Discovering Relational Emerging Patterns

Annalisa Appice, Michelangelo Ceci, Carlo Malgieri, and Donato Malerba

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

Abstract. The discovery of emerging patterns (EPs) is a descriptive data mining task defined for pre-classified data. It aims at detecting patterns which contrast two classes and has been extensively investigated for attribute-value representations. In this work we propose a method, named Mr-EP, which discovers EPs from data scattered in multiple tables of a relational database. Generated EPs can capture the differences between objects of two classes which involve properties possibly spanned in separate data tables. We implemented Mr-EP in a pre-existing multi-relational data mining system which is tightly integrated with a relational DBMS, and then we tested it on two sets of geo-referenced data.

1 Introduction

The discovery of emerging patterns (EPs) is a descriptive data mining task aiming at the detection of significant differences between objects belonging to separate classes. EPs are introduced in [4] as a particular kind of patterns (or multi-variate features) whose support significantly changes from one data class to another: the larger the difference of pattern support, the more interesting the patterns. Due to the sharp change in support, EPs can be used to characterize object classes. For example, EPs have been used to predict the likelihood of diseases such as acute lymphoblastic leukemia [13] and to explore high-dimensional data such as gene expression data [12].

Several algorithms [18,4,10] have been proposed to discover EPs from data belonging to separate classes (data populations) and stored in a single relational table. Independent units of each data population D_i are described by a fixed vector S of explanatory attributes $X_1, X_2, \ldots, X_m$ and are tagged with a class label $Y = C_i$. The EPs which distinguish a target data population D_i from the background data D_j are in the form $P(GR^{D_j \to D_i}(P))$, where P is a set of items ($P \subseteq S$) and $GR^{D_j \to D_i}(P)$ is the support ratio (or *growth rate*) of P over D_j to D_i. Formally $GR^{D_j \to D_i}(P) = \frac{s_{D_i}(P)}{s_{D_j}(P)}$, where $s_{D_i}(P)$ ($s_{D_j}(P)$) is the support of P on D_i (D_j). Since an item refers to an attribute-value pair, the *itemset* P can be interpreted as a conjunction of attribute values. Formally, given a growth rate threshold $minGR \geq 1$, an EP from D_j to D_i is an itemset P whose growth rate from D_j to D_i is greater than $minGR$.

Although research on EPs has reached relative maturity over the last years, there is still a number of interesting issues which remain open. One issue concerns

R. Basili and M.T. Pazienza (Eds.): AI*IA 2007, LNAI 4733, pp. 206–217, 2007.

the need to face the challenges of real-world data mining tasks involving complex and heterogeneous data with different properties which are modeled by as many relations as the number of object types. Mining data scattered over the multiple tables of a relational database (*relational data*) poses the problem of taking into account attributes of related (i.e. *task-relevant*) objects when investigating properties of some *reference* objects which are the main subject of analysis. Classical EPs discovery methods do not distinguish task-relevant from reference objects, nor do they allow the representation of any kind of interaction. Therefore, we propose to resort to a Multi-Relational Data Mining (MRDM) approach [6] in order to deal with both relational data and relational patterns.

In this paper, we propose a novel method, called Mr-EP (*M*ulti-*R*elational *E*merging *P*atterns), which discovers EPs from relational data and is capable to capture the change in properties of separate classes of data spanned in multiple data tables. The class variable is associated with the reference objects, while explanatory attributes refer to either the reference objects or the task-relevant objects which are someway related to the reference objects. The structural information required to mine such *relational* EPs can be automatically obtained from the database schema by navigating foreign key constraints. For each class, relational EPs are expressed as SQL queries stored in XML format.

The paper is organized as follows. In the next section the background of this research and related works are discussed, while in Section 3 the problem of EPs discovery is formalized in the multi-relational framework. Relational emerging patterns discovery is described in Section 4. Lastly, experimental results are reported in Section 5 and then some conclusions are drawn.

2 Related Works

The combination of relational representation with pattern discovery has been deeply investigated for several data mining tasks. Data mining research has provided several solutions for the task of frequent pattern and association rule discovery both in a propositional and a relational setting, but, at the best of our knowledge, this work represents the first attempt at extracting relational EPs.

