Discovering Associations between Spatial Objects: An ILP Application

Donato Malerba and Francesca A. Lisi

Dipartimento di Informatica, Università degli Studi di Bari,
Via Orabona 4, 70126 Bari, Italy
{malerba, lisi}@di.uniba.it

Abstract. In recent times, there is a growing interest in both the extension of data mining methods and techniques to spatial databases and the application of inductive logic programming (ILP) to knowledge discovery in databases (KDD). In this paper, an ILP application to association rule mining in spatial databases is presented. The discovery method has been implemented into the ILP system SPADA, which benefits from the available prior knowledge on the spatial domain, systematically explores the hierarchical structure of task-relevant geographic layers and deals with numerical aspatial properties of spatial objects. It operates on a deductive relational database set up by selecting and transforming data stored in the underlying spatial database. Preliminary experimental results have been obtained by running SPADA on geo-referenced census data of Manchester Stockport, UK.

1 Introduction

One of the great challenges for the near future is knowledge discovery in ever growing spatial sets [4]. Nevertheless, most work in the KDD community up to now has been almost exclusively focused on pattern discovery in relational and transaction databases. Only in recent times, data mining methods and techniques have been proposed for *the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases* [8]. Peculiarity of the spatial domain is that the attributes of the neighbors of some spatial object of interest may have an influence on it and therefore have to be considered as well [6]. Thus, spatial data mining algorithms cannot neglect the implicit relations of spatial neighborhood (e.g. topological relations) that are defined by the explicit location and extension of spatial objects.

As the interest in KDD is generally increasing, many recent applications of ILP methods and techniques to KDD have also emerged [3]. We claim that spatial data mining is a promising ILP application domain for two main reasons. First, ILP relies on the theory of computational logic which supplies representation and reasoning means appropriate for the spatial domain where relations among objects play a key role and are often inferred by qualitative reasoning. Second, ILP offers an elegant solution to multi-relational mining whereas traditional approaches to spatial data mining usually solve the problem by collapsing multiple relations into the universal

C. Rouveirol and M. Sebag (Eds.): ILP 2001, LNAI 2157, pp. 156–163, 2001.

relation [9]. To the best of our knowledge, very few contributions from ILP to knowledge discovery in spatial databases have been reported in the literature. GwiM [14] is a general-purpose ILP system that can solve several spatial data mining tasks, though no insight in the algorithmic issues has been provided. INGENS [11] is an inductive GIS with learning capabilities that currently support the classification task.

In this paper, we focus our attention on the task of mining *spatial* association rules, namely the detection of associations between spatial objects, and propose to accomplish the task by means of a novel special-purpose ILP system, called SPADA (Spatial PAttern Discovery Algorithm) [12]. It benefits from the available prior knowledge on the spatial domain, systematically explores the hierarchical structure of task-relevant geographic layers and deals with numerical aspatial properties of spatial objects. Furthermore, it operates on a *deductive relational database* (DDB) set up by selecting and transforming data stored in the underlying spatial database. The analysis of geo-referenced census data have been chosen as an application domain. Indeed, the advances in the practice of geo-referencing socioeconomic phenomena allow census data to be conceptualized as spatial objects with numerical aspatial properties.

The paper is organized as follows. Section 2 introduces the spatial data mining problem solved by SPADA. Experimental results on geo-referenced census data of Stockport, one of the ten Metropolitan Districts of Greater Manchester, UK, are reported in Section 3. Conclusions are given in Section 4.

2 Mining Spatial Association Rules with SPADA

The discovery of spatial association rules is a descriptive mining task aiming at the detection of associations between *reference objects* and *task-relevant objects*, the former being the main subject of the description while the latter being spatial objects that are relevant for the task at hand and spatially related to the former. For instance, we may be interested in describing a given area by finding associations among large towns (reference objects) and spatial objects in the road network, hydrography and administration layers (task-relevant objects). Some kind of taxonomic knowledge on task-relevant geographic layers may also be taken into account to get descriptions at different concept levels (*multiple-level association rules*). As usual in association rule mining, we search for associations with large support and high confidence (*strong rules*). Formally, SPADA can solve the following spatial data mining problem:
Given
- a spatial database (SDB),
- a set of reference objects S,
- some task-relevant geographic layers R_k, $1 \leq k \leq m$, together with spatial hierarchies defined on them,
- two thresholds for each level l in the spatial hierarchies, *minsup*[l] and *minconf*[l]

Find strong multiple-level spatial association rules.

