Comparative Study of Spatial Data Mining Techniques

Kamalpreet Kaur Jassar
Research Scholar
BBSBEC, Dept. Of CSE,
Fatehgarh Sahib, Punjab, India

Kanwalvir Singh Dhindsa,Ph.D
Associate Professor
BBSBEC, Dept. Of CSE,
Fatehgarh Sahib, Punjab, India

ABSTRACT
Spatial data mining is a mining knowledge from large amounts of spatial data. Spatial data mining algorithms can be separated into four general categories: clustering and outlier detection, association and co-location method, trend detection and classification. All these methods have been compared according to various attributes. This paper introduces the fundamental concepts of widely known spatial data mining algorithms in a comparative way. It focuses on techniques and their unique features.

General Terms
Spatial data mining, spatial database, Clustering, geographic data

Keywords
Spatial clustering, Clustering and Outlier Detection, Association and Co-Location, Classification, Trend-Detection, Clustering algorithms, Knowledge Discovery in Database

1. INTRODUCTION
Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets [2]. Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial Databases [5]. It is a highly demanding field because large amounts of spatial data have been collected in various applications i.e. ranging from remote sensing to geographical information systems (GIS), environmental assessment, computer cartography, and planning. Recent studies on spatial data mining have extended the scope of data mining from relational and transactional databases to spatial databases. This paper summarizes comparison of spatial data mining techniques. It shows that spatial data mining using clustering is a promising field also gives fruitful research results and many challenging issues. The explosive growth of spatial data and extensive use of spatial databases emphasize the need for the computerized discovery of spatial knowledge [9]. Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters .A clustering procedure could group the customers in such a way that customers with similar buying patterns are in the same cluster. Thus, the main concern in the clustering process is to reveal the organization of patterns into "sensible" groups, which allow us to discover similarities and differences, as well as to derive useful conclusions about them. This idea is applicable in many fields, such as life sciences, medical sciences and engineering. Clustering may be found under different names in different contexts, such as unsupervised learning (in pattern recognition), numerical taxonomy (in biology, ecology), typology (in social sciences) and partition (in graph theory)[3].

2. RELATED WORK
Sumathi and Geetha et al. [2] presented the techniques of spatial data mining in the following four categories: Clustering and Outlier Detection, Association and Co-Location, Classification and Trend-Detection. It also discussed some trends and applications of spatial data mining. **Sisodia et al. [4]** explained about Clustering that it is an unsupervised learning task where one seeks to identify a finite set of categories termed clusters to describe the data . Various clustering algorithms of data mining have been considered and it also focuses on the clustering basics, requirement, classification, problem and application area of the clustering algorithms. Spatial data mining as a knowledge discovery process has been explained by **Ng and Han [5]**. Thus, it plays an important role in a) extracting interesting spatial patterns and features; b) capturing intrinsic relationships between spatial and non-spatial data; c) presenting data regularity concisely and at higher conceptual levels; and d) helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance.

Maria and Yannis Batistakis et al. [3] explained the fundamental concepts of clustering while it surveys the widely known clustering algorithms in a comparative way. Moreover, it addresses an important issue of clustering process regarding the quality assessment of the clustering results. This is also related to the inherent features of the data set under concern. Various clustering algorithms have been compared according to different attributes.**P. Murugavel et al. [6]** proposed three clustering algorithms (PAM, CLARA and CLARAN) for outlier detection. Outliers detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data. It has many uses in applications like fraud detection, network intrusion detection and clinical diagnosis of diseases. Using clustering algorithms for outlier detection is a technique that is frequently used. The clustering algorithms consider outlier detection only to the point they do not interfere with the clustering process. In these algorithms, outliers are only by-products of clustering algorithms and they cannot rank the priority of outliers.**Dr. M. Hemalatha et al. [9]** focuses on the sole features that distinguish spatial data mining from traditional data mining. Major activities and research needs in spatial data mining research are discussed. And it gives the Applications and Techniques, Issues and challenges on spatial data mining. It shows that spatial data mining is a promising field with rich research results and many challenging issues.

