A Decision Tree for Multi-Layered Spatial Data

Nadjim Chelghoum[1], Karine Zeitouni[1], and Azedine Boulmakoul[2]

PRISM Laboratory, University of Versailles, 45 avenue des Etats-Unis, 78035 Versailles Cedex, France, Nadjim.Chelghoum@prism.uvsq.fr, Karine.Zeitouni@prism.uvsq.fr[1]
LIST Laboratory, Science and Techniques Faculty of Mohammedia, B.P. 146 Mohammedia, Morocco, boul@uh2m.ac.ma[2]

Abstract

Spatial data mining fulfils real needs of many geomatic applications. It allows the geomatics community to take advantage of the growing availability of geographically referenced data and benefit from this rich information resource. This article addresses spatial data classification for using decision trees. A new method called SCART which differs from conventional decision trees by considering the specifics of geographical data, namely their organisation in thematic layers, and their spatial relationships is proposed. SCART is an extension of CART methods in two ways. On the one hand, the algorithm considers several thematic layers as in the so-called relational data mining area, and on the other hand, it extends discriminating criteria to address concerns about the neighbourhood. As such, the algorithm determines which combination of attribute values and spatial relationships of neighbouring objects provide the best criterion.
Keywords: spatial data mining, classification rules, decision tree, spatial relationship, spatial database

1 Requirements

The growing development of automatic mapping results in the production of large spatial databases. More and more applications require access to large data volumes, however, the complexity and size of these databases exceed our capacity to effectively analyse them. It thus seems appropriate to develop and apply techniques in automatic knowledge extraction through processes referred to as data mining.

The domain of interest for this particular paper is in traffic risk analysis (Huguenin 00). Traffic risk analysis requires the identification of road safety

problems in an effort to propose appropriate safety measures. This project aims at identifying relevant risk models to help in various traffic safety tasks. The risk assessment is based on an analysis of information about previous injury accidents collected by police forces. Currently, however, , this analysis has been based on statistics with no consideration about the various spatial relationships that are associated with the accidents. This work aims at identifying risky road sections and analysing and explaining those risks with respect to the geographic context.

The risk analysis presented in this study combines accident information with thematic information about the road networks, the population census, the buildings, and other geographic neighbourhood detail. The paper presents details about the classification task and builds a decision tree that integrates the spatial features of the thematic layer in this case, accident information. Through the decision tree along with the spatial assessment of accidents, one can explain and predict the danger of roads by their geographic context.

Thus, it appears that decision trees can be effectively extended through an integrated assessment of the properties of neighbouring objects. As such potential exists for the development of explanations about analysed phenomena. Two technical problems arise:

1. Neighbouring objects could belong to thematic layers other than the theme analysed.. Yet decision trees consider only one table (theme attributes) where each row represents a learning example. In these cases , a multi-table decision tree is needed;
2. Many definitions of neighbours exist, giving rise to confusion. Indeed, a spatial relationship could be topologic when the objects touch each other, or metric when they are close. In this case, each separating distance represents a particular spatial relationship. Consequently, multi-layered spatial decision trees are more than a multi-table decision tree. The multi-layered spatial decision tree should support the automatic filtering of the multiple and even infinite number of spatial criteria.

The remainder of this paper is organised as follows: section 2 gives a summary of the state of the art in spatial data mining; section 3 presents the proposed method and specifies the algorithm and section 4 gives the prototype architecture, test results, and discusses the implementation issue. Our conclusions are presented in section 5.

2 Background

This section links this work to general research in spatial data mining, highlights the support of spatial relationships and describes other works on decision trees.

2.1 Spatial Data Mining

The goal of spatial data mining is to discover hidden knowledge from spatial databases by combining spatial and non-spatial properties. The spatial data mining methods are usually an extension of those used in conventional data mining (Fayyad 96). Spatial data mining consists of two functions (Zeitouni 00a). The first function addresses a spatial phenomenon by exploring data, for example identifying risky zones by viewing the spatial distribution of the accident locations. The second function explains or even predicts the phenomena while looking for some association or relationship with properties of the geographic environment. For instance, accidents could be "explained" by the state of the road or the surrounding urban density. The spatial classification clarifies these explanatory methods.

