Spatial Data Preparation for Knowledge Discovery

Vania Bogorny[1], Paulo Martins Engel[1], Luis Otavio Alvares[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
{vbogorny,engel,alvares}@inf.ufrgs.br

Abstract. *There is a well known necessity to extract knowledge from spatial databases. Dozens of algorithms for data mining and knowledge discovery are reported in the specific literature to supply this necessity. However, these algorithms have some general drawbacks. Some consider only spatial data and others, only non-spatial data. Most are pseudo-codes, which are usually not implemented in toolkits, and need inputs in a specific representation. These inputs are often in unusual formats, and if these formats are easily obtainable in real databases is normally not mentioned nor considered. Therefore, outside academic institutions, data mining practice in real spatial databases is rare. This research proposes a methodology to prepare spatial data for data mining, as well as an interoperable software architecture to implement the methodology. Experiments with different input data formats were achieved with different data mining toolkits in order to validate the methodology. The software architecture was implemented in a data preparation toolkit and applied to artificial and real spatial databases.*

1. Introduction

Spatial data are used more and more in areas such as urban planning, transportation, telecommunication, and so on. These data are stored and manipulated by Spatial Database Management Systems (SDBMS) and Geographic Information Systems (GIS) is the technology which provides a set of operations and functions for the analysis and visualization of spatial data. Among the large amount of data stored in spatial databases there is implicit, nontrivial and previously unknown knowledge unable to be captured by GIS operations. Specific techniques are necessary to find this kind of knowledge, which is the objective of the Knowledge Discovery in Databases (KDD) research area.

Knowledge discovery is an interactive and iterative process, depending on many user decisions. Fayyad in (Fayyad et al 1996) classifies the KDD process for non-spatial databases in five main steps: selection, preprocessing, transformation, data mining and interpretation/evaluation. The selection, preprocessing and transformation steps refer to data preparation for data mining. Data mining is the step of applying discovery algorithms which produce an enumeration of patterns over the data. Interpretation is the step where the patterns discovered by data mining algorithms are visualized and analyzed.

Data mining algorithms should be designed to directly access databases, but a large time-consuming exercise is involved in transforming the database into a data mining algorithm-compatible format, which is usually a single table or a single file. This limitation causes a *gap* between databases and data mining algorithms.

Classical data mining algorithms are implemented in many toolkits, but these toolkits do not provide spatial data preparation functions. *Spatial* data mining algorithms are not implemented in many available toolkits, and data preparation in both approaches is basically a manual step. This burden is carried by the end user whom guides the KDD process, which is usually an expert in the application domain, but not an expert in spatial databases.

Data preparation is an important and repetitive step and the discovered patterns depend on the type and the quality of the input dataset. It is stated that this step consumes as much as 60 to 80 percent of the whole KDD process (Adriaans 1996).

Although some techniques to simplify data preparation have been proposed for non-spatial databases, almost no research has focused in reducing the time and work needed to prepare spatial databases, despite advances in spatial data mining algorithms.

Based on the necessity to preprocess large spatial databases for the practice of data mining and knowledge discovery, the first objective of this research is to develop a methodology for spatial data preparation for spatial and classical data mining algorithms (in this paper only data preparation for classical data mining is considered). A second objective is to provide an interoperable framework (software architecture) to implement the methodology for Open GIS-based SDBMS. A third objective is to implement the software architecture as an open source data preparation toolkit, for relational databases. More specifically, this research aims to:

- select the relevant set of spatial features on which discovery will be performed;
- reduce the amount of geographic data with filters on non-spatial attributes;
- transform spatial data in spatial predicates and transpose them into the single table/file format to apply *classical* data mining algorithms;
- transform spatial data into the required geometric format into the single table/file for *spatial* DM algorithms;
- provide an open source data preparation toolkit to automate data preparation steps;
- reduce spatial data preparation time.

