Transactions on Information and Communications Technologies vol 19 © 1998 WIT Press, www.witpress.com, ISSN 1743-3517

Spatial Data Mining with Fuzzy Decision Trees

C. Marsala & N. Martini Bigolin*
LIP6, Université Pierre et Marie Curie,
4 place Jussieu, 75252 Paris cedex 05, FRANCE.
EMail: {Christophe.Marsala,Nara.Martini-Bigolin}@lip6.fr

Abstract

In this paper, an approach is presented to search for useful patterns and discover hidden information in Spatial Object-Oriented Databases (SOODB). Although many approaches of knowledge discovery for relational spatial databases exist, there is a growing interest in mining SOODB. Indeed, object-oriented databases are well-suited to represent complex spatial information. Moreover, a very large number of existing spatial databases are ready to be mined. We propose an algorithm to mine a SOODB. After a spatial object query and a mathematical and fuzzy preprocessing, we apply decision tree based techniques and fuzzy set theory to discover knowledge. An experiment on a region of France to discover classification rules related to houses and urban area is conducted with this algorithm to validate the interest of the approach.

1 Introduction

Knowledge discovery in databases (KDD) [1] has been used to extract implicit information from vast amounts of data. Recently, this technology has attracted the interest of researchers in several fields such as databases, statistics, machine learning, data visualization and information theory.

There already exist many approaches for mining relational databases [2], [3]. These approaches are used to discover several kind of rules (association [4], classification [5], discrimination [6]) by means

*Supported by CNPq - Brazil

of machine learning techniques (decision trees based techniques [7], clustering techniques [8],...).

However, object-oriented databases (OODB) [9] have become popular and influential in the development of new generations of database systems. This has motivated the research on techniques for data mining in object-oriented databases. Han et al [10] overview the mechanisms for knowledge discovery in object-oriented databases with an emphasis on the techniques for generalization of complex data objects, methods and class hierarchies. Yoon and Henschen [11] extracted knowledge from large data sets in object-oriented databases to facilitate semantic query processing in database systems. Nevertheless, the research on knowledge discovery in object-oriented databases is still shallow, since object structures are complex and difficult to process.

Researchers in Geographic Information Systems (GIS) [12] have also shown interest in knowledge discovery in spatial databases, called Spatial Data Mining. Spatial Data Mining has been defined as *the extraction of interesting spatial patterns and features, general relationships between spatial and non-spatial data, and their general data characteristics not explicitly stored in spatial databases* [13]. This technology is becoming more and more important in spatial databases, because a tremendous amount of spatial and non-spatial data have been collected and stored in large spatial databases using automatic data collection tools.

More significant spatial data mining works have been developed to discover knowledge from relational databases [14], [15], [16]. However, knowledge discovery in SOODB is still an open search area with large potential.

In this paper, we present an approach to discover knowledge from a spatial object-oriented database (SOODB). A training set is generated through a spatial object query and a preprocessing. This preprocessing is done by means of background knowledge given as mathematical functions and fuzzy set operators. Afterwards, decision tree based techniques and fuzzy set theory enables us to discover knowledge. A decision tree summarizes the knowledge lying in a set of data into a set of decision rules. Spatial data are handle by fuzzy set theory. This leads us to the construction of *fuzzy decision trees.* With this method, spatial data are represented as fuzzy modalities during both the construction of a tree and its use.

This paper is composed as follows: in Section 2, we introduce SOODB and fuzzy decision trees. In Section 3, our approach is presented to discover knowledge from a SOODB. An application of this approach is given in Section 4. Finally, we conclude and present some directions for future research.

2 Data structures and learning techniques

2.1 Object structures and spatial data

2.1.1 Object-Oriented Databases

An OODB is composed by a set of *objects*. An object is associated with an *object-identifier* and a *value*. A value possesses a type either *atomic* (string,...) or *structured*. The structure can be a *collection* (a list, a set,...) or a *tuple* (a set of typed attributes).

Objects are grouped into *classes* which are organized in a *hierarchy*. The object's behavior is determined by a set of methods. The instances of a class are defined as a set of *objects*. Each object is associated with a *name* that references it in the database.