2.2 Spatial Relationships

As emphasised above, the main considerations in spatial data mining is that it considers the spatial relationships among objects (Egenhofer 93). Unlike the relational data model, spatial relationships are implicit. Computing them requires many spatial join operations, which can be computationally burdensome. .

In a recent article, (Zeitouni 00b) a method to simplify this process using a secondary structure has been presented. This structure is called spatial join index (SJI), and is an extension of the well-known join indices introduced by (Valduriez 87) in the relational database framework. It pre-computes the exact spatial relationships between objects of two thematic layers. As shown in Fig. 1, a SJI is a secondary table that references matching objects from thematic layers R and S and stores their spatial relationships. In case this relationship is topological (such as

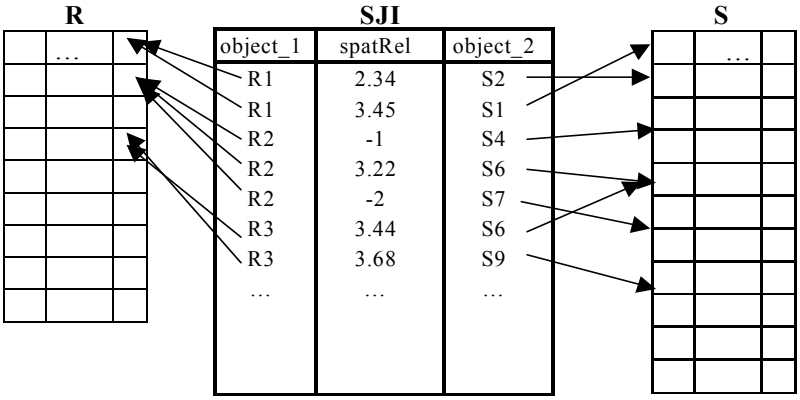

Fig. 1. Spatial join index

inclusion, adjacency or overlapping), the *spatRel* attribute will contain a negative code, such as *(R2, S4)*. Otherwise, it will store the exact distance value. For performance reasons, the distance calculation is limited to a given useful perimeter.

Unlike join indices for a relational model, this extension optimises join operations for multiple criteria (topological and metric). Indeed, objects will match if *spatRel* fulfils certain criteria. This corresponds to a basic relational selection on the SJI table.

2.3 Spatial Decision Tree

A decision tree is an hierarchical knowledge structure that corresponds to a sequence of decision rules. This method aims to determine which attributes (called explanatory) or which criteria of these attributes provide the best distribution of the actual dataset regardless of the given attribute values (called classes). The tree is built recursively by testing and applying subdivision criteria on a training dataset. The test for criteria is based on statistical computation of entropy or information gain. Subdivision criteria are determined at attribute level in the ID3 (Quinlan 86) method while they operate on attribute values in the CART method (Breiman 84). The decision rule sequences are composed of criteria of tree paths starting from the root to the leaves. The main advantage of this technique is its simplicity for decision-makers and people who are not well versed in the complexities of the data analysis domain. It may be less powerful, however, in terms of quality of prediction, than some of the more complex tools such as neural networks.

As emphasised above, unlike conventional decision trees, a spatial decision tree uses data from several tables. One approach consists in using predicate logic instead of attribute values. However, this approach requires that all relational data be transformed into a predicate set. Recently, a new field called relational data mining has been developed. It addresses notably the extension of decision trees for multiple relational tables (Knobbe 99). This more recent method, however, does not solve the problem of spatial relationship determination.

Ester et al. (Ester 97) proposes an algorithm dealing with spatial databases based on ID3. They use the concept of a neighbourhood graph to represent the spatial relationships. This algorithm considers the properties of neighbouring objects in addition to those of the actual object. In the traffic accident example, each object could have many neighbours (e.g., an accident could be near a school and a bus stop). As a result, spatial criteria are not sufficiently discriminating and thus the segmentation may be incorrect. Moreover, this method is limited to only one given relationship. Finally, it does not support the concept of thematic layers

which is an essential component in geographical databases. An additional classification method has been proposed in (Koperski 98). In this case, data are first generalised, then all "attribute = value" are transformed into logic predicates. Such transformations are computationally costly and are limited to a few spatial relationships. In previous work, a two-step solution was implemented to address some of the above shortcomings (Zeitouni 01). The first step computes the spatial join between the target object collection and other themes, while the second step builds a conventional decision tree on the join result. Since spatial criteria are a many-to-many relationship, join operations could result in some target objects being duplicated and give rise to being classified into incorrect classes. As in (Ester 97), the results were shown to be problematic. .