1.2. Outline

The remainder of the paper is organized as follows: in Section 2 we describe some related work. Section 3 presents the methodology adopted to develop this research. Section 4 presents a methodology to prepare spatial data for classical data mining, independent of SGBD and data mining toolkit. Section 5 presents a software architecture for spatial data preparation for relational databases. Section 6 concludes the paper and shows the directions of future work.

2. Related Work

There is no specific related work in spatial data preparation for data mining. Most researches in this area define operations or query languages to preprocess spatial databases and extract knowledge. (Sattler and Schallehn 2001) proposed new operations for a query language to select, to integrate, to transform, to clean, to reduce and to transpose data into a single table representation for data mining. However, this approach is for non-spatial data, since no spatial aspects are described in this approach.

Han (Han et al 1997) proposed a geo mining query language (GMQL) for spatial data mining. The drawback in this approach is that the proposed language is not available in commercial or open source SDBMS, only in the GEOMINER software prototype developed by the authors, and that is no longer available.

Ester (Ester et al 1996) defined a set of basic operations and indexes which should be implemented by a SDBMS for KDD. In 2000, Ester (Ester et al 2000) introduced some neighborhood graphs and paths, designed to compute spatial relations for data mining. These approaches have the same drawback: the proposed operations are not implemented by most commercial and open-source SDBMS. Their contributions are theoretical, not practical.

Lazarevic (Lazarevic 2000) proposed a software system for spatial data analysis and modeling with some data preparation steps to detect geographic data inconsistencies. In our point of view this approach is useful to construct cartographic databases, and not to mine geographic databases, where it is supposed that digital maps are previously validated.

Malerba (Malerba et al 2002) proposed an object-oriented data mining query language for classification and association rules, which is implemented by the INGENS (Malerba et al 2003) software prototype. The first problem in this approach is that it works with object-oriented databases, while most SDBMS are relational or object-relational. The second problem is that only the INGENS software prototype implements the proposed language, and for the ATRE algorithm.

Appice et al (2003) defined a spatial features extractor named FEATEX, to select features from a spatial database and to create an output for the SPADA algorithm. The drawbacks of this approach are that most SDBMS do not implement FEATEX's functions and its output is a format for one specific algorithm.

3. Research Methodology

The methodology of work for this research follows a set of steps, described bellow:

1. Study all spatial aspects to be considered in the knowledge discovery process.
2. Study the input data format required by spatial and classical data mining algorithms.
3. Adapt and to transform spatial data for data mining.
4. Create artificial databases with implicit patterns.
5. Prepare and preprocess the artificial databases in different formats and perform experiments with data mining algorithms until finding the implicit rules.
6. Apply the same data preparation steps used with the artificial dataset, with a real database.
7. Study filters to reduce the amount of data and functions to optimize the data preparation process.
8. Define a methodology for data preparation, independent of SDBMS and data mining algorithms.
9. Create an interoperable software architecture and spatial data mining toolkit.
10. Apply the methodology and the software to real data in order to validate and refine the methodology.

4. Spatial Data Preparation

Data mining in conventional databases basically differs from data mining in spatial databases in the spatial relations between spatial features. Classical DM algorithms are unable to interpret the meaning of geographic coordinates (x,y). The coordinates will be considered as two non-spatial variables (e.g. *age* or *gender*) and nothing useful will be achieved. To apply classical DM, spatial data have to be defined in terms of spatial predicates rather than items (Shekhar and Chawla 2003). A spatial association rule for example can be of form $P_1 \wedge P_2 \wedge .... \wedge P_n \rightarrow Q_1 \wedge Q_2 \wedge .... \wedge Q_{m,}$ where at least one of *P* or *Q* is a spatial predicate. For example: *age=old* ∧ *antenna=CLOSE* → *disease=C32*, i.e., if *age* is 60 and *antenna* is *close* than the disease is *cancer*. The spatial predicate is the materialization of a spatial relation.

There are basically three spatial relations between two spatial features to consider: topological, distance and orientation. Topological relations characterize the kind of intersection between two geographic features, such as *crosses, contains, inside, covers, coveredBy, equal, disjoint, overlaps*.

