A relational perspective on spatial data mining

Donato Malerba

Dipartimento di Informatica,
Università degli Studi di Bari,
Via Orabona 4, I-70126 Bari, Italy
Fax: +39-0805443269
E-mail: malerba@di.uniba.it

Abstract: Remote sensing and mobile devices nowadays collect a huge amount of spatial data, which have to be analysed in order to discover interesting information about economic, social and scientific problems. However, the presence of a spatial dimension adds some problems to data mining tasks. The geometrical representation and relative positioning of spatial objects implicitly define spatial relationships, whose efficient computation requires a tight integration of the data mining system with the spatial DBMS. The interaction between spatially close objects causes different forms of autocorrelation, whose effect should be considered to improve the predictive accuracy of induced models and patterns. Units of analysis are typically composed of several spatial objects with different properties and their structure cannot be easily accommodated by classical double entry tabular data. In the paper, a way is shown to face these problems when a (multi-)relational data mining approach is considered for spatial data analysis. Moreover, the challenges that spatial data mining poses on current relational data mining methods are presented.

Keywords: spatial data mining; multi-relational data mining; MRDM; geographic knowledge discovery.

Reference to this paper should be made as follows: Malerba, D. (2008) 'A relational perspective on spatial data mining', *Int. J. Data Mining, Modelling and Management*, Vol. 1, No. 1, pp.103–118.

Biographical notes: Donato Malerba is a Full Professor of Computer Science at the University of Bari, Italy. His research activity concerns theory and applications of machine learning and data mining. He has published more than 150 papers in international journals and conference proceedings. He is/has been in many European and national projects on data mining and machine learning. He is/has been member of the program committees of many international conferences. He was involved in the organisation of several workshops and conferences on data mining and machine learning. He acted as guest-editor of special issues of six international journals.

1 Introduction

In a large number of application domains, such as traffic and fleet management, environmental and ecological modelling, robotics, computer vision and, more recently, computational biology and mobile computing, collected data present a spatial dimension. Indeed, they are measurements on one or more attributes of the objects, which occupy

specific locations. These (spatial) objects are characterised by a geometry (e.g., a line or an area) which is formulated by means of a reference system. This geometry implicitly defines both spatial properties, such as orientation, and spatial relationships of a different nature, such as topological (e.g., intersects), distance or direction (e.g., north-of) relations. A 'geographical' object represents a special case of a spatial object whose relative position is specified with respect to the physical earth.

Studies in spatial data structures (Güting, 1994), spatial reasoning (Egenhofer and Fransoza, 1991) and computational geometry (Preparata and Shamos, 1985) have paved the way for the investigation of 'spatial data mining', which is related to the extraction of interesting and useful but implicit spatial patterns (Koperski et al., 1996). A spatial pattern expresses a spatial relationship among (spatial) objects and can take different forms, such as classification rules, association rules, regression models, clusters and trends.

Spatial data mining has received considerable attention in the recent years (Roddick and Spiliopoulou, 1999; Roddick et al., 2001). However, most of the works in this area are simple adaptations of conventional data mining tools and techniques, which do not recognise the uniqueness of the spatial dimension. Unfortunately, conventional data mining algorithms perform poorly on spatial data since they are based on the assumption that data samples are independently generated. Moreover, they generally use very simple representations of spatial objects and spatial relationships (Buttenfield et al., 2000). The former are uniformly represented as vectors of common features, which do not suitably express the diversity of spatial objects. The latter are often limited to Euclidean distances, although the spatial dependency can also manifest itself across non-Euclidean distances, topological and directional relationships. Han et al. (2001) review many spatial clustering algorithms, which deal with points in a d-dimensional space and consider only one spatial relationship, the distance. A similar considerations also applies to other tasks, such as spatial regression and outlier detection (Gao et al., 2006; Shekhar et al., 2001).

In the recent years, we are also assisting to a growing attention for a class of data mining algorithms, known as multi-relational (or simply relational), which operate on data scattered through multiple tables (relations) of a relational database and discover relational patterns that involve multiple relations and are typically stated in a more expressive language (e.g., predicate calculus and SQL) than patterns defined on a single data table. Many data mining tasks (e.g., classification, clustering, association analysis) have already been adapted to a multi-relational setting.

In this position paper we argue that the multi-relational setting is the most suitable for spatial data mining problems, since it can deal with the heterogeneity of spatial objects, it can distinguish their different role (reference or task-relevant), it can naturally represent a large variety of spatial relationships among objects and it can accommodate different forms of spatial autocorrelation. We also mention the most significant attempts to design multi-relational data mining (MRDM) systems, which discover relational patterns from spatial data, we illustrate the many challenges that must be overcome, and issues that must be resolved before the relational approach can be effectively applied to spatial data.

The paper is organised as follows. Peculiar issues of spatial data mining tasks are introduced in the next section, and then, based on these considerations, the multi-relational approach to spatial data mining is motivated in Section 3. Section 4 closes with some open problems and a list of challenges for researchers interested in developing MRDM methods for the analysis of spatial data.

2 What is special about spatial data mining

In this section, the four main issues that characterise spatial data mining tasks are reported.

