Mining and Filtering Multi-level Spatial Association Rules with ARES

Annalisa Appice, Margherita Berardi, Michelangelo Ceci, and Donato Malerba

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

Abstract. In spatial data mining, a common task is the discovery of spatial association rules from spatial databases. We propose a distributed system, named ARES that takes advantage of the use of a multi-relational approach to mine spatial association rules. It supports spatial database coupling and discovery of multi-level spatial association rules as a means for spatial data exploration. We also present some criteria to bias the search and to filter the discovered rules according to user's expectations. Finally, we show the applicability of our proposal to two different real world domains, namely, document image processing and geo-referenced analysis of census data.

1 Introduction

Spatial data mining investigates the problem of extracting pieces of knowledge from data describing spatial objects, which are characterized by a geometrical representation (e.g. point, line, and region in a 2D context) and a position with respect to some reference system. The relative positioning of spatial objects defines implicitly spatial relations of different nature, such as directional and topological. The goal of spatial data mining methods is to extract spatial patterns, that is, patterns involving spatial relations between mined objects such that they are certain, previously unknown, and potentially useful for the specific application [10].

In [13] the authors have proposed a spatial data mining method, named SPADA (Spatial Pattern Discovery Algorithm), that discovers spatial association rules, that is, association rules involving spatial objects and relations. It is based on an Inductive Logic Programming (ILP) approach to (multi-) relational data mining [5] and permits the extraction of multi-level spatial association rules, that is, association rules involving spatial objects at different granularity levels. For each granularity level, SPADA operates in three different phases: i) pattern generation; ii) pattern evaluation; iii) rule generation and evaluation.

In this paper, we describe the integration of SPADA in a full-fledged spatial data mining system, named ARES (Association Rules Extractor from Spatial data), that assists data miners in the complex process of extracting the units of analysis from the spatial database, specifying the background knowledge on the application domain and defining some form of search bias. The last aspect is particularly relevant, since the

M.-S. Hacid et al. (Eds.): ISMIS 2005, LNAI 3488, pp. 342 – 353, 2005.

number of discovered patterns or association rules is usually high and the interest of most of them does not fulfill user expectations. The new spatial data mining tool has been applied to two different domains, namely, document image processing and geo-referenced analysis of census data, thus proving the generality of the proposed solution.

The paper is organized as follows. The problem of mining spatial association rules is reported in Section 2. In Section 3, some related works are presented. Section 4 describes the ARES distributed architecture that supports the interface of SPADA with a spatial database by generating high-level logic descriptions of spatial data. Some filtering mechanisms implemented in SPADA are described in Section 5. Finally, the application of ARES to two case studies is described in Sections 6.

2 Mining Spatial Association Rules

The discovery of spatial association rules is a descriptive mining task aiming to detect associations between *reference objects* (*ro*) and some *task-relevant objects* (*tro*). The former are the main subject of the description, while the latter are spatial objects that are relevant for the task in hand and are spatially related 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 and Q are a set of literals, called *items*, such that $P \cap Q = \varnothing$, 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*. A *spatial pattern* expresses a spatial relationship among spatial objects and can be expressed by means of predicate calculus. A *spatial association rule* is an association rule whose corresponding pattern is spatial. An example of spatial association rule is:

is_a(X, large_town), intersects(X,Y), is_a(Y, road) $\Rightarrow$
intersects(X, Z), is_a(Z, road), Z$\neq$*Y (91%, 100%)*

to be read as "If a large town X intersects a road Y then X intersects a road Z distinct from Y with 91% support and 100% confidence." By taking into account some kind of taxonomic knowledge on task-relevant objects 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, large_town), intersects(X, Y), is_a(Y, regional_road) $\Rightarrow$
intersects(X, Z), is_a(Z, main_trunk_road), Z$\neq$*Y (65%,71%)*

The problem of mining association rules solved by SPADA can be formally stated as follows:

Given a spatial database (SDB), a set of reference objects S, some sets R_k, $1 \leq k \leq m$, of task-relevant objects, a background knowledge *BK* including some hierarchies H_k on objects in R_k , M granularity levels in the descriptions (1 is the highest while M is the lowest), a language bias *LB* that constrains the search space and a couple of thresholds *minsup*[l] and *minconf*[l] for each granularity level;

Find strong multi-level spatial association rules. Each R_k is typically a layer of the spatial database while hierarchies define *is-a* (i.e., taxonomical) relations of spatial objects in the same layer.

