International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:12, 2014

Spatial Data Mining by Decision Trees

S. Oujdi, H. Belbachir

***Abstract*—**Existing methods of data mining cannot be applied on spatial data because they require spatial specificity consideration, as spatial relationships.

This paper focuses on the classification with decision trees, which are one of the data mining techniques. We propose an extension of the C4.5 algorithm for spatial data, based on two different approaches *Join materialization* and *Querying on the fly the different tables.*

Similar works have been done on these two main approaches, the first - Join materialization - favors the processing time in spite of memory space, whereas the second - Querying on the fly different tables- promotes memory space despite of the processing time.

The modified C4.5 algorithm requires three entries tables: a target table, a neighbor table, and a spatial index join that contains the possible spatial relationship among the objects in the target table and those in the neighbor table. Thus, the proposed algorithms are applied to a spatial data pattern in the accidentology domain.

A comparative study of our approach with other works of classification by spatial decision trees will be detailed.

***Keywords*—**C4.5 Algorithm, Decision trees, S-CART, Spatial data mining.

I. Introduction

REFERENCE [9] shows that 80% of an organization's data are spatial and can be localized. They follow the first law of geography introducing the notion of neighborhood that connects this data to each other. It is this feature that distinguishes them from traditional data, but it also leads to the fact that classical data mining techniques become obsolete hence the need for adaptation to this type of data. This is where the spatial data mining has emerged, combining several approaches to the confluence of several fields such as geographic information systems, statistics, spatial analysis, databases and classical data mining.

Spatial data mining can be considered as a multi-relational data mining, where each table represents a category of spatial data representing in its turn a certain phenomenon that is called a thematic layer. The difference is that the spatial data mining takes into account the spatial relationships between the data, which makes it a full-fledged field of data mining. This has attracted the interest of many researchers who have proposed various techniques divided into two categories: based on monothematic approaches such as segmentation techniques, and Co-location, as well as techniques based on multi-thematic approaches such as association rules and decision trees.

Among the techniques of data mining, we are interested in decisions trees with C4.5 algorithm that we adapted to the spatial data mining, by implementing it with two different approaches. The first is to make a join between tables when necessary, whereas the second materializes all joins once and for all.

This article includes four sections. The first is the introduction. The second section outlines the main work in the field of spatial data mining. The third section presents the modified algorithm C4.5 with two approaches: Querying on the fly the different tables and the Join materialization. The final section presents the results of experiments of this algorithm on a sample of spatial data in the domain of road accidents.

S. Oujdi and H. Belbachir are with the Faculty of Mathematics and Computer-science, Department of Computing, LSSD Laboratory, Science and Technology University of Oran, Algeria (e-mail: sihem.oujdi@univ-usto.dz, h_belbach@yahoo.fr).

II. Related Works

The spatial data mining's techniques are used in different domains: geographic information systems, geoscience, meteorology, medicine and epidemiology, economics… etc.

We can distinguish two types of approaches in spatial data mining: monothematic approaches and multi-thematic approaches.

A. Monothematic Approaches

They are often related to statistics and data analysis. The idea is to incorporate a contiguity parameter into the model or to weight the variables by the values of the neighborhood. This is feasible, because in the case of a single theme, the data are described with the same variables and are comparable.

Since we will focus on the multi-thematic approaches, we will just mention monothematic ones.

1. Analysis of Localizations without Attributes

Among the monothematic approaches, some ones are only based on localizations. They tend to explore a set of locations (points set) to reveal trends or concentrations. Among major works, we have: trend analysis by the method of density [10], and Clustering [8].

2. Analysis of Localizations Provided with Numerical Measures

This category focuses on measurements taken on a spatial domain, often covering space by a surface cutting. It is frequently a single numeric attribute. The analysis aims to characterize the spatial variation of this or these measures. Among major works, we have: the overall and local spatial autocorrelation [6] and trend analysis by linear regression [14].

3. Analysis of Localizations with Provided Categories

Localizations are assumed to be described by categorical attributes. The analysis focuses on the simultaneous presence of categories in space or on the characteristic properties extending neighborhood. Among major works, we have: Co-

localization [6], [11], [13], and characterization [7].

B. Multi-Thematic Approaches

The multi-thematic data mining methods are generally based on spatial predicates interpreted as properties to be considered in the model to induce. To do this, these methods distinguish a target theme of analysis and explore other themes (or phenomenons) that may influence it. There are several methods of multi-thematic approaches; the most common are the association rules and supervised classification. In our works, we are interested in the classification by decision trees.