2.1 Spatial information modelling

At the 'conceptual' level, the two main approaches to modelling spatial information are 'field-based' and 'object-based' (Shekhar and Chawla, 2003). In field-based models, the world is seen as a continuous surface over which features vary. Spatial variation is defined by a number of field functions, of the form:

$$f : R^n \mapsto Attribute\ domain.$$

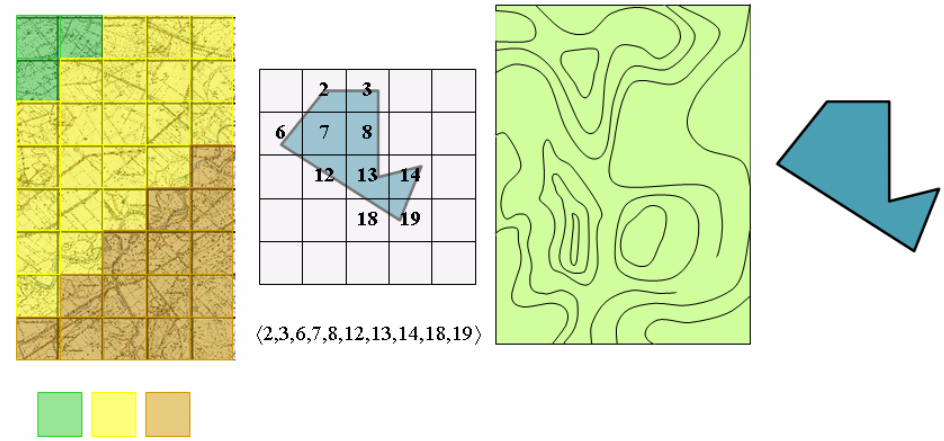
The elevation, precipitation or temperatures of a given geographic area are three examples of field functions. Interactions between spatial objects are defined by combining field functions with field operators, such as addition and composition:

$f + g : x \to f(x) + g(x)$ (addition),
$f \circ g : x \to f(g(x))$ (composition).

In object-based modelling, the world is seen as a surface littered with distinct, identifiable and relevant objects which can be zero-dimensional (or punctual), one-dimensional (or linear) or two-dimensional (or surfacic).[1] Interactions between spatial objects are described by means of topological, directional and distance-based operators.

The formulation of a spatial data mining must consider the representation of the spatial information at the 'logical' level. Two types of data structures have been reported in the literature: 'tessellation' and 'vector' (Laurini and Thompson, 1992). The tessellation model partitions the space into a number of cells, each of which is associated with a value of a given attribute. No variation is assumed within a cell and values correspond to some aggregate function (e.g., average) computed on original values in the cell. A grid of square cells is a special tessellation model called 'raster'. In the vector model, the geometry is represented by a vector of coordinates, which define points, lines or polygons. The tessellation and vector models can be equally used to represent spatial information modelled by either field-functions or spatial objects at the conceptual level (see Figure 1). The tessellation model is simple and frequently used (e.g., in remote sensing), nevertheless, it requires large storage capabilities, the operations on objects are time-consuming and the geometry of a spatial object is imprecise. The vector model is a concise and precise representation, easy to scale (although some spatial operations, such as intersection, remain computationally demanding) and is well supported by spatial database management systems (DBMS). Henceforth, we will refer to the vector model as the logical model adopted for the representation of spatial information. In particular, we assume that a field function or a spatial object is represented by one or more tuples of a 'layer', that is, a database relation R_i with a number of elementary attributes $A_1^i, \ldots, A_{m_i}^i$ and possibly a geometry attribute G represented in the vector mode. A spatial pattern, which expresses the interaction among spatial objects, will be defined by means of topological, directional and distance-based relationships.

Figure 1 (a) Field-based data represented in tessellation mode (b) object-based data represented in tessellation mode (c) field-based data represented in vector mode (d) object-based data represented in vector mode (see online version for colours)



2.2 Heterogeneity of spatial objects

Spatial patterns often describe interactions between objects of different types, such as a town and a highway. In spatial databases, objects of different types are organised in separate layers, each of which has a distinct set of attributes. For instance, a town can be described in terms of economic and demographic factors, as well as a polygon corresponding to its administrative boundary, while a highway is described by the average speed limit, traffic and driving safety conditions, as well as a polyline corresponding to its path. To deal with object heterogeneity in spatial patterns, the design of a spatial data mining method should not be strictly bound to process objects in one specific layer.

2.3 The implicit definition of spatial relationships among objects

Spatial objects have a locational property, which implicitly defines several spatial relationships between objects. Topological relationships are invariant under homomorphisms, such as rotation, translation and scaling. Their semantics is precisely defined by means of the nine-intersection model proposed by Egenhofer and Franzosa (1991). The distance between two points is typically computed according to the Euclidean metric, while the distance between two geometries (e.g., two areas) is defined by some aggregate function (e.g., the minimum distance between two points of the areas). Distance relationships can be non-metric, especially when they are defined on the basis of a cost function which is not symmetric (e.g., the drive time). Directional relations can be expressed by the angle formed by two points with respect to the origin of the reference system or by an extension of Allen's interval algebra, which is based on projection lines (Mukerjee and Joe, 1990). In a spatial database, implicit binary spatial relationships correspond to spatial joins $R_i \bowtie_\theta R_j$ between two layers R_i and R_j, where θ is a binary predicate (e.g., 'intersects', 'contains', 'northwest', 'adjacent') evaluated on the geometry

attributes G_i and G_j of the two layers (Shekhar and Chawla, 2003). The relational nature of spatial patterns makes the computation of these spatial joins crucial for the development of effective and efficient data analysis methods. To complicate matters, the data analyst is generally interested in spatial patterns where object interactions are abstracted from the geometry of involved spatial objects (e.g., a river crosses a city, whatever the geometric representations of rivers and cities are).

