SD-Miner: A Spatial Data Mining System

Shrikant R. Adam[1] **Vaibhav G. Khetal**[2] **Rupesh V. Kanse**[3] **Rohit B. Kamble**[4]
[1,2,3,4]Department of CSE, TKIET / Shivaji University, Kolhapur, India

Abstract — Owing to the GIS technology, a vast volume of spatial data has been accumulated, thereby incurring the necessity of spatial data mining techniques. In this paper, we propose a new spatial data mining system named SD-Miner. SD-Miner consists of three parts: a graphical user interface for inputs and outputs, a data mining module that processes spatial data mining functionalities, a data storage model that stores and manages spatial as well as non-spatial data by using a DBMS. In particular, the data mining module provides major spatial data mining functionalities such as spatial clustering, spatial classification, spatial characterization, and spatio-temporal association rule mining. SD-Miner has its own characteristics:

(1) It supports users to perform non-spatial data mining functionalities as well as spatial data mining functionalities intuitively and effectively.
(2) It provides users with spatial data mining functions as a form of libraries, thereby making applications conveniently use those functions.
(3) It inputs parameters for mining as a form of database tables to increase flexibility.

Keywords: Spatial data, Spatial Data Mining, SDMiner.

I. INTRODUCTION

Due to the development of information technology, a vast volume of data is accumulated on many fields. Since automated methods for filtering /analyzing the data and also explaining the results are required, a variety of data mining techniques finding new knowledge by discovering hidden rules from vast amount of data are developed [5].

In the field of geography, due to the development of technology for remote sensing, monitoring, geographical information systems, and global positioning systems, a vast volume of spatial data is accumulated. Also, there have been many studies of discovering meaningful knowledge from the spatial data.

Since the spatial data has its own characteristics different from the non-spatial data, direct using of general data mining techniques incurs many difficulties [9]. So there have been many studies of spatial data mining techniques considering the characteristics of the spatial data [1-3, 6-10]. However, commercial tools for spatial data mining have not been provided. Currently, many commercial data mining tools are available, but these tools support the spatial data mining functionalities. Also, while some academic spatial data mining tools such as GeoMiner [4] are available, there are almost no commercial spatial data mining tools. So, for easy using of spatial data mining for real spatial data applications, developments of spatial data mining tools are needed. In this paper, we propose a new spatial data mining system named SD-Miner. SD-Miner supports four important spatial data mining functionalities: spatial clustering [10], spatial classification [8], spatial characterization [1], and spatio-temporal association rule mining [7]. We first analyze characteristics of previous spatial data mining techniques and suggest techniques to improve their efficiency in developing SD-Miner.

The rest of this paper is organized as follows. Section 2 explains four important spatial data mining functionalities adopted in SD-Miner. Section 3 introduces the structure of SD-Miner. Section 4 discusses some our enhancements for improving the efficiency of SD-Miner. Section 5 gives the final conclusion of this paper and the future work.

II. SPATIAL DATA MINING

In this section, the concept and characteristics of spatial data mining techniques adopted in SDMiner are explained. Spatial data mining considers the characteristics of the spatial data in data mining and thus can be regarded as a spatial extension of data mining. i.e., spatial data mining discovers hidden interesting spatial relationships, various spatial patterns, and knowledge from a spatial database [5]. Spatial objects are composed of not only general attributes represented as text but also spatial attributes such as point, line, and surface of two- or three-dimensional space [3]. The spatial objects also have topological information. The spatial data mining system must have functions for extracting of patterns of spatial and non-spatial attributes.

A. Spatial clustering

Spatial clustering classifies spatial objects as multiple groups according to its positional or geographical characteristics. SD-Miner uses GDBSCAN [10] with some enhancements. GDBSCAN is a spatial extension of DBSCAN, which is a density-based clustering technique. GDBSCAN makes clusters with criteria of the minimal distance between spatial objects. If the distance between objects is smaller than some user predefined distance, the two objects are included in an *eps*-neighbor. If the number of objects in an *eps*-neighbor is more than the predefined threshold, the objects are managed in the same cluster and the base object of the neighbor is called *core* object.

GDBSCAN has the following benefits. First, it is adequate for clustering spatial objects having shapes and extensions rather than points, and produces results similar to human decisions since it employs a density-based approach. So, it makes clusters with best considering of characteristics of spatial data. Second, the effect of noises is not significant. Third, it is efficient and scalable since it produces clusters with just one scan of the entire spatial data.

