Data Mining, Springer: Berlin, pp. 74-101. 2001.