International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Review of Spatial Algorithms in Data Mining

V.Padmavati[1]

[1]School of Studies in Computer Science & IT, Pt. Ravishankar Shukla University, Raipur (Chhattisgarh), 492010
p.padmakp@gmail.com

Abstract*: Data mining,* **is the process or action of taking out the obscure and predictive information from wide range of databases,** ***is proven to be new technology with much potential to help companies focus on the most important information in their respective data wares. Mining techniques can be used frequently on existing software and hardware platforms to increase the value of existing information resources, and can be combined with new products and systems. In this study I discussed many reviews, their pros and cons features. I have proposed future work for a hierarchical based clustering algorithm known as CURE algorithm for spatial and non spatial objects. It is suitable with huge datasets and produces optimum results with minimum errors and handles outlier conditions.***

Keywords: CURE, Spatial Clustering algorithm,BANG, Pattern

1. Introduction

This Paper deals with the review of mining the spatial data but the set of techniques that I will discuss applies to many different types of dataset, including time related, spatial and textual databases.I have made further comparative studies of the various clustering algorithms for different size of clusters given by various reviewers.

2. Review of the Related Literature

2.1 Development Work for Spatial Data Mining Primitives & Algorithms

I have taken the related work of Martin ester and Alexander frommelt which reveals study about spatial data mining for conventional database, algorithmic approach and efficient dbms support in following way:-

Spatial data mining algorithms heavily depend on the efficient processing of neighborhood relations since the neighbors of many objects have to be investigated in a single run of a distinctive algorithm. Therefore, it provides general concepts for nearby relations as well as an efficient implementation of these concepts will allow a tight integration of spatial data mining algorithms with a spatial database management system. In this paper, I have taken surrounding graphs and paths and a small set of conventional databases for their manipulations. I showed that distinctive spatial data mining algorithms are well supported by the proposed basic operations.

I discussed here those points which allow only those nearby paths that will significantly minimize the searching space for spatial data mining algorithms. For this I considered nearby indices to enhance the processing of our old-fashioned databases. The potency and efficiency of the discussed approach was evaluated by using an analytical cost model and an extensive experimental study on a geographic database.

Conclusion of this review, defines surrounding graphs and paths and a small set of database primitives for spatial data mining. Research and Studies states that in typical applications the exponential number of all nearby paths can be reduced to a linear number of related nearby paths. I also revealed that spatial data mining algorithms like spatial clustering, their characterization, and classification and behavior detections are well supported by the proposed operations.

Finally, I have given support for nearby indices to speed-up the processing of our database primitives and can be easily created in a commercial DBMS by using standard functions.

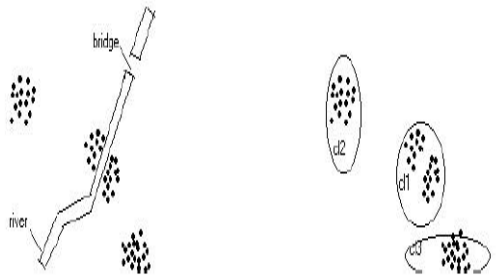
2.2 Discourse for Efficient Clustering Algorithm

Second work I have added of Mohamed el-zawawy,who has developed in the title of algorithm for spatial clustering with obstacles.In this review, I have undergone for an efficient clustering technique to solve the problem of clustering while obstacles are coming while forming clusters. The proposed algorithm divides the spatial area into rectangular shapes as cells, where each cell contains statistical information that enables us to tag the cell as dense or non-dense. Further I also mark each cell as obstructed (i.e. intersects any obstacle) or non-obstructed cells. Then the algorithm finds the regions (clusters) of connected, dense, non-obstructed cells. Finally, the algorithm finds a center point for each such region and returns results of those as centers of the relatively dense regions (clusters) in the spatial area.

Spatial databases contain spatial-related information such databases include geographical area databases, VLSI chip technology databases, and medical and satellite image databases.It can be used in many applications such as seismology (grouping earthquakes clustered along seismic faults), minefield detection (grouping mines in a minefield), and astronomy (grouping stars in galaxies). Clustering, in spatial data mining, is a useful technique for grouping a set of objects into classes or clusters such that objects within a cluster have high similarity among each other, but are dissimilar to objects in other clusters.

