Approach for Spatial Database Mining

Prof. Dr. Ala'a H. AL-Hamami

Amman Arab University, Jordan

Assest Prof. Dr. Soukaena Hassan

Dr. Mazin Sameer Al-Hakeem

University of Technology, Computer Sciences

Abstract:

Most of the previous spatial mining works are depend on strategy of organizing the huge spatial data in a suitable data structure and usually the data organized as R-Tree. The data mining algorithms then applied on each level of R-Tree. This method causes time consuming and takes huge storage area and leads to inadequate results. The proposed approach suggests the following strategy for efficient spatial mining. It collects all the spatial data and organizes it (according to normalization and generalization) to a flat data base. After that the following steps will be executed: build the proposed spatial database, apply mining algorithms on the proposed Structure of the spatial data to extract the association rules, clusters and classes. Finally analyzes the resulted patterns from the mining algorithms.

Keywords

Association rules, cluster, classes, data mining, spatial mining and spatial data mining.

1. Introduction

Spatial data mining is a special kind of data mining. The main difference between usual data mining and spatial data mining is that: in spatial data mining tasks we use not only non-spatial attributes (as it is usual in data mining in non-spatial data), but also spatial attributes. Spatial data mining methods and techniques have been proposed for the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases.

With wide applications of remote sensing technology and automatic data collection tools tremendous amounts of spatial and no spatial data have been collected and stored in large spatial databases. Traditional data organization and retrieval tools can only handle the storage and retrieval of explicitly stored data. The extraction and comprehension of the knowledge implied by the huge amount of spatial data though highly desirable pose great challenges to currently available spatial database technologies This situation demands new technologies for knowledge discovery in large spatial databases or spatial data mining that is extraction of implicit knowledge spatial relations or other patterns not explicitly stored in spatial databases. Recently there have been a lot of research activities on knowledge discovery in large databases data mining [5-7].

2. Related Works

Here we will introduce samples of previous spatial mining works. This sample has three good researches in spatial mining, the first one is a proposed method for mining spatial database using association rule, the second proposes method for mining spatial database using clustering, and the third proposed method for mining spatial database using classification.

2.1 Malerba and etal [1] proposed a method for the discovery of spatial association rules, that is, association rules involving spatial relations among (spatial) objects. The method is based on a multi-relational data mining approach and takes advantages of the representation and reasoning techniques developed in the field of Inductive Logic Programming (ILP) and using R-Tree as index structure for spatial data. Here an example of spatial association rule that can be generated:

is_a(X,large_town) ^intersects(X,Y) ^is_a(Y,road) → intersects(X,Z) ^is_a(Z, road) ^Z≠ Y (91%, 85%).

This provides more insight into the nature of the task relevant objects Y and Z, according to the spatial hierarchy reported in Fig. 1. It is noteworthy that the support and the confidence of the last rule changed. Generally, the lower the granularity level, the lower the support of association rules. Therefore, they follow Han proposal [2] to use different thresholds of support and confidence for different granularity levels.

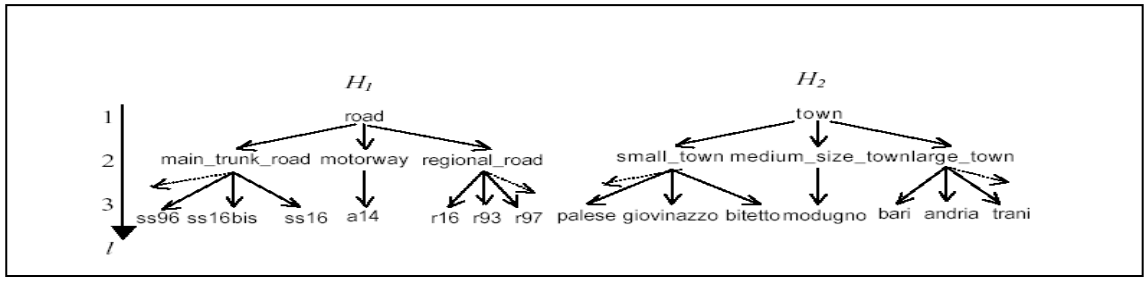

Figure (1): spatial hierarchy.