B. Spatial classification

Usually, in the spatial classification, the objects are classified with considering of spatial and nonspatial attributes [8]. The spatial classification also uses the decision tree. A big difference between the spatial classification and the classification is that the aggregation value of the spatial objects in a near region is used in the spatial classification.

We adopt the technique proposed in [8]. For making of a decision tree, the technique additionally uses predicates on relationship among spatial objects as decision criteria. For this, as the first step, the spatial attributes are represented as spatial predicates and then the possible useful predicates are extracted with the RELIEF algorithm

[8].For the second step, the decision tree is constructed with the predicates.

The benefits of this technique are as follows. Since the decision tree is constructed after discarding ineffective predicates, the tree construction cost is greatly reduced. Speedy and correct classification can be made via using the simple explainable rules by constructing a binary decision tree and minimizing computational cost of pruning with the RELIEF algorithm.

C. Spatial characterization

Spatial characterization extracts a global outline of data classes for a given spatial region by using the spatial objects of the region [1]. It gives simple and clear abstract information of the region. SD-Miner uses the technique proposed in [1].

Spatial characterization evaluates whether the characteristics of given spatial objects are expanded to near region. To do this, the objects are defined as a neighbor of each other with considering of their distance or direction. The neighbor information is managed by using the neighbor table. The region handled by spatial characterization can be expanded with a spatial expansion algorithm using the neighbor table.

D. Spatio-temporal association rule mining

By using spatial association rule mining, we can represent topological relationship and distance relationship of spatial objects via analyzing the relationship among spatial data and between spatial and non-spatial data. Also, by adding of temporal data analysis, we can use spatio-temporal association rule mining. SD-Miner uses spatiotemporal association rule mining that is a temporal extension of technique proposed in [7].

In order to use spatial association rule mining, the spatial relationships between spatial objects must be defined. The relationships are represented as spatial predicates. The predicates defined by a user are stored as *concept hierarchy* data in the database of SD-Miner. If the spatial relationship is defined as the predicates, the predicates can be regarded as non-spatial attributes. So we can extract spatiotemporal association rules by using the well-known Apriori algorithm.

Spatial association rule mining can be applied whether the predicates are spatial or non-spatial. So we can extract association rules among spatial predicates or among spatial and non-spatial predicates. This technique can be applied hierarchically by using level-based predicates if exist. So, we can extract detailed as well as abstracted association rules in this case.

III. ARCHITECTURE OF SD MINOR

SD-Miner is composed of three main parts (see Figure 1): graphical user interface (GUI), SD Miner module, and data storage module (also, called DBMS management module).

The GUI part gets user input variables needed for mining, transfers them to the SD-Miner module, and shows the mining results as table, chart, and map. Each mining function needs different input variables and has different result representations.

The SD-Miner module processes respective data mining functions and transfers the results to the data storage module. This module provides four data mining functions explained in Section 2.

The data storage module stores data using DBMS for data mining and supports easy communication between the SD-Miner module and the DBMS. We use Oracle 10g as the DBMS and use spatial functions supported in Oracle 10g for efficiency of SD-Miner. The data storage module handles four categories of data: concept hierarchies, spatial data, non-spatial data, and temporal data.

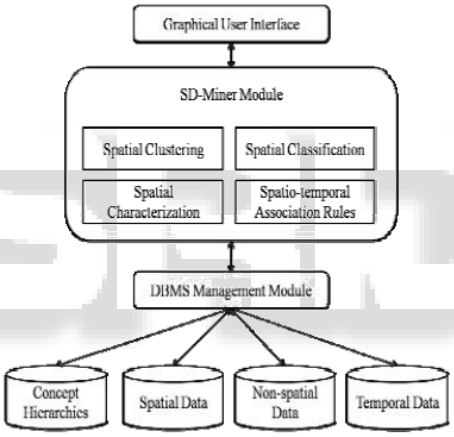

Fig. 1: Architecture of SD-Miner.

Our SD-Miner has following four benefits. First, the functions in the SD-Miner module can handle both of spatial and non-spatial data. They can automatically detect what type of data is used.

Second, since the functions are developed in a library style, they can be used in other systems that require spatial data mining functions. Also, other functions can be easily added into our SD-Miner. Third, since the user input is usually handled as a database table style, the portability is very high. Fourth, user opinion can be used for defining spatial predicates since the scale used in spatial data is different in each case.