In the last few years, many effective and differentiable clustering methods have been developed. These methods can be categorized into partitioning methods [KR90, NH94, BFR98], hierarchal methods [KR90, ZRL96, GRS98, KHK98], density based methods [EKSX96, ABKS99, HK98], grid-based methods [WYM97, SCZ98, AGG98], model-based methods [SD90, Koh82], and constrained-based

methods [THH01]. Most of these algorithms, however, provide very few solutions for users to specify real life constraints such as physical obstacles. In many applications, physical obstacles like mountains and rivers could substantially affect the result of a clustering algorithm. For example, consider a telephone-company planer who wishes to locate a suitable number of telephone connections in the area shown in Figure 1(a) to serve the customers who are represented by points in the figure. In such a situation, however, natural obstacles exist in the area and they should not be ignored. Ignoring these obstacles will result in clusters like those in Figure 1(b), which are obviously inappropriate and hence not a good practice. For example, a river splits cluster cl1, and some customers on one side of the river will have to travel a long way to reach the telephone cabinet at the other side. Thus the ability to handle such real life constraints in a clustering algorithm is very important.



(a) Customer's locations and obstacles (b) Clusters formed when ignoring obstacles

Figure 1: Planning the locations of ATMs

In this paper, I have gone with an efficient spatial clustering technique, *SCPO*, which considers the presence of obstacles. The algorithm finds all the *dense*, *non-obstructed* regions that form the clusters by a breadth-first search and determines a center for each region (cluster).

The proposed algorithm has the following advantages over the work that has done to solve the problem of clustering in the presence of obstacles [THH01].

1) It handles noise or other outlier conditions. Outliers are spatial objects, which are not covered in any cluster and should be discarded during the data mining process.
2) It does not use any randomized search.
3) Instead of specifying the number of the desired clusters in advance, it finds the natural number of clusters in the spatial area.
4) When the data is updated, we do not need to re-compute all information in the cell grid. Instead, incremental update can be provided later on.

Many studies have been conducted in cluster analysis. Methods related to my work are discussed briefly in this section and I emphasize what I believe are some limitations which are addressed by my approach.

Conclusion in this review, I discussed a new approach to spatial clustering in the presence of obstacles. The algorithm, *SCPO*, finds clusters in the spatial area along with their centers taking into consideration existing of natural obstacles. The algorithm has several advantages over existing algorithms. First, it does not use randomized search to determine the center of a cluster, therefore the quality of results is guaranteed. Second the algorithm requires minimum input. Third, outliers are efficiently handled as they are disregarded from the clusters. Extending the algorithm to work for *d* dimensions, *d* >2, is an interesting and challenging task.

2.3 Development Work for Mining Large Spatial Databases

Third review I taken of "Xiaowei, xu,hans, peter kriegel, jorg sander ref.in [4]" in the title of A distribution based clustering algorithm for mining in large spatial data bases. In this, I undergone with the new clustering algorithm as Distribution Based Clustering of Large Spatial Databases to discover clusters of this type. Experimental based study reveals that this algorithm, as contrast to partition making algorithms such as algorithm of Clustering using large applications with randomized search, finds arbitrary shape like clusters. In addition, Distribution based algorithm does not require any input parameters, as consider for the clustering algorithm. Distribution based scanning "In [4]" requires two input parameters which may be difficult to be given for large databases. Hence, the productivity of "DBCLASD In ref [4]" on large spatial databases is very impressive when considering its nonparametric nature and its good quality for clusters of arbitrary shape.

In this paper, I consider the task of clustering in spatial databases, regarding the problem of detecting clusters of points which are distributed as belonging to same group also called uniform distribution or random distribution.The problem finds many areas e.g. seismology In [4] (collection of earthquakes clustered containing seismic faults),In 4 minefield detection (grouping mines in a minefield) and astronomy In [4] (grouping stars in galaxies).Content rich spatial database applications requires the following requirements for clustering algorithms:

1. Minimal number of input parameters.
2. Discovery of clusters with arbitrary shapes..
3. Good efficiency on huge databases.

None of the well-known clustering algorithms fulfills the combination of these requirements. In this review, I discussed on the new clustering algorithm DBCLASD In [4](Distribution Based Clustering of Large Spatial Database) which is based on the assumption that the points inside a cluster are uniformly distributed and is quite reasonable for many applications.

Applications of DBCLASD:-

- It works effectively on real databases where the data is not exactly uniformly distributed.
- It dynamically determines the appropriate number and shape of clusters for a database without requiring any input parameters.

Finally, the algorithm will be efficient for large databases.

In review, I presented with the new clustering algorithm DBCLASD In[4] which is appropriate to fulfill the combination of above stated requirements where as the most concerned algorithms fails to offer solution to these.

