S-TILDE: Spatial Top- down Induction Logical DEcision tree

CHELGHOUM Nadjim
PRISMa Laboratory, Claude Bernard University of Lyon I
France
nchelgho@bat710.univ-lyon1.fr

1. Requirements

Unlike conventional data, spatial data are dependent on each other because most spatial phenomena are influenced by the neighbourhood 7, 7. This property is at the origin of the well-known 1st law in geography: "*Everything is related to everything else but nearby things are more related than distant things*" 7. For instance, the shellfish contamination in a lagoon is influenced by the neighbouring agriculture fields. Analyzing spatial data without considering this property is definitely incorrect 7.

Spatial data mining consists to mine spatial data 7, 7. According to the above requirement, it should consider the interaction between spatial objects. This comes down to consider the spatial relationships and properties of neighboring objects as potentially explanatory for the analyzed phenomena. Meanwhile, in shellfish contamination analysis, one can explain and predict the contaminated sampling points by the properties of their nearest agricultural watersheds.

Formally, spatial data mining should take into consideration not only the properties of analyzed objects but also the properties of neighboring objects and spatial relationships witch link the analyzed object and the neighboring object. Or, this requirement raises two technical problems. The first is the maladjustment of the existing methods. The second is the complexity of the spatial relationship computing.

1.1 Maladjustment of the traditional methods

Right now, data analysis in geography has been essentially based on traditional statistics and multidimensional data analysis 7, 7 and does not take into account spatial property. This analysis is performed by diverse methods, from the most basic in statistics (average, variance, histograms, etc.), to multivariate analysis, more exploratory and based on the factorial analysis, passing by correlation and regression analysis. All those methods apply to quantitative or qualitative data but not to spatial data. As they consider the individuals as independent from each other, the important feature of spatial auto-correlation is ignored.

Some Geographical Information System (GIS) 7 or statistic tools, however, include geostatistic and spatial statistic functions. This is provided notably in Splus Spatial Stat of MathSoft 7. Other tools like Spatial Analyst for ArcView of ESRI allow specialized analysis, i.e. mapping the statistic analysis results. This is not sufficient for real spatial analysis.

Furthermore, spatial database queries constitute another way to spatial analysis. Those queries can use spatial relationship predicates. For instance, the user can query the database for the countries having more than 10000 inhabitants where the accident rate is more than the average rate. One

disadvantage of this kind of analysis is that it is much confirmatory than exploratory. Another is the lack, in database systems, of advanced statistic computations and spatial statistic models -such as Moran and Geary indices. Nevertheless, this approach could serve in the filtering phase of the knowledge discovery process, i.e. when preparing the dataset on which the analysis will focus.

1.2 Spatial relationship complexity

Spatial relationships specify how some object is located in space in relation to some reference object 7. They present relationships linking at least two spatial objects and having spatial semantics explicitly defined. For example, a spatial relationship RoadCrossing may link objects of type Road and have for semantics the crossing topological relationship.

Spatial relationships are used intensively by spatial data mining because they have a great importance in the spatial analysis. They may be in different types: metric, topologic or directional and they exist explicitly or implicitly in a spatial data base. In the implicit case, in order to be exhibited, handled and integrated in spatial data mining process, they are defined using graphs, binary or weighted matrixes or they are materialized into traditional data input columns.

The problem is tat these spatial relationships are complex 7 and computing them requires many spatial join operations with complex and expensive geometric computing. So, it is necessary to optimize their handling and their uses. Otherwise, these spatial relationships are multiple (distance, inclusion…). Therefore, the choice of the relevant spatial relationship is difficult to done. The existing spatial data mining methods 7, 7, 7, 7 are limited to some relation chosen by the expert knowing the application domain. This choice becomes difficult when the potentially interesting relations are multiple. So, it is necessary to find and to propose news methods that permit to choose automatically this relevant spatial relationship.

The remainder of this paper is organized as follows . Section 2 gives some preliminary definitions. Section 3 presents the proposed approach. Section 4 describes the application of our approach to the spatial decision tree. The experimentations and the obtained results in shellfish contamination analysis in Thau lagoon (south France) are presented in Section 5, followed by a discussion and a conclusion.

2. Background

In this section, we will present briefly the spatial decision tree, inductive logic programming and TILDE method that we will use in the proposed approach. More details on spatial data mining and spatial relationship are given in 7, 7, 7, 7.

2.1 Spatial decision trees

A decision tree is a hierarchical knowledge structure that corresponds to a sequence of decision rules. This method aims to determine which attributes (called explanatory) or which criteria on these attributes provide the best distribution of the actual dataset regardless to a given attribute values (called classes). The tree is built recursively by testing and applying subdivision criteria on a training dataset. The test of criteria is based on statistical computation of entropy or information gain. There exist diver models of decision trees. Subdivision criteria are determined at attribute level in the ID3 method 7 while they operate on attribute values in CART method 7. 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 aware of data analysis domain. However, it could be less powerful, in term of quality of prediction, than complex tools such neuronal networks.