IV. IMPROVEMENTS IN SD-MINOR

In this section, problems of existing techniques explained in Section 2 are discussed and refinements applied for them in SD-Miner are also explained. The process of each function is also shown.

A. Spatial Clustering

Although GDBSCAN has many benefits, it only uses the number of objects having a smaller distance than some

predefined distance as criteria. So, it does not consider the characteristics of spatial data that have sizes and shapes. We employed another criterion *area* for deciding the core object. If the area sum of objects is larger than some predefined ratio of the total area, the main object is selected as the core object. Figure 2 shows an example of GDBSCAN clustering when the number of objects for cluster selection is 3. If the number is 4, Cluster 2 will not be formed as a cluster. If our area criterion is used, it could be selected as a cluster. So, it is preferable to use the area as another criterion. We tested this enhancement with real data and got meaningful results. Our system provides two options: one is to use both of the distance and area criteria, and another is to use only one of them.

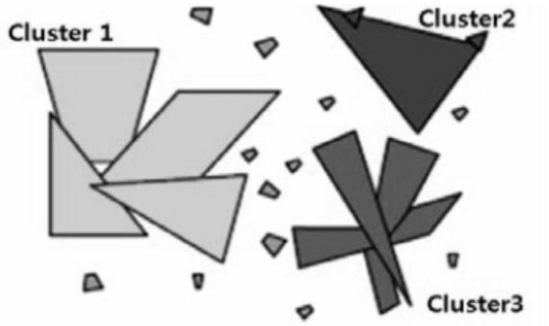

Fig. 2: An example of spatial clustering.

We also suggest another technique for processing efficiency. For selecting of a core object, we only consider the objects not yet handled for clustering.

With this technique, we perform only one scan of data for clustering. The process of our spatial clustering is depicted in Figure 3. At first, user inputs the needed parameters and the name of spatial and non-spatial objects. The parameters are the criteria used and its dependent values. And then, the analysis is performed, information for clustering is extracted, and the clustering results are created.

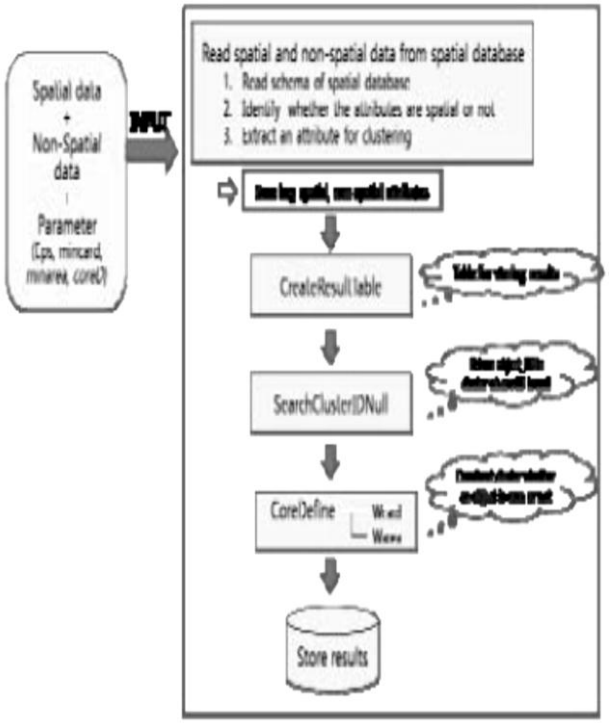

Fig. 3: Process of spatial clustering.

B. Spatial classification

As explained in Section 2.2, we use a binary decision tree for spatial classification. The spatial objects must have spatial predicates. These predicates are stored as attributes of objects. The concept of 'buffer' is used for spatial classification. Buffer means the area extended from an object with some given distance. It is used that predicates of spatial objects having a form of nonspatial attributes that are related on space (for example, the number of population in the circle having radius 1km cantered at a shopping centre).

The value is called as *aggregation value*. The value can be used in a spatial predicate. In [8], the buffer size in a predicate having greatest information gain for the predicate is selected for all candidates of user inputs. Since the same size is applied to all predicates, it may be not adequate to some predicates.

In SD-Miner, we calculate the suitable buffer size for each predicate from all candidates of user input.

As a result, a different buffer size is used for each object, and thus we can have more precious classification results.