To deal with several hierarchies at once in a uniform manner, objects in them are mapped to one or more of the M user-defined description granularity levels so that frequency of patterns as well as strength of rules depend on the level l of granularity with which patterns/rules describe data. To be more precise, 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].

3 Related Works

The problem of mining multi-level association rules has its roots in the seminal work by Han and Fu [6], where the authors proposed an improved version of the Apriori algorithm [1] to handle multiple level association rules by remapping the original database and performing the mining by a progressive deepening of the levels. A slight variant of this approach has been adopted by Koperski and Han [11] and implemented in the system GeoMiner [7] in the context of spatial data mining.

A different solution is adopted in the work by Morimoto [17], where different types of patterns, called "Frequent neighboring class sets" are extracted. In particular, first the spatial dataset is partitioned by considering closeness relations and then the patterns are discovered on single partitions.

Another solution is proposed by Ding [4], where rules are extracted from RSI (Remote Sensed Imagery) data by means of the Peano Count Tree (P-tree) structure, which is a lossless representation of the image data. By using P-trees, association rule mining algorithm with fast support calculation and significant pruning techniques is possible. Although RSI data can be considered a kind of spatial data, rules discovered are not "spatial" in a strict sense.

All solutions reported above suffer from severe limitations due to the restrictive data representation formalism, known as *single-table assumption* [5]. More specifically, it is assumed that data to be mined are represented in a single table (or relation) of a relational database, such that each tuple represents an independent unit of the sample population and columns correspond to properties of units. In spatial data mining applications this assumption turns out to be a great limitation. Indeed, different geographical objects may have different properties, which can be properly modeled by as many data tables as the number of object types. The ILP system WARMR [3] overcomes this limitation by resorting to multi-relational data mining. It supports the discovery of frequent DATALOG patterns by adapting Mannila and Toivonen's levelwise method [16] to the case of conjunctive formulas which are organized according to θ-subsumption. Although it has been presented as a system able to use *is-a* hierarchies, WARMR is not a system for mining multi-level association rules because it lacks of mechanisms for taxonomic reasoning [13]. Moreover, the transformation of frequent patterns into association rules requires the specification of a list of possible patterns that can occur in the antecedent or conseguent of a rule. Therefore, the specification of a bias for the output rules is a must and not an additional opportunity given to the user to filter some rules. Finally, WARMR is not integrated in a system for spatial data analysis that supports users to extract both spatial and apatial properties of either reference objects or task-relevant object.

Therefore, to the best of our knowledge, SPADA can be considered the only multi-relational data mining method especially conceived for spatial data mining tasks and implemented in a system, named ARES, that supports the user in all preprocessing steps.

4 The Architecture of ARES

ARES has a distributed architecture based on a client-server model (see Figure 1). SPADA is on the server side, so that several data mining tasks can be run concurrently by multiple users. SPADA is implemented in Prolog and allows to specify both the background knowledge *BK* (hierarchies are expressed by a collection of ground atoms that define the binary predicate *is_a*, while domain specific knowledge is expressed as sets of definite clauses) and a language bias *LB* that constrains the search for patterns.

On the client side, the system includes a Graphical User Interface (GUI) implemented as Java application, which provides the user with facilities for controlling all parameters of the data mining process as well as the module RUDE (relative unsupervised discretization algorithm), which discretizes a numerical attribute of a relational database in the context defined by other attributes [14].