In [1], a frequent pattern is defined as an itemset whose support is greater than a predefined minimum threshold value (minimum support), while an association rule is an implication in the form $A \Rightarrow C(s, c)$, where A and C are itemsets and $A \cap C = \oslash$. The support s provides an estimate of the probability $p(A \cup C)$, while the confidence c provides an estimate of the probability $p(C|A)$. An association rule $A \rightarrow C$ ($s\%$, $c\%$) is *strong* if the pattern $A \cup C$ ($s\%$) is frequent and the confidence of the rule is greater than a predefined minimum threshold value (minimum confidence).

The two best known association rule discovery methods defined in the relational framework are WARMR [3] and SPADA [14]. They are both based on an Inductive Logic Programming (ILP) approach, where both data and background (or domain) knowledge is represented in a first-order logic formalism, such as Horn clausal logic. Relational frequent patterns are discovered according to the

levelwise method described in [16], which consists in a level-by-level exploration of the lattice of patterns ordered by θ-subsumption [17]. Strong association rules are then generated from frequent patterns.

Unlike frequent patterns and/or association rules which capture regularities in data describing unclassified objects, EPs capture changes in data describing objects of different classes. This adds one main source of complexity to the learning task, since the monotonicity property does not hold for EPs. Suppose a pattern P is not an EP from D_j to D_i, that is, the growth rate of P from D_j to D_i is not greater than the user-defined threshold. For any super-pattern Q of P ($P \subseteq Q$), its support is less than or equal to that of P for both classes C_i and C_j, while its growth rate (the support ratio) is free to be any real value between 0 and ∞. Therefore, a superset of a non-EP may or may not be an EP.

In the seminal work by Dong and Li [4], EPs are discovered by assuming that each data set is stored in a single data table. A border-based approach is adopted to discover the EPs discriminating between separate classes. Borders are used to represent both candidates and subsets of EPs; the border differential operation is then used to discover the EPs. Zhang et al. [18] have described an efficient method, called ConsEPMiner, which adopts a level-wise generate-and-test approach to discover EPs which satisfy several constraints (e.g., growth-rate improvement). Finally, Fan and Ramamohanarao [7] have proposed a method which improves the efficiency of EPs discovery by adopting a CP-tree data structure to register the counts of both the positive and negative class.

A further direction of research concerns the usage of EPs in learning accurate data classifiers [5,8,11,9]. EP-based classification is related to the associative classification framework [15] where classifiers are built by carefully selecting high quality association rules. The advantage of EPs over association rules is that EPs provide features which better discriminate objects of distinct classes.

3 Problem Definition

In this work we assume that both reference objects and task-relevant objects are tuples stored in tables of a relational database D according to a schema S. The set R of reference objects is the collection of tuples stored in a table T of D called *target* table. Similarly, each set R_i of task-relevant objects corresponds to a distinct table of D. The inherent "structure" of data, that is, the relations between reference and task-relevant objects, is expressed in the schema S by foreign key constraints (FK). Foreign keys make it possible to navigate the data schema and retrieve all the task-relevant objects in D which are related to a reference object and, thus, are capable of discriminating between the values of the target attribute Y.

Before providing a formal definition of the problem to be solved, some other definitions need to be introduced.

Definition 1 (Key predicate). *Let S be a database schema and T be a table of S representing the target table for the task at hand. The "key predicate"*

associated with T in S is a first order unary predicate $p(t)$ such that p denotes the table T and the term t is a variable that represents the primary key of T.

Definition 2 (Structural predicate). *Let S be a database schema and $\{T_i, T_j\}$ be a pair of tables in S such that there exists a foreign key FK in S between T_i and T_j. A "structural predicate" associated with the pair of tables $\{T_i, T_j\}$ in S is a first order binary predicate $p(t, s)$ such that p denotes FK and the term t (s) is a variable that represents the primary key of T_i (T_j).*

Definition 3 (Property predicate). *Let S be a database schema, T_i a table of S and ATT be an attribute of T_i which is neither primary key nor foreign key for T_i in S. A "property predicate" associated with the attribute ATT of the table T_i is a binary predicate $p(t, s)$ such that p denotes the attribute ATT, the term t is a variable representing the primary key of T_i and s is a constant which represents a value belonging to the range of ATT in T_i.*

A relational pattern over S is a conjunction of predicates consisting of the key predicate and one or more (structural or property) predicates over S. More formally, a relational pattern is defined as follows:

Definition 4 (Relational pattern). *Let S be a database schema. A "relational pattern" P over S is a conjunction of predicates:*

$$p_0(t0_1), p_1(t1_1, t1_2), p_2(t2_1, t2_2), \ldots, p_m(tm_1, tm_2)$$

where $p_0(t0_1)$ is the key predicate associated with the target table of the task at hand and $\forall i = 1, \ldots, m$ $p_i(ti_1, ti_2)$ is either a structural predicate or a property predicate over S.