2.4 Spatial autocorrelation

By picturing the spatial variation of some observed variables in a map, we may observe regions where the distribution of values is smoothly continuous with some boundaries possibly marked by sharp discontinuities. In this case, a variable is correlated with itself through space. Formally, spatial autocorrelation is defined as the property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly associated pairs of observations (Legendre, 1993). Informally, spatial positive (negative) autocorrelation occurs when the values of a given property are highly uniform (different) among similar spatial objects in the neighbourhood. In geography, spatial autocorrelation is justified by Tobler's (1970) first law of geography, according to which 'everything is related to everything else, but near things are more related than distant things'. However, spatial autocorrelation occurs in many other disparate fields, such as sociology (e.g., social relations affect social influence), web mining (e.g., hyperlinked web pages typically share the same topic) and bioinformatics (e.g., proteins located in the same place in a cell are more likely to share the same function than randomly selected proteins). In statistics, spatial autocorrelation is divided into two primary types: spatial error (correlations across space in the error term), and spatial lag (the dependent variable in space i is affected by the independent variables in space i, as well as those, dependent or independent, in space j). Most statistical models are based on the assumption that the values of observations in each sample are independent of one another, but spatial autocorrelation (or spatial dependence, as it is typically called in statistics) clearly indicates a violation of this assumption. As observed by LeSage and Pace (2001), 'anyone seriously interested in prediction when the sample data exhibit spatial dependence should consider a spatial model', since this can take into account different forms of spatial autocorrelation. In addition to predictive data mining tasks, this consideration can also be applied to descriptive tasks, such as spatial clustering or spatial association rule discovery. More in general, the analysis of spatial autocorrelation is crucial and it can be fundamental for building a spatial component into (statistical) models for spatial data. The inappropriate treatment of sample data with spatial dependence could obfuscate important insights and observed patterns may even be inverted when spatial autocorrelation is ignored (Kühn, 2007).

2.5 Limits of current solutions

Traditional data mining algorithms do not offer adequate solutions to all these issues. They do not deal with spatial data characterised by geometry, do not handle observations of different types, do not naturally represent spatial relationships between observations nor take them into account when mining patterns.

To overcome some of these limitations, several extensions have been investigated in spatial statistics, where spatial dependence is typically modelled by the following linear models (LeSage and Pace, 2001):

$$y = X\alpha + \beta Dy + DX\gamma + \in$$

where:

1 y is the $n \times 1$ vector of observations of the dependent (or response) variable

2 α considers the influence of the independent (or explanatory) variables observed in i on the response variable in i

3 D is an $n \times n$ matrix, called spatial weight matrix, which defines the neighbourhood, i.e., $D_{ij} > 0$ for observations j sufficiently close (as measured by some metric) to observation i, $D_{ij} > 0$ otherwise

4 β reflects the strength of the spatial dependence on the response variable of the neighbours

5 γ reflects the strength of the spatial dependence on the explanatory variables of the neighbours

6 $\in$ reflects the 'noise' or a stochastic disturbance in the spatial dependence relation.

However, the application of these spatial models still presents some problems. First, the spatial weight matrix D has to be carefully defined in order to specify to what extent a spatially close observation in space j can affect the response observed in i. With a proper choice of D, the residual error should, at least theoretically, have no systematic variation. Second, it is unclear how D can express the contribution of different spatial relationships, such as a polluting industry in an 'adjacent' area and a highway 'crossing' the same area. Third, spatial dependencies are all handled in a pre-processing or feature extraction step, which typically ignores the subsequent data mining step. In principle, a data mining method, which can handle spatial dependencies directly, presents the advantage of considering only those dependencies that are really relevant to the task at hand. Fourth, all spatial objects involved in spatial phenomena (rows of matrix X) are uniformly represented by the same set of attributes. This can be a problem when spatial objects are of different types and are characterised by different properties. Fifth, there is no clear distinction between the reference (or target) objects, which are the main subject of analysis and the task-relevant objects, which are spatial objects 'in the neighbourhood' that can help to account for the spatial variation.

3 Opportunities for a relational approach

The problems reported above are due to the fact that in spatial data mining, the units of analysis are typically composed of several spatial objects with different properties and

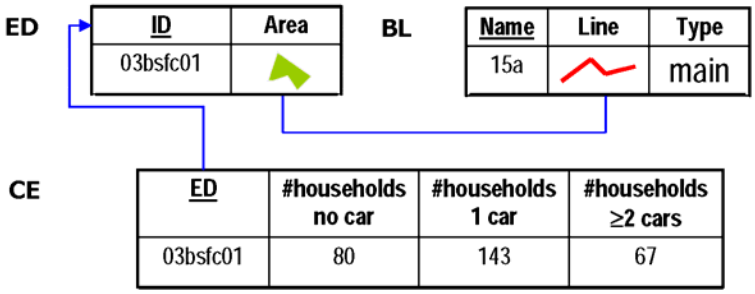
that their structure cannot be easily accommodated into a classical double-entry table, whose columns correspond to elementary (nominal, ordinal or numeric) single-valued attributes. In fact, spatial datasets can be naturally modelled as a set of relations $R_1,\ldots,R_n$, such that each relation R_i has a number of elementary attributes $A_1^i,\ldots,A_{m_i}^i$ and possibly a geometry attribute G_i (in which case, it is a layer). Relationships are either explicitly modelled by means of foreign key constraints or implicitly represented by spatial joins.