The SDB (Oracle Spatial) can run on a third computation unit. Many spatial features (relations and attributes) can be extracted from spatial objects stored in the SDB. The feature extraction requires complex data transformation processes to make spatial relations explicit and representable as ground Prolog atoms. Therefore, a middle layer module is required to make possible a loose coupling between SPADA and the SDB by generating features of spatial objects. The module, named FEATEX (Feature Extractor), is implemented as an Oracle package of procedures and functions, each of which computes a different feature [2]. According to their nature, features extracted by FEATEX can be distinguished as geometrical, directional and topological features. Geometrical features (e.g. area, length) are based on principles of Euclidean geometry, directional features (e.g. north, south,) regard relative spatial orientation in 2D, while topological features (e.g. crosses, on top) are relations preserving themselves under topological transformations such as translation, rotation, and scaling. In addition, hybrid features (e.g. roughly parallel), which merge properties of two or more feature categories, can be also extracted by FEATEX.

On the client side, the system WISDOM++ [15] can be used to extract spatial data from document image and store them in the SDB. The process performed by WISDOM++ consists of the preprocessing of the raster image of a scanned paper document, the segmentation of the preprocessed raster image into basic layout components, the classification of basic layout components according to the type of content (e.g., text, graphics, etc.), the identification of a more abstract representation of the document layout (layout analysis), the classification of the document in one of predefined categories (e.g. business letter, scientific paper) on the basis of its layout and content, and the identification of semantically relevant layout components (e.g. title, abstract of a scientific paper) called logical components (*document image understanding* [15]). The final representation includes both layout structure (extracted in the layout analysis) and logical structure (semantic information extracted by means

of document classification and understanding) computed on the original image. A further processing step stores the output structures in the SDB. WISDOM++ makes use of an Oracle Database to store intermediate data.

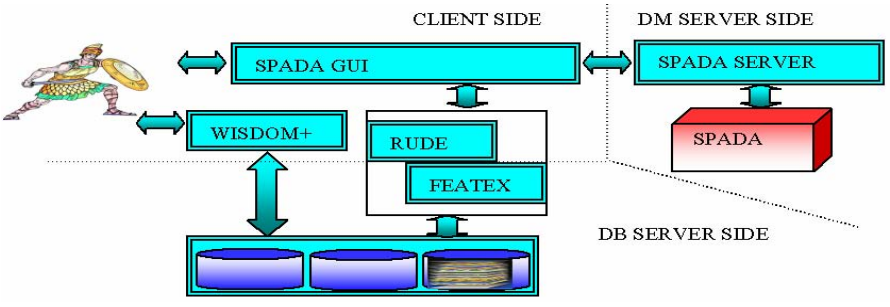

Fig. 1. ARES architecture

In order to handle spatial data provided by WISDOM++, FEATEX has been extended to allow features of layout components to be extracted (see section 5.1).

5 Filtering Patterns and Association Rules

The efficiency of the data mining process is very important to tackle real-world problems. In order to improve the efficiency of the search process, SPADA associates each candidate pattern with backward pointers to parent patterns both at the same granularity level (intra-space parenthood) and at higher granularity levels (inter-space parenthood). Backward pointers are profitably exploited in the pattern generation phase to prevent the generation of some infrequent patterns [12]. In a more recent release of SPADA (3.0), backward pointers are also exploited in the pattern evaluation phase. Indeed, by associating each pattern with the list of support objects, it is possible to perform the evaluation of each pattern solely on the support objects of its intra-space parenthood instead of the whole set *S* of reference objects. An additional caching technique compensates the overhead in looking for the parenthood of each pattern, since it has a cost, which increases with the number of stored patterns.

