A Survey on Data Mining using Clustering Techniques

T.Revathi, Dr.P.Sumathi

Abstract-Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).Basically there are different types related to data mining like Text Mining, Web Mining, Multimedia Mining, Spatial Mining, Object Mining etc. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Data mining is evolved in a multidisciplinary field, including database technology, machine learning, artificial intelligence, neural network, information retrieval, and so on. In principle data mining should be applicable to the different kind of data and databases used in many different applications, including relational databases, transactional databases, data warehouses, object- oriented databases, and special application- oriented databases such as spatial databases, temporal databases, multimedia databases, and time- series databases. Spatial data mining, also called Spatial mining, is data mining as applied to the spatial data or spatial databases. Spatial data are the data that have spatial or location component, and they show the information, which is more complex than classical data. A spatial database stores spatial data represents by spatial data types and spatial relationships and among data. Spatial data mining encompasses various tasks. These include spatial classification, spatial association rule mining, spatial clustering, characteristic rules, discriminant rules, trend detection. This paper presents how spatial data mining is achieved using clustering.

Keywords - *Clustering, Data mining, Spatial data, Text Mining, Web Mining, Multimedia Mining, Object Mining.*

I. INTRODUCTION

Data mining techniques are used in a many research areas, including mathematics, cybernetics, genetics and marketing. Data mining is sorting through data to identify patterns and establish relationships. Large amounts of data has been collected and stored in large data bases by database technologies and data collection techniques. For some applications only a small amount of the data in the databases is needed. This data is called knowledge or information. Data mining is the process of extracting knowledge from these large databases. Data mining is also called knowledge discovery in databases or KDD process. Although there have been many studies of data mining in relational and transaction databases .

Data mining is in great demand in other applicative databases, including spatial databases, temporal databases, object-oriented databases, multimedia databases, etc. The aim of this paper is on spatial data mining. Spatial data mining is the process of extracting interesting knowledge from spatial databases. The spatial databases contain objects that represent space. The spatial data represents topological and distance information. This spatial objects is organized by spatial indexing structures. Spatial data mining, or knowledge discovery in spatial database, refers to the extraction of implicit knowledge, spatial relocations, or other patterns not explicitly stored in spatial databases.

Spatial data mining methods can he applied to extract interesting and regular knowledge from large spatial databases. This knowledge can be used for understanding spatial and non spatial data and their relationships. This knowledge is very useful in Geographic Information Systems (GIS), image processing, remote sensing and so on. Knowledge discovered from spatial data can be of various forms, like characteristic and discriminant rules, extraction and description of prominent structures or clusters, spatial associations, and others. The purpose of this paper is to provide an overall picture of the spatial data mining, and how spatial data mining is achieved through clustering process.

II .TYPES OF DATA MINING

Basically there are different types related to data mining mainly

A Text Mining

Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. The discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources.

B.Web Mining

Web mining, a type of data mining used in customer relationship management (CRM), takes advantage of the huge amount of information gathered by a Web site to look for patterns in user behavior.

C.Multimedia Mining

Multimedia Mining is discovering knowledge from large amounts of different types of multimedia data. It involves the extraction of implicit knowledge, multimedia data

- *T.Revathi is currently pursuing Ph.D in Manonmaniam Sundaranar University,Tirunelveli. E-mail: trevathi_psg@yahoo.co.in*
- *Dr.P.Sumathi,Asst.Prof, PG & Research Department of Computer Science,Govt.Arts College,Coimbatore.*

relationships, or other patterns not explicitly stored in multimedia databases.

D.Object Mining

Object Mining is meant to focus on the use of data/text mining and knowledge discovery to produce software components.

III. CLUSTERING

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

A.Basic Types of Clustering

Partitional Clustering

A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

Hierarchical clustering

A set of nested clusters organized as a hierarchical tree.

B.Clustering Algorithms

Partitional Clustering Algorithm:

A.K-means algorithm:It is an iterative clustering algorithms in which items are moved among the sets of cluster until the desire set is reached.

B.PAM(Partitioning Around Medoids): It is also called the k-medoids algorithm represents a cluster by a medoid.

Hierarchical clustering Algorithm:

A.Agglomerative Algorithm : It start with each individual items in its own cluster and iteratively merge cluster until all items belong in one cluster.

B.Divisive Cluster Algorithm: With divisive clustering, all items are initially placed in one cluster and cluster are repeatedly split in two until all items are in their own cluster.