3 The Proposed Method

The proposed classification algorithm is based on two ideas. The first is the utilisation of the spatial join index presented in the section 2.2 and the second is the adaptation of relational data mining methods.

Since the spatial join index formalises neighbourhood links within thematic layers and represents them using relational tables, the classification can directly use the relational schema instead of a predicate set. Indeed, the method uses a target table, the join index tables, and neighbour tables describing other theme attributes. The algorithm details are given in section 3.2 below.

This approach is an extension of the CART method (Breiman 84) that we call SCART and includes the concept of Spatial CART. The information gain is computed using the Twoing expression. The difference with CART is that a node may be partitioned according to a criterion resulting from neighbouring objects, which may have a particular spatial relationship with the target objects. To avoid duplications, the right son of a node is defined as the complement of the left son (*right_son = node – left_son*). The originality of our method, regardless of relational decision trees, is to precisely qualify the neighbourhood relationship. Thus, computing the information gain combines the neighbours' attributes and their distance or their topological relationships with target objects.

3.1 The Method Concepts

Information gain: This is a measurement used to split a node in the CART algorithm. This measurement relates the gain of class homogeneity in case the node splits according to a particular criterion – such as (*attribute = value*) or (*attribute < value*). The "best" split would be the one maximising the information gain. A number of formulas exist for information gain such as Gini. The proposed algorithm uses the Twoing indice that is more suitable for multi-class target attributes.

Saturation conditions: These are the criteria under which the node split terminates. Usually the user specifies these criteria.. The node split is stopped when all objects in the node are in the same target attribute class. In this case, the node is referred to as a pure node. It will also stop when no criteria exist that improves the information gain. The other possible criteria may be a minimal occupation of the node, a maximal depth of the tree or a threshold value for the information gain.

Node encoding: Since the decision tree is binary, an encoding technique has been adopted to identify each tree node. The root has a value code of 1. A node code is then defined recursively by:

*left_son_code = 2 * father_code*

and

*right_son_code = 2 * father_code + 1*

Assignment procedure: A decision tree is a progressive partitioning technique. In the splitting process, objects will be assigned to a left or right son. We propose a virtual representation of partitions by dynamic assignment of target objects to a node (a leaf of the tree). The object will be assigned to the node code.

3.2 Algorithm of a Spatial Decision Tree

The following provides details about the algorithm.

Input parameters:

- Target_table: the analysed objects (i.e. the analysed thematic layer),
- Neighbor_table: thematic layer objects (neighbors of analysed objects),
- Spatial_join_index: the join index table,
- Target_attribute: the attribute to predict (i.e. class labels),
- Predictive_attributes: attributes from a target table or neighbour table that could be used to predict the target attribute,
- Saturation_condition: condition under which the split is considered invalid..

Output:

A binary decision tree

Step 1:

Initially, assign all target objects to the root (i.e. to node number 1)

Step 2:

```
Best_gain = 0
For each predictive_attribute
    For each attribute_value
      If the predictive_attribute belongs to the target_table
            Info_gain = compute information gain        -- as in CART
      Else If the predictive_attribute belongs to neighbor_table
            For each spatial relationship spatRel
                  Info_Gain = Compute information gain for the split criterion
```

"exists neighbours (by mean of *spatRel*) having such attribute_value".

If Info_gain > Best_gain

Save the split criterion

-- The retained split criterion is the one maximising the information gain for all predictive attributes.

Step 3:

If the current leaf is not saturated

Perform the node split

Assign its objects to the left son or to the right son accordingly.

Step 4:

Replace the current node by the following according to the code number.

Iterate step 2 and step 3.

--The algorithm will stop when all the leaves of the lower level are saturated.

Fig. 2. SCART Algorithm

SCART is an extension of CART in two ways. First it uses several tables and attributes of a complex relational database. Second, it may combine some attribute values in the split criterion (neighbour predictive attributes and spatial relationships). For simplification, this description is limited here to one neighbour table.