The extension of decision trees to spatial data consists to consider, not only the properties of the analyzed objects, but also the properties of the neighbouring objects and their spatial relationship. There exist several methods of spatial decision tree 7, 7, 7.

Ester et al. in 7 propose an algorithm dealing with spatial databases based on ID3. They use the concept of neighborhood graph to represent the spatial relationships. This algorithm considers the properties of neighboring objects in addition to those of the actual object. But, each object could have many neighbors (e.g. an accident could be near a school and a bus stop). So, spatial criteria are not discriminative and the segmentation is wrong. Moreover, this method is limited to only one given relationship. Finally, it does not support the concept of thematic layers while this is essential in geographical databases.

Koperski et al. in 7 propose another method that considers the spatial predicates (like the adjacency), the spatial functions (such as the distance) as well as the non spatial values of other objects having a spatial relationship with the actual object (like the population living at a given distance from the stores). The originality of this method is that it automatically determines relevant predicates and functions. The relevance of the distance, in other words, the maximum size of the geographical extensions either is determined by an expert, or computed starting from a given maximum distance and decreasing it in the way to maximize the informational gain. However, this algorithm necessitates transforming all attribute values into predicates, which is fastidious. Another limit is that only one property of neighboring objects is checked (for instance park type in close_to (X, park)). This is why it was not adopted here.

In our previous work 7, we have implemented a two-step method. The first step computes the spatial join between the target object collection and other themes, while the second step build a conventional decision tree on the join result. Since spatial criteria are a many-to-many relationships, join operation could make some target objects duplicated and risks to be classified in different classes. As in Ester et al.'s method 7, the result is incorrect.

2.2 Inductive logic programming

Inductive logic programming (ILP) denomination is due to Muggleton 7. It is a research area at the intersection of machine learning and logical programming. Unlike deductive logical programming that derives consequences from theories, inductive logic programming aims at finding some hypotheses H from a set of observations E. It generalizes from individual instances/observations in the presence of background knowledge, finding regularities/hypotheses about yet unseen instances. It realizes the same task than the traditional data mining. The difference is that data mining operates only on data organized in a unique table with ″attribute=value″ format, while inductive logic programming applies to both data and their relationships. This is allowed by the ability of ILP to handle the input data as well as the extracted models in first order logic -named also predicates logic- 7.

Formally, the inductive logic programming is defined as follows 7:

Input data: Three sets of clauses: B, P and N with

- B: background knowledge expressed with Horn clauses format
- P: positive examples expressed with Horn clauses format,
- N: negative examples expressed with Horn clauses format,

Output data: Find a hypothesis H such that $\forall e \in P$: $H \cup B \models e$ (H is complete) and $\forall e \in N$, $H \cup B$ (non $\models$) e (H is consistent), where $\models$ stands for logical implication or entailment.

So, we look to find a hypothesis H that explains more positive examples and less negative examples. This research is done by inversing the deductive reasoning and often based on propositionnalization[1] 7 or on upgrading data mining methods to deal with data expressed in first order logic 7. The definition of language, alphabet and concepts used in ILP are presented in 7.

2.3 TILDE (Top- down Induction Logical DEcision tree)

TILDE 7 is a decision tree classification method based on the first order logic. It is an upgrade of a well-known data mining method: C4.5 of Quinlan 7. It generates a binary decision tree according to logical decision tree definition.

To build a decision tree, TILDE use the same principle that other classical decision tree techniques: successive applications of subdivision criteria on a learning population in order to access to sub-populations that maximize the number of objects in one class. The premise of the decision rule is the conjunction of literal and the conclusion is disjunction of literal. Because of the fact that the subdivision criteria are based on a simple or derived predicate, TILDE can consider the relations between tables, expert's rules or predicates expressed by conjunction of simple predicates. So, it generates a less deep tree than C4.5. The following figure describes this algorithm.

Input parameters: T: Tree, E: Set of example, B: background knowledge
Procedure Build_Tree (T, E, B, True); // Class is a predicate in E. Initially, T is empty and Q = true
Output parameters: T: binary decision tree

Procedure Build_Tree
Input: N: node, E: Set of examples in the node N, Q: premises of the node N

IF (E is homogeneous) **THEN**

1. K ← Majority_Class; N : leaf (info (E)) ;

ELSE

2. L ← set of the specializations of Q in E
3. Q_b ← The best condition that segments E /* is determined by using heuristic: gain ratio*/
4. Conj ← $Q_b \wedge Q$;
5. E1 = {e ∈ E/ e is true in Conj}; E2 = {e ∈ E/ e is false in Conj};

[1] Transforming the ILP problems to propositional form

6. Build_Tree (left, E1, B, Q_b};
7. Build_Tree (right, E2, B, Q} ;
8. N = node (Conj, left, right) ;

END

Output: Tree T;