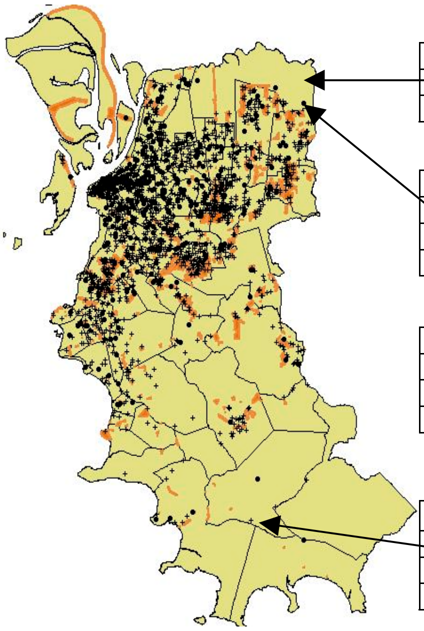
Distance relations are based on the Euclidean distance between two spatial features. Let *dist* be a distance function, *p* be an arithmetic predicate <, >, >=, <= or =, let *d* be a real number and let *A* and *B* be spatial features. The distance relation between *A* and *B* is expressed as *dist* (A, B) *p d*.

Direction relations deal with the order as spatial features are located in space in relation to each other or in relation to a reference object.

4.1. A Spatial Data Preparation for Classical Data Mining

Most data mining algorithms have as an input a single table or single file. For classical data mining, the single table represents the *target feature* on which discovery will be performed. Each row in the single table is an independent unit, i.e. a different instance of *target feature* and each column is an item characterizing this unit.

Figure 1 shows a sample of a spatial database used in case studies. On the left side is a map, which graphically represents spatial data. On the right side, a set of database tables (district, cellular antenna, factory and patient) with spatial and non-spatial attributes.



(a) District

Gid	Name	Area	Perimeter	Shape
1	Mario Quintana	6732056.03	13575.73	Polygon [$(x_1,y_1),(x_2,y_2)$,..]
2	Protasio Alves	8255365.88	23120.24	Polygon [$(x_1,y_1),(x_2,y_2)$,..]

(b) Cellular Antenna

Gid	Shape
1	Polygon [$(x_1,y_1),(x_2,y_2)$,..]
2	Polygon [$(x_1,y_1),(x_2,y_2)$,..]
3	Polygon [$(x_1,y_1),(x_2,y_2)$,..]

(c) Factory

Gid	Type	Impact_Degree	Shape
1	Chemical	High	Point[(x_1,y_1)]
2	Textile	Low	Point[(x_1,y_1)]
3	Metallurgical	High	Point[(x_1,y_1)]

(d) Patient

Gid	Age	Gender	Disease	Data	Address	Shape
1	34	M	I63	03/04/1999	Ipiranga	Point[(x_1,y_1)]
2	60	M	F19.2	04/11/1999	V. Aires	Point[(x_1,y_1)]
3	45	F	C32	01/12/1999	V. Jardim	Point[(x_1,y_1)]

Figure 1. Spatial and non-spatial data representation in spatial databases

Let's consider that *Patient* is the target feature type and the objective is to figure out possible relations among cancer and cellular antennas or cancer and factories. Each row of the single table should be a different hospitalized patient. Factories and antennas will be the *relevant feature types*, which in addition to the non-spatial attributes will characterize the *target feature*. The spatial relations among the *target feature type* (patient) and the *relevant feature types* (cellular antennas and factories) should be materialized as columns in the single table. The materialization depends on the granularity level, which varies according to the objective of the KDD process.

For example, if the objective is to investigate possible relations between *cancer* and *cellular antennas*, than the feature type granularity level is considered, as shown in Table 1. In this case, only the type *antenna* is relevant, and not *which antenna*. When the granularity level is feature type, than a dominant spatial relation is defined. In this example, if at least one cellular antenna is close, than the relation *CLOSE* is dominant over *FAR*.