Henceforth, we will also use the set notation for relational patterns, that is, a relational pattern is considered a set of atoms.

Definition 5 (Key linked predicate).
Let $P = p_0(t0_1), p_1(t1_1, t1_2), p_2(t2_1, t2_2), \ldots, p_m(tm_1, tm_2)$ be a relational pattern over the database schema S. For each $i = 1, \ldots, m$, the (structural or property) predicate $p_i(ti_1, ti_2)$ is "key linked" in P if

- *$p_i(ti_1, ti_2)$ is a predicate with $t0_1 = ti_1$ or $t0_1 = ti_2$, or*
- *there exists a structural predicate $p_j(tj_1, tj_2)$ in P such that $p_j(tj_1, tj_2)$ is key linked in P and $ti_1 = tj_1 \vee ti_2 = tj_1 \vee ti_1 = tj_2 \vee ti_2 = tj_2$.*

Definition 6 (Completely linked relational pattern). *Let S be a database schema. A "completely linked" relational pattern is a relational pattern $P = p_0(t0_1), p_1(t1_1, t1_2), \ldots, p_m(tm_1, tm_2)$ such that $\forall i = 1 \ldots m$, $p_i(ti_1, ti_2)$ is a predicate which is key linked in P.*

Definition 7 (Relational emerging patterns). *Let D be an instance of a database schema S that contains a set of reference objects labeled with $Y \in \{C_1, \ldots, C_L\}$ and stored in the target table T of S. Given a minimum growth*

rate value (minGR) and a minimum support value (minsup), P is a "relational emerging pattern" in D if P is a completely linked relational pattern over S and some class label C_i exists such that $GR^{\overline{D_i}\to D_i}(P) > minGR$ and $s_{D_i}(P) > minsup$, where:

- *D_i is an instance of database schema S such that $D_i.T = \{t \in D.T | D.T.Y = C_i\}$ and $\forall T' \in S, T' \neq T$: $D_i.T' = \{t \in D.T' |$ all foreign key constraints FK are satisfied in $D_i\}$.*
- *$\overline{D_i}$ is an instance of database schema S such that $\overline{D_i}.T = \{t \in D.T | D.T.Y \neq C_i\}$ and $\forall T' \in S, T' \neq T$: $\overline{D_i}.T' = \{t \in D.T' |$ all foreign key constraints FK are satisfied in $\overline{D_i}\}$.*

The support $s_{D_i}(P)$ of P on database D_i is computed as follows:

$$s_{D_i}(P) = \frac{|O_P|}{|O|}, \tag{1}$$

where O denotes the set of reference objects stored as tuples of $D_i.T$, while O_P denotes the subset of reference objects in O which are covered by the pattern P. The growth rate of P for distinguishing D_i from $\overline{D_i}$ is the following:

$$GR^{\overline{D_i}\to D_i}(P) = \frac{s_{D_i}(P)}{s_{\overline{D_i}}(P)} \tag{2}$$

As in [4], we assume that $GR(P) = \frac{0}{0} = 0$ and $GR(P) = \frac{>0}{0} = \infty$.

The problem of discovering relational EPs can now be formalized as follows. *Given*:

- a relational database D with a data schema S,
- a set R of reference objects tagged with a class label $Y \in \{C_1, C_2, \ldots, C_L\}$,
- some sets R_i, $1 \leq i \leq h$ of task-relevant objects,
- a pair of thresholds, that is, the minimum growth rate ($minGR \geq 1$) and the minimum support ($minsup > 0$).

Find:
the set of *relational emerging patterns* which discriminate between reference objects belonging to distinct classes in D.

In this work, we resort to the relational algebra formalism to express a relational emerging pattern P by means of an SQL query. The SELECT statement selects primary key values for distinct reference objects of the task at hand. The FROM statement describes the joins between all the tables of S which are involved in P (i.e., the target table associated with the key predicate and the tables which are included in separate structural predicates of the pattern P). A structural predicate is translated into a join condition. The property of linkedness in relational patterns guarantees the soundness of joins. The WHERE statement describes the conditions expressed in the property predicates. Reference objects contributing to the support of the EP on each D_i are obtained as result set by running this SQL query on the database instance D_i.