The above mentioned efficiency improvements are based on the monotonicity property of the generality order that is defined for spatial patterns with respect to the support of the patterns themselves. This is a nice example of an "intelligent" exploitation of general properties to prune the search space and reduce the number of expensive tests. However, this approach does not take into account user preferences and expectations. In real-world applications, such as urban site accessibility [2], a large number of spatial patterns can be generated even for a few hundred spatial objects. Nevertheless, most of discovered patters are useless for the application at hand. Therefore, it is important to allow the user to specify his/her bias for interesting solutions, and then to exploit this bias to improve both the efficiency of the system and the quality of the discovered rules. In SPADA 3.0, the language bias LB is

expressed as a set of constraint specifications for either patterns or association rules. Users may specify the following pattern constraint:

pattern_constraint(AtomList, Min_occur, Max_occur)

where *AtomList* is a list of atoms (for atomic constraints) or a list of atom lists (for conjunctive constraints), while *Min_occur* (*Max_occur*) is positive number which specifies the minimum (maximum) number of constraints in the *AtomList* that must be satisfied. When Max_occur = '_' no limitation is imposed on the maximum number of constraints. For instance, the following:

pattern_constraint(crossed_by_green_area(_,_), crossed_by_urban_area(_,_)],1,_)

specifies that at least one of the binary spatial predicates *crossed_by_green_area* and *crossed_by_urban_area* must occur in the patterns filtered by SPADA, while the following:

pattern_constraint([[crossed_by_green_area(_,_), crossed_by_urban_area (_,_)], [crossed_only_by_road(_)]], 1, _).

specifies that at least one of either the spatial predicates *crossed_by_green_area* and *crossed_by_urban_area* or the spatial predicate *crossed_only_by_road* must occur in the patterns filtered by SPADA. It is noteworthy that this specification allows users to define both conjunctive and disjunctive constraints.

During the rule generation phase, patterns that do not satisfy a pattern constraint are filtered out. This means that they are generated and evaluated anyway. This late exploitation of pattern constraints is due to the fact that if a pattern *P* does not satisfy a constraint (e.g. the lack of the predicate *crossed_by_green_area*), it is still possible that *P* descendants (i.e., more specific patterns) satisfy it. Therefore, pattern constraints do not prune the pattern space, but improve the efficiency of the mining process, since they prevent the generation of useless rules, and hence their evaluation.

A further pattern constraint takes into account the typing mechanism of the variables to be included in the rules. A variable *X* is untyped when it does not appear as first argument of a binary is-a atom in the rule. In some applications, the occurrence of untyped variables in a rule is undesirable. Therefore, users can specify the constraint *max_rules_untyped_vars(n)*, where *n* denotes the maximum number of untyped variables in the rules being generated. As in the previous case, the specification of this constraint affects the rule generation phase.

Users may specify constraints either on the antecedent or on the consequent of spatial association rules through one of the following facts:

body_constraint(AtomList, Min_occur, Max_occur).
head_constraint(AtomList, Min_occur, Max_occur).

where *AtomList*, *Min_occur* and *Max_Occur* have the same meaning as in the pattern constraint described above. For instance, the constraint *head_constraint([mortality_ rate(_)], 1, 1)* specifies that a single occurrence of the unary predicate *high_mortality* must be in the head of the rules. As for pattern constraints, head and body constraints affect the rule generation phase. The main property of all described constraints is that they do not prevent the generation of candidate rules but only the evaluation of their confidence.

In addition to constraints above, SPADA 3.0 users can specify the fact: *rule_head_length (Min_occur, Max_occur)* in order to fix the minimum (*Min_occur*) and the maximum number (*Max_occur*) of predicates to be included in the head of generated rules. For instance, by combining the rule filters *head_constraint([mortality_rate(_)], 1, 1)* and *rule_head_length(1, 1)* the user can ask for the generation of rules containing only the predicate *mortality_rate* in the head. Rules in this form may be employed for spatial subgroup mining that is the discovery of interesting group of spatial objects with respect to a certain property of interest, as well as for classification purposes.