Table 1. Feature type granularity level

Patient	Age	Gender	Disease	Antenna
1	34	M	I63	CLOSE
2	60	M	F19.2	CLOSE
3	45	F	C32	FAR

On the other hand, if the objective is to investigate possible relations between *cancer* and *factories,* in order to discover *which* factory is polluting the environment and might be related to *cancer*, than the factory instances are considered, as shown in Table 2. Details about the input data format are in (Bogorny et al 2005a).

Table 2. Feature instance granularity level

Patient	Age	Gender	Disease	Antenna_1	Antenna_n	Factory_1	Factory_n
1	34	M	I63	CLOSE	CLOSE	CLOSE	FAR
2	60	M	F19.2	CLOSE	FAR	CLOSE	FAR
3	45	F	C32	FAR	FAR	FAR	FAR

4.2. A Spatial Data Preparation Methodology

Spatial data preparation for classical DM can be performed in three main steps, as shown in Figure 2. The selection step is composed of two sub-steps, *data definition* and *non-spatial filter*, which respectively define and retrieve from the spatial database all relevant information for the KDD process. Besides reducing the amount of data, the selection step characterizes the spatial features through non-spatial attributes, retrieving only data which satisfy the specified conditions. Noise and irrelevant data which do not fulfill the requirements of the definition step will be eliminated.

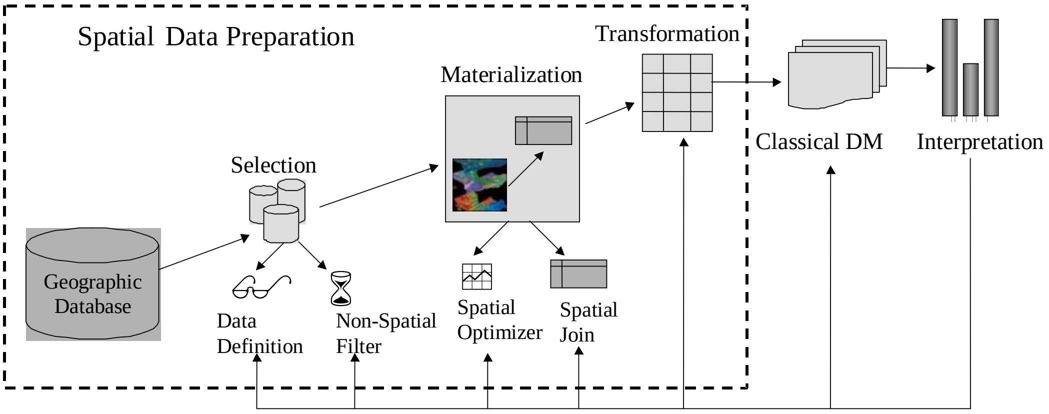


Figure 2. Method of spatial data preparation for classical data mining

Materialization is the step composed of two sub-steps, *spatial optimizer* and *spatial join*. In this step all relevant spatial relations, defined in the selection step, are computed and converted into spatial predicates. Some optimization functions and indexes to speed up the spatial neighborhood computation are defined. The optimization sub-step is not necessarily part of the spatial data preparation process, but in real spatial databases this will significantly reduce the time to compute spatial relations.

Transformation is the step where all materialized relations are transposed into the single table format required by data mining toolkits. The resultant single table will have the target feature non-spatial attributes and all materialized spatial relations with the relevant features.

5. Interoperable Software Architecture for Spatial Data Preparation

Spatial data can be stored in different databases and in different formats. In this section, we present an interoperable software architecture to implement the spatial data preparation steps, presented in the previous section, for classical data mining and relational databases. The software architecture is based on the OGC (Open GIS Consortium) database schema and OGC spatial operations (Open GIS Consortium 1999), becoming interoperable with any SDBMS constructed under these specifications. The OGC is an organization which defines standards for Geographic Information Systems.