To make a binary decision tree of a proper size, we use the concept of majority voting. So, the size of tree may be controlled by a user. With this idea, we can reduce the processing time of classification considerably.

Figure 4 shows the process of our spatial classification. At first, the training data and input data are selected for processing. And then, spatial predicates are defined by using concept hierarchy data. The predicates useful for classification are selected by using the RELIEF algorithm. The binary decision tree is constructed by using this predicates. So, we can enhance the correctness of spatial classification and reduce the time complexity.

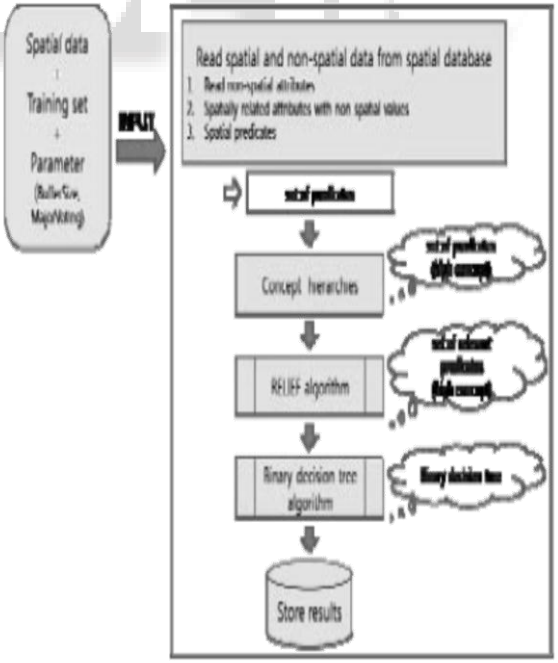

Fig. 4: Process of spatial classification.

C. Spatial characterization

According to [1], it requires O (n_2) times for making of neighbor table explained in Section 2.3. To reduce the processing time, we only handle objects in a specific not total area in building of the neighbor table. Although, this

method may make multiple neighbor tables for multiple specific regions, it is more effective in terms of the processing time. The process of our spatial characterization is depicted in Figure 5. At first, spatial and nonspatial attributes of spatial objects in a specific area are taken. Second, the targets are defined by using the input data and concept hierarchy data and frequent patterns for non-spatial data are discovered. Spatial attributes are represented as spatial relationship with spatial predicates. For pattern analysis of objects, non-spatial attributes are generalized as a hierarchical structure. Third, the neighbor table is constructed for target objects and test for applying the patterns for non-spatial attributes is performed. If the pattern is occurred at neighbor objects, we can expand the objects as target objects. This expansion process is performed for maximum expansion of the objects. Fig. 5. Process of spatial characterization.

D. Spatio-temporal association rule mining

Although spatial association rule mining has many benefits, there are difficulties on defining spatial predicates and setting of ratio values for multilevel association rule mining. It also may have different results according to input methods of non-spatial attributes. So, in SD-Miner, we suggest following methods to solve the difficulties. First, automatic calculating of minimum support and confidence ratio for multilevel association rule mining is used. The lower level values can be generated by using upper level input data with the distribution of the data. For spatial association rule mining, the spatial data for most upper level must be defined. So, input of support and confidence ratio of the most upper level is needed. For the lower levels, support and confidence ratio must be lower than its upper level since its number of objects in a lower level is smaller than those in a higher level. With the conditions, we can calculate the lower level values.

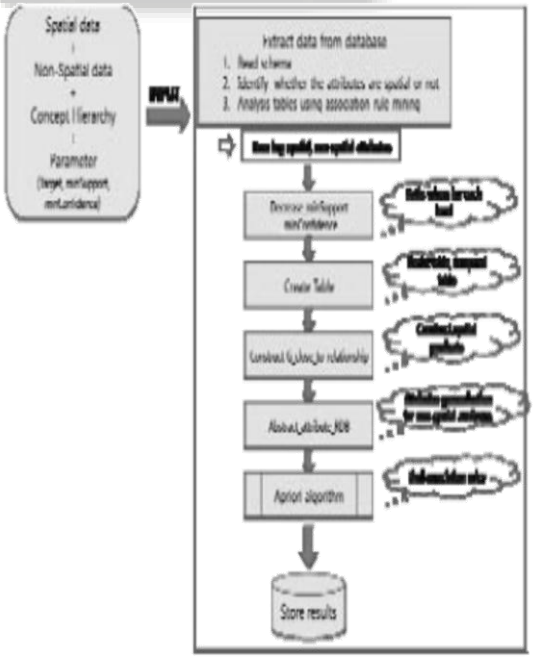

Fig. 5: Process of spatio-temporal association rule mining.

