Clustering Algorithm for Spatial Data Mining: An Overview

A.Padmapriya,
M.C.A.,M.Phil.,Ph.D
Department of Computer Science and Engineering
Alagappa University
Karaikudi

N.Subitha
Research scholar
Department of Computer Science and Engineering
Alagappa University, Karaikudi

ABSTRACT
Spatial data mining practice for the extraction of useful information and knowledge from massive and complex spatial database. Most research in this area has focused on efficient clustering algorithm for spatial database to analyze the complexity. This paper introduces an active spatial data mining approach that extends the current spatial data mining algorithms to efficiently support user-defined triggers on dynamically evolving spatial data. It shows that spatial data mining is a promising field, with fruitful research results and many challenging issues.

Keywords
Spatial data mining, Spatial database, K-mean, Spatial relationship, Datamining.

1. INTRODUCTION
Data mining is the process of discovering interesting, knowledge such as patterns, associations, changes, anomalies and significant structures, from large amount of data stored in database, data warehouses or other information repositories [6]. Due to the wide availability of huge amounts of data in electronic forms, the imminent need for turning such data into useful information and knowledge for broad application including market analysis, business management, and decision support, data mining has attracted a great deal of attention in information industry in recent years[7]. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include

- association-patterns where one event is connected to another event, such as purchasing a pen and purchasing paper,
- sequence or path analysis -patterns where one event leads to another event, such as the birth of a child and purchasing diapers,
- classification-identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases,
- clustering-finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences,
- forecasting-discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes

Data mining has been popularly treated as a synonym of knowledge discovery in database although some researchers view data mining as an essential step of knowledge discovery.

In general, a knowledge discovery process consists of an iterative sequence of the following step

1. Data cleaning, which handles noisy, erroneous, missing, or irrelevant data.

2. Data integration, where multiple, heterogeneous data source may be integrated into one.

3. Data selection, where data relevant to the analysis task are retrieved from the database.

4. Data transformation, where data are transferred or consolidated into forms appropriate for mining by performing summary or aggregation operations.

5. Data mining, which is an essential process where intelligent methods are applied in order to extract data patterns.

6. Pattern evaluation which is to identify the truly interesting patterns representing knowledge based on some interestingness measures.

7. Knowledge presentation, where visualization and knowledge representation technique are used to present the mined knowledge to the user.

With the widely available relational database system and data warehouses, the four processes: data cleaning, data integration, data selection, and data transformation, can be performed by constructing data warehouses and performing some OLAP operations on the constructed data warehouses. The data mining, pattern evaluation and knowledge presentation processes are sometimes integrated into one (possibly iterative) process, referred as data mining [5].