To solve the problem Koperski and Han propose a top-down, progressive refinement method which exploits taxonomies both on topological relations and spatial objects [9]. The method has been implemented in the module Geo-associator of the

spatial data mining system GeoMiner [7]. We propose an upgrade of Geo-Associator to first-order logic representation of data and patterns. The approach is inspired to the work on multi-relational data mining reported in [2] and operates on a DDB set up by a preliminary feature extraction step from SDB and denoted $D(S)$. In particular, we resort to Datalog [1], whose expressive power allows us to specify also prior knowledge (BK) such as spatial hierarchies, spatial constraints and rules for spatial qualitative reasoning. Given a set of Datalog atoms A, a *spatial association rule* in $D(S)$ is an implication of the form $P \rightarrow Q$ (s%, c%), where $P \subseteq A$, $Q \subseteq A$, $P \cap Q = \emptyset$, and at least one atom in $P \cup Q$ represents a spatial relationship. The percentages s and c are called the support and the confidence of the rule respectively. An example of spatial association rule in our framework is:

is_a(A, large_town), intersects(A,B), intersects(A,C), is_a(C, regional_road), intersects(D,C), D\=A, C\=B $\rightarrow$ is_a(B, main_trunk_road), is_a(D, large_town) (54%, 86%)

"GIVEN THAT *54% of large towns intersect both a main trunk road and a regional road the latter intersecting a large town distinct from the previous one*, IF *a large town A intersects two spatial objects the former being an unknown B while the latter being a regional road which in turn intersects some spatial object D distinct from A* THEN WITH CONFIDENCE 86% *B is a main trunk road and D is a large town*".

The choice of an ILP algorithm to accomplish the mining task at hand heavily affects the whole KDD process. Indeed, $D(S)$ is obtained by selecting and transforming the portion of *SDB* that concerns the set of reference objects S and adding it to the *BK*. Data selection encompasses the retrieval of spatial objects eventually together with their spatial and aspatial properties and the extraction of spatial relationships between reference objects and task-relevant objects. In particular, SPADA can extract topological relations whose semantics has been defined according to the 9-intersection model [5]. It is noteworthy that finding the right compromise between on-line computation (time-consuming solution) and materialization (space-consuming solution) of spatial relations is a hot topic in spatial data mining. More sophisticated computational solutions are reported in [6, 9]. Once selected, this data needs to be transformed in a suitable format. For instance, numerical properties of spatial objects with a large domain must be discretized in order to be handled by logic-based data mining methods. SPADA currently implements an adaptation of the relative unsupervised discretization algorithm RUDE [10] to the first-order case.

The spatial data mining step requires the solution to two sub-tasks: 1) Find large (or frequent) spatial patterns; 2) Generate strong spatial association rules. The reason for this decomposition is that frequent patterns are commonly not considered useful for presentation to the user as such. They can be efficiently post-processed into association rules that exceed given threshold values of support and confidence. It is noteworthy that SPADA, analogously to Geo-Associator but differently from WARMR [2], exploits is-a taxonomies for extracting multiple-level patterns and association rules. Thus, *largeness* and *strength* depend on the level currently explored in the hierarchical structure of task-relevant geographic layers. To be more precise, a pattern P is *large* (or frequent) at level l if $\sigma(P) \geq minsup[l]$ and all ancestors of P with respect to the spatial hierarchies are large at their corresponding levels. A spatial association rule $P \rightarrow Q$ is *strong* at level l if the pattern $P \cup Q$ is large and the confidence is high at

level l, namely $\varphi(Q|P) \geq minconf[l]$. In SPADA, the counting procedures for support and confidence are based on the coverage test of spatial observations, being it the ILP counterpart of counting the number of reference objects that satisfy a certain spatial pattern. Indeed, the s*patial observations* are portions of $D(S)$, each of which concerns one and only one reference object. Thus, the two percentages associated to $P \rightarrow Q$ mean that s% of spatial observations in $D(S)$ are covered by $P \cup Q$ and c% of spatial observations in $D(S)$ that are covered by P are also covered by $P \cup Q$ respectively.

Further details about representation and algorithmic issues can be found in [12].