Figure 3 shows the architecture design, which has three abstraction levels: *data repository, data preparation* and *data mining*. Data repository is the spatial database level, which can be any SDBMS implemented under OGC specifications. Data mining is the level which represents any classical data mining toolkit, and data preparation is the level which covers the gap between spatial databases and classical data mining toolkits.

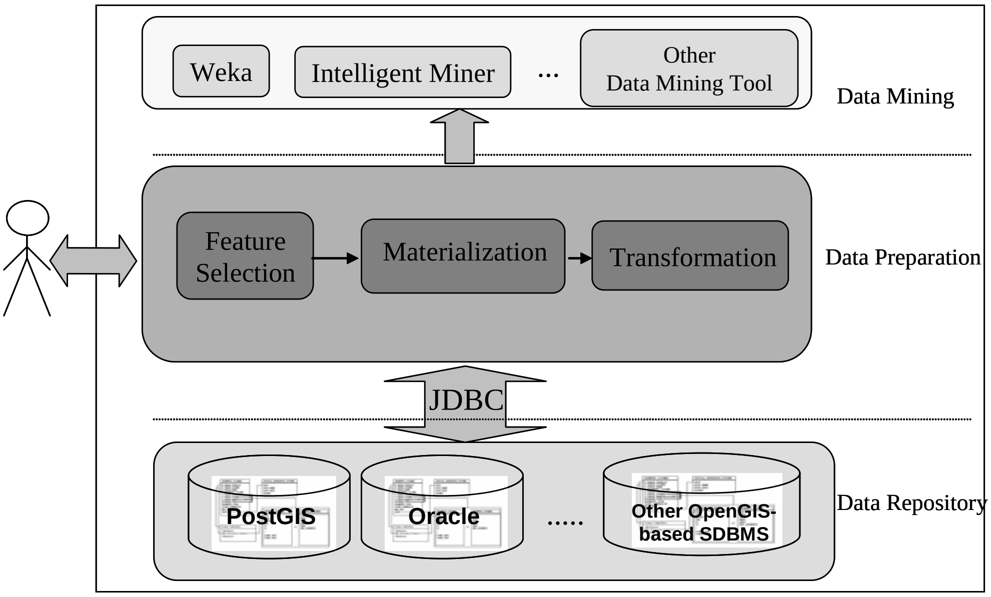


Figure 3. Data Preparation Software Architecture for Classical Data Mining

The data preparation level consists of three main modules, each one implementing the steps presented in the previous section: *selection, materialization,* and *transformation.* All modules can be implemented with standard SQL statements. SQL is the standard data manipulation language implemented by most SDBMS.

The *selection* module implements the first step in the data preparation process. This module is based on the OGC database schema and retrieves all spatial feature types (database tables) in the specified schema.

Materialization is the module which computes the spatial predicates. This module performs with the OGC functions. The topological relations are performed with SQL statements and the topological functions *touches, contains, equal, covers, crosses,* and *inside.* Distance relations are computed with the function *distance.* Buffer areas are computed with the *buffer* and *intersection* operation.

The optimization step is implemented with the *envelope* function, in order to store the minimum boundary rectangle (MBR) of spatial features represented as lines or polygons. Experiments performed before and after the MBR showed a reduction in computation time around 80%.

The *transformation* module implements the transformation step of the data preparation process and is performed with traditional SQL statements.

The proposed software architecture can be implemented according to three use-cases:

- an independent middleware between spatial databases and data mining toolkits;
- a special module in a Geographic Information System (GIS); or
- a data preparation module of a data mining toolkit.

So far, we implemented the proposed software architecture as a middleware for OpenGIS-based SDBMS and the Weka (Witten and Frank 2000) data mining toolkit (Bogorny et al 2005b). Experiments were performed with real and artificial databases.

6. Conclusion and Future work

Although a large number of spatial data mining papers are available in the literature, knowledge discovery in real spatial databases is an arduous task. Some reasons are that only a few toolkits are available for spatial data mining and the input is in a restrictive format. Another problem is that data preparation is basically a manual task, and for classical data mining algorithms, many data preparation steps are required.