2.2 Ester M., and etal [3] proposal is to select a relatively small number of representatives from the database and to apply the clustering algorithm only to these representatives. Their method called *focusing on representatives* which makes use of a spatial index structure, e.g. an R-tree. From each data node of an R-tree, see figure (2), one or several representatives are selected. Clearly, the clustering strategy of the R-tree, which minimizes the overlap between directory rectangles, yields a well- distributed set of representatives.

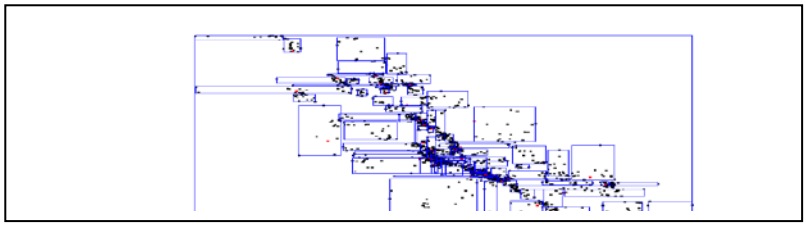

Figure (2): using R-Tree in spatial mining (clustering).

2.3 Ester and etal [4] extends the ID3 algorithm for spatial databases, see figure (3). It does not only consider the attributes of the object O to be classified but also considers the attributes of neighboring objects. Since the influence of neighboring objects and their attributes decreases with increasing distance, the length of the relevant neighborhood paths is limited by the input parameter max-length. For each path in paths, the attr of the index-th object of path and class_attr of the first object of path (i.e.

One of the objects to be classified) are considered for the calculation of the information gain.

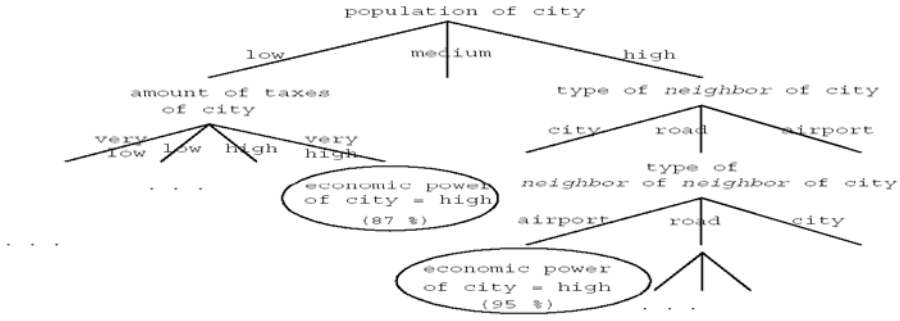

Figure C 5.8.6: Spatial decision tree

Figure (3): spatial mining using decision tree (classification).

3. The proposed system

To explain the proposed system, we will introduce it in details in the following steps:

First step: Suppose we have any of the following area of earth; see figure (4), and the GIS for it. We want to extract both explicit and implicit relations and patterns among spatial objects which are represent the desired area. So we must get the spatial and nonspatial attributes for all spatial objects.

Figure (4): area of earth

Second step: The following non spatial attributes (type, size, population, employment rate, etc. …..) are presented in the spatial database.

Third step: The following spatial attributes are presented in the spatial database:

- The first attribute will represent the spatial objects (O1, O2, O3, ……, On), this attribute represents the identification of the transaction.
- The second attribute will represent the type of the spatial object such as (town, road, ….). In the proposed spatial database we will represent these spatial attributes as in the following codes:

(Code of the type of the spatial object)

1= town, 2 = road, 3 = river, 4 = sea, 5 = lake, 6 = mine, 7 = forest, 8 = bridge, 9 = highway,….etc.

- The third attribute will represent the size of the spatial object. In the proposed spatial database these spatial attributes will have the following codes:

(Code of the size of the spatial object)

1 = large, 2 = medium, 3=small.

- The fourth attribute will represent the shape of the spatial object. These spatial attributes will have the following codes:

(Code of the shape of the spatial object)

1 = point, 2 = line, 3=polygon.

