International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 2, ISSN 2249 –071X

Spatial Data Mining Through Cluster Analysis

A .Santhi Latha
Head,
Department for Basic Science and Humanities in Vignan's Lara Institute of Technology and Science, Vadlamudi, Guntur, A.P., India

J. Swapna Priya
Asst. Professor,
Department of Information Technology, Vignan's Lara Institute of Technology and Science, Vadlamudi, Guntur, A.P., India

Sk. Abdul Kareem
Asst. Professor,
Department of Information Technology, Vignan's Lara Institute of Technology and Science, Vadlamudi, Guntur, A.P., India

M. Pavani Devi
Asst. Professor,
Department of Computer Science & Engineering, Sri Sai Madhavi Institute of Science & Technology, A.P., India

Abstract **— Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. The main objective of the spatial data mining is to discover hidden complex knowledge from spatial and not spatial data despite of their huge amount and the complexity of spatial relationships computing. However, the spatial data mining methods are still an extension of those used in conventional data mining. Spatial data is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. The collected data far exceeded human's ability to analyze. Recent studies on data mining have extended the scope of data mining from relational and transactional databases to spatial databases. In this paper we discuss how cluster analysis can be helpful for mining spatial data. Cluster analysis divides data into meaningful or useful groups (clusters). If meaningful clusters are the goal, then the resulting clusters should capture the "natural" structure of the data. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, and to provide a grouping of spatial locations prone to earthquakes. However, in other cases, cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbors of points.**

Key Words — Cluster Analysis, Data Mining, Spatial data, Spatial data mining.

I. INTRODUCTION

Spatial data mining is the application of data mining techniques to spatial data. Data mining in general is the search for hidden patterns that may exist in large databases. Spatial data mining is the discovery of interesting the relationship and characteristics that may exist implicitly in spatial databases. Because of the huge amounts (usually, terabytes) of spatial data that may be obtained from satellite images, medical equipments, video cameras, etc. It is costly and often unrealistic for users to examine spatial data in detail. Spatial data mining aims to automate such a knowledge discovery process. Thus it plays on important role in

1. Extracting interesting spatial patterns and features
2. Capturing intrinsic relationships between spatial and non spatial data
3. Presenting data regularity concisely and at higher conceptual levels and
4. Helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance.

Spatial database stores a large amount of space related data, such as maps, preprocessed remote sensing or medical imaging data and VLSI chip layout data. Spatial databases have many features distinguishing them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multi dimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

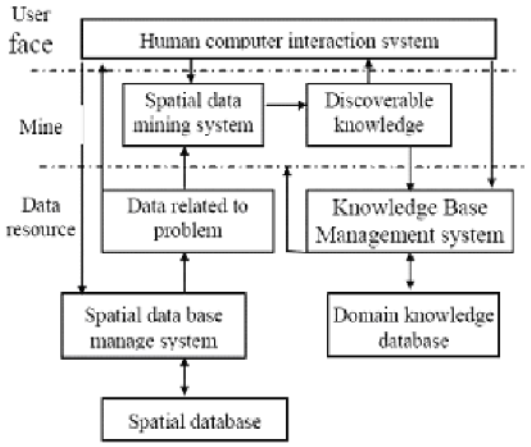

Fig.1. The systematic structure of spatial data mining

The spatial data mining can be used to understand spatial data, discover the relation between space and the non space data, set up the spatial knowledge base, excel the query, reorganize spatial database and obtain concise total characteristic etc.. The system structure of the spatial data mining can be divided into three layer structures mostly, such as the Figure 1 show [1].The customer interface layer is mainly used for input and output, the miner layer is mainly used to manage data, select algorithm and storage the mined knowledge, the data source layer, which mainly includes the spatial database (camalig) and other related data and knowledge bases, is original data of the spatial data mining.

Cluster analysis divides data into meaningful or useful groups (clusters). If meaningful clusters are the goal, then the resulting clusters should capture the "natural" structure of the data. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, and to provide a grouping of spatial locations prone to earthquakes. However, in other cases, cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbors of points. Cluster Analysis is a branch of statistics that in the past three decades has been intensely studied and successfully applied to many applications. To the spatial data mining task at hand, the attractiveness of cluster analysis is its ability to find structures or clusters directly from the given data, without relying on any hierarchies. In this paper we discuss how cluster analysis can be helpful for mining spatial data.