The manipulation of a database is done with a query language. This language supports the extraction of data from the current base. The answer of this query is a set of objects or a set of values for these objects.

2.1.2 Spatial Data

Geographical data is composed of non-spatial and spatial description of these objects. Non-spatial data is information of the kind: name, population of town and etc... Spatial data specifies the localization of non-spatial data. It can be represented by three spatial primitives: points, lines, and areas.

A point represents (the geometric aspect of) an object for which only its location in space, but not its extent, is relevant. For example, a house can be a point in a large geographic area (a large scale map) (Figure 1). A region is the abstraction for something having an extent in 2D-space, e.g. a country, a lake, a national park or a house in small scale map. A line is the basic abstraction for facilities for moving through space, or connections in space (roads, rivers, cables for phone, electricity, etc) [17].

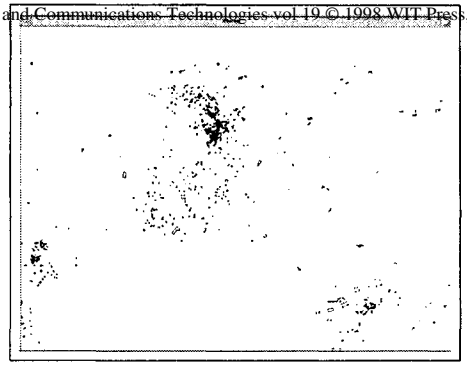

Figure 1: A region of France (1:125000)

2.2 Fuzzy set theory

In classical theory, given a set $\mathcal{X}$ and a subset $U \subseteq \mathcal{X}$, each element $x \in \mathcal{X}$ either belongs to U or does not belong to U. It can be summarized as follows: Given a set $\mathcal{X}$ and a subset $U \subseteq \mathcal{X}$, let μ_U be the *characteristic function* such that: $\mu_U : \mathcal{X} \longrightarrow \{0, 1\}$, and $\forall x \in \mathcal{X}$, if $x \in U$ then $\mu_U(x) = 1$, otherwise $\mu_U(x) = 0$.

In *fuzzy set theory* [18], the *membership degree* of an element x can vary from 0 to 1. The membership function μ_U of the *fuzzy set* U is defined as: $\mu_U : \mathcal{X} \longrightarrow [0, 1]$.

Fuzzy set theory leads to take into account *numerical-symbolic* attributes. Such an attribute has values that can be fuzzy sets. Thus, given a numerical attribute U that takes numerical values in $\mathcal{X}$, a *fuzzy partition* can be defined on $\mathcal{X}$ by means of a set of fuzzy sets of $\mathcal{X}$. And therefore, U can be considered as a numerical-symbolic attribute.

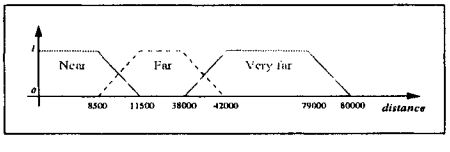

Figure 2: Fuzzy modalities for the distance

For instance, the distance between two points can be considered as a numerical-symbolic attribute: either a numerical value, as "172 meters", or a numerical-symbolic value such as "far" is associated with it. In Figure 2, the attribute *distance* and its fuzzy values (*near*,

far, very far) forming a fuzzy partition of its universe $\mathbb{R}^+$ are represented.

2.3 A learning technique

2.3.1 Fuzzy Decision Trees

A decision tree is a natural structure of knowledge. Each node in such a tree is associated with a test on the values of an attribute, all edges from a node are labeled with values of the attribute belonging to a fuzzy partition of its universe, and each leaf of the tree is associated with a value of the class. Let A be a set of attributes and A_d the particular attribute representing the decision (also called *class*).

Edges can also be labeled by numerical-symbolic values. Such kind of values leads to the generalization of decision trees into *fuzzy decision trees* [19]. An example of fuzzy decision tree is given in Figure 5. Fuzzy decision trees handle numeric-symbolic values either during their construction or when classifying new cases. The use of fuzzy set theory enhances the understandability of decision trees when considering numerical attributes. Moreover, it has been proven in [20] that the fuzziness leads to a better robustness when classifying new cases.

