Spatial Data Mining using Cluster Analysis

Ch.N.Santhosh Kumar[1], V. Sitha Ramulu[2], K.Sudheer Reddy[3], Suresh Kotha[4], Ch. Mohan Kumar[5]

[1]Assoc. Professor, Dept. of CSE, Swarna Bharathi Inst. of Sc. & Tech, Khammam, A.P, India.
[2]Assoc. Professor, Dept. of CSE, Swarna Bharathi Inst. of Sc. & Tech, Khammam, A.P, India.
[3]Research Scholar, Dept. of CSE, Acharya Nagarjuna University, Guntur, A.P, India.
[4]Asst. Professor, Dept. of CSE, Swarna Bharathi Inst. of Sc. & Tech, Khammam, A.P, India.
[5]Asst. Professor, Dept. of CSE, Swarna Bharathi Inst. of Sc. & Tech, Khammam, A.P, India.
santhosh_ph@yahoo.co.in, vsitaramu.1234@gmail.com, sudheercse@gmail.com, sureshk543@gmail.com, mail2mohan.ch@gmail.com

Abstract

Data mining, which is refers to as Knowledge Discovery in Databases(KDD), means a process of nontrivial exaction of implicit, previously useful and unknown information such as knowledge rules, descriptions, regularities, and major trends from large databases. 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.

Index Terms

Clustering, Database, Data mining, Spatial data.

I. INTRODUCTION

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[1,2]. 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 [1,3],

DOI : 10.5121/ijcsit.2012.4407

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[4] **.** 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[5]

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. 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.

Data mining methods [6] are not suited to spatial data because they do not support location data nor the implicit relationships between objects. Hence, it is necessary to develop new methods including spatial relationships and spatial data handling. Calculating these spatial relationships is time consuming, and a huge volume of data is generated by encoding geometric location. Global performances will suffer from this complexity. Using GIS, the user can query spatial data and perform simple analytical tasks using programs or queries. However, GIS are not designed to perform complex data analysis or knowledge discovery. They do not provide generic methods for carrying out analysis and inferring rules. Nevertheless, it seems necessary to integrate these existing methods and to extend them by incorporating spatial data mining methods. GIS methods are crucial for data access, spatial joins and graphical map display. Conventional data mining can only generate knowledge about alphanumerical properties.

III. 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 of properties 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.

IV. CLUSTER ANALYSIS

Cluster analysis[7] divides data into meaningful or useful groups (clusters). Cluster analysis is very useful in spatial databases. For example, by grouping feature vectors as clusters can be used to create thematic maps which are useful in geographic information systems.

Types of Clustering:

The collection of clusters is known as clustering. There are various types of clustering as follows.

A. Hierarchical versus Partitional

Partitional clustering is the process of dividing the data objects in to non overlapping subsets or clusters. Each object must belong to a subset. If these clusters are divides into sub clusters, then it is called hierarchical clustering. It is in the form of a tree.

B. Exclusive versus Overlapping versus Fuzzy

Exclusive clustering assign each object to a single cluster. If an object is assigned to more than one cluster then it is called non exclusive clustering. This is also called overlapping clustering. Fuzzy

clustering is defined as every object is a member of every cluster.Each object has membership weight. It is in between 0 and 1.

C. Complete versus Partial

In Partial clustering, only some objects are assigned to clusters, the remaining are un assigned. But in complete clustering, each object must be assigned to a cluster.

V. CLUSTERING METHODS FOR SPATIAL DATA MINING

A. Partitioning Around Medoids(PAM)

PAM is similar to K- means algorithm. Like k- means algorithm, PAM divides data sets in to groups but based on mediods. Whereas k- means is based on centroids. By using mediods we can reduce the dissimilarity of objects within a cluster. In PAM, first calculate the mediod, then assigned the object to the nearest mediod, which forms a cluster.

Let 'i' be the object, 'v_i ' be a cluster. Then the object i is nearer to the mediod mv_i than m_w

$d(i, mv_i) \leq d(i, m_w)$ for all $w = 1,..., k$.