Example 1. Let us consider a set of molecules (reference objects) described in terms of the "logP" property and the "mutagenicity" class. Each molecule is composed by one or more atoms (task-relevant objects) and each atom is described by the "charge". An example of relational pattern P is the following:

molecule(MolID), logPInMolecule(MolId,[5..10[),atom(MolId,AtomId1),
chargeInAtom(AtomId1,[3.2,5.8[),atom(MolID,AtomId2),
chargeInAtom(AtomId2,[5.0,7.1[)

P can be expressed by means of the SQL query:

```
SELECT distinct M.MolID
FROM (Molecule M INNER JOIN Atom A1 on M.MolId=A1.MolId)
     INNER JOIN Atom A2 on M.MolId=A2.MolId
WHERE M.logP>=5 AND M.logP<10 AND
     A1.Charge >=3.2 AND A1.Charge<5.8 AND
     A2.Charge >=5.0 AND A2.Charge<7.1
```

4 Relational EPs Discovery

We address EP discovery by adapting the algorithms proposed for frequent pattern discovery to the special case of EPs. The blueprint for the frequent patterns discovery algorithms is the levelwise method [16] that explores level-by-level the lattice of patterns ordered according to a generality relation ($\geqslant$) between patterns. Formally, given two patterns $P1$ and $P2$, $P1 \geqslant P2$ denotes that $P1$ ($P2$) is more general (specific) than $P2$ ($P1$). The search proceeds from the the most general pattern and iteratively alternates the candidate generation and candidate evaluation phases.

In this paper, we propose an enhanced version of the aforementioned levelwise method which works on EPs rather than frequent patterns. The space of candidate EPs is structured according to the θ-subsumption generality order [17].

Definition 8 (θ-subsumption). *Let $P1$ and $P2$ be two relational patterns on a data schema S such that both $P1$ and $P2$ are key completely linked patterns with respect to a target table T in S. $P1$ θ-subsumes $P2$ if and only if a substitution θ exists such that $P2\ \theta \subseteq P1$.*

Having introduced θ-subsumption, we now go to define generality order between completely linked relational patterns.

Definition 9 (Generality order under θ-subsumption). *Let $P1$ and $P2$ be two completely linked relational patterns. $P1$ is more general than $P2$ under θ-subsumption, denoted as $P1 \geqslant_\theta P2$, if and only if $P2$ θ-subsumes $P1$.*

θ-subsumption defines a quasi-ordering, since it satisfies the reflexivity and transitivity property but not the anti-symmetric property. The quasi-ordered set spanned by $\geqslant_\theta$ can then be searched according to a downward refinement operator which computes the set of refinements for a completely linked relational pattern.

Definition 10 (Downward refinement operator under θ-subsumption). *Let $\langle G, \geqslant_\theta \rangle$ be the space of completely linked relational patterns ordered according to $\geqslant_\theta$. A downward refinement operator under θ-subsumption is a function ρ such that $\rho(P) \subseteq \{Q \in G | P \geqslant_\theta Q\}$.*

We now define the downward refinement operator ρ' to explore the space of candidate EPs for distinguishing D_i from $\overline{D_i}$.

Definition 11 (Downward refinement operator for EPs). *Let P be a relational EP for distinguishing D_i from $\overline{D_i}$. Then $\rho'(P) = \{P \cup \{p(t1, t2)\} | p(t1, t2)$ is a structural or property predicate key linked in $P \cup \{p(t1, t2)\}$ and $P \cup \{p(t1, t2)\}$ is an EP for distinguishing D_i from $\overline{D_i}\}$.*

The downward refinement operator for EPs is a refinement operator under θ-subsumption. In fact, it can be easily proved that $P \geqslant_\theta Q$ for all $Q \in \rho'(P)$. This makes Mr-EP able to perform a levelwise exploration of the lattice of EPs ordered by θ-subsumption. More precisely, for each class C_i, the EPs for distinguishing D_i from $\overline{D_i}$ are discovered by searching the pattern space one level at a time, starting from the most general EP (the EP that contains only the key predicate) and iterating between candidate generation and evaluation phases. In Mr-EP, the number of levels in the lattice to be explored is limited by the user-defined parameter $MAX_M \geq 1$. In other terms, MAX_M limits the maximum number of structural predicates (joins) within a candidate EP. Since joins affects the computational complexity of the method, a low value of MAX_M guarantees the applicability of the algorithm to reasonably large data The monotonicity property of the generality order $\geqslant_\theta$ with respect to the support value (i.e., a superset of an infrequent pattern cannot be frequent) is exploited to avoid the generation of infrequent relational patterns. In fact, an infrequent pattern on D_i cannot be an EP for distinguishing D_i from $\overline{D_i}$.

