A Brief on Spatial Data Mining

Kanak Meena
Reseacrh Scholar (CSE), IGDTUW, Delhi
Kanak556@gmail.com

Nikita Jain
Student (CSE), IGDTUW, Delhi
Nikita97.jain@gmail.com

Abstract: **Spatial data mining is the process of mining the spatial and spatiotemporal data, in order to discover otherwise hidden and unknown patterns and trends in data. Its application areas include business prospecting, store prospecting, hospital prospecting and automobile insurance etc. This paper includes a survey of spatial data mining, its types, techniques and roles in the field of research. Clustering, Classification, Cloropeth display have been the main focus of the paper. All kinds of clustering such as partitioning based, hierarchical based, grid based, model based and density based clustering techniques have been applied here. The outcome of the survey is a consolidated study of all the above techniques and methods of spatial data mining including a detailed summary, advantages, disadvantages, comparison and distinguishes among several techniques. Such type of survey has not been existing till date. So, our main objective behind the paper was to bring a standardized and uniform platform for all the techniques of spatial data mining.**

Keywords: Spatial Data, Data Mining, Classification, Big Data.

I. INTRODUCTION

Data mining refers to the science of discovering patterns and trends from the given databases. Spatial Data Mining is the branch of data mining which deals with the analysis of spatial databases in order to obtain useful information and patterns. This process [1] is largely related to spatial data, spatial databases and mining in detail. Spatial data, also known as geospatial data or geographic information is the data or information that identifies the geographic location of features on Earth, such as oceans, land, plateaus and more. This spatial data is stored in the spatial databases. A Spatial database is a relational database that is optimized for storing and querying data that is represented altogether in clusters of spatial data. They also represent simple geometric objects like points, line and polygons [2]. Spatial data mining differs from data mining only in a way that in the former, apart from the non-spatial attributes (such as height, weight, etc.), spatial attributes are also taken into consideration.

The spatial information in data mining have contributed to various applications in the field such as business prospecting, store prospecting, hospital prospecting and automobile insurance [3]. In Business prospecting spatial data mining can be used to determine if the sales of a business can be improved by collocating it with some other franchise (For example, bringing a pizza store and a movie theatre together). Finding huge applications in store prospecting, spatial data mining can be used to find an appropriate location for a store, in order to boost its sales (For example, within 50 miles of a major city, in a state with no sales tax). Automobile insurance can be put into use as – How to determine if the given home or work location of a customer is in a region with high or low rates of accident claims or auto thefts.

Other applications include Spatial classification based on region or personalization that can help us analyze if northeastern US customers, in a given category of age or income, are more inclined towards "soft" or "hard" rock music [3]. It is also applied in property analysis which uses rules of colocation to explore otherwise unseen associations between closeness to a highway and the cost of a house, or the sales of a store. The most valuable and important application is Property assessment which can help in analyzing the value of a house, assess the values of similar houses in an area, and obtain an approximate based on such spatial correlations.

The main objective behind this survey paper was to explore and extract knowledge about spatial data mining, its types, role, uses and its contribution in betterment of today's industry [4]. Comparison between different techniques of spatial data mining have been studied and finally being concluded to a reliable method. Due to the increased demand of spatial data mining in data science, this paper has made an attempt to understand and implement the ideas of different scientists regarding the spatial data processes and thus dig deeper into the world of spatial data. Since a few years, statistical spatial analysis has been widely used for examining spatial data. The enormous growth of spatial data and increase in the usage of spatial databases highlight the requirement for an automated method of discovering spatial knowledge. The core goal of spatial data mining is to be able to extract actionable patterns from given information.

The subjects that have been covered in this paper include Spatial data mining challenges techniques such as SBD technique [5], Clustering algorithms [6] such as Grid based methods, Hierarchical methods, Partitioning methods, Constraint based clustering, Methods based on co-occurrence of categorical data, Scalable Algorithm, Algorithm for High-D data, Clustering used in ML, K-means clustering [4], SD (CLARANS Algorithm, NSD (CLARANS) Algorithm, CRH Algorithm, DBSCAN algorithms [7]. Several choropleth display techniques [8] have been included such as Class based Choropleth display, spatial clustering, Integrating space and attribute characteristics, I/O requirement analysis.

II. Analysis of various techniques of spatial data mining with their advantages and disadvantages

1. Techniques related to Spatial Big Data Challenges

In this, tradition v/s spatial data ways to collect mobility and cloud computing services. Traditional [9] and big data techniques are used in this paper. Traditional techniques include road maps, tabular representations, and graph representations. Big data techniques include GPS trace, historical speed profiles and fuel consumption via engine measurement.

