Discovering Emerging Patterns in Spatial Databases: A Multi-relational Approach

Michelangelo Ceci, Annalisa Appice, and Donato Malerba

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

Abstract. Spatial Data Mining (SDM) has great potential in supporting public policy and in underpinning society functioning. One task in SDM is the discovery of characterization and peculiarities of communities sharing socio-economic aspects in order to identify potentialities, needs and public intervention. Emerging patterns (EPs) are a special kind of pattern which contrast two classes. In this paper, we face the problem of extracting EPs from spatial data. At this aim, we resort to a multi-relational approach in order to deal with the degree of complexity of discovering EPs from spatial data (i.e., (i) the spatial dimension implicitly defines spatial properties and relations, (ii) spatial phenomena are affected by autocorrelation). Experiments on real datasets are described.

1 Introduction

Spatial data are collected in a spatial database at a rate which requires automated data analysis methods to extract implicit, unknown, and potentially useful information. Data mining technology provides several data analysis tools for a variety of tasks. However, the presence of a spatial dimension in the data adds complexity to the data mining tasks. First, geometrical representation and positioning of spatial objects implicitly define spatial properties and relations. Second, spatial phenomena are characterized by autocorrelation (observations of spatially distributed variables are not location-independent). This means that when attributes of some units of analysis are investigated, attributes of *any* spatial object in the neighborhood of the 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.

In this work, we consider the spatial descriptive task of emerging patterns (EPs) discovery. Initially introduced in [3], EPs are kind of patterns (or multivariate features) whose support significantly changes from one class of data to another: the larger the difference of pattern support, the more interesting the patterns. Due to this sharp change in support, EPs can be used to characterize object classes. Several algorithms [8,3,4] have been proposed to discover EPs from data belonging to separate classes (data populations) and stored in a single relational table. But, the challenges posed by spatial dimension in the data makes

J.N. Kok et al. (Eds.): PKDD 2007, LNAI 4702, pp. 390–397, 2007.

necessary to resort to a powerful data representation in order to model properties and interactions possibly involving several spatial object types. In a recent work [5], the system SPADA has been proposed to deal with challenges posed by spatial dimension in the task of association rule discovery. But, although association rules and EPs are both descriptive patterns, they are significantly different: association rules capture *regularities* in data belonging to the same class, while EPs capture *changes* from a data class to another. This adds one source of complexity to the EPs discovery task, i.e., the fact that differently from association rules monotonicity property does not subsist for EPs.

We propose a Multi-Relational Emerging Patterns discovery (Mr-EP) algorithm that deals with the challenges posed by spatial dimension in the data. 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.

2 Problem Definition

We assume that a spatial database reduces to a relational database D with schema S once implicit spatial relationships between reference objects and task relevant objects have been extracted and stored in separate tables. In this perspective, reference objects, task-relevant objects and (spatial) interactions among them are tuples stored in tables of D. The set R of reference objects is the collection of tuples stored in a table T of D called target table. Each set R_i of task-relevant objects corresponds to a distinct table of D. Reference objects and task-relevant objects are described by means of both spatial and aspatial attributes. Similarly, the (spatial) interactions between different sets of spatial objects (reference objects and task-relevant objects) are stored in tables of D. The inherent "structure" of data, i.e., the (spatial) relations between reference objects and task-relevant objects is expressed in the schema S by the foreign key constraints (FK). By this mapping, the discovery of spatial EPs can be reformulated as the task of discovering (multi-)relational EPs. Before providing a formal definition of the problem to be solved, some definitions need to be introduced.

Definition 1 (Key, Structural and Property predicate)
Let S be a database schema.
– The "key predicate" associated with the target table for the task at hand 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.
– A "structural predicate" associated with the pair of tables $\{T_i, T_j\}$ in S such that there exists a foreign key FK in S between T_i and T_j, 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).
– A "property predicate" associated with the attribute ATT of the table T_i (which is neither primary nor foreign key) 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 representing a value belonging to the range of ATT.

Structural predicates are used to represent spatial relations between spatial objects. 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 2 (Relational pattern). *Let S be a database schema. A "relational pattern" P over S is a conjunction of predicates $p_0(t0_1), p_1(t1_1, t1_2), \ldots, p_m(tm_1, tm_2)$, 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.*

A spatial pattern is a relational pattern that involves objects and relations which have a spatial nature. Henceforth, we will also use the set notation for relational patterns, that is, a relational pattern is considered a set of atoms.

Definition 3 (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 4 (Completely linked relational pattern). *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 5 (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 (i) 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 $D_i\}$ and (ii) $\overline{D_i}$ 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 $s_{D_i}(P) = |O_P|/|O|$, 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 $GR^{\overline{D_i} \to D_i}(P) = s_{D_i}(P)/s_{\overline{D_i}}(P)$. As in [3], we assume that $GR(P) = \frac{0}{0} = 0$ and $GR(P) = \frac{>0}{0} = \infty$.

The problem of discovering spatial EPs can be formalized as follow.
Given: (i) A spatial database SDB to be reduced to a relational database D, (ii) a set R of reference objects tagged with a class label $Y \in \{C_1, \ldots, C_L\}$, (iii) Some

sets R_i, $1 \leq i \leq h$ of task-relevant objects, (iv) a pair of thresholds, that is, the minimum growth rate ($minGR \geq 1$) and the minimum support ($minsup > 0$). The goal is to discover the set of the *relational emerging patterns* to discriminate between reference objects belonging to contrasting classes in SDB.