6 The Application: Two Case Studies

In this section, we describe the application of SPADA to two distinct real-world problems, namely mining document images and mining geo-referenced census data. In the former problem, spatial objects are layout components extracted by means of WISDOM++. Layout components are in the same page of a document and have a common geometrical representation: they are all rectangles with edges parallel to the axes associated to the left and top border of a page. As result of the document understanding process, layout components may be associated with some components of the document logical structure, whose hierarchical organization defines the hierarchy of task-relevant objects. Discovered spatial association rules can be used in a generative way. For instance, if a part of the document is hidden or missing, strong spatial association rules can be used to predict the location of missing layout/logical components [9]. This problem is also related to document reformatting [8].

In the second problem, the goal is to perform a joint analysis of both socio-economic factors represented in census data and geographical factors represented in topographic maps. The discovery of interesting association rules on geographically distributed socio-economic phenomena can be a valuable support to good public policy. In this case, spatial objects are territorial units for which census data are collected as well as entities of the transport network (roads and rails), while the hierarchies are either based on layers of the topographic map or defined on the basis of a conceptual categorization or urban areas.

6.1 Document Image Processing

In this application SPADA takes as input a collection of ground facts describing both the layout and the logical structures of the documents processed by WISDOM++. Spatial features (relations and attributes) are used to describe the logical structure of a document image. In particular, we use FEATEX to extract locational features such as the coordinates of the centroid of a logical component, geometrical features such as the dimensions of a logical component, and topological features such as relations between two components. We use the *aspatial* feature *type_of* that specifies the content type of a logical component (e.g. *image*, *text*, *horizontal line*). In addition there are other aspatial features, called *logical* features which define the label associated to the logical components. They are: *affiliation, page_number, figure,*

caption, index_term, running_head, author, title, abstract, formulae, subsection_title, section_title, biografy, references, paragraph, *table, undefined.*

The specification of the domain specific knowledge allows SPADA to associate information on page order to layout components, since the presence of some logical components may depend on the order page (e.g. *author* is in the first page). An example related to the first page is

at_page(X,first) :- part_of(Y,X), page(Y,first).

The specification of the hierarchy (Figure 2) allows the system to extract spatial association rules at different granularity levels.

In this task, the *ro* are the logical components associated with logical features different from *undefined*. The *tro* are all the logical components. This is specified by means of the language bias *LB*. In particular, we ask for rules containing at least one binary spatial predicate:

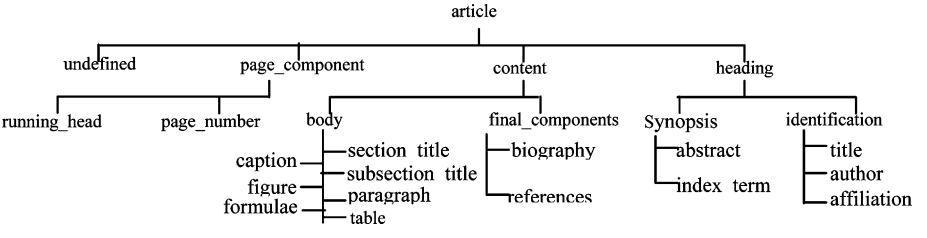

Fig. 2. Hierarchy of logical components

pattern_constraint([only_middle_coumnl(_,_),only_left_row(_,_), ... , on_top(_,_)],1).

Furthermore, we are interested in rules containing the *ro* in the antecedent. For instance, if we use *abstract* as *ro* the constraint is:

body_constraint([abstract(_)], 1).