Example

To investigate the social effects of public transportation in a British city, a spatial data set made of three relations is considered (see Figure 2). The first relation *ED* contains information on enumeration districts, which are the smallest area units for which census data are published in UK. In particular, *ED* has two attributes, the identifier of an enumeration district and a geometry attribute (a closed polyline), which describes the area covered by the enumeration district. The second relation *BL* describes all the bus lines, which cross the city. In this case, relevant attributes are: the name of a bus line, the geometry attribute (a line) representing the route of a bus and the type of bus line (classified as main or secondary). The third relation *CE* contains some census data relevant for the problem, namely, the number of households with 0, 1, or '≥ 2' cars. This relation also includes the identifier of the enumeration district, which is a foreign key for the table *ED*. A unit of analysis corresponds to an enumeration district (the reference object), which is described in terms of the number of cars per household and crossing bus lines (bus lines are the task-relevant objects). The relationship between reference objects and task-relevant objects is established by means of a spatial join, which computes the intersection between the two layers *ED* and *BL*. This relationship allows us to discover truly relational patterns, such as 'the enumeration districts with a high percentage of households which own less than two cars, are served by at least two bus lines, one of which is a main bus line'. Here, the verb 'served' is purposely introduced, to show that spatial patterns of interest may not necessarily be expressed in terms of the original spatial predicates used in the spatial join operations. The most obvious interpretation of this verb can be the topological relation 'intersect' between the area of an enumeration district and the bus line, although other more sophisticated interpretations are possible (e.g., on the basis of the length of the intersected line). However, it may well be the case that an enumeration district with few households owning less than two cars is not actually crossed by a bus line, but rather it is spatially surrounded by several other enumeration districts where all conditions above hold. To take this spatial autocorrelation into account, a spatial join between *ED* and itself can be computed and the relational patterns can be searched across the units of analysis.

The previous example shows that MRDM offers the most suitable setting for spatial data mining tasks. Indeed, MRDM tools can be applied directly to data distributed over several relations to find relational patterns, which involve multiple relations (Džeroski and Lăvrac, 2001). Relational patterns can be expressed not only in SQL, but also in first-order logic (or predicate calculus), which explains why many MRDM algorithms originate from the field of inductive logic programming (ILP) (Muggleton, 1992; De Raedt, 1992; Lăvrac and Džeroski, 1994).

Figure 2 Relational representations of data on the social effects of public transportation in a British city (see online version for colours)



Notes: *ED* and *BL* are two layers since each of them has a geometric attribute. *CE* has a foreign key for *ED*.

Upgrading a classical data mining algorithm devised for double-entry tabular data to a relational setting is not a trivial task (Van Laer and De Raedt, 2001). For instance, it may be necessary to extend the definition of distance measure to data distributed among several tables. For propositional patterns expressed as pure conjunctive (queries) or pure disjunctive (clauses) formulae, the generality order $\succeq$ coincides with the subset ($\subseteq$) relation, while for relational patterns it is necessary to consider different generality orders (e.g., θ-subsumption), whose computation is NP-complete (Gottlob and Leitsch, 1985). Consequently, search efficiency is a concern for MRDM algorithms, which use some form of 'declarative bias' to limit the search space for interesting patterns. An exhaustive list of theoretical results and techniques that have been developed to improve the efficiency and scalability of MRDM approaches is reported in Blockeel and Sebag (2003).

The handling of spatial data adds difficulties to the upgrading of classical data mining algorithms. Indeed, it is necessary to define a representation of spatial objects, to define operators for spatial joins, to optimise the computation of spatial joins with spatial indexes, to distinguish reference from task-relevant objects and to devise some visualisation techniques of discovered patterns on maps.

4 Challenges for a relational approach

Although the MRDM setting seems the most suitable for spatial data mining, there are still several challenges that must be overcome and issues that must be resolved before the relational approach can be effectively applied to spatial data mining. Some of them are reported in the following.

4.1 Spatial relationships are many and not explicitly modelled

Many MRDM methods take advantage of knowledge on the data model (e.g., foreign keys), which is obtained free of charge from the database schema, in order to guide the search process. However, this approach does not suit spatial databases, since the database navigation is also based on the spatial relationships, which are not explicitly modelled in