Construction of a Fuzzy Decision Tree

A decision tree can be constructed from a set of examples by *inductive learning*. Inductive learning is a process to generalize knowledge from the observation of a given phenomenon. It is based on a set of *examples*, called the *training set*. Each example is a case already solved or completely known, associated with a pair [description, class] where the *description* is a set of pairs [attribute, value] which is the available knowledge.

A decision tree is built from its root to its leaves. The training set is successively split by means of a test on the value of a chosen attribute (the divide and conquer strategy). Each test is associated with a node in the tree. This process is resumed until the current training set fullfils a given criterion. In this case, this set will label a leaf of the decision tree.

Transactions on Information and Communications Technologies vol 19 © 1998 WIT Press, www.witpress.com, ISSN 1743-3517

A path of the tree is equivalent to an IF...THEN rule $R : Pr \longrightarrow Co$. The premise Pr for such a rule is composed by tests on values of attributes, and the conclusion Co is the value of the decision that labels the leaf of the path.

Therefore, the whole fuzzy decision tree is equivalent to a fuzzy rule base $RB = \{R_1, ..., R_N\}$.

2.3.3 Construction of Fuzzy Partitions

The process of construction of fuzzy decision trees is based on the knowledge of a fuzzy partition for each numerical attribute. However, it can be difficult to possess such a fuzzy partition. An automatic method of construction of a fuzzy partition from a set of values for a numeric-symbolic attribute is proposed in [21]. With this method, a fuzzy partition is generated automatically from a set of numerical values.

3 Discovery knowledge in SOODB

To discover knowledge from an SOODB, both the object-oriented nature of the data and their spatial specificity need to be taken into account simultaneously. However, no data mining techniques that can handle these two data properties exists yet. Thus, to solve this problem, we propose to split the process of knowledge discovery into several steps (Figure 3): *selection*, *preprocessing*, *data mining* and *interpretation*. Each step performs a transformation of the spatial data stored in the SOODB into another representation more appropriate to the next step.

These transformations make use of a knowledge base. This base is composed of background knowledge such as the semantic linked to the object structure, the spatial object topology and the expert knowledge.

3.1 Selection: Spatial OO query

The data selection is made through a spatial query Q. This query is run on the $SOODB$ to extract a data set relevant to the data mining

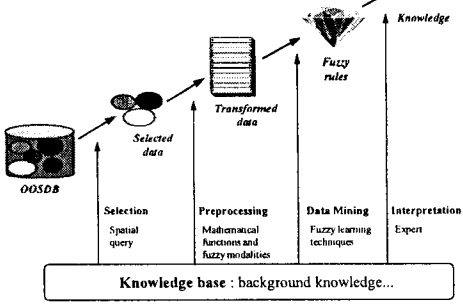

Figure 3: Knowledge discovery process

task. The answer of this query is a limited set of spatial objects SO. This query is represented by: $Q : SOODB \longrightarrow SO$.

For instance, the answer of a query performed on an area of the map is the set of spatial objects related to this area.

3.2 Preprocessing: Mathematical functions and fuzzy set theory

This step processes object oriented spatial data. Preprocessing is a function $\sigma : SO \longrightarrow TS$ which transforms a set of spatial objects SO into a training set TS.

3.2.1 Mathematical functions

The transformation of a spatial object $O_S \in SO$ into a data set $D \subseteq TS$ is performed by means of a set $\mathcal{S}$ of mathematical functions f. We have $\mathcal{S} = \{f \mid f : SO \longrightarrow \mathcal{X}_f\}$ where $\mathcal{X}_f$ is the set of values computed by the function f.

For instance, a *line* l is a spatial object composed by a set of points $\{p_{l_1}, ..., p_{l_n}\}$. A particular function f is the Euclidean distance: $f : SO \longrightarrow \mathbb{R}$ defined as $f(l) = ||p_{l_1} - p_{l_n}||$.

3.2.2 Fuzzy set theory

For a given function f, if $\mathcal{X}_f$ is a set of continuous values (*ie.* $\mathcal{X}_f \subseteq \mathbb{R}$), a transformation $T_{\mathcal{X}_f}$ of these values into fuzzy values can be done according to a fuzzy partition of $\mathcal{X}_f$ into m fuzzy sets $U_1, ..., U_m$. Thus, we have $T_{\mathcal{X}_f} : \mathcal{X}_f \longrightarrow [0.1]^m$.