Figure 1: TILDE Algorithm

Initially, the tree is empty, the premise Q = true and all observations E are in the root node. We start by verifying if this root node is homogeneous or not. If that is the case, we denote this node as a leaf (saturated node) and we recover all the information concerning this node (line 1). Otherwise, we compute the set of the specializations[2] of the premise Q in E (line 2). Among these specializations, we keep the one that gives the best segmentation of E. This best segmentation is chosen according to a criterion used in C4.5 and based on the gain ratio (line 3). We add this specialization to the premise of Q and we split the father node into two nodes: left son that contains the observations that verifies the condition and right son that contains the observations that don't verify the condition (line 5). We reiterate the tree building procedure on each of the left and right sons (line 6 and 7) and we insert the father node in the tree (line 8). The process stops when all the nodes are saturated.

3. The proposed approach

As emphasized above, spatial data mining uses intensively spatial relationships because they have a great importance in the spatial analysis. These relationships are often implicit and, to be exhibited, they require costly joins on spatial criteria. Zeitouni et al. in 7 proposed to materialize them by using a secondary structure called spatial join index. The idea is to calculate the exact spatial relationship between the locations of two collections of spatial objects and to store it in a table according to the following schema (ID1, spatial-relationship, ID2). We propose to exploit this structure and integrate it in spatial data mining process. In addition that the join via this index is more effective than a spatial join, this relational organization offers us a big advantage: it reduces the spatial data mining to relational data mining 7.

Henceforth, all spatial data mining problem can be reduced to relational data mining problem and the use of spatial relationship becomes possible because they would be considered by the analysis methods like an attribute to analyze as other attributes. So, the choice of the relevant spatial relationship can be done automatically by the analysis methods answering thus the second problem highlighted previously. Nevertheless, this organization cannot be analyzed directly by the conventional data mining methods because these methods consider that the input data is in a unique table and each row in this table is an observation or an individual object to analyze. So, we are faced by a problem related to the fact that we cannot exploit directly the data organized in several tables. It is possible to have one table by joining the different initial tables. However, this operation can duplicate some rows because the observations to analyze are in N-M link with the neighbouring objects (see the Figure 2). This leads to wrong results when we use the conventional data mining methods because of the multiple counting of these observations. For example, rectangle R1 (see the Figure 2) is duplicated as many as the existing neighbouring objects C_i. The

[2] The specialization consists in adding a literal to the premise of a clause or to substitute a variable by one term

same object will be counted several times and risk to be classified in different classes if we apply a classical decision tree algorithm, generating thus non-discriminative rules. The existing works circumvent this problem by generalizing the duplicated data like in 7.

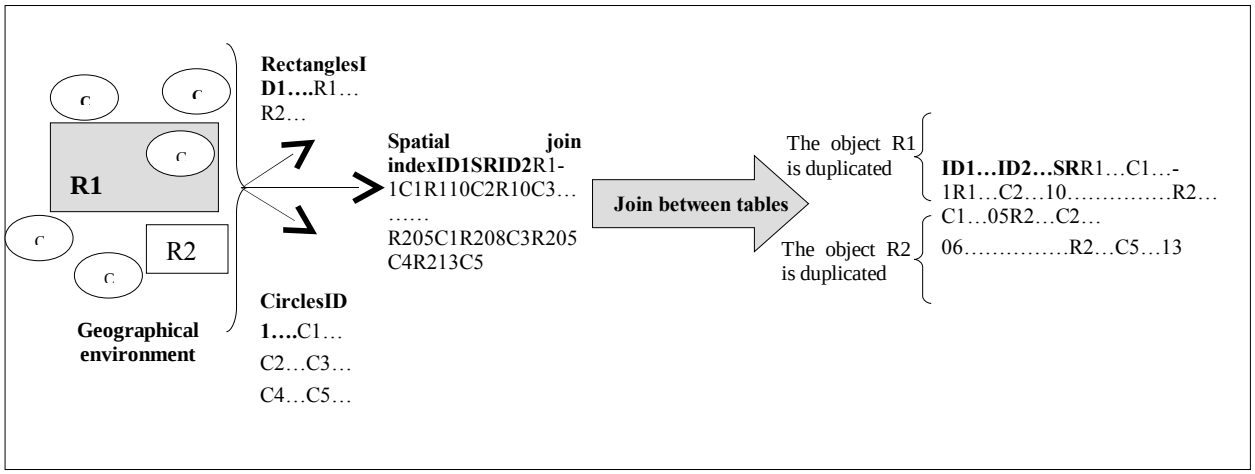

Figure 2: Spatial join index (at the left) and join problem (at the right)

To solve this multi-tables problem, we propose an alternative: we transform multi-tables data into predicates logic and we apply the advanced techniques based on inductive logic programming (ILP) to extract the knowledge. This permits us to benefit from ILP progress in terms of algorithms, simplicity of its models and the possibility to integrate in analysis the implicit knowledge. A panorama of ILP algorithms is presented in 7. The transformation of the relational data in first order logic is made according to the rules given below (see Tableau 1) and described in 7. Figure 3 defines in predicates logic the example given in Figure 2.