The works presented in [1]-[5], and described in [15] are interested by classification of spatial data by decisions trees. These methods have been implemented in the context of the risk analysis of road accidents. Three alternatives were explored that we present in the following.

1. Approach 1 - Querying on the Fly Different Tables

It consists of taking as input three tables: table of objects to be analyzed, neighborhood objects table and spatial join index table. Whenever the attribute to be analyzed is an attribute of neighborhood, the algorithm uses a double join between the target table, the spatial join index table and the neighbors table; it is here where modification is to be made to existing algorithms. Otherwise, we apply the classical algorithm without modification. As advantages, the data can be used without prior treatment and no extra memory space is required. As disadvantage, execution time degrades quickly with the increase of data volume.

2. Approach 2 - Join Materialization

It consists in materializing joins between the three tables of analysis: target table, the neighbors table and join index once and for all to avoid recalculating them each time. The join leads to duplication of analysis objects; the method of data mining has been modified on the result of the join in order to take account of this duplication. As advantage, execution time is improved compared to the first approach. As disadvantage, there is high consumption of memory space.

3. Approach 3 - Reorganization of Data in a Single Table

This approach consists of using the COMPLETE operator, whose role is to join different input tables of analysis process into a single table without duplication of objects, the idea is to complement and not join the target table by the other two tables. Its principle is to generate for each attribute value of the linked table an attribute in the result table. As advantage, data are brought to a single table without duplication. As disadvantage, there is loss of information.

These three approaches were applied to the same dataset treating the accidentology domain. The obtained results are shown below in Fig. 1.

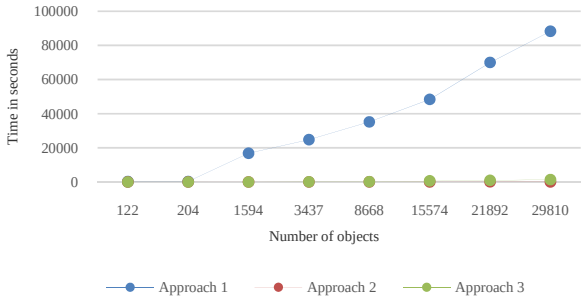

Fig. 1 Execution time according to the size of the target table for the three algorithms

Approach 2- join materialization - is better in terms of execution time compared to the other two approaches.

4. Extended ID3 Decision Tree Algorithm for Spatial Data

The work presented in [12] proposes a new algorithm for spatial decisions trees based on the ID3 algorithm for discrete characteristics represented with points, lines and polygons. The proposed ID3 algorithm uses the informational gain for the selection of attributes; the proposed algorithm uses the spatial informational gain to choose the best division layer from a set of explanatory layers. The new formula of spatial informational gain is proposed using spatial measurements for the points, lines and polygons characteristics.

III. Proposed Approaches

Based on decision trees works that we have just presented, we have proposed an approach that overcomes the following observed limitations:

1) S-CART: Generates binary trees evolving in depth only.
2) ID3: Do not support the continuous data, often found, especially in spatial databases (Measurements, meteorological data...).

To go beyond these limits, we have chosen C4.5 algorithm that supports a maximum data types, and even the management of missing data, which improves this point compared to ID3 algorithm, and by the same occasion, generating N-ary trees, unlike the S-CART algorithm.

Obviously, an adaptation of C4.5 algorithm to this type of data should be done, for this, we have adapted this algorithm to spatial data mining using two different approaches based on the works of [15] that we described in the related works: Approach 1 - Querying on the fly different tables, Approach 2 - join materialization.

The mainly modification of the C4.5 algorithm lies in the calculation of the informational gain, given that the spatial data mining differs from the classical one in the fact that the data is distributed over several tables. In addition to multi-table data mining, it must be taken into account the spatial relationships that are precalculated in a spatial join index.

A. Approach 1: Querying on the Fly Different Tables

In terms of adaptation of the C4.5 algorithm with - Querying on the fly different tables- approach, we take the same steps as for other works with S-CART in [15] and ID3 in

[12]. Here is the adapted C4.5 algorithm to spatial data mining according to the -querying on the fly- approach.

The algorithm takes as input three tables: table of objects to be analyzed (target table), table of objects of the neighborhood (neighbors table) and the spatial join index, for which it makes a join when it is necessary (When calculating informational gain concerns an attribute that does not belong to the target table). The general algorithm is described in Fig. 2.