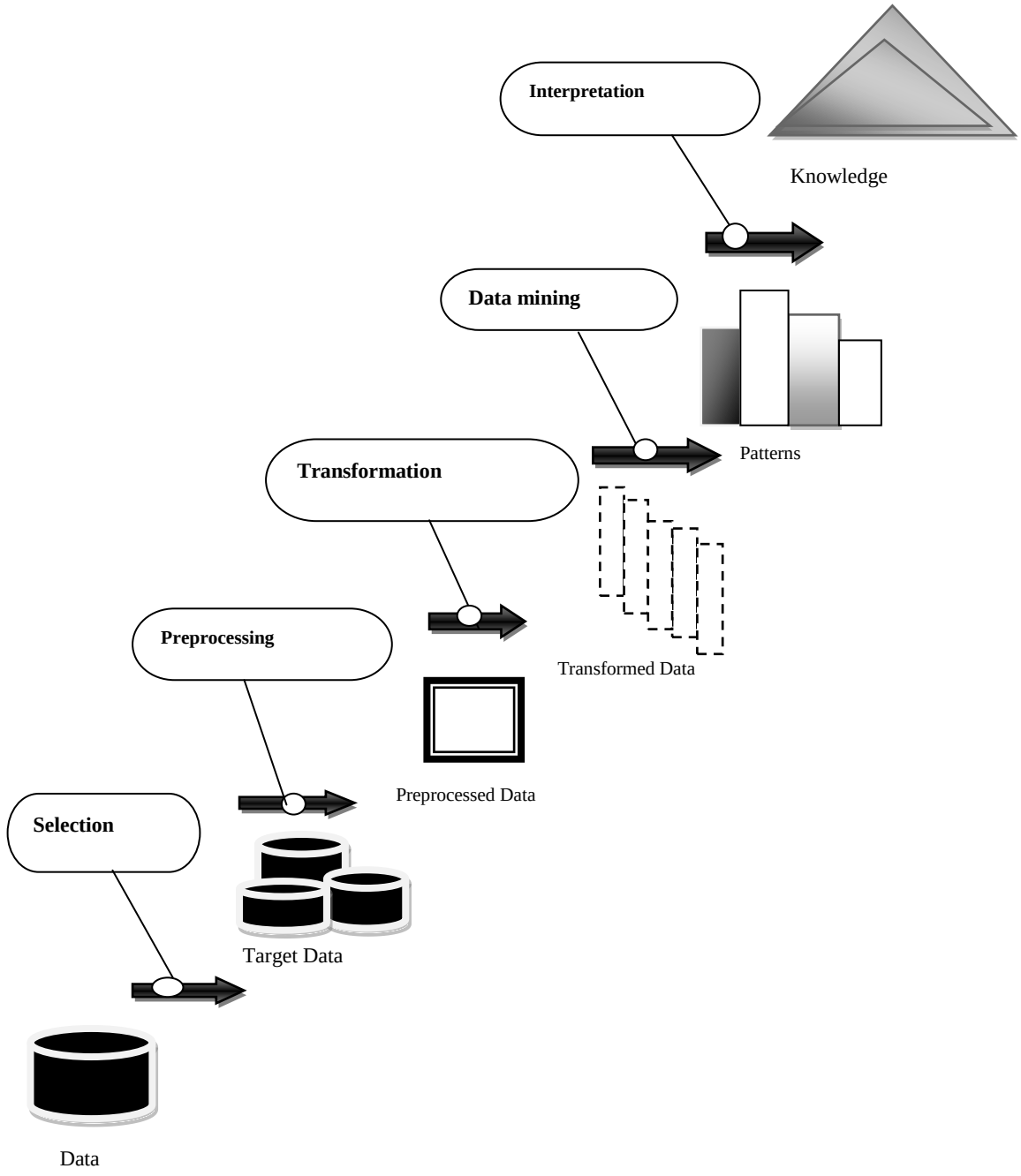

Figure l: An overview of the steps comprising the KDD process

2. BACKGROUND STUDY

In general, data mining tasks can be classified into two categories: descriptive data mining and predictive data mining. The former describes the data set in a concise and summary manner and presents interesting general properties of the data whereas the latter construct one or a set of models, performs inference on the available set if data, and attempts to predict the behavior of new data sets.

A data mining system may accomplish one or more of the following data mining tasks [1, 4].

1. **Class description**. Class description provides a concise and succinct summarization of a data and distinguishes it from others .The summarization of a collection of data is called class characterization; whereas the comparison between two or more collections of data is called class comparison or discrimination. Class description should cover not only its summary properties, such as count,

sum, and average, but also its properties on data dispersion, such as variance, quartile, etc.

For example, class description can be used to compare European versus Asian sales of a company, identify the important factors which discriminate the two classes, and present a summarized overview.

2. **Association**. Association is the discovery of association relationships or correlations among a set of items. They are often expressed in the rule form showing attribute-value conditions that occur frequently together in a given set of data. An association rule in the form of X→Y is interpreted as "database tuples that satisfy X are likely to satisfy Y".
   Association analysis is widely used in transaction data analysis for directed marketing, catalog design, and other business decision making process.

Substantial research has been performed recently on association analysis with efficient algorithms proposed, including the level-wise Apriori search, mining multiple-level, multi-dimensional associations, mining associations for numerical, categorical, and interval data, meta-pattern directed or constraint-based mining, and mining correlations.

3. **Classification.** Classification analyzes a set of training data (i.e., a set of objects whose class label is known) and constructs a model for each class based on the features in the data. A decision tree rules is generated by such a classification process, which can be used for better understanding of each class in the database and for classification of future data [1]. For example, one may classify diseases and help predict the kind of diseases based on the symptoms of patients.