II. Cluster Analysis

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the "better" or more distinct the clustering. The definition of what constitutes a cluster is not well defined, and, in many applications clusters are not well separated from one another. Nonetheless, most cluster analysis seeks as a result, a crisp classification of the data into non-overlapping groups.

Fig. 2a. Initial Points

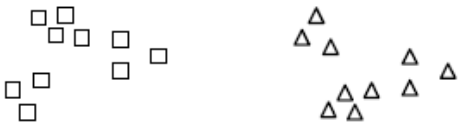

Fig. 2b. Two Clusters

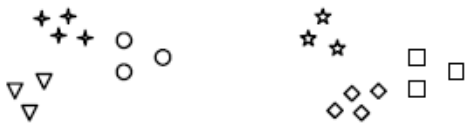

Fig. 2c. Six Clusters

Fig. 2d. Four Clusters

To better understand the difficulty of deciding what constitutes a cluster, consider figures 2a through 2d, which show twenty points and three different ways that they can be divided into clusters. If we allow clusters to be nested, then the most reasonable interpretation of the structure of these points is that there are two clusters, each of which has three subclusters. However, the apparent division of the two larger clusters into three subclusters may simply be an artifact of the human visual system. Finally, it may not be unreasonable to say that the points form four clusters. Thus, we stress once again that the definition of what constitutes a cluster is imprecise, and the best definition depends on the type of data and the desired results.

Cluster analysis is a classification of objects from the data, where by classification we mean a labeling of objects with class (group) labels. As such, clustering does not use previously assigned class labels, except perhaps for verification of how well the clustering worked. Thus, cluster analysis is distinct from pattern recognition or the areas of statistics know as discriminant analysis and decision analysis, which seek to find rules for classifying objects given a set of pre-classified objects.

III. Spatial Data Mining Primitives

Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. The primitives of spatial data mining are:

A. Rules

There are several kinds of rules can be discovered from databases in general. For example characteristic rules, discriminant rules, association rules, or deviation and evalution rules can be mined [1]. A ***Spatial characteristic rule*** is a general description of the spatial data**.** For example, a rule describing the general price range of houses in various geographic regions in a city is a spatial characteristic rule. A *discriminant rule* is general description of the features discriminating or contrasting a class of spatial data from other class(es) like the comparison of price ranges of houses in different geographical regions. A *spatial association rule* is a rule which describes the implication of one a set of features by another set of features in spatial databases. For example, a rule associating the price range of the houses with nearby spatial features, like beaches, is a spatial association rule.

B. Thematic Maps

Thematic map is map primarily design to show a theme, a single spatial distribution or a pattern, using a specific map type. These maps show the distribution of features over limited geography areas [1]. Each map defines a partitioning of the area into a set of closed and disjoint

regions; each includes all the points with the same feature value. Thematic maps present the spatial distribution of a single or a few attributes. This differs from general or reference maps where the main objective is to present the position of the object in relation to other spatial objects.

Thematic maps may be used for discovering different rules. For example, we may want to look at temperature thematic map while analyzing the general weather pattern of a geographic region. There are two ways to represent thematic maps: *Raster*, and *Vector*. In the *raster image* form thematic maps have pixels associated with the attribute values. For example, a map may have the altitude of the spatial objects coded as the intensity of the pixel (or the color). In the *vector representation*, a spatial object is represented by its geometry, most commonly being the boundary representation along with the thematic attributes. For example, a park may be represented by the boundary points and corresponding elevation values.

IV. Clustering Techniques For Spatial Data Mining

Cluster analysis is a branch of statistics that has been studied extensively for many years. The main advantage of using this technique is that interesting structures or clusters can be found directly from the data without using any background knowledge, like concept hierarchy. A similar approach in machine learning is known as *unsupervised* learning. Clustering algorithms used in statistics, like PAM or CLARA [2], are reported to be inefficient from the computational complexity point of view. As per the efficiency concern, a new algorithm called CLARANS (Clustering large Applications based upon RANdomized Search), was developed for cluster analysis. Experimental evidence showed that CLARANS outperforms the two existing cluster analysis algorithms, PAM (Partitioning Around Medoids) and CLARA (Clustering LARge Applications).