For instance, the Euclidean distance $||p_1 - p_2||$ between two points is a real value from $\mathbb{R}^+$. Given the set $\{near, far, very\ far\}$ of fuzzy sets describing the distance (see Figure 2), the real value $||p_1 - p_2||$ is converted into a set of three membership degrees to each fuzzy set: $\mu_{near}(||p_1 - p_2||)$, $\mu_{far}(||p_1 - p_2||)$ and $\mu_{very\ far}(||p_1 - p_2||)$.

3.3 Data mining: Fuzzy decision trees

The data mining step transforms a training set TS into a set RB of rules.

Given a training set TS, obtained from the previous step, and a particular attribute $A_d \in \mathcal{A}$ (the decision), provided by the expert, the data mining step is a function $\Theta : TS \times \mathcal{A} \longrightarrow RB$.

This function generates a set $RB = \{R_1, ..., R_N\}$ of rules $R_i : Pr_i \longrightarrow Co_i$. The premise Pr_i of rule R_i is a conjunction of tests $A_k = a_{k_j}$ on values a_{k_j} of attributes $A_k \in \mathcal{A} - \{A_d\}$. The conclusion Co_i of a rule is a value d_i for A_d.

For instance, the function Θ can be an algorithm to generate decision trees. From a training set, it produces a decision tree equivalent to a rule base.

In order to take into account the fuzzy sets introduced in the previous step, an appropriate algorithm has to be used. In this case, Θ produces a fuzzy rule base RB.

For instance, the function Θ is an algorithm to generate fuzzy decision trees.

3.4 Interpretation

To summarize, the process of knowledge discovery generates a set RB of rules from an SOODB. It can be interpreted by an expert. The introduction of fuzzy set theory produces fuzzy rules that are more understandable than classical rules in presence of numerical values for attributes.

4 Application

An application of our method is presented in this section.

Given a database, a selection, a preprocessing and a data mining are performed. The output result is a set of rules to classify houses as urban or non-urban. The SOODB used was provided by the French

"Institut Géographique National" (IGN) and was implemented on top of the O_2 DBMS [22].

The whole process is described in the following sequence.

4.1 Selection

A query made on the SOODB is used to extract a relevant set of data. This data set $\mathcal{H}$ consists of all houses pertaining to an area (see Figure 4).

The selection of these data is performed with the following query:

```
select x.house from x in DataBase1
where x.house->inArea(CoordPtMin, CoordPtMax);
```

where the method `inArea(CoordPtMin, CoordPtMax)` determines whether an object `house` is in the area defined by the two points `CoordPtMin` and `CoordPtMax`. This area is defined by the user with the graphical interface system. A point is selected with the mouse. It is associated with the center of the area and `CoordPtMin` and `CoordPtMax` are the boundaries of this area, computed from this selected point.

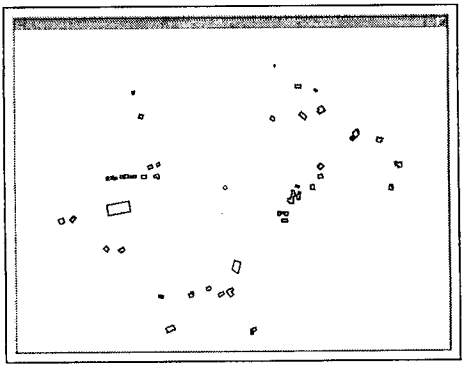

Figure 4: Result of a query (1:31860)

4.2 Preprocessing

Each object from $\mathcal{H}$ is associated with additional information to make up an example of a training set.

This information is computed by means of knowledge related to the application. An instance of such knowledge, is the mathematical function $d(p_1, p_2)$ given to value the Euclidean distance between the two points p_1 and p_2.