There have been many classification methods developed in the fields of machine learning, statistics, database, neural network, rough sets, and others. Classification has been used in customer segmentation, business modeling, and credit analysis.

4. **Prediction**. This mining function predicts the possible values of some missing data or the value distribution of certain attributes in a set of objects. It involves the finding of the set of attributes relevant to the attribute of interest (e.g., by some statistical analysis) and predicting the value distribution based on the set of data o f data similar to the selected objects. For example, an employee's potential salary can be predicted based on the salary distribution of similar employees in the company. Usually, regression analysis, generalized linear model, correlation analysis and decision trees are useful tools in quality prediction. Genetic algorithms and neural network models are also popularly used in prediction.

5. **Clustering**. Clustering analysis is to identify clusters embedded in the data, where a cluster is a collection of data objects that are "similar" to one another. Similarity can be expressed by distance functions, specified by users or experts. A good clustering method produces high quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high [10]. For example, one may cluster the houses in an area according to their house category, floor area, and geographical locations.

Data mining research has been focused on high quality and scalable clustering methods for large databases and multidimensional data warehouses.

6. **Time-series analysis**. Time-series analysis is to analyze large set of time-series data to find certain regularities and interesting characteristics, including search for similar sequence or subsequence patterns, periodicities, trends and deviations. For example, one may predict the trend of the stock values for a company based on its stock history, business situation, competitor's performance, and current market.

There are also other data mining tasks, such as outlier analysis, etc. Identification of new data mining tasks to make better use of the collected data itself is an interesting research topic.

Applications

Data mining is a young discipline with wide and diverse applications, there is still a nontrivial gab between general principles of data mining tools for particular applications.

1. Biomedical and DNA Data Analysis.
2. Financial Data Analysis.
3. Retail Industry.
4. Telecommunication Industry.

3. SPATIAL DATA MINING

Spatial data are the data that have spatial or location component, and they the information, which is more complex than classical data. A spatial database stores spatial data represents by spatial data types and spatial relationship and among data [6, 8].
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[8] etc.

Data Attributes

DATA = the (WHAT) dimension determines an attribute of an object.
SPATIAL DATA = (WHERE) & (WHAT) denotes attribute data referenced to a specific location.
The Attributes of spatial objects are highly dependent on location and often influenced by neighboring objects.

Spatial Database

A 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 database carry topological and or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques [12].

Spatial Data mining

Spatial data mining is the process [18] of discovering interesting and previously un-known, 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. Spatial data mining , i.e., mining knowledge from large amounts of 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 system(GIS), computer cartography, environmental assessment and planning[8] 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.

(A) Spatial Data Mining Methods

Spatial data mining has to perform various methods some of them are mentioned below

1. Generalization Based Knowledge Discovery
2. Clustering Methods
3. Aggregate Proximity Measuring
4. Spatial Association Rules

Among the four methods the research is based on clustering method.

Goals

There are different goals of spatial data mining are ordered below,

- Understanding spatial data
- Discovering spatial relationships and relationships between spatial and non-spatial data
- Constructing spatial knowledge bases
- Reorganizing spatial databases
- Optimizing spatial queries

(B)Challenges of Spatial Data Mining

Spatial Data mining must efficiently overcome the following challenges [11, 14]:

1. A crucial challenge to spatial data mining is the exploration of efficient spatial data mining techniques.
2. Huge amount of spatial data.
3. Complexity of spatial data types and spatial access methods.