Proposition 1 (Property of θ-subsumption monotonicity). *Let $\langle G, \geqslant_\theta \rangle$ be the space of relational completely linked patterns ordered according to $\geqslant_\theta$. $P1$ and $P2$ are two patterns of $\langle G, \geqslant_\theta \rangle$ with $P1 \geqslant_\theta P2$ then $O_{P1} \supseteq O_{P2}$.*

Therefore, when $P1 \geqslant_\theta P2$, we have $s_{D_i}(P1) \geq s_{D_i}(P2)$ and $s_{\overline{D_i}}(P1) \geq s_{\overline{D_i}}(P2)$ $\forall i = 1, \ldots, L$. This is the counterpart of one of the properties exploited in the family of the Apriori-like algorithms [1] to prune the space of candidate patterns. To efficiently discover relational EPs, Mr-EP prunes the search space by exploiting the θ-subsumption monotonicity of support (*prune1* criterion). Let P' be a refinement of a pattern P. If P is an infrequent pattern on D_i ($s_{D_i}(P) < minsup$), then P' has a support on D_i that is lower than the user-defined threshold ($minsup$). According to the definition of EP, P' cannot be an EP for distinguishing D_i from $\overline{D_i}$, hence Mr-EP does not refine patterns which are infrequent on D_i.

Unluckily, the monotonicity property does not hold for the growth rate: a refinement of an EP whose growth rate is lower than the threshold $minGR$ may or may not be an EP. Anyway, as in the propositional case [18], some mathematical considerations on the growth rate formulation can be usefully exploited to define two further pruning criteria.

First (*prune2* criterion), Mr-EP avoids generating the refinements of a pattern P in the case that $GR^{\overline{D_i}\to D_i}(P) = \infty$ (i.e., $s_{D_i}(P) > 0$ and $s_{\overline{D_i}}(P) = 0$). Indeed, due to the θ-subsumption monotonicity of support $\forall P' \in \rho'(P)$: $s_{\overline{D_i}}(P) \geq s_{\overline{D_i}}(P')$ then $s_{\overline{D_i}}(P') = 0$. Thereby, $GR^{\overline{D_i}\to D_i}(P') = 0$ in the case that $s_{D_i}(P') = 0$, while $GR^{\overline{D_i}\to D_i}(P') = \infty$ in the case that $s_{D_i}(P') > 0$. In the former case, P' is not worth to be considered (*prune1*). In the latter case, $P \geqslant_\theta P'$ and $s_{D_i}(P) \geq s_{D_i}(P')$. Therefore, P' is useless since P has the same discriminating ability than P' ($GR^{\overline{D_i}\to D_i}(P) = GR^{\overline{D_i}\to D_i}(P') = \infty$). We prefer P to P' based on the Occams razor principle, according to which all things being equal, the simplest solution tends to be the best one.

Second (*prune3* criterion), Mr-EP avoids generating the refinements of a pattern P which add a property predicate in the case that the refined patterns have the same support of P on $\overline{D_i}$. We denote by:

$SameSupport(P)_{\overline{D_i}} = \{P' \in \rho'(P) | s_{\overline{D_i}}(P) = s_{\overline{D_i}}(P'),\ P' = P \wedge p(t1,t2),$
$p(t1,t2)$ is a property predicate$\}$.