IV. SPATIAL DATA MINING DEFINITION

Spatial data mining (SDM) consists of extracting knowledge, spatial relationships and any other properties which are not explicitly stored in the database. SDM is used to find implicit regularities, relations between spatial data and/or non-spatial data. The specificity of SDM lies in its interaction in space. In effect, a geographical database constitutes a spatio-temporal continuum in which properties concerning a particular place are generally linked and explained in terms of the properties of its neighborhood. We can thus see the great importance of spatial relationships in the analysis process. Temporal aspects for spatial data are also a central point but are rarely taken into account.

V. SPATIAL DATA MINING TASKS

Basic tasks of spatial data mining are:

A. Classification

An object can be classified using its attributes. Each classified object is assigned a class.Classification is the process of finding a set of rules to determine the class of an object.

B. Association Rules

Find (spatially related) rules from the database. An association rule has the following form: A ->B(s%; c%), where s is the support of the rule (the probability, that A and B hold together in all the possible cases) and c is the confidence (the conditional probability that B is true under the condition of A e. g. ”if the city is large, it is near the river (with probability 80%)” or ”if the neighboring pixels are classified as water, then central pixel is water (probability 80%).”

C. Characteristic Rules

The characterization of a selected part of the database has been defined in as the description ofproperties that are typical for the part in question but not for the whole database. In the case of a spatial database, it takes account not only of the properties of objects, but also of the properties of their neighborhood up to a given level.

D. Discriminant Rules

Describe differences between two parts of database e. g. find differences between cities with high and low unemployment rate.

E. Clustering

Clustering means it is the process of grouping the database items in to clusters. All the members of the cluster has similar features. Members belong to different clusters has dissimilar features.

F. Trend Detection

Finds trends in database. A trend is a temporal pattern in some time series data. Spatial trend is defined as follows: consider a non spatial attribute which is the neighbor of a spatial data object. The pattern of changes in this attribute is called spatial trend.

VI. CLUSTERING METHODS FOR SPATIAL DATA MINING

A.Partitioning Around Medoids(PAM)

The PAM k-medoids clustering algorithm, for example, evaluates a set of k objects considered to be representative objects (medoids) of k clusters within T objects such that the non-selected objects are clustered with the medoid to which it is the most similar (i.e. closest in terms of the provided distance metric).The process operates by swapping one of the medoids with one of the objects iteratively such that the total distance between non-selected objects and their medoid is reduced.

The algorithm can be depicted as follows:

Step 1: Initialization - choose k medoids from T objects randomly
Step 2: Evaluation - calculate the cost D't − Dt for each swap of onemedoid with one object, where Dt is the total distance before the swap and D't is the total distance after the swap.
Step 3: Selection - accept the swap with the best cost and if the cost is negative, go to step 2; otherwise record the medoids and terminate the program.
The computational complexity of the PAM algorithm is O((1 + β)k(T − k)2) which is based on the number of partitions per object, where β is the number of successful swaps. It can also be expressed as O'((1 + β)k2(T − k)2) based on the number of distance calculations, i.e., one partition per object is equivalent to k distances calculations. Clearly, this is time consuming even for the moderate number of objects and a small number of medoids.
The algorithm proceeds in two steps:

- 1.BUILD-step: This step sequentially selects k "centrally located" objects, to be used as initial medoids
- SWAP-step: Swap a selected object and unselected object. This is done if this process can decrease the objective function .

B.Clustering LARgeApplications(CLARA)
Compared to PAM, CLARA can deal with much larger data sets. Like PAM CLARA also finds objects that are centrally located in the clusters. The main problem with PAM that it finds the entire dissimilarity matrix at a time. So for n objects, the space complexity of PAM becomes O(n2). But CLARA avoid this problem.
CLARA accepts only the actual measurements (i.e,. n ´ p data matrix). CLARA (Clustering LARge Applications) (Kaufman & Rousseeuw 1990) shown in Figure 1 reduces the computational complexity by drawing multiple samples of the objects and applying the PAM algorithm on each sample. The final medoids are obtained from the best result of these multiple passes as below of these multiple passes as in figure 1.