In this work, we resort to the relational algebra formalism to express a relational emerging pattern P by means of an SQL query.

3 Relational EPs Discovery

We have adapted 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 [6] 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 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 [7].

Definition 6 (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 7 (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 be searched according to a downward refinement operator which computes the set of refinements for a completely linked relational pattern.

Definition 8 (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' for EPs.

Definition 9 (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 [8], 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:
$SameSupp(P)_{\overline{D_i}} = \{P' \in \rho'(P) | s_{\overline{D_i}}(P) = s_{\overline{D_i}}(P'),\ P' = P \wedge p(t1,t2),$
$p(t1,t2)$ is a property predicate$\}$.

For the monotonicity property, $\forall P' \in SameSupp_{\overline{D_i}}(P)$: $s_{D_i}(P) \geq s_{D_i}(P')$. 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 $SameSupp_{\overline{D_i}}(P)$. 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 $P' \in SameSupp_{\overline{D_i}}(P)$ 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 $P''' = P \cup \{q(t3,t4)\} \notin SameSupp_{\overline{D_i}}(P)$.*

The proof is reported in [2]. 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*).

4 Experimental Results

Spatial EPs have been discovered in two spatial databases named North-West England (NWE) Data and Munich Data. EPs have been discovered with *min* $GR = 1.1$, $minsup = 0.1$. MAX_M is set to 3 for NWE Data and to 5 for Munich Data.

NWE data (provided in the European project *SPIN!*) concern both 214 census sections (wards) of Greater Manchester and digital maps data. Census data describe the mortality percentage rate and four deprivation indexes: Jarman (need for primary care), Townsend and Carstairs (health-related analyzes) and DoE (for targeting urban regeneration funds). The higher the index value the more deprived the ward. The mortality rate (target attribute) takes values in the finite set $\{low, high\}$. 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). The number of "non disjoint" relationships is 5313. Mr-EP discovers 60 EPs to discriminate high from low mortality rate wards and 55 EPs to discriminate low from high mortality rate wards. In the following, some EPs are reported. For the class mortality_rate=high:

$wards(A) \wedge wards\_rails(A, B) \wedge wards\_doeindex(A, [6.5..9.2])$

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 $wards\_doeindex(A, [6.5..9.2])$ (i.e. A is a deprived zone to be considered as target zone for regeneration fundings) is a property predicate. This pattern has 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 Oracle Spatial 10g query:

```
SELECT distinct W.ID FROM Wards W, Rails R
WHERE RELATE(W.Geometry,R.Geometry)='INTERSECTS'
   AND W.DoEIndex between 6.5 and 9.2
```

An example of EP discovered for the class mortality_rate=low:

$wards(A) \wedge wards\_townsendidx(A, [-3.8.. - 2.01] \wedge wards\_greenareas(A, B)$

This pattern has a support of 0.113 and a growth rate of 2.864. It captures the event that a ward with a low townsend deprivation level (i.e., A is not deprived from the point of view provided by health-related analysis) which overlaps at least one green area discriminates wards with low mortality rate from the others.

Munich data describe the level of monthly rent per square meter for flats in Munich expressed in German Marks. Data describe 2180 flats located in the 446 suquarters 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 U-Bahn stops, 15 S-Bahn stops and 1 railway station) within Munich are available for this study.The "area" of subquarters is obtained by the spatial dimension of this data. 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) have been discretized into the two intervals $low = [2.0, 14.0]$ or $high =]14.0, 35.0]$. The "close to" relation between districts (autocorrelation on districts) and the "inside" relation between apartments and districts have been considered. Mr-EP discovers 31 (31) EPs to discriminate the apartments with high (low) rent rate per square meters from the class of apartments with low (high) rent rate per square meters. In the following, some EPs are reported. For rate_per_squaremeters=high:

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

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 that 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. This pattern 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:

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

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 form the others. It can be motivated by considering that old buildings do not offer the same facilities of a recently built apartment.

5 Conclusions

In this paper, we present a spatial data mining method that resorts to a MRDM approach to discover a characterization of classes in terms of EPs involving spatial objects and relations thus providing a human-interpretable description of the differences between separate classes of spatially referenced data. The method is implemented in a system that is tightly integrated with a Oracle $10g$ DBMS. The tight-coupling with the database makes the knowledge on data structure available free of charge to guide the search in the pattern space by taking into account spatial interaction implicit in spatial dimension. Spatial EPs have been used to capture data (spatial) changes among several populations of geo-referenced data.

Acknowledgments

This work is supported by "ATENEO-2007" project "Metodi di scoperta della conoscenza nelle basi di dati: evoluzioni rispetto allo schema unimodale".

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. Appice, A., Ceci, M., Malgieri, C., Malerba, D.: Discovering relational emerging patterns. In: Basili, R., Pazienza, M. (eds.) AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, LNAI. Springer, (to appear)
3. 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)
4. Li, J.: Mining Emerging Patterns to Construct Accurate and Efficient Classifiers. PhD thesis, University of Melbourne (2001)
5. Lisi, F.A., Malerba, D.: Inducing multi-level association rules from multiple relations. Machine Learning 55, 175–210 (2004)
6. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 1(3), 241–258 (1997)
7. Plotkin, G.D.: A note on inductive generalization. 5, 153–163 (1970)
8. Zhang, X., Dong, G., Ramamohanarao, K.: Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In: Knowledge Discovery and Data Mining, pp. 310–314 (2000)