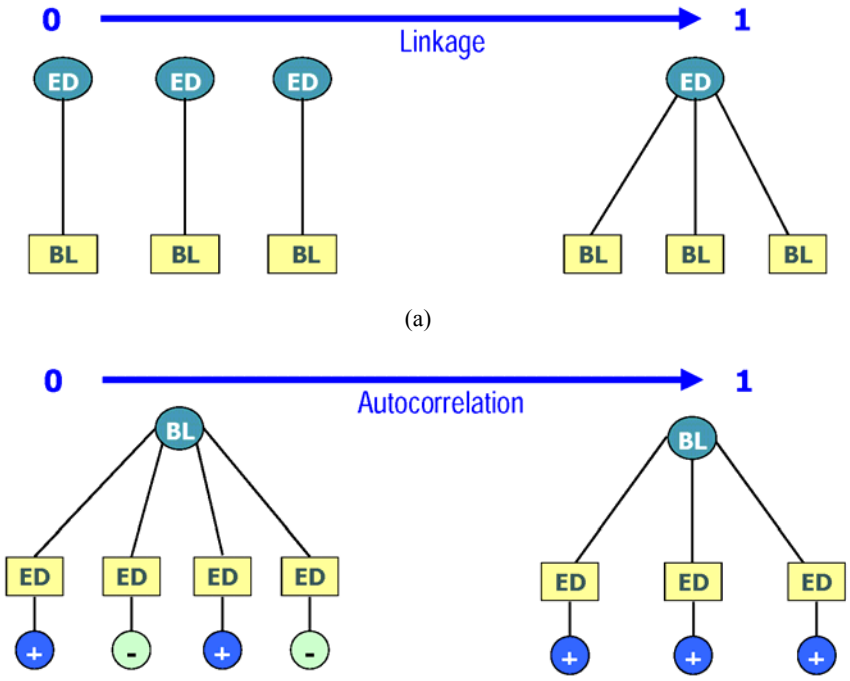
the schema. To solve this problem, spatial relationships can be computed and explicitly represented during the pre-processing step of the knowledge discovery process. This approach is typically applied by statisticians before computing the spatial weight matrix *D*. It has also been adopted in the GeoMiner system (Han et al., 1997), whose data mining algorithms, though, operate on a single database relation obtained from the pre-processing step. Also Ester et al. (1999) propose to pre-compute distance, direction and topological relations and to materialise (i.e., store) them into some database relations (called neighbourhood indices), which are then used by data mining algorithms to efficiently retrieve all neighbours (with respect to some spatial relation) of a given spatial object. A feature extraction module is implemented into the ARES system (Appice et al., 2005) to pre-compute spatial relationships, which are converted into Prolog facts used by the ILP system SPADA (Malerba and Lisi, 2001) to generate spatial association rules. The pre-computation is justified by the fact that spatial databases are rather static, since there are not many updates on objects such as geographic maps. However, the number of spatial relationships between two layers can be very large and many of them might be unnecessarily extracted. The alternative is to dynamically perform spatial joins only for the part of the hypothesis space that is really explored during the search by a data mining algorithm. This approach has been implemented in two MRDM systems, namely SubgroupMiner for subgroup mining (Klosgen and May, 2002) and Mrs-SMOTI for regression analysis (Malerba et al., 2005). Both systems achieve a tight integration with a spatial DBMS (namely, Oracle Spatial) and have been applied to datasets where few spatial relationships are actually computed. Spatial index structures, such as R-trees (Guttman, 1984), are used to speed up the processing of spatial joins. However, scalability remains a problem when many spatial predicates have to be computed. A scalability issue arises also in spatial statistics, since the spatial weight matrix *D* can be very large and sparse (LeSage and Pace, 2001).

4.2 Spatial autocorrelation can bias feature selection

Although the presence of autocorrelation in spatial phenomena strongly motivates a MRDM approach to spatial data mining, it also introduces additional challenges. In particular, it has been proven that the combined effect of autocorrelation and concentrated linkage (i.e., high concentration of objects linked to a common neighbour) (see Figure 3) can bias feature selection in relational classification (Jensen and Neville, 2002). In particular, the distribution of scores for features formed from related objects with concentrated linkage presents a surprisingly large variance when the class attribute has a high autocorrelation. This large variance causes feature selection algorithms to be biased in favour of these features, even when they are not related to the class attribute, that is, they are randomly generated. Conventional hypothesis tests, such as the χ^2-test for independence, which evaluate statistically significant differences between proportions for two or more groups in a dataset, fail to discard uninformative features. Indeed, they are based on the i.i.d. assumption, while observations drawn from a relational data set may not be independent. Most MRDM algorithms do not account for this bias, a notable exception being a relational probability tree-learning algorithm that uses a randomisation test to adjust for feature selection bias (Neville et al., 2003). Pseudo samples are generated from the relational data by retaining the linkage present in the original sample and the autocorrelation among the class labels, and, at the same time, by destroying the

correlation between the original attributes and the class labels. Therefore, pseudo samples appropriately conform to the null hypothesis and can be used to estimate a p-value for the actual data.

Figure 3 (a) Examples of low and high linkage between *ED*'s and bus lines (b) examples of low and high autocorrelation between *ED*'s crossed by the same bus line (see online version for colours)



4.3 Learning from unlabeled spatial data

Inductive learning algorithms designed for predictive tasks may require large sets of labelled data. However, the common situation is that only few labelled training data are available for mining, although a very large test set must be classified. This is especially true in geographical data mining, where large amounts of unlabeled geographical objects (e.g., map cells) are available and manual annotation is fairly expensive. Inductive learning algorithms would actually use only the few labelled examples to build a prediction model, thus, discarding a large amount of information potentially conveyed by the unlabeled instances. The idea of transductive inference (or transduction) (Vapnik, 1998) is to analyse both the labelled (training) data and the unlabeled (working) data to build a classifier and classify (only) the unlabeled data as accurately as possible. Transduction is based on a (semi-supervised) smoothness assumption according to which, if two points x_1 and x_2 in a high-density region are close, then the corresponding outputs y_1 and y_2 should also be close (Chapelle et al., 2006). However, in spatial domains, where

closeness of points corresponds to some spatial distance measure, this assumption is implied by (positive) spatial autocorrelation. Therefore, the transductive setting seems especially suitable for spatial classification and regression, and more in general, for those relational learning problems characterised by autocorrelation on the dependent variables. Only recently, a work on transductive relational learning has been reported in the literature (Malerba et al., 2009): some preliminary results on spatial classification tasks show the effectiveness of the transductive approach.