3 An Application to Stockport Census Data

In some works on spatial representation from the social scientist's perspective, socio-economic phenomena have been conceptualized as *spatial objects* in the sense of entities having both spatial location and spatially independent attribute characteristics [13]. Population data are among the potentially spatial socioeconomic data. They are usually geo-referenced with respect to areal spatial objects such as census zones, electoral constituencies, local government areas, or regular grid squares. In the UK, for instance, the geo-referencing areal units are ED (enumeration district), Ward, District, and County. They form a hierarchy based on the *inside* relationship among locations. Thus the ED is the smallest unit for which census data are published nowadays. Furthermore, the digital ED boundaries produced for the 1991 UK census enable the spatial representation of census data in the computer databases. Generally speaking, population censuses of the 1990s provided an added impetus to the application of GIS to socioeconomic uses. One of the most interesting topic areas for identifying potential users of such GIS applications is the public debate over Unitary Development Plans (UDP) in the UK. The district chosen for investigation is Stockport, one of the ten Metropolitan Districts of Greater Manchester, UK. It is divided into twenty-two wards for a total of 589 EDs. The case study is expected to show the potential benefit of data mining methods and techniques to one or more potential users. In particular, census data are extremely important for policy analysis and, once geo-referenced and conceptualized as spatial objects with numerical aspatial properties, supply a good test-bed to SPADA. Thus census data (89 tables, each with 120 attributes in average) and digital ED boundaries have been loaded into an Oracle-Spatial database, i.e. a relational DBMS extended with spatial data handling facilities. The ED code allows the joining of the two kinds of data and the generation of test data.

We have focused our attention on transportation planning, which is one of key issues in the UDP. Let us suppose that some decision-making process about the motorway M63 is ongoing. Describing the area of Stockport served by M63 (i.e. the wards of Brinnington, Cheadle, Edgeley, Heaton Mersey, South Reddish) may be of support to the planners. In this paper we report the preliminary results obtained by applying SPADA to the task of discovering multiple-level spatial association rules relating EDs intersected by the motorway M63 (S) and all EDs in the area served by M63 (R) to be characterized with respect to data about commuting.

This spatial data mining query raises several application issues for SPADA. First, census data are available at the ED level. Thus, an is-a hierarchy for the Stockport ED layer has been obtained by grouping EDs on the basis of the ward they belong to (see Figure 1) and expressed as Datalog facts in BK. Indeed, the current version of SPADA deals only with *is-a* hierarchies where the is-a relationship is overloaded, i.e. it may stand for *kind-of* as well as for *instance_of* depending on the context. Further is-a hierarchies could be derived by resorting to clustering algorithms.

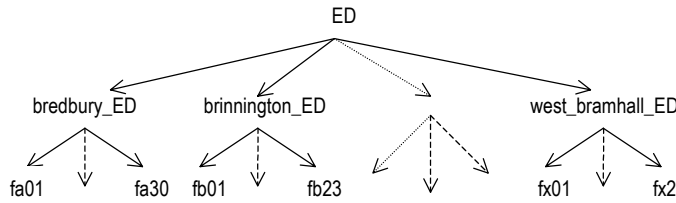

Fig. 1. An is-a hierarchy for the Stockport ED layer

Second, census data are all numeric (more precisely, integer values). The attributes that we have selected for this experiment (see Table 1) refer to residents aged 16 and over, thus they have been normalized with respect to the total number of residents aged 16 and over (s820001). Each couple of consecutive cut points *a* and *b* has generated an interval of the kind [a..b].

Last, some spatial computation is necessary. In particular, the relations of intersection (EDs-motorways) and adjacency (EDs-EDs) have been extracted as concerns the area of interest and transformed into Datalog facts of $D(S)$. It is noteworthy that the relations of accessibility and closeness have been defined by means of spatial qualitative reasoning:

```
linked_to(X, Y) :- intersect(X, m63), intersect(Y, m63), Y\=X.
close_to(X, Y) :- adjacent_to(X, Z), adjacent_to(Z, Y), Y\=X.
```

These rules have been added to BK together with the aforementioned spatial hierarchies and also the spatial constraint:

```
ed_on_M63(X) :- intersect(X, m63).
```

which defines the instances of S.

SPADA has been run on the obtained $D(S)$ with thresholds $min\_sup[1]=0.7$ and $min\_conf[1]=0.9$ at the first level, and $min\_sup[2]=0.5$ and $min\_conf[2]=0.8$ at the second level. The whole discovery process has taken 490.21 sec on a PC Pentium III with 128 Mb RAM (37.84 sec for level 1, and 452.31 sec for level 2). It has returned 744 frequent patterns out of 17619 candidate patterns and 24964 strong rules out of 40465 generated rules. Some interesting patterns have been discovered. For instance, at level $l=2$ in the spatial hierarchies, the following candidate P:

```
ed_on_M63(A), close_to(A,B), is_a(B, south_reddish_ED), linked_to(A,C), C\=B,
        s820161(C, [52.632..54.167]), is_a(C, cheadle_ED)
```

has been generated after $k=6$ refinement steps and has been evaluated with respect to $D(S)$. Since six of ten spatial observations ($|S|=10$) are covered and all the ancestor