This research presents a methodology for the end user of any application domain to easily prepare large amounts of spatial data for classical data mining.

A software architecture was also presented and implemented in order to provide a data preparation toolkit. Experiments with the Weka data mining toolkit and real spatial databases stored in PostGIS were performed. Experiments with other databases and data mining toolkits will be performed in future work.

We also intend to study and evaluate the granularity level and dominance concepts for each spatial relation. The optimization steps will also be evaluated.

The methodology and the software architecture will be extended to prepare spatial data for spatial data mining algorithms.

Acknowledgments

This research partially supported by CAPES and CNPq.

References

Adriaans, P. and Zantinge, D. *Data mining.* Addison Wesley Longman, Harlow, England, 1996.

Appice A, Ceci M, Lanza A, et al (2003) Discovery of spatial association rules in geo-referenced census data: a relational mining approach, *Intelligent Data Analysis* (Software & Data) **7**, 6.

Bogorny, V., and Alvares, L. O. (2005a) *Geographic Data Representation for Knowledge Discovery.* Technical Report UFRGS-TR-349, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

Bogorny, V., Engel, P. M. and Alvares, L. O. (2005b) A reuse-based spatial data preparation framework for data mining. To appear in: *Proceedings of the SEKE 17 th international Conference on Software Engineering and Knowledge Engineering(SEKE'2005)*(July 14-16, 2005). Taiwan, China.

Ester M., Kriegel H.-P., Sander J. and Xu X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the KDD 2th international conference on Knowledge Discovery and Data Mining (KDD'96*)(August 2-4, 1996). AAAI Press, Portland, OR, 226-231.

Ester, M., Frommelt, A., Kriegel, H.-P., and Sander, J. (2000) Spatial data mining: database primitives, algorithms and efficient DBMS support. *Journal of Data Mining and Knowledge Discovery*, 4, 2-3(Jul. 2000), 193-216.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) Knowledge discovery and data mining: towards a unifying framework. In *Proceedings of the KDD 2th international conference on Knowledge Discovery and Data Mining (KDD'96)* (August 2-4, 1996). AAAI Press, Portland, OR, 1996, 82-88.

Han, J., Koperski, K., Stefanvic, N. (1997) GeoMiner: a system prototype for spatial data mining. In *Proceedings of the ACM-SIGMOD international conference on Management Of Data (SIGMOD'97)* (May 13-15,1997). ACM Press, Tucson, AR, 553-556.

Lazarevic, A., Fiez, T. and Obradovic, Z. (2000) A software system for spatial data analysis and modeling. In *proceedings of the HICSS Hawaii International Conference on System Sciences (HICSS'00)* (Jan. 04–07, 2000). IEEE Computer Society Press, Hawaii.

Malerba, D., Appice, A. and Vacca N. (2002) SDMOQL: an OQL-based data mining query language for map interpretation tasks. In *Proceedings of the DTDM workshop on Database Technologies for Data Mining (DTDM'02)*(March 25-27, 2002). Springer, Prague, Czech Republic.

Malerba, D., Esposito, F., Lanza, A., Lisi, F.A. and Appice, A. (2003) Empowering a GIS with inductive learning capabilities: the case of INGENS. *Journal of Computers, Environment, and Urban Systems*, 27, 3 (May 2003), 265-281.

Open GIS Consortium (1999) *Open GIS simple features specification for SQL*. In URL: http://www.opengeospatial.org/docs/99-054.pdf.

Sattler, K. and Schallen, E. (2001) A data preparation framework based on a multi-database language. In *Proceedings of the IDEAS 5th International Database Engineering and Applications Symposium (IDEAS'01)*(July 16-18, 2001). IEEE Computer Society, Grenoble, France, 2001, 219-228.

Shekhar, S., Chawla, S. *Spatial databases: a tour*. Prentice Hall, Upper Saddle River, NJ, 2003.

Witten, I. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA, 2000.