A. PAM (Partitioning Around Medoids)

Assuming that there are *n* objects, PAM finds *k* clusters by first finding a representative object for each cluster. Such a representative, which is the most centrally located point in a cluster, is called a *medoid*[2]. After selecting *k* medoids, the algorithm repeatedly tries to make a better choice of medoids analyzing all possible pairs of objects such that one object is a medoid and other is not. The measure of clustering quality is calculated for each such combination. The best choice of points in one iteration is chosen as the medoids for the next iteration. The cost of a single iteration is *O(k(n-k)2)*. It is therefore computationally quite inefficient for large values of *n* and *k*.

B. CLARA (Clustering LARge Applications)

The difference between the PAM and CLARA algorithms is that the later one is based upon *sampling*. Only a small portion of the real data is chosen as a representative of the data and medoids are chosen from this sample using PAM[2]. The idea is that if the sample is selected in a fairly random manner, then is correctly represents the whole dataset and therefore, the representative objects (medoids) chosen will be similar as if chosen from the whole dataset. CLARA draws multiple samples and outputs the best clustering out of these samples. CLARA can deal with larger dataset then PAM. The complexity of each iteration now becomes *O(kS2+k(n-k))*, where S is the size of the sample.

C. CLARANS (Clustering Large Applications based upon RANdomized Search)

CLARANS algorithm mix both PAM and CLARA by searching only the subset of the dataset and it does not confine itself to any sample at any given time [2]. While CLARA has a fixed sample at every stage of the search, CLARANS draws a sample with some randomness in each step of the search. The clustering process can be presented as searching a graph where every node is a potential solution, i.e, a set of k medoids. The clustering obtained after replacing a single medoids is called the *neighbor* of the current clustering. The number of *neighbors* to be randomly tried is restricted by the parameter *maxneighbor*. If a better *neighbor* is found CLARANS moves to the neighbor's node and the process is started again, otherwise the current clustering produces a local *optimum*. If the local optimum is found CLARANS starts with new randomly selected node in search for a new local optimum. The number of local optima to be searched is also bounded by the parameter *numlocal*. CLARANS also enables the detection of outliers, e.g.. points that do not belong to any cluster.

CLARANS Algorithm:

1) Randomly pick *K* candidate medoids.
2) Randomly consider a swap of one of the selected points for a non-selected point.
3) If the new configuration is better, i.e., has lower cost, then repeat step 2 with the new configuration.
4) Otherwise, repeat step 2 with the current configuration unless a parameterized limit has been exceeded. (This limit was set to max (250, *K* *(*m* - *K*)).
5) Compare the current solution with any previous solutions and keep track of the best.
6) Return to step 1 unless a parameterized limit has been exceeded. (This limit was set to 2.)

Based upon CLARANS, two spatial data mining algorithms were developed: *Spatial dominant approach*, SD (CLARANS) and *non-spatial dominant approach*, NSD (CLARANS).

D. Spatial dominant approach SD (CLARANS)

There are different approaches to spatial data mining. Assume that a spatial database consists of both spatial and non-spatial attributes, and that non spatial attributes are stored in relations [3, 4] The general approach here is to use clustering algo rithms to deal with the spatial attributes, and use other learning tools to take care of the non-spatial counterparts.

DBLEARN is a tool for mining non-spatial attributes [5]. It takes as inputs relational data, generalization hierarchies for attributes, and a learning query specifying the focus of the mining task to be carried out. From a barning re quest, DBLEARN first extracts a set of relevant tuples via SQL

queries. Then based on the generalization hierarchies of attributes, it, iteratively generalizes the tuples. The algorithm below, called SD (CLARANS), combines CLARANS and DBLEARN in a spatial dominant fashion. That is, spatial clustering is performed first, followed by nonspatial generalization of every cluster.

SD CLARANS Algorithm:

1) Given a learning request, find the initial set of relevant tuples by the appropriate SQL queries.
2) Apply CLARANS to the spatial attributes and find the most natural number knelt of clusters.
3) For each of the k….t clusters obtained above,
(a) Collect the non-spatial components of the tuples included in the current cluster, and
(b) Apply DBLEARN to this collection of the non-spatial components.