This idea of using the inductive logic programming for spatial data mining is also used by Malerba et al. 7, 7 for the extraction of spatial association rules. Their approach consists in adapting Koperski's algorithms 7 to the spatial data expressed in first order logic. The advantage of these works is that they benefit from the expressive power offered by the predicates logic. However, as Koperski's method, they don't explore all spatial relationships and all possible distances because the spatial relationship is limited to predicates evaluated as true or false for a predefined distance. Besides, they generalize all data, which leads to a loss of detailed information.

- Each table T becomes a predicate P
- Each attribute Att of the table T becomes an argument Arg of the predicate P,
- Each tuple (Att_1,…, Att_n) of the table T becomes a fact or a model P (Arg_1,…., Arg_n)

Tableau 1: Data transformation rules into predicates logic

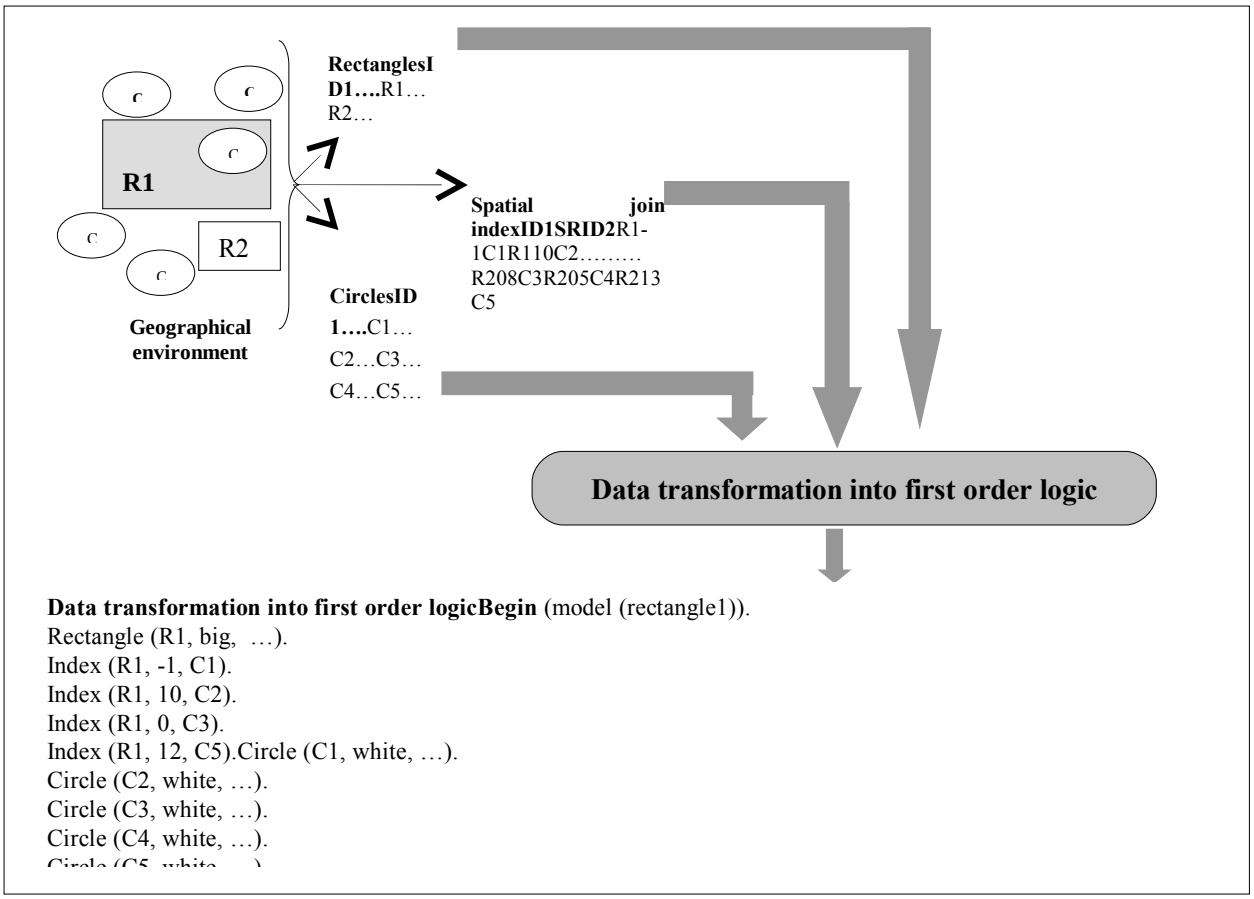

Figure 3: Example of data transformation into predicates logic

4. Application to spatial decision tree

We propose in this article an algorithm, called S-TILDE, to build spatial decision tree. It is based on our proposed approach previously presented. This algorithm is described below. We limit our description to the case of three tables: target table, spatial join index and neighborhood table.