The k representative objects should minimize the objective function, which is the sum of the dissimilarities of all objects to their nearest medoid:

Objective function = S $d(i, mv_l)$

The algorithm proceeds in two steps:

- 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 LARge Applications(CLARA)

Compared to PAM, CLARA[8] 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 is that it finds the entire dissimilarity matrix at a time. So for n objects, the space complexity of PAM becomes $O(n^2)$. But CLARA avoid this problem. CLARA accepts only the actual measurements (i.e,. $n \acute{} p$ data matrix).

CLARA assigns objects to clusters in the following way:

- 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.

- 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.

B. Clustering large Applications based upon RANdomized Search(CLARANS)

CLARANS[9] 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.

Algorithm CLARANS

1. Input parameters numlocal and maxneighbor. Initialize i to 1, and mincost to a large number.
2. Set current to an arbitrary node in Gn;k .
3. Set j to 1.
4. Consider a random neighbor S of current, and based on 5, calculate the cost differential of the two nodes.
5. If S has a lower cost, set current to S, and go to Step 3.
6. Otherwise, increment j by 1. If j maxneighbor, go to Step 4.
7. Otherwise, when j > maxneighbor, compare the cost of current with mincost. If current< mincost: mincost= cost of current, bestnode=current.
8. Increment i by 1. If i > numlocal, output bestnode and halt. Otherwise, go to Step 2.

Steps 3 to 6 above search for nodes with progressively lower costs. But, if the current node has already been compared with the maximum number of the neighbors of the node (specified by maxneighbor) and is still of the lowest cost, the current node is declared to be a “local” minimum. Then, in Step 7, the cost of this local minimum is compared with the lowest cost obtained so far. The lower of the two costs above is stored in mincost. Algorithm CLARANS then repeats to search for other local minima, until numlocal of them has been found.

As shown above, CLARANS has two parameters: the maximum number of neighbors examined (maxneighbor) and the number of local minima obtained (numlocal). The higher the value of maxneighbor, the closer is CLARANS to PAM, and the longer is each search of a local minima. But, the quality of such a local minima is higher and fewer local minima needs to be obtained. Based upon CLARANS, Ng and Han[10] further develope two spatial data mining algorithms: spatial dominant approach, SD(CLARANS) and nonspatial dominant approach, NSD(CLARANS).

C. 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, knat, of 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 detemining k_{nat} , which uses silhoutee coefflients , introduced by Kaufman and Rousseeuw [8]. 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:

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.

D. 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 generallized 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:

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.

VI. 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. 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] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, CA, 1996.

[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] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In Proc. th Int'l Symp. on Large Spatial Databases (SSD '95), pp. 47 66, Portland, Maine, August 1995

[6] Fayyad et al., "Advances in Knowledge Discovery and Data Mining", AAAI Press / MIT Press, 1996

[7] Richard C. Dubes and Anil K. Jain, (1988), Algorithms for Clustering Data, Prentice Hall.

[8] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.

[9] 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.

[10] 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.

AUTHORS PROFILE

Ch.N.Santhosh Kumar, working as Assoc.Professor Department of CSE at Swarna Bharathi Institute of Science & Technology (SBIT), Khammam. A.P., India. His Research Interests includes Database Management Systems, Data Mining and Network Security.

V. Sitha Ramulu, working as Assoc.Professor Department of CSE at Swarna Bharathi Institute of Science & Technology (SBIT), Khammam. A.P., India. His Research Interests includes Database Management Systems, Data Mining and Network Security.

K.Sudheer Reddy, Research Scholar, Dept. of CSE at Acharya Nagarjuna University, Guntur, A.P, India. His Research Interests includes Database Management Systems, Data Mining and Network Security.

Suresh Kotha, working as Asst.Professor Department of CSE at Swarna Bharathi Institute of Science & Technology (SBIT), Khammam. A.P., India. His Research Interests includes Database Management Systems, Data Mining and Network Security.

Ch. Mohan Kumar, working as Asst.Professor Department of CSE at Swarna Bharathi Institute of Science & Technology (SBIT), Khammam. A.P., India. His Research Interests includes Database Management Systems, Data Mining and Network Security.