3. SPATIAL DATA MINING ARCHITECTURE
Various architectures (models) have been proposed for data mining. . They include Han's architecture for general data mining prototype DBLEARN/DBMINER , Holsheimer *et al*'s parallel architecture Matheus *et al.*'s[15] multi component architecture. Almost all of these architectures have been used or extended to handle spatial data mining. Matheus *et al.*'s architecture seems to be very general and has been used by

other researchers in spatial data mining. This architecture comparable to others is presented in Figure 1. In this architecture, the user may control every step of the mining process. The spatial data mining can be used to understand spatial data, discover the relation between space and the nonspace data, set up the spatial knowledge base, excel the query, reorganize spatial database and obtain concise total characteristic etc .The systematic structure of the spatial data mining[9] can be divided into three layer structures mainly, as shown in Fig 1. The first layer i.e. customer interface layer is used for input and output. This layer is also known as DB interface or user face. Data is fetched from the storage using the *DB interface* which enables optimization of the queries. The second layer miner layer is mainly used for management of data, selection of algorithm and to store the mined knowledge. For example, it may decide that only some attributes are relevant to the knowledge discovery task, or it may extract objects whose usage promises good results. Third and last data resource layer, which includes the spatial database and other related data and knowledge bases which is original data of the spatial data mining. Data inputs of spatial data mining are more complex than the inputs of classical data mining because they include extended objects such as lines, polygons and points. The data inputs of spatial data mining have two distinct types of attributes: spatial attribute and non-spatial attribute. Spatial attributes are used to define the spatial location and extent of spatial objects. Spatial attributes of a spatial object include information related to spatial locations, such as elevation, latitude and longitude, as well as shape. Non-spatial attributes are used to characterize non-spatial features of objects, e.g. name, population, and unemployment rate of city. These attributes are the same as the attributes used in the data inputs of classical data mining [9].