S-TILDE: Spatial Top- down Induction Logical DEcision tree

This method is an extension of TILDE method into spatial data. This extension consists in modifying the division criterion of a node. More precisely, in Spatial TILDE, this criterion integrates the properties of neighbouring objects and their spatial relationship with the object to classify. The combination of these properties and the relevant spatial relationship will be considered to determinate the best partition. The right son is the complement of the left son.

The algorithm takes as input: (i) target table which contains the objects to analyze, (ii) neighbourhood table which contains the neighbors objects, (iii) spatial join index table, (iv) explanatory attributes that may belong to target table or neighbourhood table, (v) target attribute (class) that belong only to the target table and finally (vi) the saturation conditions under which the split is considered useless. To build the tree, the algorithm deals with two steps. The first step consists in transforming data into first order logic (see the step 1 in the Figure 5). The second apply TILDE method adapted to the spatial data (see the step 2 in the Figure 5).

The data transformation step is done according to the rules presented previously in Tableau 1. These rules are general and don't take into account the specificities of such or such method. In the case of our proposed method, these rules are insufficient for three reasons. The first one is that they don't distinguish the class values. In order to remedy this problem, we propose to add the following rule « R1: each class value becomes a predicate with as argument the classified object identifier. » The second reason is that they cannot deal with the node division criterion mentioned above and defined as a combination of neighbourhood relationship and neighbour's properties. To exceed this limit, we define the following rule « R2: we substitute the transformation of the spatial index and the neighbourhood table by the generation of the predicate NEIGHBOURHOOD (Id, spatial relationship, attributes of the neighbourhood) where Id is the identifier of the classified object ». Indeed, if the target table in Spatial TILDE describes the classified objects, what is interesting in the analysis is the type of neighbours rather than the neighbour's identity. Finally, it is necessary to add the domain rules. Here, we know that if an object V_j is at a distance Rel from O_i then it is also at a distance r such that r > Rel. For example, if an object of circular form is 50 cm far from a rectangle then it is also 80 cm far from that rectangle. To take into consideration this characteristic, we suggest to add the following rule « R3: NEIGHBOURHOOD (id, Rel, X, Y,…, Z) $\wedge$ (Rel <r) $\Rightarrow$ NEIGHBOURHOOD (id, r, X, …, Z) ». The table below presents in first order logic the example given in Figure 2.

Transformation data of the previous example into 1st order logic	
Begin (model (rectangle1)).	**Begin** (model (rectangle2)).
Rectangle (R1, …).	Rectangle (R2, , …).
Big (R1) *« derived from R1 »*	Small(R2) *« derived from R1»*
Neighbourhood (R1, -1, …).	Neighbourhood (R2, 05, …).
Neighbourhood (R1, 10, …).	Neighbourhood (R2, 06, …).
Neighbourhood (R1, 0,…).	Neighbourhood (R2, 08,…).
Neighbourhood (R1, 02, …).	Neighbourhood (R2, 05,…).
Neighbourhood (R1, 12, …).	Neighbourhood (R2, 13,…).
« derived from R1 »	*« derived from R1 »*
End	**End**
Neighbourhood (R,T)$\wedge$(R < r) $\Rightarrow$ Neighbourhood (r,T) *« derived from R3»*	

Figure 4: Example of input data expressed in the 1st order logic

Henceforth, the decision tree construction is based on data expressed in 1st order logic and the principle decision tree construction is typically the same that the one used in TILDE algorithms. Initially, the tree contains only one node with all observations and the premise Q = true. We start by verifying if this node is saturated (step 2.1). If it is the case, the development of the tree is stopped. Otherwise, we compute the set of the specializations of the premise Q and their informational gain (step 2.2). We keep the specialization that returns the best value of the informational gain and we split the current node into two nodes: left son and right son. The observations of the current node are assigned to the left son or to the right son according to their segmentation condition. We iterate this process on all nodes until saturation of all nodes.

Input parameters

- Target_table: the analyzed objects
- Neighbor_table: neighbors of analyzed objects,
- Spatial_join_index: the join index table,
- Target_attribute: the attribute to predict (i.e. class labels),

- Predictive_attributes: attributes from target table or neighbor table that could be used to predict the target attribute,
- Saturation_condition: condition under which the split is considered useless.