(C) Applications

Some of the applications of spatial data mining are listed below,

- Geographic information systems,
- Geo marketing
- remote sensing
- image database exploration
- medical imaging
- navigation
- traffic control
- environmental studies

(D) Clustering Methods

The collection of clusters is known as clustering.
Goal: like Generalization, to reveal relationships between spatial and non-spatial attributes

There are various types of clustering as follows

1. Hierarchical Methods

It can have two types of algorithms they are [9],

- Agglomerative Algorithm
- Divisive Algorithm

2. Partitioning Methods

It can contain many types of algorithms they are [10],

- Nearest Neighbor Algorithm
- Density Based Algorithm
- K-Medoids Methods
- K-Mean Methods

3. Grid Based Methods

4. Methods Based on Co-occurrence of Categorical Data.

5. Density Based methods.

4. ENHANCED K-MEANS ALGORITHM ON SPATIAL DATASET

K-Means algorithm introduced by J.B. Mac Queen in 1967, is one of the most common clustering algorithms and it is considered as one of the simplest unsupervised learning algorithms that partition feature vectors into ***k*** clusters so that the within group sum of squares is minimized.

There are several variants of the k-means clustering algorithm, but most variants involve an iterative scheme that operates over a fixed number of clusters, while attempting to satisfy the following properties: Each class has a center which is the mean position of all the samples in that class.

PROCEDURE OF K-MEAN ALGORITHM

Step 1: Place randomly initial group centroids into the ***2d*** space.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: Recalculate the positions of the centroids.

Step 4: If the positions of the centroids didn't change

go to the next step,

Else go to Step 2.

Step 5: End

(A) Working

It accepts the number of clusters to group data into, and the dataset to cluster as input values. It then creates the first ***K*** initial clusters (***K***= number of clusters needed) from the dataset by choosing ***K*** rows of data randomly from the dataset.

For Example, if there are 10,000 rows of data in the dataset and 3 clusters need to be formed, then the first ***K***=3 initial clusters will be created by selecting 3 records randomly from the dataset as the initial clusters [14, 15]. Each of the 3 initial clusters formed will have just one row of data.

The K-Means algorithm calculates the Arithmetic Mean of each cluster formed in the dataset. The Arithmetic Mean of a cluster is the mean of all the individual records in the cluster. In each of the first K initial clusters, there is only one record [16]. The Arithmetic Mean of a cluster with one record is the set of values that make up that record.

For Example if the dataset we are discussing is a set of Height, Weight and Age measurements for students in a University, where a record P in the dataset S is represented by a Height, Weight and Age measurement, then

P = {Age, Height, Weight).

Then a record containing the measurements of a student John, would be represented as

John = {20, 170, 80}

Where
John's Age = 20 years,
Height = 1.70 meters and
Weight = 80 Pounds.

Since there is only one record in each initial cluster then the Arithmetic Mean of a cluster with only the record for John as a member = {20, 170, 80}.

It Next, K-Means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity like the Euclidean Distance Measure or Manhattan/City-Block Distance Measure.

We have to re-assigns each record in the dataset to the most similar cluster and re-calculate the arithmetic mean of all the clusters in the dataset. The arithmetic mean of a cluster is the arithmetic mean of all the records in that cluster.

For Example, if a cluster contains two records where the record of the set of measurements for

John = {20, 170, 80} and
Henry = {30, 160, 120},

Then the arithmetic mean ***P*** mean is represented as

P mean = {Age mean, Height mean, Weight mean).

Age mean = (20 + 30)/2,
Height mean = (170 + 160)/2 and
Weight mean = (80 + 120)/2.

The arithmetic mean of this cluster = {25, 165, 100}.

This new arithmetic mean becomes the center of this new cluster. Following the same procedure, new cluster centers are formed for all the existing clusters.

It K-Means re-assigns each record in the dataset to only one of the new clusters formed. A record or data point is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity like the Euclidean Distance Measure or Manhattan/City-Block Distance Measure. The preceding steps are repeated until stable clusters are formed and the K-Means clustering procedure is completed [17]. Stable clusters are formed when new iterations or repetitions of the K-Means clustering algorithm does not create new clusters as the cluster center or Arithmetic Mean of each cluster formed is the same as the old cluster center. There are different techniques for determining when a stable cluster is formed or when the k-means clustering algorithm procedure is completed.

(A) **Computational complexity**

NP-hard in general Euclidean space d even for 2 clusters. NP-hard for a general number of clusters ***k*** even in the plane. If k and ***d*** are fixed, the problem can be exactly solved in time O (***n*** ***dk***+1 log ***n***), where ***n*** is the number of entities to be clustered.