4.4 Spatially lagged prediction models demand for collective inference

In predictive data mining tasks, the generation of patterns, which express the spatial autocorrelation of the dependent variable, raises the issue of how inference on new cases should be performed. Indeed, these patterns take the form:

$$y_i = f(x_i, x_{N(i)}, y_{N(i)}),$$

where $y_i(x_i)$is the value of the dependent (independent) variable in space i, while $y_{N(i)}(x_{N(i)})$ represents the value(s) of the dependent (independent) variable for i's neighbour(s). For instance, the price level for goods at a retail outlet in a city depends on the price for the same goods in the vicinity. In order to predict y_i it is necessary to know the value(s) of $y_{N(i)}$, which might be unavailable (the related values of the dependence variable are to be inferred as well). In this case, both y_i and all unknown values $y_{N(i)}$ have to be inferred collectively. A possible approach to collective inference combines locally learned individual inference models with a joint inference procedure (e.g., relaxation labelling) to make an inference. An example is iterative classification (Neville and Jensen, 2000), which dynamically updates the attributes of some objects as inferences are made about related objects. Inferences made with high confidence in initial iterations are fed back into the data and are used to inform subsequent inferences about related objects. Iterative classification works well when the classification model allows us to make initial inferences accurately, otherwise all subsequent predictions will be misled due to a ripple effect. An alternative approach is to use joint relational models, which first estimate the joint probability distribution over the variables of objects both in i and in $N(i)$ and then jointly infer the values of both y_i and $y_{N(i)}$. In particular, probabilistic relational models can be used to represent a joint probability distribution over the attributes of a relational dataset (Getoor et al., 2001; Neville and Jensen, 2003). By making inferences about multiple instances simultaneously, joint inference can exploit autocorrelation in the data to improve predictions (Jensen et al., 2004). Therefore, this inference procedure seems particularly suitable for spatial data sets and should be better investigated in the context of spatial data mining.

4.5 Spatial patterns can be discovered at various levels of granularity

Spatial objects are often organised in hierarchies. By descending/ascending through a hierarchy, it is possible to view the same spatial object at different levels of abstraction (or granularity). Spatial patterns involving the most abstract spatial objects can be well

supported but at the same time, they are the less confident. Therefore, spatial data mining methods should be able to explore the search space at different granularity levels in order to find the most interesting patterns (e.g., the most supported and confident). In the case of granularity levels defined by a containment relationship, this corresponds to exploring both global and local aspects of the underlying phenomenon. Very few data mining techniques automatically support this multiple-level analysis. In general, the user is forced to repeat independent experiments on different representations and results obtained for high granularity levels are not exploited to control search at low granularity levels (or vice versa). Two noticeable exceptions are represented by Geo-associator (Koperski and Han, 1995), a module of GeoMiner which mines spatial association rules from data represented in a single relation (table) of a relational database, and SPADA (Malerba and Lisi, 2001), which discovers multi-level spatial association rules from relational data. SPADA has also been used in an associative classification framework: once strong spatial association rules with only the class label in the consequent are extracted for each granularity level, they are used to mine either propositional or structural spatial classifiers (Ceci et al., 2004; Cesi and Appice, 2006).

4.6 Automatically exploiting background knowledge on spatial phenomena

A large amount of knowledge is often available on spatial phenomena. This is particularly true in the special case of geographic knowledge discovery, where relations among spatial objects express natural geographic dependencies (e.g., a port is adjacent to a water body). These dependences are expressed in non-novel or uninteresting patterns, but with very high support and confidence. If this geographic knowledge were used to constrain the search for new patterns, the scalability of the spatial data mining algorithms would greatly increase. Actually, these dependencies are represented either in geographic database schemas through one-to-one and one-to-many cardinality constraints or in geographic ontologies. Therefore, they can be used at no additional cost in a MRDM perspective, thus, moving a step forward toward knowledge-rich data mining (Domingos, 2007). In the context of spatial data mining, both Appice et al. (2005) and Bogorny et al. (2006) explain how to use knowledge to constrain the search space for spatial association rules.

4.7 Embedding spatial reasoners in spatial data mining systems

Spatial reasoning is the process by which information about objects in space and their relationships are gathered through measurement, observation or inference and used to reach valid conclusions regarding the objects' relationships. For instance, in spatial reasoning, the accessibility of a site A from a site B can be recursively defined on the basis of the spatial relationships of adjacency or contiguity. Principles of spatial reasoning have been proposed for both quantitative and qualitative approaches to spatial knowledge representation. Quantitative spatial reasoning deals with exact numerical values, such as coordinates and distances, and are more akin to machine reasoning, while qualitative spatial reasoning (Freksa, 1991) deals with abstract representations (e.g., 'northwest' and 'far') and is more closely related to the way humans reason. Qualitative spatial reasoning is arguably efficient and can deal to some extent with imprecision, uncertainty and incompleteness, which quantitative reasoning cannot. Embedding spatial reasoning in spatial data mining is crucial to make the right inferences

either when patterns are generated or when patterns are evaluated. Surprisingly, there are few examples of data mining systems, which support some form of spatial reasoning. In SPADA, a limited form of spatial inference is supported if rules of spatial reasoning are encoded in the background knowledge (Malerba et al., 2002). In particular, SPADA applies an ILP technique, known as 'saturation', to make explicit those pieces of information that are implicit in the spatial units of analysis, given the background knowledge. However, although a general-purpose theorem prover for predicate logic can be used for spatial reasoning (as in SPADA): constraints which characterise the spatial problem solving have to be explicitly formulated, in order to make the semantics consistent with the target domain 'space'. Therefore, embedding specialised spatial inference engines in the spatial data mining systems seems to be the most promising, but still unexplored, solution.