patterns are large at their level ($l \leq 2$), the pattern is a large one at level l=2 with 60% support. For the sake of clarity, the following pattern

ed_on_M63(A), close_to(A,B), is_a(B, ed_in_M63_area), linked_to(A,C), C\=B,
s820161(C, [52.632..54.167]), is_a(C, ed_in_M63_area)

is one of the large ancestors for the pattern *P*. It has been generated after k=6 refinement steps at level l=1 and is supported by 90% EDs intersected by M63. Such way of taking the taxonomies into account during the pattern discovery process implements what we refer to as the systematic exploration of the hierarchical structure of task-relevant geographic layers. Furthermore, the use of both variables and atoms of the kind \= allow SPADA to distinguish between multiple instances of the same class of spatial objects (e.g. the class ed_in_M63_area).

Table 1. Numerical attributes in the application to Stockport census data.

Attribute	Description	Cut points in the attribute domain
s820161	Persons who work out of the district of usual residence and drive to work	0.0, 6.25, 8.333, 12.973, 17.241, 19.048, 20.943, 23.529, 25.0, 25.926, 27.586, 29.032, 29.865, 31.25, 33.333, 34.375, 36.182, 38.235, 40.0, 42.105, 45.455, 46.667, 48.194, 50.0, 51.515, 52.632, 54.167, 56.0, 57.143, 58.333, 58.824, 60.0, 60.714, 61.538, 63.889, 65.217, 66.667, 67.742, 69.565, 71.429, 72.902, 100.0
s820213	Employees and self-employed who reside in households with 3 or more cars and drive to work	0.0, 2.222, 15.385, 28.0, 29.521, 31.034, 33.333, 35.068, 37.5, 38.095, 38.889, 41.043, 42.857, 48.387, 72.727
s820221	Employees and self-employed who reside in households with 3 or more cars and work out of the district of usual residence	0.0, 2.222, 4.762, 9.091, 10.345, 13.636, 18.182, 19.355, 21.131, 23.529, 25.0, 28.571

One of the strong rules that have been derived from the frequent pattern *P* is:

ed_on_M63(A), close_to(A,B), is_a(B,south_reddish_ED)
→ linked_to(A,C), is_a(C,cheadle_ED), B\=C, s820161(C,[52.632..54.167]) (60%, 100%)

"GIVEN THAT *60% of EDs intersected by M63 are close to a South Reddish ED and are linked via M63 to a Cheadle ED where 52-54% residents aged 16 and over work out of the district of usual residence and drive to work,* IF *an ED intersected by M63 is close to a South Reddish ED* THEN WITH CONFIDENCE 100% *it is linked via M63 to a Cheadle ED where …"*.

Other examples of strong rule at the second level are:

ed_on_M63(A), close_to(A,B), s820221(B,[10.345..13.636])
→ linked_to(A,C), is_a(C,brinnington_ED), B\=C (60%, 86%)

"GIVEN THAT *60% of EDs intersected by M63 are close to an ED - where 10-13% residents aged 16 and over are employees and self-employed who reside in households with 3 or more cars and work out of the district of usual residence - and are linked via M63 to a Brinnington ED distinct from the previous one,* IF *an ED intersected by M63 is close to an ED where* … THEN WITH CONFIDENCE 86% *it is linked via M63 to a Brinnington ED distinct from the previous one"*.

ed_on_M63(A), close_to(A,B), s820221(B,[19.355..21.131])
→ is_a(B,heaton_mersey_ED) (70%, 100%)

"GIVEN THAT *70% of EDs intersected by M63 are close to a Heaton Mersey ED where 19-21% residents aged 16 and over are employees and self-employed who reside in households with 3 or more cars and work out of the district of usual residence* IF *an ED intersected by M63 is close to an ED where …* THEN WITH CONFIDENCE 100% *the latter ED belongs to the ward of Heaton Mersey"*.

One may wonder whether these frequent patterns and strong rules convey novel knowledge and, in positive case, what kind of knowledge. The evaluation of data mining results is beyond the scope of this paper. Nevertheless a naive interpretation of results in our application might lead us to state that the motorway M63 intersects an area of Stockport which is characterized by a high percentage of commuters by car who may benefit from some improvement of the road network.

4 Conclusions and Future Work

The work presented in this paper reports an ILP application to spatial association rule mining. Experimental results obtained by applying the novel special-purpose ILP system SPADA to geo-referenced census data of Manchester Stockport show that the expressive power of first-order logic enables us to tackle applications that cannot be handled by the traditional approach to spatial data mining. Furthermore, DDBs offer effective representation means for domain knowledge, constraints and qualitative reasoning. In particular, we can embed rules for the inference of implicit spatial relationships that are too numerous to be either stored in the spatial database or computed by computational geometry algorithms.