A particular kind of such knowledge consists in fuzzy modalities (*near*, *far* and *very far*) on the numerical universe of values distance (Figure 2). Given a house h, the number of houses in the three fuzzy area defined by these fuzzy modalities is valued. For each house h' different from h, the distance $d(p_h, p_{h'})$ is evaluated. This distance is transformed into membership degrees $\mu_{near}(h')$, $\mu_{far}(h')$ and $\mu_{very\ far}(h')$. Thus, the number of houses in the area defined by the modality *near* is given by the fuzzy measure of cardinality:

$$Nr\ near = \sum_{h' \in \mathcal{H}} \mu_{near}(h')$$

And so on, for the other modalities.

Some examples of the obtained training set is shown in Table 1.

Table 1: Examples from the Training set

House	*Nr near*	*Nr far*	*Nr very far*	*Urban*
h1	0.1	4.2	5.7	No
h2	2.0	2.0	6.0	No
h3	3.3	3.8	7.9	No
h4	0.0	4.0	1.0	No
h5	16.3	15.7	12.0	Yes
h6	11.0	24.2	9.7	Yes
h7	7.7	20.2	16.1	Yes
h8	14.9	12.1	9.0	Yes

4.3 Data Mining

A fuzzy decision tree is constructed from this training set with the Salammbô system [20] chosen as the algorithm (*cf* Figure 5). The input of the system is the training set and the output is a set of classification rules.

As mentioned, this system generates automatically a fuzzy decision tree and determines fuzzy modalities for the universe of values

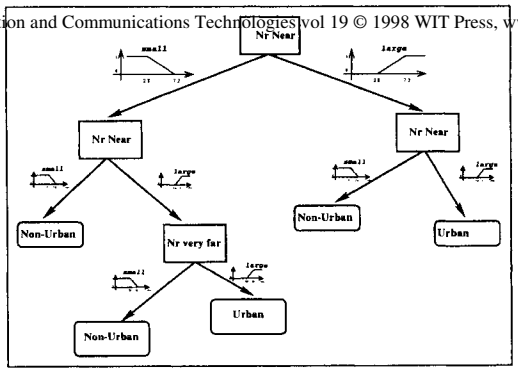

Figure 5: Fuzzy decision tree

of each numerical attribute. During the construction of the fuzzy decision tree, a fuzzy partition is generated upon the universe of values of the three attributes (*Nr near*, *Nr far* and *Nr very far*). The obtained fuzzy modalities on the number of houses in area are given in Figure 2. These fuzzy modalities will label the premises of the induced classification rules.

4.4 Validation

This set of rules, induced from houses belonging to a zone around a particular town, has been tested with other houses pertaining to a zone around another town.

Step 4.1 and step 4.2 were resumed with another center for a zone to generate a set (the *test set*) to check the fuzzy rule base obtained in step 4.3.

For instance, in our application, the studied region (Figure 1) is composed of three towns T_1, T_2 and T_3. A training set was generated from a zone around the town T_1 (Figure 4) and a test set was generated from a zone around the two other towns.

The average error rate when classifying houses around towns T_2 and T_3 with the obtained fuzzy decision tree is 89.9%. In other words, given 413 houses from the new zone, 371 houses are perfectly classified as *urban* or *non-urban* with the induced set of fuzzy rules.

5 Conclusion

In this paper, an approach has been presented to discover hidden information in a Spatial Object-Oriented Databases (SOODB).

There exists many approaches that focused on knowledge discovery either in spatial relational databases or in object-oriented databases. However, the use of knowledge discovery techniques in SOODB is still a challenge.

We introduce an algorithm to mine a SOODB. After a spatial object query and a mathematical and fuzzy preprocessing, we apply decision tree based techniques and fuzzy set theory to discover knowledge. This introduction of preprocessing and fuzzy decision trees enables us to handle spatial data related to a geographical region. Our algorithm has been applied and tested on a region of France to discover characterization rules related to houses and urban area.

In future work, we plan to automate the process of building the queries used to construct the training set. The preprocessing step is being integrated as operators of the query language in order to be apply directly on the data of the SOODB.

Acknowledgments

This work has been made possible thanks to the database provided by IGN (French National Geographic Institute).

The authors express their thanks to Bernadette Bouchon-Meunier and Anne Doucet for their guidance and their helpful comments.

References

[1] Fayyad U. M., Piatetsky-Shapiro G., & Smyth P. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, Fall 1996.