As an example, the target table may be the accident thematic layer; neighbour table may be the building thematic layer; the target attribute may be the accident gravity or the involved category; a predictive attribute may belong to the target table such as speed, or to the neighbour table such as the building category. Note that when the split condition uses an attribute of another thematic layer, the semantics of the partitioning is somewhat different. It means that the existence of neighbouring objects with such neighbourhood relationships, fulfils a condition such as (*attribute compared_to value*) and gives the best information gain along with the best partitioning of the actual node.

4 Implementation and Discussion

This method has been implemented and tested on real data sets for an application in road transport safety. An example of the results is given in Fig. 3. It classifies accidents according to the involved categories (pedestrians, two wheels – bicycles and motorcycles – or others – vehicles –). As shown here, the first split criterion relates the closeness of a "walkway" within a particular distance (100 m). This criterion leads to more pedestrian accident categories. The right son is partitioned again into the left part close to schools where the pedestrian rate increases and the vehicle rate decreases, and conversely for the right son. The third level shows a case where the algorithm chooses a predictive attribute belonging to the target table. The last leaf on the bottom of the tree could be interpreted such as "when

accidents are far from “walkways”, schools and administration, then they involve fewer vulnerable categories such as pedestrians and two wheels.

This implementation was made using the Oracle 8i DBMS and the Java language environment (see Fig. 4). It allows a first validation of the proposed method. More work is required to validate this method at two levels. An operational level of validation is required and needs domain expert input both for the procedures and for prototyping. Additionally, tests need to be extended to other datasets and other geographical areas. Finally, a performance evaluation and optimisation are necessary especially since large volumes of data may effect the behaviour of the algorithm.

Some optimisation techniques have already been implemented such as the direct object reference (ROWID in Oracle). Other techniques have been considered such as reducing the scan of tables by prior implementation of join operations and database schema transformation. Zhe, (98) explores fast joins and will be assessed in a forthcoming study.