- The fifth attribute will represent the directions state: north of, south of, east of, west of, north west of, north east of, south west of, south east of. For more explanation see figure (1). These spatial attributes will have the following codes:

(Code of the direction, Code of the related spatial object)

(A, Oi) = (north of, Oi)
(B, Oi) = (south of, Oi)
(C, Oi) = (east of, Oi)
(D, Oi) = (west of, Oi)
(E, Oi) = (north east of, Oi)
(F, Oi) = (north west of, Oi)
(G, Oi) = (south east of, Oi)
(H, Oi) = (south west of, Oi)

- The sixth attribute will represent the Position state: disjoint, overlap, meet, covers, covered by. For more explanation see figure (5). These spatial attributes will have the following codes:

(Code of the direction, Code of the related spatial object)

(I, Oi) = (overlap, Oi)
(II, Oi) = (meet, Oi)
(IIIC, Oi) = (covers, Oi)
(IV, Oi) = (covered by, Oi)
(V, Oi) = (disjoint, Oi)

- The seventh attribute will represent the distance between spatial objects.

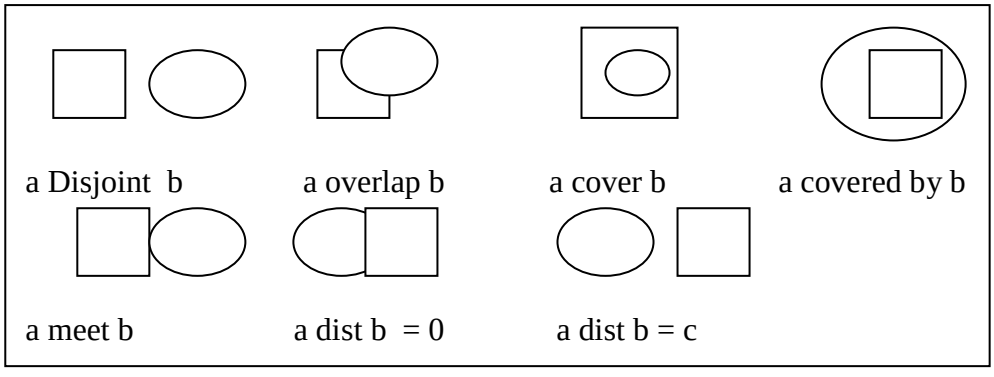

Figure (5): the position state spatial attribute.

Fourth step: In the proposed system the proposed spatial database will be in the following design (6).

Spatial object	Type	Size	Direction	Position	Dist	Population	Employment
O1	1	1	(a, o2)	(I, o3)	O3 <50 km	High	high

Figure (6): the design of the proposed spatial database.

Fifth step: After completing the build of spatial database we will begin the spatial mining algorithms to extract the implicit patterns and relationships among spatial objects. The first mining algorithm will be the association rule mining algorithm. The proposed spatial database is as usual relied upon alphanumerical and often transaction-based. The problem of discovering association rules is to find relationships between the existence of a spatial object (spatial or non spatial attributes) and the existence of other spatial objects (spatial or non spatial attributes) in a large repetitive collection.

Association rules would give the probability that some objects attributes appear with others based on the processed transactions, for example large town ^ near to water →high population [90%], meaning that there is a probability of 0.9 that high population is found when the town is large and near water. Essentially, the problem consists of finding objects attributes that frequently appear together, known as frequent or large objects attributes-sets.

The problem is stated as follows: Let $I = \{i1, i2, ...im\}$ be a set of literals, called attributes. Let D be a set of transactions, where each transaction T is a set of attributes such that $T \subseteq I$. A unique identifier *TID* is given to each transaction. A transaction T is said to contain X, a set of attributes in I, if $X \subseteq T$. An *association rule* is an implication of the form "$X \Rightarrow Y$", where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ has a *support* s in the transaction set D is s% of the transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set D with *confidence* c if c% of transactions in D that contain X also contain Y . In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a *support* and *confidence* greater that given thresholds. These rules are called *strong rule*. *Now in this work (association rules) we* find (spatially related) rules from the database. Association rules describe patterns, which are often in the database.

- If type O1 = 1 (town) and size O1 = 1 and direction O1 (north, o3 (type = 3)) then position O2 (type = 2) (intersect, O1) (s = 50%, c = 80%).

- If type O1 = 1 (town) and employment O1= high then dist between O1 and (O3 (type =3) or O7 (type 4)) < 50 km and position O7 (type =9) (overlap, O1) (s = 50%, c = 90%).