It has some of the advantages are **r**elatively efficient: O (***tkn***), where n is the number of instances, c is the number of clusters, and t is the number of iterations. Normally, ***k, t*** << ***n***. Often terminates at a local optimum. The global optimum may be found using techniques such as: simulated annealing or genetic algorithms

Also has some disadvantages it's applicable only when mean is defined.

- Need to specify ***c***, the number of clusters, in advance.
- Unable to handle noisy data and outliers.
- Not suitable to discover clusters with non-convex shapes.

5. CONCLUSION

Data mining/ Knowledge Discovery of spatial Data is a large, active field of research with wide application in GIS, remote sensing, medical imaging, traffic control, environmental studies etc. Although, the field is quite young, a number of algorithms and techniques have been proposed to discover various kinds of knowledge from spatial data with the help of K-means clustering algorithm. This work motivated us and gives future direction towards designing an efficient clustering algorithm for spatial database with reduced complexity. The variety of yet unexplored topics and problems makes knowledge discovery in spatial database an attractive and challenging research field.

6. REFERENCES

[1] R. Agrawal, M. Mehta, J. Shafer, R. Srikant, A. Arning, T. Bollinger. The Quest Data Mining System. Proceedings of 1996 International Conference on Data Mining and Knowledge Discovery(KDD'96), Portland, Oregon, pp. 244-249, August 1996.

[2] K. Alsabti, S. Ranka, and V. Singh, ªAn Efficient k-means Clustering Algorithm,º Proc. First Workshop High Performance Data Mining, Mar. 1998

[3] P. S. Bradley, U. Fayyad, and C. Reina, "Scaling Clustering Algorithms to Large Databases", Proc. 4 th International Conf. on Knowledge Discovery and Data Mining (KDD-98). AAAI Press, Aug. 1998

[4] M. S. Chen, J. Han, and P.S.Yu. Data Mining: An Overwiew from a Database Perspective. IEEE Transcations on Knowledge and Data Engineering, 8(6):883, 1996.

[5] Dan Pelleg and Andrew W. Moore. Accelerating exact k-means algorithms with geometric reasoning. In KDD, pages 277–281, 1999.

[6] Ester M., Kriegel H.-P., and Sander J. 1997 "Spatial Data Mining: A Database Approach", Proc. 5th Int. Symp. on Large Spatial Databases, Berlin, Germany, pp. 47-66.

[7] U. M. Fayyades, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[8] W. Lu, J. Han, and B. C. Obi. Discovery of General Knowledge in Large Spatial Databases. In Proc. Far East Workshop on Geographic Information Systems pp. 275-289, Singapore, June 1993

[9] G. Karypis, E.-H. Han, and V. Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling," Computer, vol. 32, no. 8, pp 68–75, Aug. 1999

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

[11] Koperski K. Adhikary J., Han J. 1996 "Knowledge Discovery in Spatial Databases: Progress and Challenges", Proc. SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Technical Report 96-08, University of British Columbia, Vancouver,Canada.

[12] 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

[13] Krzysztof Koperski, Junas Adhikary, JiaweiHan. Spatial Data Mining: Progress and Challenges Survey Paper. Workshop on Research Issues on Data Mining and Knowledge Discovery, 1996

[14] G. Milligan and M. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set,"Psychometrika, vol. 50, pp. 159–179, 1985

[15] Paul S. Bradley and Usama M. Fayyad. Refining initial points for k-means clustering. In Jude W. Shavlik, editor, ICML, pages 91–99. Morgan Kaufmann, 1998.

[16] Raymond T. Ng and Jiawei Han, CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE TRANSACTIONS ON KNOWLEDGE and DATA ENGINEERING, Vol. 14, No. 5,

[17] Shai Ben-David, D´avid P´al, and Hans Ulrich Simon. Stability of k-means clustering. Lecture Notes in Computer Science, 4539:20–34, 2007

[18] Shekhar, S., and Chawla, S. 2003. Spatial Databases A Tour. Prentic e Hall (ISBN 0-7484-0064-6).

[19] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. An efficient k-means clustering algorithm Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell., 24(7):881–892, 2002.