For the monotonicity property, $\forall P' \in SameSupport_{\overline{D_i}}(P)$: $s_{D_i}(P) \geq s_{D_i}(P')$. This means that $GR^{\overline{D_i}\to D_i}(P) \geq GR^{\overline{D_i}\to D_i}(P')$. P' is more specific than P but, at the same time, P' has a lower discriminating power than P. This pruning criterion prunes EPs that could be generated as refinements of patterns in $SameSupport_{\overline{D_i}}(P)$. However, it is possible that some of them may be of interest for our discovery process. Their identification is guaranteed by the following:

Proposition 2. *Let $P' \in SameSupport_{\overline{D_i}}(P)$ such that $P' = P \cup \{p(t1,t2)\}$ with $p(t1,t2)$ being a property predicate. Let $P'' \in \rho'(P')$ such that $P'' = P' \cup \{q(t3,t4)\}$ with $q(t3,t4)$ being a property predicate. If P'' is an EP discriminating D_i from $\overline{D_i}$ and $s_{\overline{D_i}}(P'') \neq s_{\overline{D_i}}(P)$ then $P''' = P \cup \{q(t3,t4)\} \notin SameSupport_{\overline{D_i}}(P)$.*

Proof: *Let $O_P^{\overline{D_i}}$ denote the set of reference objects in $\overline{D_i}$ covered by a pattern P. By construction, $P'' \in \rho'(P') \cap \rho'(P''')$ and $O_{P''}^{\overline{D_i}} = O_{P'}^{\overline{D_i}} \cap O_{P'''}^{\overline{D_i}}$. Since $P' \in SameSupport_{\overline{D_i}}(P)$, we have $s_{\overline{D_i}}(P) = s_{\overline{D_i}}(P')$, that is, $O_P^{\overline{D_i}} = O_{P'}^{\overline{D_i}}$. Moreover, for the θ-subsumption monotonicity property, we have $O_{P'''}^{\overline{D_i}} \subseteq O_P^{\overline{D_i}}$. Therefore, we have: $O_{P''}^{\overline{D_i}} = O_{P'}^{\overline{D_i}} \cap O_{P'''}^{\overline{D_i}} = O_P^{\overline{D_i}} \cap O_{P'''}^{\overline{D_i}} = O_{P'''}^{\overline{D_i}}$. Since $s_{\overline{D_i}}(P'') \neq s_{\overline{D_i}}(P)$ by hypothesis, then it is also true that $s_{\overline{D_i}}(P''') \neq s_{\overline{D_i}}(P)$. Therefore, $P''' \notin SameSupport_{\overline{D_i}}(P)$.*

According to proposition 2, we can prune P' (but not P''') without preventing the generation of EPs more specific than P'. It is noteworty to observe that this pruning criterion operates only when $p(t1,t2)$ is a property predicate. Differently, pruning of structural predicates would avoid the introduction of a new variable thus avoiding the discovery of further EPs obtained by adding property or structural predicates involving such variable.

Finally, additional candidates not worth being evaluated are those equivalent under θ-subsumption to some other candidate (*prune4*).

5 Experimental Results

Mr-EP has been implemented as a module of the MRDM system MURENA (MUlti RElational aNAlyzer) which interfaces the Oracle 10g DBMS. We tested the method on two real world geo-referenced data sets: the North-West England Census Data and the Munich Census Data. Both data sets include numeric attributes, which are handled through an equal-width discretization to partition the range of values into a fixed number of bins. EPs have been discovered with $minGR = 1.1$, $minsup = 0.1$. MAX_M is set to 3 for North-West England Census Dataset and to 5 for Munich Census Dataset. In this work, we present only a qualitative interpretation of EPs. Each EP is analyzed in terms of a human interpretable pattern that is descriptive of characteristics discriminating between separate classes of relational data.

The North-West England Census Data. Data were obtained from both census and digital maps provided by the European project SPIN! (http://www.ais.fraunhofer.de/KD/SPIN/project.html). They concern Greater Manchester, one of the five counties of North West England (NWE). Greater Manchester is divided into into 214 census sections (wards). Census data are available at ward level and provide socio-economic statistics (e.g. mortality rate) as well as some measures of the deprivation of each ward according to information provided by Census combined into single index scores. We employed the Jarman score that estimates the need for primary care, the indices developed by Townsend and Carstairs to perform health-related analyses, and the DoE index which is used in targeting urban regeneration funds. The higher the index value the more deprived the ward. In this application, the mortality percentage rate (target attribute) takes values in the finite set $\{low = [0.001, 0.01], high =]0.01, 0.18]\}$. The analysis we performed was based on deprivation factors and geographical factors represented in topographic maps of the area. Vectorized boundaries of the 1998 census wards as well as of other Ordnance Survey digital maps of NWE are available for several layers such as urban area (115 lines), green area (9 lines), road net (1687 lines), rail net (805 lines) and water net (716 lines). Objects of each layer are stored as tuples of relational tables including information on the object type (TYPE). For instance, an urban area may be either a "large urban area" or a "small urban area". Topological relationships between wards and objects in these layers are materialized as relational tables expressing non-disjoint relations. The number of materialized "non disjoint" relationships is 5313.