Output parameters

A binary decision tree

Algorithm

Step 1

Step 1.a

Materializes the joins between Neighbor_table and Spatial_join_index

// Attributes of this table are the predictive attributes of Neighbor_table, Spatial_Relation and the identifier of the analyzed objects

```
CREATE TABLE Neighborhood_Table AS
SELECT I.ID1, V.Predictive_Attributes, I.Spatial_Relation
FROM Neighbor_table V, Spatial_join_index I
WHERE I.ID2 = V.ID2);
```

Step 1.b

Transform data in first order logic according to the following rules

- R1: Traget_Table and Neighborhood_Table becomes predicates P and V,
- R2: Predictives attributes of Traget_Table becomes arguments of predicate P,
- R3: Predictives attributes of Neighborhood_Table becomes argument of predicate V,
- R4: Each class value becomes a predicate. Its argument is the identifier of analysed object.
- R5: Each tuple (a_1, …, a_n) of Traget_Table becomes fact or model P (a_1, a_2, …, a_n),
- R6: Each tuple (b_1, …, b_n) of Neighborhood_Table becomes fact or model V (b_1, b_2, …, b_n),
- R7: Set of observations E contains only predicate derived from R4. The knowledge database B contains other predicates.
- R8: $V(b_1, \ldots, b_n, Rel) \wedge (Rel < r) \Rightarrow V(b_1, \ldots, b_n, r)$

Step 2

// Build decision tree using TILDE method

Build_Tree (N: node, E: Set of examples in the node N, Q: premises of the node N)

Begin

If (E is homogeneous) **Then**

2.1. K ← Majority_Class; N : leaf (info (E)) ;

Else

2.2. L ← set of the specializations of Q in E

2.3. Q_b ← The best condition that segments E /* is determined by using heuristic: gain ratio*/

2.4. Conj ← $Q_b \wedge Q$;

2.5. E1 = {e ∈ E/ e is true in Conj}; E2 = {e ∈ E/ e is false in Conj};

2.6. Build_Tree (left, E1, B, Q_b};

2.7. Build_Tree (right, E2, B, Q} ;

2.8. N = node (Conj, left, right) ;

End

End of algorithm

Figure 5: Spatial CART algorithm using the third alternative

5. Experimentation and results

Our approach has been tested in shellfish contamination analysis in Thau lagoon (south France). Because of its position between lands and sea, the water quality of Thau lagoon has been decreased as a consequence of inputs from agricultural, industrial and urban watersheds affecting the shellfish farming activities. The high anthropogenic pressures to which it is subjected leads regularly to crisis and to chemical or microbiological contaminations of shellfish reducing its economical activities and often destroying the whole of the shellfish livestock. Our objective is to identify and to predict this shellfish contamination knowing the description of the lagoon and its geographical neighborhood. It means to apply a supervision classification by spatial decision tree.

5.1 Tests of spatial decision tree

These tests are based on a real database provided by Ifremer[3], collected in situ and describing the Thau lagoon and its geographical neighborhood. We find information that concern the sampling points in the lagoon such as depth, water temperature, saltiness degree or information that concern the watersheds such as agricultural field, urban field or industrial field. An example of result is given in the Figure 7. It is obtained by using ACE system[4]. The input parameters of this test are summarized in the Figure 6.

INPUT PARAMETER		
Target objects	Sampling points (2128 sampling points)	
Neighboring objects	Watersheds (30 watersheds)	
Explanatory predicates	Neighbourhood (distance, Nature)	
Class	Contaminated	Uncontaminated
saturation criteria	Confidence >= 0.25	

Figure 6: Input parameters of test 1

Spatial decision tree expressed in prolog language
Class (uncontaminated) :— Neighborhood (1788, agricultural watershed)!. (≈53.3%) Class (contaminated) : — Neighborhood (3332, urban watershed)!. (≈56%) Class (contaminated) :—Neighborhood (2137, agricultural watershed)!. (≈50%) Class (uncontaminated). (≈ 52%) **Note:** the rules writing order is important. To determine the class of a new example, we test first the rule 1, and then in failure case, we test the rule 2. In failure case with the rules 1 and 2, we test the Rule 3 and so on.

Figure 7: Spatial decision tree

In this tree, the left son of the root corresponds to the sampling points near the agricultural watershed (distance ≤1788m) affected to the uncontaminated class. The affectation to a class means that this one is more frequent in the node than on the whole of the done sampling. The right son is the complement of the left son. It is segmented on its turn to two nodes: the left one contains sampling points that are near to the urban watershed (distance ≤ 3332m) affected to the

[3] French Research Institute for Exploitation of the Sea (www.ifremer.fr)

[4] http://www.cs.kuleuven.ac.be/~dtai/ACE/

contaminated class and the right one contains sampling points that are far from the urban watershed and so on. The development of the tree stops when all nodes are saturated.

These rules show that the agricultural watersheds are not the reason of the lagoon contamination because in their around there are more uncontaminated sampling points, whereas it is possible that the urban watershed are the reason of this contamination because there are relatively more contaminated sampling points in their around. These new knowledge remain to validate by experts.

5.2 Cartographic visualization

The ultimate aim of this process is to localize on the cards the zones corresponding to the discovered rules. Here, this would permit for example to underline the links between the contaminated sampling points and a perimeter of 3332m around the inlets urban watershed (rule 2 of the Figure 7).