5 Conclusions

In this paper, some important issues concerning the discovery of patterns and models from spatial data are presented and discussed. The main specificity of spatial data mining is due to the implicit definition of spatial relationships between objects. We advocate a multi-relational approach to spatial data mining in order to properly deal with these spatial relationships. This approach is promising but it poses several challenges to current MRDM systems, namely:

1 the absence of an explicit modelling of the task-relevant spatial relationships

2 the bias caused by spatial autocorrelation on feature selection

3 the exploitation of the many unlabeled spatial objects in a semi-supervised or transductive setting

4 the demand for collective inference in spatially lagged prediction models

5 the discovery of spatial patterns at various levels of granularity

6 the automated exploitation of background knowledge on spatial phenomena

7 the integration of spatial reasoners into spatial data mining systems.

Obviously, this list of challenges is not exhaustive, but rather it is indicative of the necessity for developing synergies between researchers interested in spatial statistics, MRDM, visualisation, spatial databases and GIS. It is hard to envisage whether the different communities of researchers will actually join forces. Nonetheless, there is good cause for optimism: real applications, such as sales prediction of individual shops, urban data analysis, location-based services, cry out for this collaboration.

Acknowledgements

This paper is based on an invited talk by the author given at the IASC International Conference on Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal, 30 August to 1 September 2007. The author gratefully acknowledges Dr. Annalisa Appice and Dr. Michelangelo Ceci for their useful suggestions on earlier versions of this paper and Lynn Rudd for reading the final version.

References

Appice, A., Berardi, M., Ceci, M. and Malerba, D. (2005) 'Mining and filtering multilevel spatial association rules with ARES', in Hacid, M.S., Ras, Z.W. and Tsumoto, S. (Eds.): *15th International Symposium On Methodologies For Intelligent Systems, ISMIS 2005*, *LNAI*, Springer, Vol. 3488, pp.342–353.

Blockeel, H. and Sebag, M. (2003) 'Scalability and efficiency in multi-relational data mining', *SIGKDD Explorations*, Vol. 5, No. 1, pp.17–30.

Bogorny, V., Camargo, S., Engel, P. and Alvares, L.O. (2006) 'Mining frequent geographic patterns with knowledge constraints', in *Proc. of the 14th annual ACM Int. Symposium on Advances in Geographic Information Systems*, pp.139–146.

Buttenfield, B., Gahegan, M., Miller, H.J. and Yuan, M. (2000) 'Geospatial data mining and knowledge discovery', Technical report, University Consortium for Geographic Information Science Research White Paper, Washington, D.C., available at http://www.ucgis.org/emerging/gkd.pdf), 2000.

Ceci, M. and Appice, A. (2006) 'Spatial associative classification: propositional vs. structural approach', *Journal of Intelligent Information Systems*, Vol. 27, No. 3, pp.191–213.

Ceci, M., Appice, A. and Malerba, D. (2004) 'Spatial associative classification at different levels of granularity: a probabilistic approach', in Boulicaut, J.F., Esposito, F., Giannotti, F. and Pedreschi, D. (Eds.): *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2004*, *LNAI*, Springer-Verlag, Vol. 3202, pp.99–111.

Chapelle, O., Schölkopf, B. and Zien, A. (Eds.) (2006) *Semi-supervised Learning*, MIT Press, Cambridge, MA.

De Raedt, L. (1992) *Interactive Theory Revision*, Academic Press, London.

Domingos, P. (2007) 'Toward knowledge-rich data mining', *Data Mining and Knowledge Discovery*, Vol. 15, No. 1, pp.21–28.

Džeroski, L. and Lavrăc, N. (2001) *Relational Data Mining*, Springer-Verlag, Berlin.

Egenhofer, M.J. and Franzosa, R. (1991) 'Point-set topological spatial relations', *International Journal of Geographical Information Systems*, Vol. 5, No. 2, pp.161–174.

Ester, M., Kriegel, H-P. and Sander, J. (1999) 'Knowledge discovery in spatial databases', in Burgard, W., Christaller, T. and Cremers, A.B. (Eds.): *KI-99: Advances in Artificial Intelligence, 23rd Annual German Conference on Artificial Intelligence*, *LNCS*, Springer, Vol. 1701, pp.61–74.

Freksa, C. (1991) 'Qualitative spatial reasoning', in *Proceedings of the Conference on Cognitive and Linguistic Aspects of Geographic Space*, pp.361–372.

Gao, X., Asami, Y. and Chung, C-J.F. (2006) 'An empirical evaluation of spatial regression models', *Computers & Geosciences*, Vol. 32, No. 8, pp.1040–1051.

Getoor, L., Friedman, N., Koller, D. and Pfeffer, A. (2001) 'Learning probabilistic relational models', in Džeroski, S. and Lavrăc, N. (Eds.): *Relational Data Mining*, Springer-Verlag, pp.307–335.

Gottlob, G. and Leitsch, A. (1985) 'Point-set topological spatial relations', *Journal of the ACM*, Vol. 32, No. 2, pp.280–295.

Güting, R.H. (1994) 'An introduction to spatial database systems', *VLDB Journal*, Vol. 4, No. 3, pp.357–399.

Guttman, A. (1984) 'R-trees: a dynamic index structure for spatial searching', in *Proceedings of the ACM SIGMOD Int. Conf. on Management of Data*, pp.47–57.

Han, J., Kamber, M. and Tung, A.K.H. (2001) *Spatial Clustering Methods in Data Mining: A Survey*, Taylor and Francis, pp.188–217.

Han, J., Koperski, K. and Stefanovic, N. (1997) 'Geominer: A system prototype for spatial data mining', in *Proceedings of the ACM SIGMOD Int. Conf. on Management of Data*, pp.553–556.