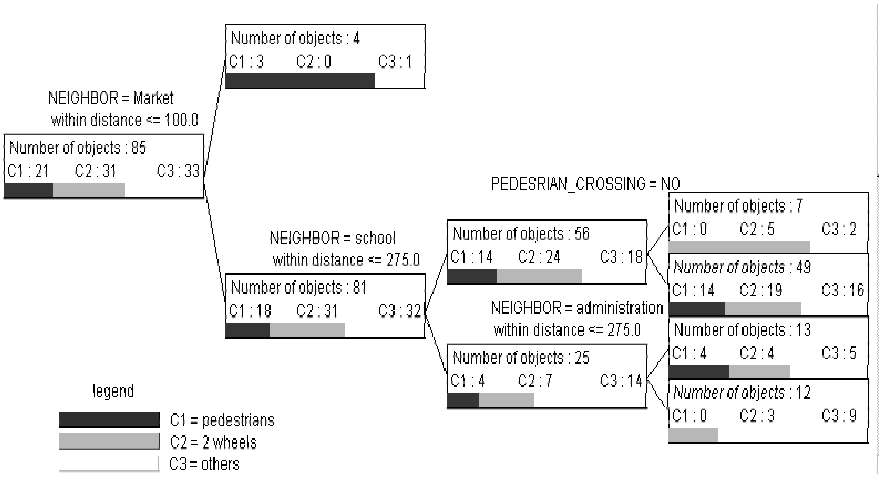

Fig. 3. A spatial decision tree example

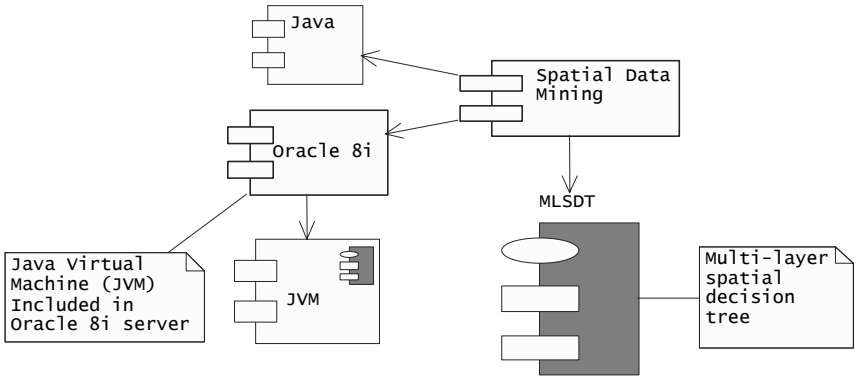

Fig. 4. The software architecture

5 Conclusion and Future Work

This article proposes SCART, a new classification method for spatial data. Two main requirements have been considered in effectively using spatial data in decision trees: multiple layers and automatic filtering, have been briefly assessed and the calculation of neighbourhood relationships have been considered. The intent is to facilitate the use of all spatial relationships through a relational table, and then use and extend the relational data mining methodology.

The spatial join index is merely a correspondence table between two relational tables and in this was corresponds to relational data mining methods. In exploring SCART, it is apparent that it supports the classification of spatial objects according to both their attributes and their neighbours' attributes. It also determines the relevant neighbourhood relationship. Moreover, the organisation of thematic layers has been completely integrated.

From a spatial data mining stand point, the general approach of representation of the spatial relationships as tabular data is very promising. This *a priori* structure could be used in other methods such as spatial clustering or spatial association rules.

Research investigations are required for the algorithmic performance and optimisation, however, other decision tree methods that are disk oriented also require further assessment (Mehta 96), (Gehrke 98). Their application needs to be assessed in an effort to improve the algorithm cost relative to large volumes of data. The second orientation will be the extension to spatio-temporal data and multimedia data that also have complex structures.

References

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Ed: Wadsworth & Brooks. Montery, California

Egenhofer MJ, Sharma J (1993) Topological Relations Between Regions in R2 and Z2. In: 5th International Symposium, SSD'93. Singapore, Springer-Verlag, pp 316-331

Ester M, Kriegel HP, Sander J (1997) Spatial Data Mining: A Database Approach. In: Proceedings of 5th Symposium on Spatial Databases, Berlin, Germany

Fayyad et al (1996) Advances in Knowledge Discovery and Data Mining. AAAI Press / MIT Press

Gehrke J, Ramakrishnan R, Ganti V (1998) RainForest – A Framework for Fast Decision Tree Construction on Large Datasets. In: Proceedings of the 24 the Annual International Conference on Very Large Data Bases (VLDB). New York, pp 416 - 427

Huguenin-Richard F (2000) Approche géographique des accidents de la circulation : proposition de modes opératoires de diagnostic, application au territoire de la métropole lilloise. Ph.D. thesis, Université de Franche-Comté

Knobbe AJ, Siebes A, Wallen V, Daniel MG (1999) Relational Decision Tree Induction. In: Proceedings of PKDD' 99. Prague, Czech Republic

Koperski K, Han J, Stefanovic N (1998) An Efficient Two-Step Method for Classification of Spatial Data. In: Proceedings of International Symposium on Spatial Data Handling (SDH'98). Vancouver, Canada, pp 45-54

Mehta M, Agrawal R, Rissanen J (1996) SLIQ: A Fast Scalable Classifier for Data Mining. In: Proc. of Int. Conf. On Extending Database Technology (EDBT'96). Avignon, France, pp 18-32

Quinlan JR (1986) Induction of Decision Trees. Machine Learning 1: 82 - 106

Valduriez P (1987) Join indices. ACM Transactions on Database Systems 12(2): 218-246

Zeitouni K (2000) A Survey on Spatial Data Mining Methods Databases and Statistics Point of Views. In: IRMA 2000, Information Resources Management Association International Conference, Data Warehousing and Mining. Anchorage, Alaska.

Zeitouni K, Yeh L, Aufaure MA (2000) Join indices as a tool for spatial data mining. In: International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lecture Notes in Artificial Intelligence, no. 2007. Lyon, France, Springer, pp 102-114

Zeitouni K, Chelghoum N (2001) Spatial Decision Tree - Application to Traffic Risk Analysis. In: ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon

Zhe L, Kenneth AR (1998) Fast joins using join indices. The VLDB Journal 8:1-24