Cards presented below shows the contaminated and uncontaminated sampling points and the around of agricultural watershed (distance≤1788m) and urban watershed (distance≤3332m). In the first card which present the first rule, the around of the agricultural watershed are presented by empty circles. In the second card, the around of the agricultural and urban watershed are presented respectively by full and empty circles. The contaminated and uncontaminated sampling points are presented respectively by black and grey points. According to the computing done by the algorithm, we know that the uncontaminated sampling point proportion is more important around the agricultural watershed (empties circles in the first card) and the contaminated sampling point proportion is more important around the urban watershed (empties circles in the second card). These confirm ours rules (1 and 2) extracted using our algorithm S-TILDE and the link hypothesis between the urban watershed and the contamination must be studied because there are more uncontaminated point around these watersheds. This type of rules would have been difficult to discover visually because of the sampling points superposition- each point on card covers approximately 270 sampling points-.

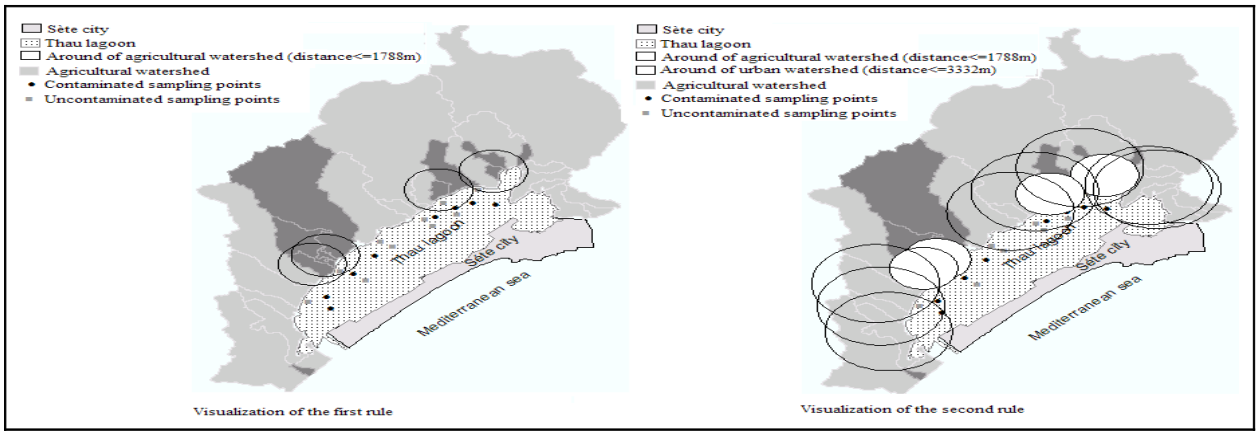

Figure 8: Cartography of the first and second rules

6. Conclusion

The main specificity of the spatial data mining is that it integrates, in the analysis, the spatial relationships. For the implementation of its methods, we have proposed, in this paper, an approach

with two steps. The first one consists to materialize these spatial relationships and to store them in spatial join index reducing spatial data mining problem to the multi-tables data mining problem. The second step proposes original solution to multi-tables data mining problem in spatial data mining setting. It consists to transform data in first order logic and to apply the advanced techniques based on inductive logic programming (ILP) to extract the knowledge. This approach is attractive and very promising. It permits to analyze spatial objects according to both their attributes and their neighbors' attributes and it determine automatically the relevant neighborhood relationship. Moreover, the organization in thematic layers is totally integrated.

The application of this approach to the spatial decision tree method has been described in this paper and an algorithm called S-TILDE has been proposed. The obtained results on shellfish contamination are presented and confirm the efficiency of our approach and our algorithm. Compared to the existing algorithms, S-TILDE offers us many advantages: (it) it guarantees a correct classification (contrary to Ester et al.'s algorithm), (ii) the classification of spatial object takes into account the spatial (spatial relations) and non spatial attributes (iii) it considers, not only the properties of the analyzed objects, but also the properties of the neighbours objects and their spatial relationship, (iv) it choose automatically the relevant spatial relationship, (v) it makes a distinction between themes, (vi) it is not limited to only one neighbour. It is applicable also in the case of several linked neighbours. The table bellow compares our algorithm S-TILDE with the two main methods of spatial decision tree: 7 and 7.

	Ester et al.	Koperski et al.	S-TILDE
It guarantees a correct classification	No	Yes	Yes
It considers the neighbors' properties and spatial relationships	Yes	Yes	Yes
It choose automatically the relevant spatial relationship	Non	No	Yes
Neighborhood degree	= 1	=1	≥ 1
Distinction between themes	Non	Non	Yes

7. References

[1] Anselin L., D.A. Griffith (1988), Do spatial effect really matter in regression analysis? Regional Science Association 65, 11-34.