Sixth step: The second algorithm *which is used in the proposed spatial mining is the classification* mining algorithm, which is find a set of rules that determines the class of the classified object according to its attributes e. g." If population of town = high and the size of the town = large then the town will be classified as a high employment town".

Decision Tree Classifiers (DTC's) are used successfully in many diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition, to name only a few. Perhaps, the most important features of DTC's is their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret. The decision tree classifier is one of the possible approaches to multistage decision making; table look-up rules, decision table conversion to optimal decision trees, and sequential approaches are others. The basic idea involved in any multistage approach is to break up a complex decision into a union of several simpler decisions, hoping the final solution obtained this way would resemble the intended desired solution.

The proposed spatial database has one global scheme which includes the following attributes: (spatial object id, type, size, position, state distance, population and employment). Now we start building the decision tree classifier for two purposes: the first one is to delete all the available classes of the spatial objects, and the second is to classify the town to three classes; these are (high employment town, medium employment town, and low employment town). DTC is a tree as declared previously. So the most important step is how to choose the attribute to be the root node, then how to choose each one of the internal nods to complete the splitting and built this classifier. In this research the classifier will be built without the need to measure the entropy of each one of these attributes to decide which one is representing the root node and then which one represents the more powerful attribute to be the internal node to complete the splitting. Because the power of each attribute is very clear in the proposed spatial database, so:

- The root node would be the type of spatial object so, by this attribute the tree would be split into four classes: the first one is the land of earth, the second is water (which implicit sub classes river, sea and lake), the third is road (which implicit sub classes roads, highway and bridges) and the fourth class is mines and forest, see the following figure (7).

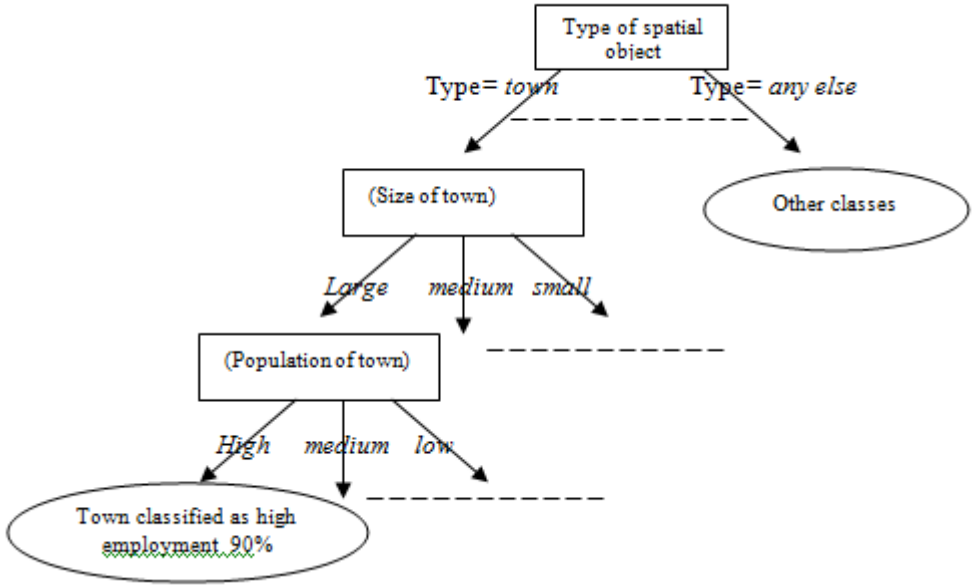

Figure (7) built the DTC which includes the choice of the best attribute to classify towns with spatial mining depending on the proposed spatial database.

***Seventh step*: T**he third algorithm *which is used in the proposed spatial mining is the clustering* mining algorithm. It groups the object from database into clusters in such a way that objects in one cluster are similar and objects from different clusters are dissimilar e. g. we can find clusters of towns with similar level of employment. Now by using *characteristic rules* (which describe some part of database e. g. "city is an object in the place has type = 1 (town), and has population (don't matter with any measure high, medium or low)".

The steps of clustering the towns to three clusters (high employment, medium employment and low employment) will be summarized by the following steps:

- Convert all the value of attributes to numerical value.
- Get the summation of all the values of the attributes for each town.
- The Euclidian distance law $\sum(x - yi)$ is used to take the minimum distance between the town and all the towns of one cluster.