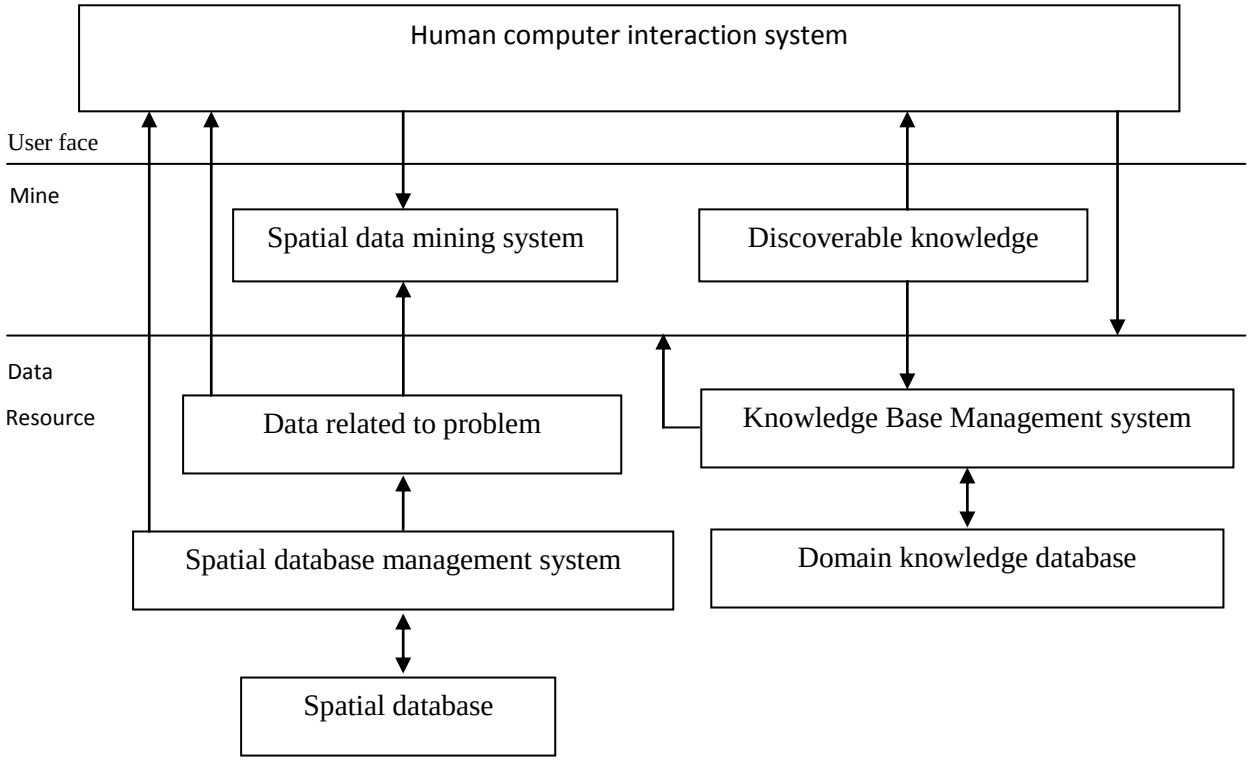

Fig. 1 Structure of spatial data mining [9]

4. SPATIAL MINING TECHNIQUES

There are various kinds of patterns that can be discovered from databases and can be presented in many different forms. Based on general data mining it can be classified into four main categories: clustering and outlier detection, association and co-location method, classification and trend detection [2]

4.1 Clustering and Outlier Detection

Spatial clustering is a process of grouping a set of spatial objects into groups, these groups are called clusters. Objects within a one cluster show a high degree of similarity, whereas the objects present in another clusters are as much non-similar as possible [2]. Clustering is a very well known technique to deal with the large geographical datasets. Clustering algorithms can be divided into four general categories: hierarchical method, partitioning method, grid-based method and density-based method [4].

4.2 Association and Co-Location

When clustering methods are performed on the data, we can find only characteristic rules, which describe spatial objects according to their non-spatial attributes. There are many situations in which we want to discover spatial rules that associate one or more spatial objects with others. One of the biggest research challenges is the development of methods for selecting potentially interesting rules from among the mass of all discovered rules in mining association rules [2].

4.3 Classification

Every data object i.e. stored in a database is characterized by its attributes. Therefore Classification is a technique, which is used to find rules that describe the partition of the database into an explicitly given set of classes. It is considered as predictive spatial data mining, because a model is created first according to which the whole dataset is analyzed.

4.4 Trend Detection

A spatial trend is defined as a regular change of one or more than one non-spatial attributes when spatially moving away from a start object. Therefore, spatial trend detection technique used to find patterns of the attribute changes with respect to the neighborhood of some spatial object.

5. COMPARISON AND DISCUSSION

This section offers an overview of the main characteristics of the clustering algorithms presented in a comparative way.

Table 1. Comparison of various Spatial Mining Methods

Characteristics	**Clustering And Outlier Detection**	**Association And Co-Location**	**Classification**	**Trend-Detection**
Summarization	No pre-defined classification required.	Pre-defined classification	Predefined classification	No predefined classification required
Machine Learning	Unsupervised	Association rules	Supervised	Trend rules
Performance	High	Medium	Low	Medium
Mining Class	Descriptive	Descriptive	Predictive	Predictive
Robust	Yes	No	No	No
Tool	Geo-SOM, SDmine	Co-Location Miner	WEKA	Fuzzy Logic
Algorithm	K-means, PAM, CLARA, CLARAN etc.	Apriori Algorithm etc.	k-nearest neighbor classifier, Apriori SVM etc	Global Trend, Local Trend etc

Table 1 depicts the main concepts and the characteristics of the most representative algorithms. First characteristic is Summarization of above techniques. Next characteristic is Machine Learning which means to get knowledge from study, experience or being taught. This is of two types: supervised and unsupervised. Unsupervised learning model is not provided with the correct results during training. In supervised model Training data includes both the input and the desired results. Another characteristic is performance which is high in clustering technique as compared to others. Mining class may be classified into predictive and descriptive. Predictive analysis is to analyze current and historical facts to make predictions about future events. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Another characteristic is robust which means to avoid unsatisfactory results. As shown in Table 1. Only clustering technique is able to handle noise and outlier while others are not. This is the main advantage of clustering method which also leads to raise the level of performance.