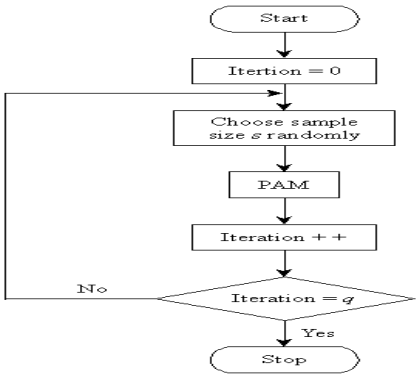

Figure 1CLARA Algorithm

The algorithm can be depicted as follows:
Repeat the following steps q times
Step 1: Call the PAM algorithm with a random sample, s objects from the original set of T objects.
Step 2: Partition the T objects based on the k medoids obtained from previous step. Update the better medoids based on the average distance of the partition.
The computational complexity of the CLARA algorithm is O(q(ks2 + (T −k)) + βks2) based on the number of partitions per object or O'(q(k2s2 + k(T −k)) + βk2s2) based on the number of distance calculations, where q, s, k, β and T are the number of samples, object size per sample, number of medoids,the number of successful swaps for all samples tested and the total number of objects, respectively. Clearly, the CLARA algorithm can deal with a larger number objects than can PAM algorithm if s << T.
If the sample size s is not large enough, the effectiveness (ie. the average distance) of the CLARA algorithm is reduced. However, the efficiency (in termsof computation time) is impaired if the sample size is too large. There is tradeoff between the effectiveness and efficiency in CLARA algorithm. The best clustering cannot be obtained in CLARA if one of the best medoids is not included in the sample objects. In order to achieve both efficiency and acceptable performance (in terms of average distance per object) the CLARANS (Clustering Large Applications based on RANdomized Search) algorithm (Ng & Han 2002) was proposed.
CLARA assigns objects to clusters in the following way:

Step 1:BUILD-step: Select k "centrally located" objects, to be used as initial medoids. Now the smallest possible average distance between the objects to their mediods are selected, that forms clusters.

Step 2: SWAP-step: Try to decerase the average distance between the objects and the mediods. This is done by replacing representative objects. Now an object that does not belong to the sample is assigned to the nearest mediod.

C. Clustering large Applications based upon RANdomized Search(CLARANS)
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. One key difference between CLARANS and PAM is that the former only checks a sample of the neighbors of a node. But, unlike CLARA, each sample is drawn dynamically in the sense that no nodes corresponding to particular objects are eliminated outright. In other words, while CLARA draws a sample of nodes at the beginning of a search, CLARANS draws a sample of neighbors in each step of a search. This has the benefit of not confining a search to a localized area.
The clustering process in CLARANS (Ng & Han 2002) is formalized as searching through a certain graph where each node is represented by a set of k medoids in which two nodes are neighbors if they differ by one medoid. Each node has k(T – k) neighbors, where T is the total number of objects. CLARANS starts with a randomly selected node. It moves to a 3etermin node if a test for the maxneighbour number of neighbours is successful; otherwise it records the current node as a local minimum. If the node is found to be a local minimum, it restarts with new randomly selected

node and repeats the search for a new local minimum.The procedure continues until some threshold numlocal of local minima have been found a returns the best node. As shown in Figure 2

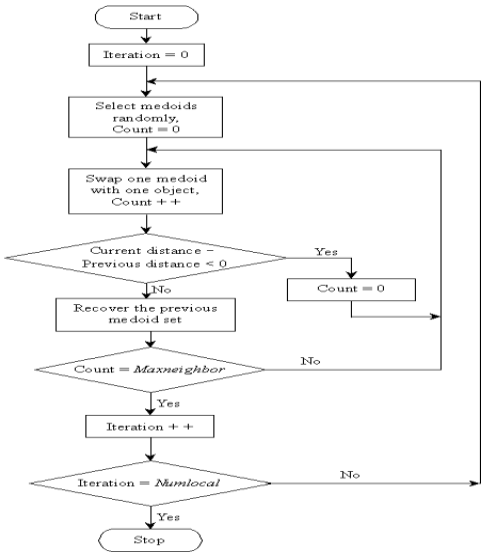

Figure 2CLARANS algorithm

The CLARANS algorithm can be 4eterminin as below:
Repeat the following steps numlocal times.
Step 1: Select a current node randomly and calculate the average distance of this current code, where node is the collection of k medoids.
Step 2: Repeat the following max neighbour times.

- Select a neighbor node randomly and calculate the average distance of this node. If the average distance is lower, set current node to be the 4neighbour node.

The computational complexity is O((β + numlocal)(T – k)) based on the number of partitions per object or O'((β + numlocal)k(T – k)) based on the number of distance calculations, where β is the number of test moves between nodes.