Advantages

1. Very well differentiation between traditional and big data techniques.
2. Historical profiles can be used to gather information about suicide cases.
3. GPS Trace and engine measurement- modern technology [10].

Disadvantages

1. Requires change in frame of reference perspective to individual travelling through transport.
2. SBD technique increases computational cost and ambiguity of traditional routing techniques.
3. Increase need for diverse solution method in SBD [11] [12].
4. By using GPS Trace, privacy issues come up.

Table 1 : Differentiate between statistics and hybrid based method

Name	Method	Merits	De-Merits	Remarks
Statistics-based	By means of statistical features such as packet size, packet arrival time and flow duration.	More uniqueness.	As no. of features increases, mapping becomes difficult.	Inefficient as no. of features increases.
Hybrid method	By combining any of the above methods	More accurate	Only 2-class classifier is implemented till date.	Scope for UDP needs to be determined.

2. Techniques related to Flow Mapping and Multivariate visualization of large spatial interaction data

It focuses on three main aspects regarding the spatial data interaction. First, [13] Natural regions with spatial interactions with spatially constrained graph partitioning. Second, [14] Generalization of flow patterns into higher levels of abstraction. Third, performing multivariate clustering of multivariate data.

Advantages

1. Good, efficient method-based framework for explanatory analysis, examination and visualization of voluminous spatial interaction data.
2. Supports variety of user interaction features.
3. Easily processed larger sets of data.
4. A method for flow mapping and visualization together has been explained properly.

Disadvantages

1. New, advanced visual interface and user interaction strategies are needed for the complexity of spatial interaction patterns.
2. Currently, examination of multiple levels simultaneously is not possible; only one at a time can be assessed using the specifications (such as number of regions).

3. **Clustering Techniques**

It provides a detailed review of various clustering methods and techniques used in data mining. [15] It makes use of the clustering algorithms such as:

1. Hierarchical methods
2. Partitioning methods
3. Constraint based clustering
4. Grid based methods
5. Methods based on co-occurrence of categorical data
6. Scalable Algorithm
7. Algorithm for High-D data

8. Clustering used in ML.

Incremental clustering algorithm introduced based on clustering algorithm DBSCAN which can be used with any database consisting of data in metric space are being discussed in this paper. The techniques include:

1. DBSCAN Algorithm
2. Incremental DBSCAN Algorithms.
3. Performance evaluation log database using www log base algorithm.
4. Speed up factors using cost DBSCAN Algorithm.

Advantages

1. All clustering methods discussed effectively.
2. Spatial Data Mining techniques involve less complexity and cost friendly.
3. DBSCAN requires only a distance function; hence, it can be applied on any data base that contains data from metric space.

Disadvantages

1. How to use different clustering methods simultaneously (together) is not discussed.
2. Very vague idea of ML thru ML and Grid-based methods.
3. It is assumed here that values and minpoints of DBSACAN do not change significantly while inserting or deleting objects. However, this may always not hold true.

Table 2: Differentiate between Clustering Techniques

Algorithm/ Criteria	**Hierarchical Clustering**	**K-means clustering**	**K-medoid Clustering**	**DBSCAN method**	**Denclue method**	**Self-Organizing map**
Initial condition	No	Yes	Yes	Yes	Yes	Yes
Termination condition	Not precise	Precise	Precise	Precise	Precise	Precise
Granularity	Flexible	K and initial point	K and initial point	Threshold	Threshold	Parameter
Arbitrary value	No requirement	Numeric attribute	Numeric attribute	Numeric attribute	Numeric attribute	N.A.
Shape of data set	Arbitrary	Convex	Convex	Arbitrary	Arbitrary	N.A.
Effect on size of data sets	Not good	good	Not good	Not good	Not good	Good
Implementation	Simple	Simple	Complicated	Simple	Simple	Simple

4. Techniques for Algorithm and applications of spatio-temporal data mining

In this, basically reviews the spatial data mining algorithm by computational and I/O requirements and a brief introduction to some of its applications have been given. The techniques applied are [16]: Spatial Autoregressive Mode (SAR)

1. Gaussian process learning.
2. Biomass Monitoring.
3. Complex Object Recognition.
4. Social Media mining.
5. Climate change coupled models.

Advantages

1. Spatial data mining workflow is integrated with advanced computing framework such as cloud computing.
2. Biomass monitoring allows despicable data to entroupe into the mining details efficiently.

Disadvantages

1. New advanced approaches are needed to sort the computational and I/O challenges faced and efficient modelling of spatial and temporal constraints [17].
2. Compression and sampling demand further advanced research.

