Application of Spatial Data Mining for Agriculture

D.Rajesh
AP-SITE, VIT University, Vellore-14

ABSTRACT
The research of spatial data is in its infancy stage and there is a need for an accurate method for rule mining. Association rule mining searches for interesting relationships among items in a given data set. This paper enables us to extract pattern from spatial database using k-means algorithm which refers to patterns not explicitly stored in spatial databases. Since spatial association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly. An interesting mining optimization method called progressive refinement can be adopted in spatial association analysis. The method first mines large data sets roughly using a fast algorithm and then improves the quality of mining in a pruned data set. The k-means algorithm randomly selects k number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges. The above concept is applied in the area of agriculture where giving the temperature and the rainfall as the initial spatial data and then by analyzing the agricultural meteorology for the enhancement of crop yields and also reduce the crop losses.

Keywords: Spatial Mining, K-means, agriculture

1. INTRODUCTION
Spatial Data Mining is the discovery of interesting patterns from large geospatial databases. It refers to the extraction of knowledge, spatial relationships or other interesting patterns not explicitly stored in spatial databases. Since spatial association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly. An interesting mining optimization method called progressive refinement can be adopted in spatial association analysis. The method first mines large data sets roughly using a fast algorithm and then improves the quality of mining in a pruned data set [5]. The k-means algorithm randomly selects *k* number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges. This paper deals with getting the temperature and the rainfall as the initial spatial data and there by analyzing the Agricultural meteorology for the enhancement of crop yields and the reduction of crop losses based on the k-means algorithm [2]. The data objects can be considered of any dimensions, for the simplicity here we have considered two dimensions. The data objects are clustered or grouped based on the principle of maximizing intra class similarity and minimizing interclass similarity. Each cluster can be viewed as class of objects from which rules can be derived. All classes and methods in this paper are built according to the java coding style and naming convention [3]. The function names and class names describes themselves. The paper has been modularized to various modules in order to have a clear design, understanding and efficient coding.

2. ESSENCE OF APPROACH

2.1. Basic Facts:

Data mining has attracted a great attention in the information industry and in society as a whole in recent years, due to wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for application ranging from market analysis, fraud detection, to production control, disaster management and science exploration.

Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of various functionalities: data collection and database creation, database management (including data storage and retrieval, and database transaction processing and advance data analysis
Knowledge discovery as a process consists of an iterative sequence of following steps:

1. *Data cleaning,* that is, to remove noise and inconsistent data.
2. *Data integration,* that is, where multiple data sources are combined.
3. *Data selection,* that is, where data relevant to the analysis task are retrieved from the database.
4. *Data transformation*, that is, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. *Data mining,* that is, an essential process where intelligent methods are applied in order to extract the data patterns.
6. *Knowledge presentation,* that is, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2.2. Spatial Data Mining
Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering of

discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns.

Efficient tools for extracting information from geo- spatial data are crucial to organizations which make decisions based on large spatial datasets, including NASA, the National Imagery and Mapping Agency (NIMA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT). These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology. Specific features of geographical data that preclude the use of general purpose data mining algorithms are: i) rich data types(e.g., extended spatial objects) ii) implicit spatial relationships among the variables, iii) observations that are not independent, and iv) spatial autocorrelation among the features.

. Association rule mining searches for interesting relationships among items in a given data set. The discovery of association relationships among huge amount of data is useful in many applications. Association rule mining consists of first finding frequent itemsets, from which strong association rules in the form A => B are generated. These rules also satisfy a minimum confidence threshold.

2.3. Cluster analysis:

The finding of frequent itemsets is done by Cluster Analysis. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects [4]. One such clustering method is partitioning method, in which it creates an initial set of k partitions, where parameter k is the number of partitions to construct; then it uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

It is the most well known commonly used centroid based technique that takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is but intercluster similarity is low. Clustering similarity is measured in regard to the mean value of the objects in cluster which can be viewed cluster's centroid or center of gravity.

The kmeans algorithm which is used in this paper, randomly selects *k* number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges.

The algorithm attempts to determine k partitions that minimize the squared error functions. The method is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is O(nkt), where n is the total number of objects, k is the number of clusters, and t is the number of iterations. The method often terminates at the local optimum. K-means is the most well known commonly used centroid based technique that takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is but intercluster similarity is low. Clustering similarity is measured in regard to the mean value of the objects in cluster which can be viewed cluster's centroid or center of gravity [2].

The algorithm randomly selects *k* number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges.