2.4 Development Work for Production of Useful Patterns

Fourth review I added here are the works given by "shashi shekar, pushing zhang, yan huang, ranga, raju vatsavai In[7]" who have developed the research in title of Trends in spatial data mining as follows:-
Spatial data mining is the process of discovering interesting and previously un-known, but potentially useful patterns from large spatial datasets. They stated about how to get useful patterns from spatial datasets which is more difficult than getting the corresponding patterns from traditional, numeric and categorical data due to the complexity of spatial data types, relationships and autocorrelation features.

2.4.1 Description

The incremented growth of spatial data and widespread use of spatial databases focuses on the need for the automated discovery system of spatial knowledge. The inherent complexity of spatial data and their relationships limits the use of typical data mining techniques for getting spatial patterns.For extracting information some effective tools from geo-spatial data are sensitive to organizations which take decisions based on huge spatial datasets, like "In [7]NASA, the United States Department of Transportation (USDOT).The National Imagery and Mapping Agency (NIMA), the National Cancer Institute (NCI)."They consist of many application domains consisting of ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology.

Clementine, See5/C5.0, and Enterprise Miner, all these data mining tools were designed to reveal facts for massive commercial databases. To know customer-buying patterns strategies and techniques those tools were mainly designed according to based on market data, in analyzing scientific, engineering data, astronomical data, multi-media data, and web based data. Getting patterns from spatial data sets is more difficult than getting corresponding patterns from traditional numeric and categorical data due to the various complexity issues of spatial data types, spatial relationships, and spatial autocorrelation features

In this review I focused on the unique features that distinguish spatial data mining from classical data mining in the following four categories: data input, statistical foundation, output patterns, and computational process. Major requirements are presented by me of spatial data mining research, especially consisting of output patterns in the form of predictive models, spatial outliers, spatial correlation rules, and clusters. Finally, I got areas of spatial data mining where further research is needed.

2.4.2 Data Input

The data inputs for spatial data mining are more complicated than the input parameters for classical data mining because they include objects of type such as points, lines, and polygons, circles, ellipses etc. Spatial data mining have two broad categories of attributes: non-spatial and spatial. Non-spatial are used to refer non-spatial features of objects, such as name of any object, population data, and literacy rate for a city which are similar to attributes used in the data inputs of conventional data mining whereas spatial is about spatial locations e.g. longitude, latitude and elevation, arbitrary shapes.

2.4.3 Statistical Foundation

In this study I used statistical models to represent observations in terms of random variables. These models are preferred for spatial data estimation, description, and prediction based on probability theory. Spatial data is a resultant from observations carried on the number of random variables on a variable parameter as time like Z(s): s is an element D, where s is a spatial location and D is possibly a random set in a spatial framework.

Here I discussed some of the spatial statistical problems which can occur: point process, lattice, and geostatistics.

Point process: A structure for the spatial arrangement of points in a pattern set. Natural processes can be modeled as spatial point patterns, e.g., positions of trees in forest areas and locations of bird living areas. Spatial point patterns are grouped into random or non-random processes. For a random pattern, the average distance calculations are to be $\frac{1}{2*\sqrt{density}}$, where density is the average number of points per unit area. for real processes, the calculative distance lies within certain limits, then I conclude that the pattern is generated by a random process; otherwise it is by non-random process.

Lattice: It refers to a gridded structure in a spatial framework, which is a countable set of regular or irregular spatial sites related to each other via a neighborhood relationship.

Geostatistics: Geostatistics reveals the study of spatial continuity and weak stationary objects which is one of the characteristics of spatiotemporal data sets. Geostatistics provides a set of statistics tools, such as kriging [Cressie1993] for the interjection of attributes at unsampled locations

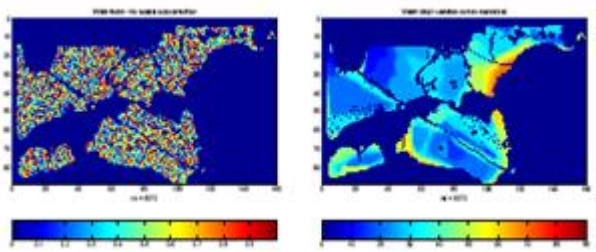


Figure 2: Attribute Values in Space with Independent Identical Distribution and Spatial Autocorrelation

The basic requirement of statistical analysis is that the data samples are independently generated e.g. successive tosses of coin, or the rolling of a die, selection of playing cards at random. But in the analysis of spatial data, this independence is mostly false. In real, spatial data continues to be highly

self-correlated. The economies of a region tend to be similar. A change in natural resources, wildlife, and temperature varies gradually from space to space. The spatial statistics study states, an area within statistics dedicated to the analysis of spatial data, then the property is called as spatial autocorrelation. For example, Figure 2 shows the value distributions of an attribute in a spatial framework for an independent identical distribution and a distribution with spatial autocorrelation.