For the near future, we plan to face the issues of efficiency and scalability in SPADA. Particular attention will be also drawn on the issue of robustness. Indeed, data pre-processing in spatial data mining is remarkably error-prone. For instance, the generation of the predicate close_to is based on the user-defined semantics of the closeness relation, which should necessarily be approximated. Further work on the data selection and transformation is expected to give some hints on noise handling in this application domain. As for the test on real-world spatial data sets, much work has still to be done. In particular, we are interested in experiments with mixed census-topographic data because they show that the interpretation of spatial relations can change as spatial objects are added.

Acknowledgements

This work is supported by the IST project SPIN! (http://www.ccg.leeds.ac.uk/spin/). We would like to thank Jim Petch, Keith Cole and Mohammed Islam (MIMAS, University of Manchester, England) and Chrissie Gibson (Department of Environmental and Geographical Sciences, Manchester Metropolitan University, England) for providing census data and digital ED boundaries of Manchester Stockport.

References

1. Ceri, S., Gottlob, G., Tanca, L.: What you Always Wanted to Know About Datalog (And Never Dared to Ask). IEEE Transactions on Knowledge and Data Engineering 1(1) (1989) 146-166.
2. Dehaspe, L., Toivonen, H.: Discovery of frequent Datalog patterns. Data Mining and Knowledge Discovery 3(1) (1999) 7-36.
3. Dzeroski, S.: Inductive Logic Programming and Knowledge Discovery in Databases. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press (1996) 117-152.
4. Egenhofer, M.J., Glasgow, J., Günther, O., Herring, J.R., Peuquet, D.J.: Progress in Computational Methods for Representing Geographic Concepts. Int. J. Geographical Information Science 13(8) (1999) 775-796.
5. Egenhofer, M.J., Herring, J.R.: Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. In Egenhofer, M.J., Mark, D.M., Herring, J.R. (eds): The 9-Intersection: Formalism and its Use for Natural-language Spatial Predicates. Technical Report 94-1, U.S. NCGIA (1994.).
6. Ester, M., Kriegel, H.P., Sander, J.: Spatial Data Mining: A Database Approach. In: Scholl, M., Voisard, A. (Eds.): Advances in Spatial Databases. LNCS 1262, Springer-Verlag, Berlin (1997) 47-66.
7. Han, J., Koperski, K., Stefanovic, N.: GeoMiner: A System Prototype for Spatial Data Mining. In: Peckham, J. (Ed.): SIGMOD 1997, Proceedings of the ACM-SIGMOD Int. Conf. on Management of Data. SIGMOD Record 26(2) (1997) 553-556.
8. Koperski, K., Adhikary, J., Han, J.: Spatial Data Mining: Progress and Challenges. In: Proc. Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada (1996).
9. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases. In: Egenhofer, M.J., Herring, J.R. (Eds.): Advances in Spatial Databases. LNCS 951, Springer-Verlag, Berlin (1995) 47-66.
10. Ludl, M.-C., Widmer, G.: Relative Unsupervised Discretization for Association Rule Mining. In D.A. Zighed, H.J. Komorowski, J.M. Zytkow (Eds.): Principles of Data Mining and Knowledge Discovery. LNCS 1910, Springer-Verlag, Berlin (2000) 148-158.
11. Malerba, D., Esposito, F., Lanza, A., Lisi, F.A.: Discovering geographic knowledge: The INGENS system. In: Ras, Z.W., Ohsuga, S. (Eds.): Foundations of Intelligent Systems, LNAI 1932, Springer-Verlag , Berlin (2000) 40-48.
12. Malerba, D., Esposito, F., Lisi, F.A.: A Logical Framework for Frequent Pattern Discovery in Spatial Data. In: Russell, I., Kolen, J. (Eds.): Proc. 14th Int. FLAIRS Conference. AAAI, Menlo Park:CA (2001) 557-561.
13. Martin, D.J.: Spatial representation: the social scientist's perspective. In: Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. (eds): Geographical Information Systems, vol.1, Principles and Technical Issues, 2nd edition. John Wiley & Sons (1999) 71-80.
14. Popelinsky, L.: Knowledge Discovery in Spatial Data by means of ILP. In: Zytkow, J.M., Quafalou, M. (Eds.): Principles of Data Mining and Knowledge Discovery. LNAI 1510, Springer-Verlag, Berlin (1998) 185-193.