K-Means Algorithm: The algorithm for partitioning, where each cluster's center is represented by mean value of objects in the cluster.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

1. arbitrarily choose k objects from D as the initial cluster centers.
2. **repeat**
3. (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster;.
4. update the cluster means,i.e. calculate the mean value of the objects for each cluster;
5. **until** no change;

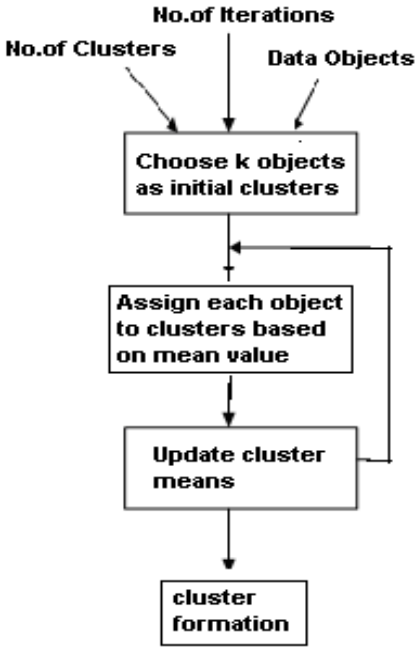

Fig.1: Cluster Formation using K-means

3. SUMMARY OF RESULTS

The algorithm attempts to determine k partitions that minimize the squared error functions. It works well clusters are compact clouds that are rather well separated from one another. The method is relatively scalable and efficient in processing large

datasets because the computational complexity of the algorithm is O(nkt), where n is the total number of objects, k is the number of clusters, and t is the number iterations. Normally, k<<n and t<<n. The discovery of association and sequential patterns in multidimensional analysis can be helpful for starting point for further exploration, making them a popular tool for understanding data.

The system will produce the Cluster formation as the output with the association rules generated for each cluster based on the threshold value. The produces six clusters. The system also uses some Graphical user interface to work on data and for the user view.

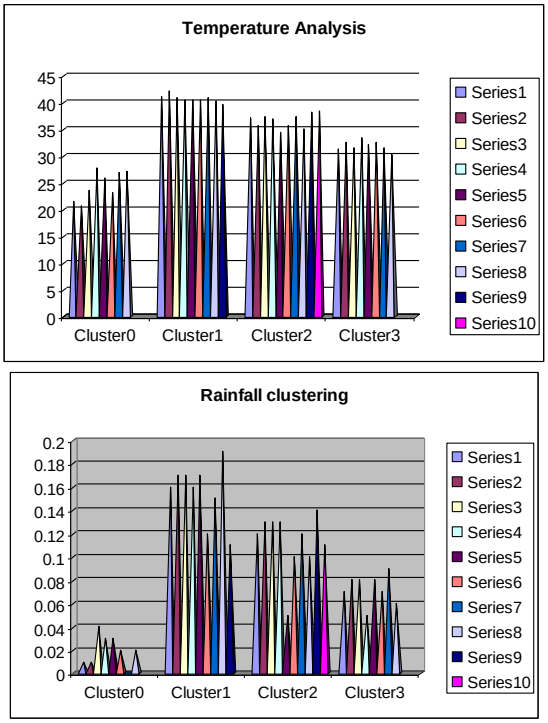

Figure 1.1 Cluster Analysis Graph

The algorithm attempts to determine k partitions that minimize the squared error functions. It works well clusters are compact clouds that are rather well separated from one another. The method is relatively scalable and efficient in processing large datasets because the computational complexity of the algorithm is O(nkt), where n is the total number of objects, k is the number of clusters, and t is the number iterations. Normally, k<<n and t<<n. The discovery of association and sequential patterns in multidimensional analysis can be helpful for starting point for further exploration, making them a popular tool for understanding data.

4. CONCLUSION AND FUTURE WORK

Association rule mining from spatial data mining is a topic of much importance and many applications. Methods of data mining are under research. This paper has provided an overview of data clustering method using cluster analysis and there by generates patterns/rules. ”Association Rule Mining Analysis” usually sounds like something very smart, difficult to understand, something that is useful only to those researchers and professors wearing thick glasses. But the reality is just opposite. Although we might not be aware of it, pattern analysis using association rule mining is present in many aspect of our everyday life.

The paper considers only two dimensions and on the basis of these two dimensions it clusterizes the data objects. In future the paper can be implemented for more than two dimensions. The current method to find the distance between the data point and the cluster is Euclidean distance. These methods give circular clusters. In future new mathematical functions can be derived and various kinds of polygon clusters can be obtained. Inputs to the paper are embedded in the source code itself. Oracle spatial architecture can also be used to input the data.

Currently the paper only deals with the text input. In future it will be enhanced to include some new algorithms to clusterize large data like graphical maps. In future the paper can be enhanced to include graphical user interface to carry out the clustering.

5. REFERENCES

[1]. M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li,. “New algorithms for fast discovery of association rules”, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, page283, 1997.

[2].Jiawei Han, Member and Yongjian Fu, Member, “Mining Multiple-Level Association Rules in Large Databases” , ieee transactions on knowledge and data engineering, vol 11, no.5, September/October, 2000.

[3]. Sam Y. Sung, Member, IEEE Computer Society, Zhao Li, Chew L. Tan, and Peter A. Ng, “Forecasting Association Rules Using Existing Data Sets”, ieee transactions on knowledge and data engineering, vol 15, no.6, November/December 2003.

[4] Chi-Farn Chen; Ching-Yueh Chang; Jiun-Bin Chen **“**Spatiial knowledge discovery using spatial datamining method”,Geoscience and Remote Sensing Symposium, ieee International Volume 8, Issue 25, Page(s): 5602 - 5605 July 2005

[5]. Eun-Jeong Son, In-Soo Kang, Tae-Wan Kim, Ki-Joune Li, “A Spatial Data Mining Method by Clustering Analysis”, Proceedings of the 6th International Symposium on Advances in Geographic Information Systems, 1998.