In this review I have presented the features of spatial data mining that distinguish it from classical data mining in the following categories: description, input and statistical foundation.

I have also discussed major research accomplishments and techniques in spatial data mining, especially those related to four important output patterns: predictive models, spatial outliers, spatial correlation rules, and spatial clusters. I have also identified research needs and requirements for spatial data mining.

References

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
[2] Zhang, J. 2004. Polygon-based spatial clustering and its application in watershed study. MS Thesis, Department of Computer Science and Engineering, University of Nebraska-Lincoln, December 2004.
[3] M. Ankerst, M. Breunig, H. -P. Kriegel, and J. Sander. OPTICS: ordering points to identify the clustering structure. In *Proc. 1999 ACM-SIGMOG Int. Conf. Management of Data (SIGMOD'99),* Philadelphia, PA, June 1999.
[4] Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)-" A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise "by Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu
[5] R. Agrawal, J. Gehrke, D.Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. 1998 ACM-SIGMOG Int. Conf. Management of Data (SIGMOD'98*, Seattle, WA, June 1998.
[6] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proc. 1998 Int. Conf Knowledge Discovery and Data Mining (KKD'98),* pages, New York, NY, Aug. 1998.
[7] *Shashi Shekhar*, Pusheng *Zhang*, *Yan Huang*, and *Ranga Raju Vatsavai* "Trends in spatial data mining" (Roddick and Spiliopoulou 1999, Shekhar and Chawla 2002)
[8] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large data bases. . In *Proc. 1998 ACM-SIGMOG Int. Conf. Management of Data (SIGMOD'98*, Seattel, WA, June 1998.
[9] Hinneburg and D. A. Kein. An efficient approach to clustering in large multimedia databases with noise. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KKD'98),* New York, NY, Aug. 1998.
[10] J. Hou, a M.Sc. thesis submitted by Jean Fen-ju Hou, in the School of Computer Science, Simon Fraser University.
[11] Y. Huang, S. Shekhar & H. Xiong, *Discovering Spatia Co-location Patterns from Spatial Datasets: A General Approach,* IEEE Transactions on Knowledge and Data Eng., 2004.
[12] L. Kaufman and P.J. Rousseeuw. Finding groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
[13] R. Ng, and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94*, Santiago, Chile, Sept. 1994.
[14] J. O'Rourke. *Computational Geometry in C (2nd Ed.).* Cambridge University Press, 1998.
[15] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. . *Very Large Data Bases, 1998*.
[16] Silicon Graphics. Inc. BSP Tree: Frequently asked questions.*http://reality.sgi.com/bspfag/index.shtml*, 1997.
[17] K. H. Tung, J. Hou, & J. Han. Spatial Clustering in the Presence of Obstacles. In *Proc. 2001 Int. Conf. Data Engineering (ICDE'01),* Apr. 2001b.
[18] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. 1996 ACM-SIGMOG Int. Conf. Management of Data (SIGMOD'96),* Montreal, Canada, June 1996
[19] Margrat M. Dunham,"Data mining introductory and advanced Topics", Pearson-2004.
[20] Ji Zhang, Wynne Hsu, Mong Lee." Clustering in Dynamic Spatial Databasses", January-2005.
[21] Emin Erkan, Ken Barker," Intelligent Data Analysis**",** March-2006.
[22] Wei-Bang Chen, "Spatial Clustering with Obs An efficient hybrid clustering Algorithm", March-2006.
[23] Chuang-Cheng Chiu,Cheh-Yuan T sai, **0"** A K-Anonymity clustering Method", Aug-2007.
[24] Lian Duan,Li da Xu,Feng Guo**, "**A Local-density based Spatial Clustering Algorithm" Nov-2007.
[25] Jens keller, 2008, "Clustering biological data using a hybrid approach".
[26] Bin Jiang, Jing Chang Pan, **"**A Data Mining Application in Stellar Spectra", Dec2008.

Author Profile

V.Padmavati received the B.Sc(IT) and M.Sc(C.S.) degrees in IT and Computer Science from MCNUJC, Bhopal in 2003 and 2006, respectively. After wards she completed M.Phil(C.S.) in 2010 from C.V. Raman University,Bilaspur (C.G.). She is now in Pt. Ravishnakar Shukla University, Dep. Of comp. sci. & IT,Raipur(C.G.) as contract Assistant Professor.