E. Non- spatial dominant approach NSD (CLARANS)

To a large extent, spatial dominant algorithms, such as SD (CLARANS), can be viewed as focusing asymmetrically on discovering non-spatial characterizations of spatial clusters. Non-spatial dominant algorithms, on the other hand, focus on discovering spatial clusters existing in groups of non-spatial data items. For example, these algorithms may find interesting discoveries based on the spatial clustering or distribution of a certain type of houses. The following algorithm, NSD (CLARANS), uses DBLEARN and CLARANS to perform data mining on non-spatial and spatial attributes respectively.

NSD CLARANS Algorithm:

1) Given a learning request, find the initial set of relevant tuples by the appropriate SQL queries.
2) Apply DBLEARN to the non-spatial attributes, until the final number of generalized tuples fall below a certain threshold .
3) For each generalized tuple obtained above,
(a) Collect the spatial components of the tuples represented by the current generalized tuple, and
(b) Apply CLARANS and the heuristics presented above to find the most natural number knot of clusters.
4) For all the clusters obtained above, check if there are clusters that intersect or overlap. If exist, such clusters can be merged. This in turn causes the corresponding generalized tuples to be combined.

V. CONCLUSION

Spatial data mining is the application of data mining techniques to spatial data. Data mining in general is the search for hidden patterns that may exist in large databases. Spatial data mining is the discovery of interesting the relationship and characteristics that may exist implicitly in spatial databases. Because of the huge amounts (usually, terabytes) of spatial data that may be obtained from satellite images, medical equipments, video cameras, etc. It is costly and often unrealistic for users to examine spatial data in detail. Another important concept of data mining is cluster analysis. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. In this paper we discuss how cluster analysis can be helpful for mining spatial data.

REFERENCES

[1] M.Hemalatha. M; Naga Saranya. N. A Recent Survey on Knowledge Discovery in Spatial Data Mining, IJCI International Journal of Computer Science, Vol 8, Issue 3, No.2, may,2011.
[2] Krzysztof Koperski.; Junas Adhikary.; and Jiawei Han. Spatial Data Mining: Progress and Challenges Survey Paper, School of Computer Science Simon Fraser University Burnaby, B.C.Canada V5A IS6.
[3] W. G. Aref and H. Samet. (1991) Optimization Strategies for Spatial Query Processing, Proc. 17th VLDB, pp. 81-90.
[4] R. Laurini and D. Thompson. (1992) Fundamentals of Spatial Information Systems, Academic Press.
[5] J. Han, Y. Cai and N. Cercone. (1992) Knowledge Discovery in Databases: an Attribute- Oriented Approach, Proc. 18th VLDB, pp. 547- 559.

AUTHOR'S PROFILE

A. Santhi Latha
is currently working as Head of the Department for Basic Science and Humanities in Vignan's Lara Institute of Technology and Science, Vadlamudi. She pursued her Masters of Technology degree from Acharya Nagarjuna University, Nambur, in CSE. Her research interests include data warehousing and data mining, image processing, and Human perception and visualization Techniques.

J. Swapna Priya
is currently working as an Assistant Professor in Department of Information Technology, Vignan's Lara Institute of Technology and Science, Vadlamudi, Guntur. She received Masters of Technology degree from JNTUA, Anantapur in Computer Science and Engineering. Her research areas include Mobile Ad-hoc networks, cryptography and network security, data warehousing, and data- mining.

Sk. Abdul Kareem
is working as an Assistant Professor in Department of Information Technology, Vignan's Lara Institute of Technology and Science, Vadlamudi, Guntur. He pursued Masters of Technology degree from Acharya Nagarjuna University, Nambur in Computer Science and Engineering. His research interests include Embedded Systems, Data Mining, Artificial Neural Networks, Network Security and Image Processing.

M. Pavani Devi
is currently working as Assistant Professor in Department of Computer Science & Engineering at Sri Sai Madhavi Institute of Science & Technology, Rajahmundry, A.P., India. She Completed her M.Tech. (CSE) from Acharya Nagarjuna University, A.P., India. Her research areas includes Database Management Systems, Computer Networks and Data Warehousing & Data Mining.