Mr-EP discovered 60 EPs to discriminate high mortality rate wards from the class of wards with low mortality rate and 55 EPs to discriminate low mortality rate wards from high mortality rate wards. An example of EP extracted for the class *mortality_rate=high* is:

$wards(A) \wedge wards\_rails(A, B) \wedge wards\_doeindex(A, [6.598..9.232])$

where $wards(A)$ is the key predicate, $wards\_rails(A, B)$ is the structural predicate representing an interaction between the ward A and a ward B (this means that A is crossed by at least one railway) and $wards\_doeindex(A, [6.598..9.232])$ (i.e. A is a deprived zone to be considered as target zone for regeneration

fundings) is a property predicate. This pattern presents a support of 0.22 and growth rate 3.77. This means that wards crossed by railways and with a relatively high doeindex value present a high percentage of mortality. This could be due to urban decay condition of the area. The pattern corresponds to the SQL query:

```
SELECT distinct W.ID
FROM (WARDS W INNER JOIN WARDS_RAILS WR on W.ID=WR.WardID)
WHERE W.DOEINDEX <= 9.232 AND W.DOEINDEX >= 6.598
```

A different conclusion can be drawn from the following relational EP extracted for the class *mortality_rate=low*:

$$wards(A) \wedge wards\_townsendidx(A, [-3.86431.. - 2.01452]) \\ \wedge wards\_greenareas(A, B)$$

This pattern has a support of 0.113 and a growth rate of 2.864. It captures the event that a ward with a relative low Townsend deprivation level (i.e., the ward A cannot be considered as deprived with respect to health-related analyses) and overlaps at least one green area (i.e., a park) discriminates wards with low mortality rate from the others.

The Munich Census Data. These data concern the level of monthly rent per square meter for flats in Munich expressed in German Marks. They have been collected on 1998 to develop the 1999 Munich rental guide and describe 2180 flats located in the 446 subquarters of Munich obtained by dividing the Munich metropolitan area up into three areal zones and decomposing each of these zones into 64 districts. The vectorized boundaries of subquarters, districts and zones as well as the map of public transport stops (56 subway (U-Bahn) stops, 15 rapid train (S-Bahn) stops and 1 railway station) within Munich are available for this study (http://www.di.uniba.it/∼ceci/mic Files/munich_db.tar.gz). The objects included in these layers are stored in different relational tables (SUBQUARTERS, TRANSPORT_STOPS and APARTMENTS). Information on the "area" of subquarters is stored in the corresponding table. Transport stops are described by means of their type (U-Bahn, S-Bahn or Railway station), while flats are described by means of their "monthly rent per square meter", "floor space in square meters" and "year of construction". The monthly rent per square meter (target attribute) has been discretized into the two intervals $low = [2.0, 14.0]$ or $high =]14.0, 35.0]$. The "close to" relation between subquarters areas and the "inside" relation between public train stops and metropolitan subquarters are materialized into relational tables (*ward_close_to_ward* and *apartment_inside_district*). Similarly, the "cross" relation between districts and public train stops is materialized into the relational table *district_crossedby_tranStop*.

Mr-EP discovered 31 (31) EPs to discriminate apartment with high (low) rent rate per square meters from the class of apartments with low (high) rent rate per square meters. An example of EP extracted for the class *rate_per_squaremeters =high* is:

$$apartment(A) \wedge apartment\_inside\_district(A, B) \wedge \\ district\_close\_to\_district(B, C) \wedge district\_ext\_19\_69(B, [0.875..1.0])$$

This pattern has a support of 0.125 and a growth rate of 1.723. It represents the event that an apartment A is inside a district B which contains a high percentage (between 87.5% and 100%) of apartments with a relatively low extension (between 19 m^2 and 69 m^2). This pattern distriminates apartments with high rate per square meters form the others. It can be motivated by considering that the rent rate is not directly proportional to the apartment extension but it includes fixed expenses that do not vary with the apartment size.
For the class *rate_per_squaremeters=low* the following EP was discovered:

$$apartment(A) \wedge apartment\_inside\_district(A, B) \wedge district\_crossedby\_tranStop(B, C) \wedge apartment\_year(A, [1893..1899])$$

This pattern has a support of 0.265 and a growth rate of 2.343. It represents the event that an apartment A built between 1893 and 1899 is inside a district B that contains a railway public stop. This pattern discriminates apartments with low rate per square meters from the others. It can be motivated by considering that old buildings do not offer the same facilities of a recently built apartment.