[2] Agrawal R. & Srikant R. Fast algorithms for mining association rules. In *Proc. of the 20th Int. Conf. on Very Large Databases (VLDB)*, pages 487–499, Santiago, Chile, September 1994.

[3] Houtsma M. & Swami A. Set-oriented mining for association rules in relational databases. In *Proc. of the Int. Conf. on Data Engineering (ICDE)*, pages 25–33, Taipei, Taiwan, March 1995.

[4] Agrawal R. & Srikant R. Mining generalized association rules. In *Proc. of the 21st Int. Conf. on Very Large Databases (VLDB)*, pages 487–499, Zurich, Switzerland, September 1995.

[5] Mehta M., Agrawal R., & Rissanen J. SLIQ: A fast scalable classifier for data mining. In *Proc. of the Fifth Int. Conf. on Extending Database Technology*, pages 18–32, Avignon, France, March 1996. Lecture Notes in Computer Science n° 1057.

[6] Han J., Cai Y., & Cercone N. Knowledge discovery in databases: An attribute-oriented approach. In *Proc. of 1992 Int. Conf. on Very Large Data Bases (VLDB'92)*, pages 547–559, Vancouver, Canada, August 1992.

[7] Kamber M., Winstone L., Gong W., Cheng S., & Han J. Generalization and decision tree induction: Efficient classification in data mining. In *Proc. of the Int. Workshop on Research Issues on Data Engineering (RIDE'97)*, pages 111–120, Birmingham, England, April 1997.

[8] Xu X., Ester M., Kriegel H.-P., & Sander J. A distribution-based clustering algorithm for maning in large spatial databases. In *14th Int. Conf. on Data Engineering*, 1998.

[9] Bancilhon F., Delobel C., & Kanellakis P. *Building an Object-Oriented Databases Systems: The story of 02.* Morgan Kaufmann, 1992.

[10] Han J., Nishio S., & Kawano H. Knowledge discovery in object-oriented and active databases. In Fuchi F. & Yokoi T., editors, *Knowledge Building and Knowledge Sharing*, pages 221–230. Ohmsha, Ltd. and IOS Press, 1994.

[11] Yoon S. & Henschen L. Mining knowledge in object-oriented frameworks for semantic query processing. In *Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada, June 1996.

[12] Fotheringham A. & Rogerson S. P. *Spatial analysis and GIS: applications in GIS.* London Washington, 1993.

[13] Koperski K. & Han J. Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int. Symp. on*

Large Spatial Databases (SSD95), pages 47–66, Maine, August 1995. Lecture Notes in Computer Science n° 951.

[14] Han J., Koperski K., & Stefanovic N. GeoMiner: A system prototype for spatial data mining. In *Proc. 1997 ACM-SIGMOD Int'l Conf. on Management of Data(SIGMOD'97)*, Tucson, Arizona, May 1997.

[15] Ester M., Kriegel H.-P., & Sander J. Spatial data mining: A database approach. In *Proc. 5th Symp. on Spatial Databases*, Berlin, Germany, 1997.

[16] Koperski K., Han J., & Adhikary J. Mining knowledge in geographic data. *Communication of the ACM*, 1998, to appear.

[17] Laurini R. & Thompson D. *Fundamentals of Spatial Information Systems.* Academic Press, 1992.

[18] Zadeh L. A. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[19] Bouchon-Meunier B., Marsala C., & Ramdani M. Learning from imperfect data. In D. Dubois H. P. & Yager R. R., editors, *Fuzzy Information Engineering: a Guided Tour of Applications*, pages 139–148. John Wileys and Sons, 1997.

[20] Marsala C. *Apprentissage inductif en présence de données imprécises : construction et utilisation d'arbres de décision flous.* Thèse de doctorat, Université Pierre et Marie Curie, Paris, France, Janvier 1998. Rapport LIP6 n° 1998/014.

[21] Marsala C. & Bouchon-Meunier B. Fuzzy partioning using mathematical morphology in a learning scheme. In *Proceedings of the 5th Conference on Fuzzy System, FUZZ'IEEE*, volume 2, pages 1512–1517, New Orleans, USA, September 1996.

[22] Adiba M. & Collet C. *Objets et bases de données. Le SGBD O2.* Hermes, 1993.