We investigate the applicability of the proposed solution on 19 real-world documents, which are scientific papers published as either regular or short in the IEEE Transactions on Pattern Analysis and Machine Intelligence in the January and February 1996 issues. Each paper is a multi-page document and has a variable number of pages and layout components per page, for a total of 179 document images and 2,998 layout components. Eight hundred and eleven layout components with no clear logical meaning are labelled as *undefined*. All logical labels belong to the lowest level of the hierarchy reported in the previous section. The number of logical components is 2,177. The number of features to describe the documents presented to SPADA is 78,789, about 440 features for each page document. Average running time per document image is 1.32 secs (237.52/179), therefore this application of SPADA to document images seems scalable to larger collections of documents. The number of mined association rules is 398 at the first level, 468 at the second level, 638 at the third level and 674 at the fourth level. Many spatial patterns involving logical components (e.g., affiliation, title, author, abstract and index term) in the first page of an article are found. SPADA

has found several spatial associations involving all logical components, references and biography excluded. This can be explained by the observation that the first page generally has a more regular layout structure and contains several distinct logical components. An example of association rule discovered by SPADA is:

is_a(author,A) $\Rightarrow$ *on_top(A,B), is_a(B,heading) , height(B,[1..174]) , type_text(A)*
(*82.6%, 82.6%*)

This means that 19 logical components which represent the *author* of some paper are textual components on top of a logical component B that is the *heading* of the paper, with height between 1 and 174. At a lower granularity level, a similar rule is found where the logical component *B* is specialized as *abstract*:

is_a(author,A) $\Rightarrow$ *on_top(A,B), is_a(B,abstract) , height(B,[1..174]), type_text(A)*
(82.6%, 82.6%)

The rule has the same confidence and support reported for the rule inferred at the first granularity level.

6.2 Geo-referenced Exploratory Data Analysis

In this study we describe how it is possible to employ ARES in performing data analysis on geo-referenced census data concerning Greater Manchester, one of the five counties of North West England, that is divided into censual sections or wards, for a total of two hundreds and fourteen wards. For this application, spatial analysis is enabled by the availability of vectorized boundary of wards as well as vector geographical data about transport network, waters, green and urban areas that allow us to investigate the mortality rate (i.e. percentage of deaths with respect to the number of inhabitants) from a spatial viewpoint according to deprivation indices. Geographical layers are taken from the Meridian product of the Ordnance Survey. In particular, we decide to mine spatial association rules relating wards, which play the role of *ro*, with topological related road network (i.e. motorways, primary roads, A- and B- roads), rail network, water network (i.e. rivers, canals and waters), green area (i.e. parks and woods) and urban area (i.e. small and large areas) as *tro*.

Therefore, by using FEATEX we extract facts concerning topological relationships between wards and roads, rails, waters, green areas and urban areas reported in the spatial database for that area. An example of fact extracted by FEATEX is *crosses(ward_135, urbareaL_151)*. The number of facts is 784,107. Despite the complexity of the spatial computation performed by FEATEX to extract these facts, the results are still not appropriate for the goals of our data analysis tasks. Therefore, a domain specific knowledge should be expressed in form of a set of rules. Some of the rules used in this data mining task are:

```
crossed_by_urbanarea(X,Y) :- crosses(X,Y), is_a(Y,urban_area).
crossed_by_urbanarea(X,Y) :- inside(X,Y), is_a(Y,urban_area).
not_crossed_by_urbanarea(X) :- is_a(X,ward), \+ crossed_by_urbanarea(X,_).
```

Here the use of the predicate *is_a* hides the fact that a hierarchy has been defined for spatial objects belonging to urban area layer (see Figure 3). Similarly, four

different hierarchies have been defined to describe road network, rail network, water network and green area. The hierarchies have depth three and are straightforwardly mapped into three granularity levels. Hence, these hierarchies are used to complete the domain specific knowledge by adding rules describing topological relationships and/or not-relationschips between wards and green area, transport and water net.