D. Spatial dominant approach SD (CLARANS)

In SDCLARANS, all the data containing spatial components are collected. After that clustering is used based on CLARANS. It should be mentioned that CLARANS is used to find the most natural number, of $k_{nat,}$ clusters. One may ask, how is k_{nat} determined in the first place. It is indeed a very difficult and open question. The authors however, adopt a heuristic of 4etermining $k_{nat,}$ which uses silhoutee coefflients , introduced by Kaufman and Rousseau .Each of the clusters thus obtained is processed by generalizing its nonspatial components using DBLEARN. Note that this algorithm differs from the spatial dominant generalization algorithm (without clustering), in that the latter requires the user to provide the spatial concept hierarchies. However, in this case, it can be said that SD(CLARANS) computes spatial hierarchy dynamically. The hierarchy thus found is more "data oriented" rather than "human oriented".

SD CLARANS Algorithm:

Step1 : Given a learning request, find the initial set of relevant tuples by the appropriate SQL queries.
Step 2: Apply CLARANS to the spatial attributes and find the most natural number knelt of clusters.
Step 3: For each of the k,,,,t clusters obtained above,

- collect the non-spatial components of the tuples included in the current cluster, and
- apply DBLEARN to this collection of the non-spatial components.

E. Non-spatial dominant approach NSD (CLARANS)

This nonspatial dominant approach first applies nonspatial generalizations and spatial clustering afterwards. DBLEARN is used to perform attribute-oriented generalizations of the nonspatial attributes and produce a number of generalized tuples. Then, for each such 4eneralized tuple, all the spatial components are collected and clustered using CLARANS to find k_{nat} clusters. In the final step, the clusters thus obtained are checked to see if they overlap with eachother. If so, then the clusters are merged, and the corresponding generalized tuples are merged as well. If the rules to find are nonspatial characterizations of spatial attributes, then SD(CLARANS) has an edge. This is because NSD(CLARANS) separates the objects into different groups before clustering which may weaken the inter object similarity, or cluster tightness. On the other hand, NSD(CLARANS) is suitable if the spatial clusters within groups of data that has been generalized nonspatially is sought. However, both algorithms arrive at the same result (or rules).

NSD CLARANS Algorithm:

Step 1:Given a learning request, find the initial set of relevant tuples by the appropriate SQL queries.
Step 2: Apply DBLEARN to the non-spatial attributes, until the final number of generalized tuples fall below a certain threshold.
Step 3:For each generalized tuple obtained above,

- Collect the spatial components of the tuples represented by the current generalized tuple, and
- Apply CLARANS and the heuristics presented above to find the most natural number knot of clusters.

Step 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.

VII. CONCLUSION

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. Spatial data mining tasks include: spatial classification, spatial association rule mining, spatial clustering, characteristic rules, discriminant rules, trend detection. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. All the members of the cluster has similar features. Members belong to different clusters has dissimilar features. Several clustering methods for spatial data mining include; PAM, CLARA, CLARANS,SD(CLARANS), NSD(CLARANS).

REFERENCES

[1] *Data Mining*: Introductory And Advanced Topics-Margaret H Dunham
[2] M. Holsheimer and A. Siebes. Data mining: The search for knowledge in databases. In CWI Technical Report CS-R906, Amsterdam, The Netherlands, 1994.
[3]R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. 199 mt. Conf. VLDB, pp. 487 499, Santiago, Chile, Sept. 1994.
[4] W. Lu, J. Han, and B. C. Ooi. Discovery of General Knowledge in Large Spatial Databases. In Proc. Far East Workshop on Geographic Information Systems pp. 275-289, Singapore, June 1993.
[5] Fayyad et al., "Advances in Knowledge Discovery and Data Mining", AAAI Press / MIT Press, 1996
[6] Richard C. Dubes and Anil K. Jain, (1988), Algorithms for Clustering Data, Prentice Hall.
[7] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
[8] 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.
[9] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. Proceedings of 1994 Int'l Conference on Very Large Data Bases (VLDB'94), September 1994.
[10]Spatial Data Mining using Cluster Analysis Ch.N.Santhosh Kumar1, V. Sitha Ramulu2, K.Sudheer Reddy3, Suresh Kotha4, Ch.
[11]Improved clustering and soft computing algorithms Shu chuan chu january 1 2004
[12]data mining concepta and techniques jiawei han,micheline kamber,jian pei 2001