[2] Anselin L. (1989), What is special about spatial data? Alternative perspectives on spatial data analysis, Technical paper 89-4. Santa Barbara, NCGIA.

[3] Breiman L., J.H. Friedman, R.A. Olshen, C.J. Stone, (1984), Classification and Regression Trees, Ed: Wadsworth & Brooks. Montery, California.

[4] Blockeel H., L. De Raedt, (1998) Top-Down induction of first order logical decision trees, Artificial intelligence, 102(2-2)/ 285-297.

[5] Ceci M., A. Appice, D. Malerba, Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach, in J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004, Lecture Notes in Artificial Intelligence, 3202, 99-111, Springer, Berlin, Germany, 2004.

[6] Chelghoum N., Fouille de données spatiales - Un problème de fouille de données multi-tables, PhD thesis, University of Versailles, 2004 (www.prism.uvsq.fr/~nchelg).

[7] Chelghoum N., K. Zeitouni, A. Boulmakoul, A decision tree for multi-layered spatial data, In 10th International Symposium on Spatial Data Handling (SDH), Edition Springer, 1-10, Ottawa, Canada, , July 8-12 2002.

[8] Cressie N.A.C, Statistics for spatial data, Edition Wiley, New York, 1993.

[9] Dzeroski S., N. Lavrac, Relational Data Mining, Springer, 2001.

[10] Egenhofer M.J., Reasoning about Binary Topological Relations, Proc. 2nd Int. Symp. on Large Spatial Databases, 143-160, Zurich, Switzerland, 1991.

[11] Ester M., H.P. Kriegel, J. Sander, Spatial Data Mining: A Database Approach, In proceedings of 5th Symposium on Spatial Databases, Berlin, Germany, 1997.

[12] Han J., M. Kamber, Data Mining. Concepts and Techniques, Academic Press Ed. 2001.

[13] Koperski K., J. Han, N. Stefanovic, An Efficient Two-Step Method for Classification of Spatial Data, In proceedings of International Symposium on Spatial Data Handling (SDH'98), p. 45-54, Vancouver, Canada, July 1998.

[14] Koperski K., J. Adhikary, J. Han, Knowledge Discovery in Spatial Databases: Progress and Challenges, Proc. SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD). Technical Report 96-08, University of British Columbia, Vancouver, Canada, 1996.

[15] Kramer S., N. Lavrac, P. Flach, Propositionalization approaches to relational data mining, in relational data mining, Dzeroski S., Springer Edition, p- 262- 291, 2001.

[16] Laurini R., D. Thompson, Fundamentals of Spatial Information Systems, Academic Press, London, UK, 3rd printing, 1994.

[17] Lavrac N., S. Dzeroski, Inductive logic programming. Techniques and applications. Edition Ellis Horwood, 3-38, New York, 1994.

[18] Longley P.A., M.F. Goodchild, D.J Maguire, D.W Rhind, Geographical Information Systems, Principles and Technical Issues, John Wiley & Sons, Inc., 2nd Edition, 1999.

[19] Malerba D., F.A. Lisi, An ILP Method for Spatial Association Rule Mining. In A. Knobbe and D. van der Wallen (Eds.), Notes of the ECML/PKDD 2001 Workshop on Multi-Relational Data Mining, 18-29, Germany Freiburg, 2001.

[20] Mathsoft Inc., "S-Plus for ArcView GIS - Users Guide Version 1.0" and "S-Plus Spatial Stat.", Data Analysis Products Division, Seattle, Washington, April 1998.

[21] Munro R., Chawla S., Sun P., Complex Spatial Relationships, Technical Reports, TR-539, IEEE ICDM'03.

[22] Muggleton S., Inductive Logic Programming, New Generation Computing Edition, 295-318, 1991.

[23] Quinlan J.R., Induction of Decision Trees, Machine Learning (1), 82 - 106, 1986.

[24] Quinlan J.R., C4.5: Programs for machine learning, Morgan Kaufmann, 1993.

[25] Shekhar S., C. Sanjay (2003), Spatial Databases: A Tour, Prentice Hall.

[26] Shaw G., D. Wheeler, Statistical Techniques in Geographical Analysis, Edition David Fulton, London, 1994.

[27] Tobler W.R., Cellular geography, In Gale S, Olsson G, In Phylosophy in Geography Edition, Dortrecht, Reidel, 379-86, 1979.

[28] Van Laer W., L. De Raedt, How to Upgrade Propositional Learners to First Order Logic: a Case Study, 2000.

[29] Zeitouni K., "A Survey on Spatial Data Mining Methods - Databases and Statistics Point of Views", Book Chapter In “Data Warehousing and Web Engineering”, Shirley Becker Editor, IRM Press, pp 229-242, 2002.

[30] Zeitouni K., L. Yeh, M.A. Aufaure, Join indices as a tool for spatial data mining, Int. Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, In Artificial Intelligence n° 2007, Springer, 102-114, Lyon, France, September 12-16, 2000.