Input
Target table: table of objects to be analyzed.
Neighbors table: table of objects of the neighborhood.
The spatial join index: contains the possible spatial relationship among the objects in the target table and those in the neighbors table.
Explanatory variables: Belong to Target table or neighbors table.
Class: belong necessarily to the target table.
Stop condition: Condition that stops the development of the tree.

Progress
Progressively in the classification, the tuples of the target table will be attributed to a current sheet of the tree. Initially, all tuples are assigned to the root node.
1) For each explanatory attribute, calculate the best informational gain. At this level, we adapt the formula of informational gain when the attribute comes from neighbors table.
   If the explanatory attribute belongs to the target table then the formula is identical to that classical C4.5 algorithm, otherwise make a join to calculate the gain.
2) If the current sheet is not saturated, assign the objects of the current sheet to sons if they satisfy the condition of segmentation.
3) Iterate step 2 to the next node if it exists. Otherwise, the algorithm stops.

Output
Decision Tree.

Fig. 2 General algorithm - Querying on the fly different tables

Input
Same as the first algorithm in Fig. 2

Progress
1) Join materialization.
2) Applying C4.5 algorithm:
   Initialization: node = 1.
   For each explanatory attribute E do
   For each value of E.Val of the attribute E do
   Gain_Info_Val = informational Gain using the formula of classic C4.5 algorithm.
   Gain_Info_E = the best informational gain for the attribute E.
   Best_Gain_Info = the best informational gain of all explanatory attributes.
   Save segmentation's criteria corresponding to the Best_Gain_Info.
3) If the current node is not saturated then
   The current node is divided into multiple nodes where each node corresponds to a modality best E attribute and we assign analysis objects to sons according whether they satisfy or not the criteria of segmentation.
4) Iterate steps 3 and 4 on the next node.

The algorithm stops when all nodes are saturated.

Output
Decision tree.

Fig. 3 General - join Materialization - algorithm

B. Approach 2: Join Materialization

As to the second approach, which is -join materialization-, the algorithm takes as input the same tables described in the previous approach. Namely: table of objects to be analyzed (target table), table of objects of the neighborhood (neighbors table) and the spatial join index, for which it stores the result of join materialization in a single table, thereby returning the spatial data mining to the classic data mining. The general algorithm is described in Fig. 3.

IV. Experiments and Results

A. Experimentations Environment

We applied our approaches under Windows environment, with Oracle 10g as database management system. The tests were performed on a machine with 3.00 GHz processor and 4 GB of memory.

B. Dataset Presentation

For the purposes of our experiments, we have used a database representing the list of schools and accidents in the state of Illinois in the United States of America. The dataset includes 579 entries for schools and 971 entries for accidents. The precalculated join index from this data contains 562209 entries. For these tests, it was considered a single spatial relationship, namely the distance.

C. Study Purpose

The study goal on this dataset is to link types of accidents to nearby schools, which could help to take decisions in order to predict and to be able to intervene efficiently in these locations by having a certain prediction about the accidents types that we can have by location.

D. Results

We have established a test plan as follows: For purposes of comparison with other works, we have tested the adapted C4.5 algorithm with two different approaches on dataset by comparing it with S-CART.

Performance comparisons concerns: necessary execution time and consumed memory space during treatments.

TABLE I
Obtained Results

Algorithm	Approach 1 : Querying on the fly different tables		Approach 2 : Join materialization			
	Duration	Memory space (Mb)	Duration			Memory space (Mb)
			Total	Step 1	Step 2	
S-C4.5	18min 32.520s	263	13min 2.311s	6.358s	12min 55.953s	1014
S-CART	24min 17.435s	220	17min 10.202s	6.340s	17min 3.862s	1029

E. Discussion

We can clearly observe that the second approach - join materialization - requires less execution time (with a performance gain of about 28%) compared to the first approach - Querying on the fly different tables -. This is due to the join of all the tables is made once and for all before the treatment, which has as effect to reduce the necessary calculating time. However, this performance gain of the second approach - join materialization - is done despite a larger memory space (approximately 4.6x times in our test) compared to the first approach - Querying on the fly different

tables -. Which justify the necessary memory space needed to store the join of all tables, while the first approach consumes memory space depending on need when the join is necessary.

Comparing with S-CART algorithm, and for the first approach - Querying on the fly-, with a slightly higher memory consumption of our algorithm comparing to S-CART, we get a better execution time. This is due to higher performance of C4.5.