6 Conclusions

In this paper, we presented a novel MRDM method, called Mr-EP, which discovers a characterization of classes in terms of relational EPs thus providing a human-interpretable description of the differences between separate classes. The method was implemented in a MRDM system which is tightly integrated with a relational DBMS. The tight-coupling with the database makes the knowledge on data structure available free of charge to guide the search in the relational pattern space. Experimental results have been obtained by running Mr-EP to capture data changes among several populations of geo-referenced data. As future work, we plan to exploit relational EPs for associative classification tasks and to compare results with those already reported in a previous study [2].

Acknowledgment

This work is partial fulfillment of the research objective of ATENEO-2007 project "Metodi di scoperta della conoscenza nelle basi di dati: evoluzioni rispetto allo schema unimodale". The authors thank Nicola Barile for his help in reading a first draft of this paper.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) International Conference on Management of Data, pp. 207–216 (1993)
2. Ceci, M., Appice, A.: Spatial associative classification: propositional vs structural approach. Journal of Intelligent Information Systems 27(3), 191–213 (2006)
3. Dehaspe, L., Toivonen, H.: Discovery of frequent datalog patterns. Journal of Data Mining and Knowledge Discovery 3(1), 7–36 (1999)

4. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: International Conference on Knowledge Discovery and Data Mining, pp. 43–52. ACM Press, New York (1999)
5. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by aggregating emerging patterns. In: Arikawa, S., Furukawa, K. (eds.) DS 1999. LNCS (LNAI), vol. 1721, pp. 30–42. Springer, Heidelberg (1999)
6. Džeroski, S., Lavrač, N.: Relational Data Mining. Springer, Heidelberg (2001)
7. Fan, H., Ramamohanarao, K.: An efficient singlescan algorithm for mining essential jumping emerging patterns for classification. In: Chen, M.-S., Yu, P.S., Liu, B. (eds.) PAKDD 2002. LNCS (LNAI), vol. 2336, pp. 456–462. Springer, Heidelberg (2002)
8. Fan, H., Ramamohanarao, K.: A bayesian approach to use emerging patterns for classification. In: Australasian Database Conference, vol. 143, pp. 39–48. Australian Computer Society, Inc. (2003)
9. Fan, H., Ramamohanarao, K.: A weighting scheme based on emerging patterns for weighted support vector machines. In: Hu, X., Liu, Q., Skowron, A., Lin, T.Y., Yager, R.R., Zhang, B. (eds.) IEEE International Conference on Granular Computing, pp. 435–440. IEEE Computer Society Press, Los Alamitos (2005)
10. Li, J.: Mining Emerging Patterns to Construct Accurate and Efficient Classifiers. PhD thesis, University of Melbourne (2001)
11. Li, J., Dong, G., Ramamohanarao, K., Wong, L.: DeEPs: A new instance-based lazy discovery and classification system. Machine Learning 54(2), 99–124 (2004)
12. Li, J., Liu, H., Downing, J.: Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia. Bioinformatics 19(1), 71–78 (2003)
13. Li, J., Liu, H., Ng, S.-K., Wong, L.: Discovery of significant rules for classifying cancer diagnosis data. In: European Conference on Computational Biology, Supplement of Bioinformatics, pp. 93–102 (2003)
14. Lisi, F.A., Malerba, D.: Inducing multi-level association rules from multiple relations. Machine Learning 55, 175–210 (2004)
15. Liu, B., Hsu, W., Ma, Y.: Integrative classification and association rule mining. In: Proceedings of AAAI Conference of Knowledge Discovery in Databases (1998)
16. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 1(3), 241–258 (1997)
17. Plotkin, G.D.: A note on inductive generalization 5, 153–163 (1970)
18. Zhang, X., Dong, G., Ramamohanarao, K.: Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In: Terano, T., Chen, A.L.P. (eds.) PAKDD 2000. LNCS, vol. 1805, pp. 310–314. Springer, Heidelberg (2000)