Until now, all extracted data and user-defined background knowledge are purely spatial. However, we can observe that the mortality rate of an area cannot be defined on the basis of the geographical environment alone. We select four deprivation indecies, namely Townsend index, Carstairs index, Jarman index and DoE index, we discretize them with RUDE and generate the following four binary predicates for SPADA: *townsend_idx, carstairs_idx, jarman_idx* and *doe_idx*. The first argument of the predicate refers to a ward, while the second argument is an interval returned by RUDE. The Townsend index is a measure of multiple deprivation that is computed at ward level according to the percentage of households that are not owner occupied, percentage of households with no car, percentage of households with more than one person per room and percentage of persons who are unemployed. Similarly, Carstairs index, Jarman index and DoE index are calculated using census data to measure socio-economical deprivation of a ward.

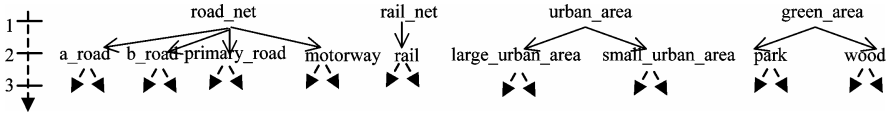

Fig. 3. Spatial hierarchies defined for four Greater Manchester layers: road net, rail net, urban area and green area

To complete the problem statement we specify a declarative bias both to constrain the search space and to filter out some uninteresting spatial association rules. In particular, we rule out all spatial relations directly extracted by means of FEATEX. Moreover, by specifying the rule filters *head_constraint([mortality_rate(_)], 1, 1)* and *rule_head_length(1, 1)* we ask for rules containing only the predicate *mortality_rate* in the head. After some tuning of the parameters *min_sup* and *min_conf* for each granularity level, we decide to run the system with the following parameter values: *min_sup*[1]=0.1, *min_sup*[2]=0.1, *min_sup*[3]=0.05, *min_conf*[1]=0.3, *min_conf* [2]=0.2, *min_conf* [3]=0.1.

Despite the above constraints, SPADA generates 413 rules from a set of 100791 candidates. A rule returned by SPADA at the first level is the following:

is_a(A, ward), crossed_by_urbanarea(A, B), is_a(B, ***urban_area****),*
townsendidx_rate(A, high) $\Rightarrow$ *mortality_rate(A, high)* (*40.72%, 72.47%*)

which states that a high mortality rate is observed in a ward *A* that includes an urban area *B* and has a high value of Townsend index. The support (40.72%) and the high confidence (72.47%) confirm a meaningful association between a geographical factor such as living in deprived urban areas and a social factor such as the mortality rate. It is noteworthy that SPADA generates the following rule:

is_a(A,ward), crossed_by_urbanarea(A,B), is_a(B, urban_area) ⇒ mortality_rate(A, high) (*56.7%, 60.77%*)

which has a greater support and a ower confidence. These two association rules show together an unexpected association between Townsend index and urban areas. Apparently, this means that this deprivation index is unsuitable for rural areas.

At a granularity level 2, SPADA specializes the task relevant object B by generating the following rule which preserve both support and confidence:

is_a(A, ward), crossed_by_urbanarea(A, B), is_a(B, ***urban_areaL****), townsendidx_rate(A,high) ⇒ mortality_rate(A, high)* (*40.72%, 72.47%*)

This rule clarifies that the urban area B is large.

Similarly, SPADA discovers association rules involving low mortality wards:

is_a(A, ward), crossed_by_urbanarea(A, B), is_a(B, urban_area), townsendidx_rate(A, low) ⇒ mortality_rate(A, low) (*21.13%, 56. 94%*)

This rule, extracted at the first granularity level, states that a low valued Townsend index ward *A* that (partly) includes an urban area *B* presents a low mortality rate.

7 Conclusions

In this paper the discovery of spatial association rules by means of ARES in two real-world case studies, namely document image analysis and geo-referenced census data analysis, is illustrated. We also present some criteria to reduce the pattern search space and to filter extracted rules in order to discover interesting association rules according to user preferences. This is achieved by exploiting the high expressive power of rule miner SPADA 3.0, integrated in ARES, as well as by endowing SPADA 3.0 of a powerful language for bias specification. Results show that ARES mines interesting rules at different granularity levels. For future work we plan to investigate the improvement of ARES in order to implement a tight-coupling between SPADA and the spatial database.