Jensen, D. and Neville, J. (2002) 'Linkage and autocorrelation cause feature selection bias in relational learning', in *Proceedings of the Nineteenth International Conference on Machine Learning*, pp.259–266.

Jensen, D., Neville, J. and Gallagher, B. (2004) 'Why collective inference improves relational classification', in *Proceedings of the 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp.593–598.

Klosgen, W. and May, M. (2002) 'Spatial subgroup mining integrated in an object-relational spatial database', in Elomaa, T., Mannila, H. and Toivonen, H. (Eds.): *Principles of Data Mining and Knowledge Discovery (PKDD), 6th European Conference*, *LNAI*, Springer-Verlag, Vol. 2431, pp.275–286.

Koperski, K. and Han, J. (1995) 'Discovery of spatial association rules in geographic information databases', in *Advances in Spatial Databases, 4th International Symposium, SSD'95*, *Lecture Notes in Computer Science*, Springer, Vol. 951, pp.47–66.

Koperski, K., Adhikary, J. and Han, J. (1996) 'Spatial data mining: progress and challenges', in *Proc. ACM SIGMOD Workshop on Research Issues on Mining and Knowledge Discovery*, pp.1–10.

Kühn, I. (2007) 'Incorporating spatial autocorrelation may invert observed patterns', *Diversity and Distributions*, Vol. 13, No. 1, pp.66–69.

Laurini, R. and Thompson, D. (1992) *Fundamentals of Spatial Information Systems*, *APIC Series*, Academic Press, London, Vol. 37.

Lavrăc, N. and Džeroski, S. (194) *Inductive Logic Programming: Techniques and Applications*, Ellis Horwood, Chichester.

Legendre, P. (1993) 'Spatial autocorrelation: trouble or new paradigm', *Ecology*, Vol. 74, pp.1659–1673.

LeSage, J.P. and Pace, K. (2001) 'Spatial dependence in data mining', in Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V. and Namburu, R.R. (Eds.): *Data Mining for Scientific and Engineering Applications*, Kluwer Academic Publishing, pp.439–460.

Malerba, D. and Lisi, F.A. (2001) 'Discovering associations between spatial objects: an ILP application', in Rouveirol, C. and Sebag, M. (Eds.): *Inductive Logic Programming, 11th International Conference, ILP 2001*, *LNAI*, Springer-Verlag, Vol. 2157, pp.156–163.

Malerba, D., Ceci, M. and Appice, A. (2009) 'A relational approach to probabilistic classification in a transductive setting', *Engineering Applications of Artificial Intelligence*, to appear, Vol. 22.

Malerba, D., Ceci, M. and Appice, A. (2005) 'Mining model trees from spatial data', in Jorge, A., Torgo, L., Brazdil, P., Camacho, R. and Gama, J. (Eds.): *Principles and Practice of Knowledge Discovery in Databases, 9th European Conference, PKDD 2005*, *LNAI*, Springer-Verlag, Vol. 3721, pp.169–180.

Malerba, D., Esposito, F., Lisi, F.A. and Appice, A. (2002) 'Mining spatial association rules in census data', *Research in Official Statistics*, Vol. 5, No. 1, pp.19–44.

Muggleton, S. (1992) *Inductive Logic Programming*, Academic Press, London.

Mukerjee, A. and Joe, G. (1990) 'A qualitative model for space', in *Proceedings of AAAI-90*, AAAI Press, pp.721–727.

Neville, J. and Jensen, D. (2000) 'Iterative classification in relational data', in *Proceedings of the AAAI-00 Workshop on Statistical Relational Learning*, pp.42–49.

Neville, J. and Jensen, D. (2003) 'Collective classification with relational dependency networks', in *Proceedings of KDD-2003 Workshop on Multi-Relational Data Mining*, pp.77–91.

Neville, J., Jensen, D., Friedland, L. and Hay, M. (2003) 'Clearning relational probability trees', in *Proceedings of the Ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp.625–630.

Preparata, F. and Shamos, M. (1985) *Computational Geometry: An Introduction*, Springer-Verlag, New York.

Roddick, J.F. and Spiliopoulou, M. (1999) 'A bibliography of temporal, spatial and spatio-temporal data mining research', *SIGKDD Explorations*, Vol. 1, No. 1, pp.34–38.

Roddick, J.F., Hornsby, K. and Spiliopoulou, M. (2001) 'An updated bibliography of temporal, spatial and spatio-temporal data mining research', in Roddick, J.F. and Hornsby, K. (Eds.): *Temporal, Spatial and Spatio-Temporal Data Mining, First International Workshop TSDM 2000 Lyon, France, September 12, 2000, Revised Papers*, *LNAI*, Springer-Verlag, Vol. 2007, pp.147–164.

Shekhar, S. and Chawla, S. (2003) *Spatial Databases: A Tour*, Prentice Hall, Upper Saddle River, NJ.

Shekhar, S., Huang, Y., Wu, W. and Lu, C.T. (2001) *What's Spatial about Spatial Data Mining: Three Case Studies*, Kluwer Academic Pub., pp.357–380.

Tobler, W. (1970) 'A computer movie simulating urban growth in the Detroit region', *Economic Geography*, Vol. 46, pp.234–240.

Van Laer, W. and De Raedt, L. (2001) 'How to upgrade propositional learners to first order logic: a case study', in Džeroski, S. and Lavrăc, N. (Eds.): *Relational Data Mining*, Springer-Verlag, pp.235–261.

Vapnik, V. (1998) *Statistical Learning Theory*, Wiley, New York1998.

Notes

1 In this paper, we consider at most two-dimensional spatial objects.