Input: Take town value from town database which obtained by characteristics rules and set it to first created cluster.
Output: Put each town into cluster similar to it, has nearest values to it.
Step1: Check the value of the desired town with each value of towns in each cluster, then record the degree of similarity with each cluster.
Step 2: Put the town in the cluster which has a minimum degree of differential with its towns.
Step 3: Take the next town and go to step 1.

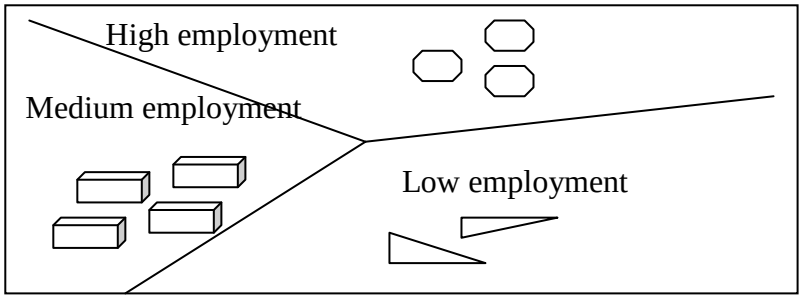

Figure (8) clustering results.

Conclusions:

In this paper we formulized a novel approach for mining spatial data type's database. We collected all the spatial data and organize it to flat database; we can conclude the following from this research:

1. The most important result we concluded in our research is: building spatial database as a flat database will make the spatial mining much more efficient that by reduce the mining to one level only so this will prevent the time and space consuming resulted in the previous work by extending the mining to multilevel.
2. By applying the new strategy we can in addition to save time and storage we could give an adequate result. These advantages can be gain due to the organizing of the spatial database and using the numerical values in representation.
3. We proposed a novel approach for building a spatial database to accommodate all the necessary requirements for applying Association rules, Clustering and Classification Algorithms.
4. With the proposed spatial database the extraction of association rule is could done by the traditional apriori algorithm without confusing, that make our proposed approach easy to use and understand by the administrators. Also the analysis step followed by extraction the rules are easy because it depended on generalization and normalization determined by the miner.
5. In this research the classifier will be built without the need to measure the entropy of each attribute to decide which one is the root node or the internal node. The decision is: the spatial object will be the root node.
6. Also the clustering in our proposed spatial database will much more easily and faster since it deals with flat database rather than R-Tree techniques.

References:

1. Malerba, D., Lisi, F. A., Appice, A., Sblendorio, F., "Mining Spatial Association Rules in Census Data: A Relational Approach", 2002.

2. Han, J., Fu, Y, " Discovery of multiple-level association rules from large databases". In U. Dayal, P.M.D. Gray, S. Nishio (eds.): VLDB'95, Proceedings of the 21st International Conference on Very Large Data Bases, Morgan-Kaufmann (1995) 420-431).

3. Ester M., Kriegel H.-P., Xu X.: "Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification", Proc. 4th Int. Symp. on Large Spatial Databases, Portland, ME, 1995, in: Lecture Notes in Computer Science, Vol. 951, Springer, 1995, pp.67-82.

4. Ester M., Kriegel H.-P., Sander J.: "Spatial Data Mining: A Database Approach", Proc. 5th Int. Symp. on Large Spatial Databases, in: Lecture Notes in Computer Science, Vol. 1262, Springer, pp. 47-66, 1997.

5. Han, J., Koperski, K., Stefanovic, N.: "GeoMiner: A System Prototype for Spatial Data Mining". In Peckham, J. (ed.): SIGMOD 1997, Proceedings of the ACM-SIGMOD International Conference on Management of Data. SIGMOD Record 26, 2 (1997) 553-556.

6. Ludl, M.-C., Widmer, G.:" Relative Unsupervised Discretization for Association Rule Mining". In D.A. Zighed, H.J. Komorowski, J.M. Zytkow (Eds.): Principles of Data Mining and Knowledge Discovery, LNCS 1910, Springer-Verlag (2000) 148-158.

7. Malerba, D., Lisi, F.A.: "An ILP method for spatial association rule mining". Working notes of the First Workshop on Multi-Relational Data Mining, Freiburg, Germany (2001) 18-29.