Acknowledgments

We thank Jim Petch, Keith Cole and Mohammed Islam (University of Manchester) for collecting data made available through Manchester Computing in the context of the IST European project SPIN (Spatial Mining for Data of Public Interest).

The work presented in this paper is partial fulfillment of the research objective set by the ATENEO-2004 project on “Metodi di Data Mining Multi-relazionale per la scoperta di conoscenza in basi di dati”.

References

1. R. Agrawal, R. Srikant: Fast algorithms for mining association rules. *In Proc. of the International Conference on Very Large Data Bases*, 1994, pp. 487-499.
2. A. Appice , M. Ceci, A. Lanza, F.A. Lisi. and D. Malerba , Discovery of Spatial Association Rules in Georeferenced Census Data: A Relational Mining Approach, *Intelligent Data Analysis,* 7, 6 2003, pp. 541-566.
3. L. Dehaspe, H. Toivonen: Discovery of frequent Datalog patterns. *Data Mining and Knowledge Discovery*, 3(1), 1999, pp. 7-36.
4. Q. Ding: Association Rule Mining on Remotely Sensed Imagery Using P-Trees. *Phd Thesis.* North Dakota State University of Agriculture and Applied Science. 2002
5. S. Džeroski and N. Lavrac, Relational Data Mining, Springer-Verlag, Berlin, 2001.
6. J. Han, Y. Fu: Discovery of multiple-level association rules from large databases., *Proc. of the 21st International Conference on Very Large Data Bases, VLDB'95*, 1995, pp. 420-431
7. J. Han, K. Koperski, N. Stefanovic: GeoMiner: A System Prototype for Spatial Data Mining. *Proceedings of the ACM-SIGMOD International Conference on Management of Data*. SIGMOD'97 Record 26, 2, 1997, pp. 553-556.
8. L. Hardman, L. Rutledge and D. Bulterman, Automated generation of hypermedia presentation from pre-existing tagged media objects, *Proc. Of the 2nd. Workshop on Adaptive Hypertext and Hypermedia*, 1998.
9. K. Hiraki., J.H. Gennari, Y. Yamamoto and Y. Anzai, Learning Spatial Relations from Images, *Machine Learning Workshop,* Chicago, 1991, pp. 407 - 411.
10. K. Koperski, J. Adhikary and J. Han, Spatial Data Mining: Progress and Challenges. Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, 1996
11. K. Koperski, J. Han: Discovery of Spatial Association Rules in Geographic Information Databases. *Advances in Spatial Databases*. LNCS 951, Springer-Verlag, 1995, pp. 47-66.
12. F.A. Lisi and D. Malerba: Efficient Discovery of Multiple-level Patterns. *Decimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati SEBD'2002*, 2002, 237-250.
13. F.A.Lisi and D. Malerba, Inducing Multi-Level Association Rules from Multiple Relations. *Machine Learning*, vol 55, pp. 175-210, 2004.
14. M.C. Ludl and G. Widmer, Relative Unsupervised Discretization for Association Rule Mining: *PKDD2000*, LNCS 1910, Springer-Verlag, 2000, pp. 148-158
15. D. Malerba, M. Ceci and M. Berardi, XML and Knowledge Technologies for Semantic-Based Indexing of Paper Documents, *DEXA 2003*, LNCS 2736, Springer, Berlin, 2003, pp. 256-265.
16. H. Mannila, H. Toivonen: Levelwise Search and Borders of Theories in Knowledge Discovery. *Data Mining and Knowledge Discovery* 1(3), 1997. pp. 241-258.
17. Y. Morimoto: Mining frequent neighboring class sets in spatial databases. *KDD '01*, 2001, pp. 353- 358.