For the second approach - join Materialization -, we find that with same memory space consumption of the two algorithms, the execution time of our one is better than S-CART. It is also due to higher performance of C4.5.

V. Conclusion

We presented our work with decision trees in which we implemented and adapted C4.5 algorithm to spatial data with two different approaches: - Querying on the fly different tables - and - Join materialization –. Similar works have been released around these two major approaches. The first, - Querying on the fly different tables- favoring memory space despite the processing time, while the second one - join materialization - favors the processing time in spite of the memory space.

Concerning decision trees, works have been implemented with S-CART and ID3 algorithms, while we opted for C4.5 algorithm, which allows overcoming certain limits compared to previous works.

In perspective, we orient our research to support multiple spatial relationships, and make tests on larger data, as a picture or a spatio-temporal data types.

References

[1] Chelghoum N, Zeitouni K, "Datamining spatial un problème de datamining multi-tables", Prism.France, Université de Versailles, 2004, Vol.14 N°02, pp.129-145.

[2] Chelghoum N, Zeitouni K., "Extension du projet TOPASE par la prise en compte des interactions entre le réseau viaire et l'environnement urbain", Convention PRISM-CERTU, Juillet 2004.

[3] Chelghoum N., Zeitouni K, Laugier T., Fiandrino A., Loubersac L., "Fouille de données spatiales - Approche basée sur la programmation logique inductive", EGC 2006, Edition CEPADUES, Lille,Janvier 2006.

[4] Chelghoum N., Zeitouni K. "Mise en œuvre des méthodes de fouille de données spatiales : Alternatives et performances", EGC 2004, Clermont-Ferrand, January 2004.

[5] Chelghoum N., Zeitouni K., Boulmakoul A., "Fouille de données spatiales par arbre de décision multi-thèmes", EGC 2002, Montpellier, January 2002.

[6] Cliff A.D., Ord J.K., "Spatial autocorrelation", Pion, London, 1973.

[7] Ester M., Frommelt A., Kriegel H.-P., Sander J., "Algorithms for Characterization and Trend Detection in Spatial Databases", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, NY, 1998.

[8] Ester M., Kriegel H.P., Sander J., Xu X., "A density-Based algorithm for discovering clusters in lager spatial databases with noise", In proceeding of second international conference on knowledge discovery and data mining, Portland, 1996, pp 226-231.

[9] Franklin, C, "An introduction to geographic information systems: linking maps to databases. Database", vol. 15, no. 2, pp.13--21, 1992.

[10] Gatrell A., Bailey T., Diggle P., Rowlingson B., "Spatial point pattern analysis and its application in geographical epidemiology", Transactions of the Institute of British Geographers, n° 21, 1996, pp. 256-274.

[11] G.Manikandan et al, "Mining of spatial co-location pattern implementation by fp growth", European Journal of Scientific Research ISSN 1450-216X Vol.68 No.3 (2012), pp. 352-366.

[12] ImasSukaesihSitanggang, RazaliYaakob, Norwati Mustapha, Ahmad Ainuddin B Nuruddin, "An Extended ID3 Decision Tree Algorithm for Spatial Data", Published in: Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on, June 29 2011 - July 1 2011, pages: 48 – 53.

[13] Shekhar Sh. and Huang Y., "Discovering Spatial Co-location Patterns: A Summary of Results", 7th Int. Symposium on Spatial and Temporal Databases (SSTD), Springer-Verlag, Lecture Notes in Computer Science, July 2001.

[14] R. Marghoubi, A. Boulmakoul, K. Zeitouni, "Utilisation des treillis de Galois pour l'extraction et la visualisation des règles d'association spatiales", Faculty of sciences and technology of Mohamadia (FTSM).

[15] Zeitouni Karine. "Mémoire d'habilitation à diriger des recherches : Analyse et extraction de connaissances des bases de données spatiotemporelles". Informatique. France : University of Versailles Saint-Quentin-en-Yvelines, December 2006.

S. Oujdi, 23- 08- 1988, Algeria. Master in computer science in 2011. Phd student since 2011 at the University of Science and Technology of Oran, Algeria. Interest in spatial data mining, member of LSSD Laboratory.

H. Belbachir, Faculty of Mathematics and Computing, Computing Department, University of Science and Technology of Oran, Algeria. PhD in computer science since 1990. Interest in advanced databases, data mining and data grid. Prof. Belbachir is co-director of Signal, Systems and Data Laboratory (LSSD). Head of the database System Group.