Second, we suggest generalization of non-spatial attributes according to the user opinion. The generalization is done by making a certain range of an attribute to a value. After doing this, we may find various association rules by using the Apriori algorithm. Third, for doing spatio-temporal association rule mining through using spatial association rule mining technique, we add temporal data as an attribute of the database. Since the value of support and confidence ratio on temporal data may be not satisfied according to the time unit, the unit must be normalized. The method is the same as the above generalization method of non-spatial attributes. The process of our spatio-temporal association rule mining is depicted in Figure 6. At first, concept hierarchies, spatial, and non-spatial data are needed. We also need a table for generalization of nonspatial attributes. The parameters such as target, support and confidence ratio for the most upper level are also needed. We automatically calculate support and confidence ratio for lower levels. The generalization values for non-spatial and temporal attributes and spatial predicates are stored at the temporary table. With the temporary table, we can process the spatio-temporal association rule mining with the Apriori algorithm.

V. CONCLUSION

Spatial data has positional and topological data that do not exist in general data, and its structure is different according to the kinds of spatial data. Also, the objects on space affect each other and the relationship of objects is also different according to the kinds of objects. There have been many researches on spatial data mining considering these characteristics of spatial data.

In this paper, we explain the concept of spatial clustering, spatial classification, spatial characterization, and spatio-temporal association rule mining. We present our experiences in developing a spatial data mining system called SDMiner that provides proper spatial data mining techniques with improved effectiveness and efficiency for real applications. SD-Miner adopts following techniques. For spatial clustering, we adapt GDBSCAN. For spatial characterization, we use a binary decision tree and the RELIEF algorithm for efficiency. For spatial characterization, we use the neighbor table for spatial extension of the current characterization method. For spatio-temporal association rule mining, we use temporal data with spatial association rule mining using the spatial concept layer. SD-Miner uses spatial data mining functions extended from general mining functions. So, it can be applied to both of spatial and non-spatial data.

Without special intervention of a user, it automatically recognizes which type of data is used.

All functions are developed in a library style, the functions can be used another system easily. Since it uses spatial functions supported by Oracle 10g for implementing of functions, it is very efficient and portable. So, SD-Miner is an efficient spatial mining system that can handle spatial and nonspatial data easily. In order to verify the practicability of our SD-Miner developed, we found meaningful results by performing spatial data mining with real-world spatial data. They cannot be presented here because of space limitations.

REFERENCES

[1] M. Ester et al., “Algorithms for Characterization and Trend Detection in Spatial Databases,” In Proc. Int'l.

Conf. on Knowledge Discovery and Data Mining, KDD, pp. 44-50, 1998.
[2] M. Ester et al., "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support," Data Mining and Knowledge Discovery, Vol. 4, pp. 193-216, 2000.
[3] M. Ester, H. Kriegel, and J. Sander, "Algorithms and Applications for Spatial Data Mining," Geographic Data Mining and Knowledge discovery, 2001.
[4] J. Han, K. Koperski, and N. Stefanovic, "GeoMiner: A System Prototype for Spatial Data Mining," In Proc. ACM Int'l. Conf. on Management of Data, ACM SIGMOD, pp. 553-556, 1997.
[5] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Academic Press,2001.
[6] E. Knorr and R. Ng, "Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining," IEEE Trans. On Knowledge and Data Engineering, IEEE TKDE, Vol. 8, pp. 884-897, 1996.
[7] K. Koperski and J. Han, "Discovery of Spatial Association Rules in Geographic Information Databases," In Proc. Int'l. Symp. on Advances in Spatial Databases, SSD, pp. 47-66, 1995.
[8] K. Koperski, J. Han, and N. Stefanovic, "An Efficient Two-Step Method for Classification of Spatial Data," In Proc. Int'l. Symp. On Spatial Data Handling, SDH, pp. 45-54, 1998.
[9] W. Lu, J. Han, and B. Ooi, "Discovery of General Knowledge in Large Spatial Databases," In Proc. Far East Workshop on Geographic Information Systems, pp. 275-289,1993.