6. CONCLUSION

Clustering or cluster analysis is one of the major tasks in various research areas. However, it may be found under different names in different contexts such as unsupervised learning [12] in pattern recognition, partition in graph theory and taxonomy in biology. The clustering aims at identifying and extract significant groups in underlying data. Thus based on a certain clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters. Clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed and are available in literature. In this paper we presented the main characteristics and applications of clustering algorithms. Moreover, the different categories in which algorithms can be classified i.e., partitioning, hierarchical, density-based, grid-based, fuzzy clustering [3]. Another important issue that we discussed in this paper is the cluster validity. This is related to the inherent features of the data set under concern. The majority of algorithms are based on certain criteria in order to define the clusters in which a data set can be partitioned. Since clustering is an unsupervised method and there is no a-priori indication for the actual number of clusters presented in a data set, there is a need of some kind of clustering results validation. We concluded the discussion on clustering algorithms by a comparative presentation and stressing the pros and cons of each category.

7. REFERENCES

[1] 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, Sept-Oct 2002.

[2] N. Sumathi, R. Geetha and Dr. S. Sathiya Bama, "spatial data mining-techniques trends and its applications", Journal of Computer Applications, Vol.1, No.4, Oct – Dec 2008.

[3] Maria Halkidi, Yannis Batistakis and Michalis Vazirgiannis, "On Clustering Validation Techniques", Journal of Intelligent Information Systems, Vol.17, pp. 107–145, 2001.

[4] Deepti Sisodia, Lokesh Singh, Sheetal Sisodia and Khushboo Saxena. "Clustering Techniques: A Brief Survey of Different Clustering Algorithms", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 1 Issue 3 Sept 2012.

[5] Raymond T. Ng and Jiawei Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", IEEE Computer Society.

[6] P. Murugavel and Dr. M. Punithavalli, "Improved Hybrid Clustering and Distance-based Technique for Outlier Removal", International Journal on Computer Science and Engineering, Vol. 3, No. 1, Jan 2011.

[7] T. Dasu and T. Johnson "Exploratory Data Mining and Data Cleaning", 2003.

[8] M.D. Boomija, "Comparison of Partition Based Clustering Algorithms", Journal of Computer Applications, Vol.1, No.4, Oct – Dec 2008.

[9] Dr. M. Hemalatha and N. Naga Saranya " A Recent Survey on Knowledge Discovery in Spatial Data Mining" , International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011.

[10] Shu-Hsien Liao, Pei-Hui Chu and Pei-Yuan Hsiao, "Data mining techniques and applications", Science Direct Expert Systems with Applications, Vol.39, Issue 12, 15 Sept 2012.

[11] Berry, M.J.A. and Linoff, G,"Data Mining Techniques For Marketing, Sales and Customer Support", 1996.

[12] D. Guo and J. Mennis, "Spatial Data Mining and geographic knowledge discovery-An introduction", Computers, Environment and Urban Systems 33, 2009.

[13] O.A.Abbas, "Comparison between Data Clustering algorithms", International Arab Journal of Information Technology, Vol.5. No.3, July 2008.

[14] Sundararajan S, Dr.Karthikeyan S," A Study On Spatial Data Clustering Algorithms In Data Mining", International Journal Of Engineering And Computer Science, Volume1 Issue 1 Oct 2012,pp. 37-41.

[15] Matheus C.J, Chan P.K, and Piatetsky-Shapiro G.1993. "Systems for Knowledge Discovery in Databeses", IEEE Transactions on Knowledge and Data Engineering 5(6):903-913.

[16] 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, Aug 1996.

[17] Davies, D.L. and Bouldin, D.W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 224–227.

[18] Han, J. and Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, USA.

[19] Gueting R.H. 1994. An Introduction to Spatial Database Systems, The VLDB Journal, pp. 357-399.

[20] A. B. M. S. Ali and K. A. Smith , On learning algorithm for classification, Applied Soft Computing, Dec 2004. pp. 119-138.

[21] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceeding of the 20th VLDB conference, pp 487–499.

[22] Jiawei Han and Michheline Kamber,Data mining concepts and techniques-a reffrence book , pp. 383-422.