5. Techniques for Choropleth display and spatial data mining

It develops iterative approach for explanatory spatial data analysis using choropleth display and spatial data mining. The techniques include:

1. Class based Choropleth display
2. Spatial clustering

3. Integrating space and attribute characteristics
4. I/O requires mental analysis.

Advantages

1. Good, efficient method-based framework for explanatory analysis, examination and visualization of voluminous spatial interaction data.
2. Supports variety of user interaction features.
3. Easily processed larger sets of data.
4. Methods for flow mapping and visualization together have been explained properly.

Disadvantages

1. The view of non -overlapping attribute classes might not be maintained.
2. No flexibility in integrating attribute similarity and spatial proximity.

6. **Spatial Data Mining Algorithm**

It is mentioned that the advances in database technologies and data collection techniques such as barcode reading, remote sensing etc. are crucial in the industry. The techniques include

1. Generalization based knowledge discovery
2. K-means clustering
3. SD (CLARANS Algorithm
4. NSD (CLARANS) Algorithm
5. CRH Algorithm.

Advantages

1. Wide GIS applications, medical imaging, robot motion planning etc. is done.
2. SD Algorithm is more verbose, application friendly as compared to NSD(CLARANS).

Disadvantages

1. Variety of unexplored topics regarding clustering and spatial databases.
2. CRH Algorithm does not work based on skeptical analysis pf spatial data sets.

Table 3. Differentiate between Algorithms on the basis of runtime, noise and shape of clusters

Algorithm	Run time	Arbitrary shape clustering	Handle noise
DBSCAN	O(nlogn)	Yes	Yes
Rough DBSCAN	O(n)	Yes	Yes
Optics	O(nlogn)	Yes	Yes
DENCLue	O(nlogn)	No	Yes
DENCOS	O(d^k)	Yes	No
Mitosis	O(nlogn)	Yes	No

d: Dimensionality of data set

n: Number of data points.

k: cardinality of data set.

7. Techniques for Spatial data infrastructure and knowledge base

The objective of this is raising awareness about different knowledge areas that are available to the people who are working towards the development of spatial data infrastructures [18]. This paper is basically a survey so no techy ques or measures are used.

Advantages

1. Alternatives for failure of SDI research and practice in order to efficiently utilize the theoretical knowledge [19].
2. Conceptual framework is designed from the expanded knowledge base of SDI.

Disadvantages

1. Factors, strategies and processes for developing SDIs are not tested or elaborated.
2. Difficulty in identifying viable motivators in public sectors.
3. NG Actors, private sectors, academia, Non-profit organization and population at large are not addressed.

8. Techniques used in Formal model for spatial data infrastructure

This deals with the Models of SDI that explain the how various sections of SDI fit together in the viewpoints in questions. The techniques used are:

1. Unified Modelling Language.

2. Five SDI Modelling viewpoints.

- Enterprise viewpoint
- Information viewpoint
- Computation viewpoint
- Engineering viewpoint
- Technology viewpoint

Advantages

1. UML Diagrams including use cases, actors and relationships are used properly.
2. Purely based on Open Distributed Processing reference model (ODRAM) of SDI and UML.

Disadvantages

1. Whether actor is passive or active is not mentioned in any UML Diagram.
2. No proper distinction between producer, provider, broker in banking system.
3. Further refinement of models required.

9. Techniques for data mining for path traversal pattern

This paper explores a different data mining technique in which path traversal patterns are mined in an environment providing distributed information, with document access [20]. The techniques include:

1. Traversal Pattern Algorithm
2. Large reference sequence determination.
3. Full Scan Algorithm.
4. Selective scan algorithm.
5. New Scan Algorithm.

Advantages

1. All mining capability algorithms which involve mining traversal paths are discussed.
2. Effects of backward references have been filtered, while that of some forward references and mining processes are mentioned.

Disadvantages

1. Not proper explanation of Full Scan, Selective and new Scan algorithm.
2. Spatial data sets in data mining algorithms missing.

III. CONCLUSION

This paper includes survey from last ten years (2000-2017) and out of all the techniques of spatial data mining discovered till now we have tried to sum up with most efficient and feasible technique.

- Among the static based and hybrid based method we have concluded that hybrid based method out of the two because it is more accurate, it has scope for (UDP) and it can be done by combining the statistic based method.
- Out of the clustering methods in which we discussed hierarchical method [21], K-means clustering, K-medoid clustering, density clustering, Denclue, self-organizing map. We have concluded that self-organizing map technique is the best because of no arbitrary value and predefined shape of data set has to be given in advance. [22] The granularity, implementation and effect on size of datasets is also upto the point. So, it is better than other techniques.
- Out of the classification methods in which we discussed rough DBSCAN, optics, DENCOS, MITOSIS we conclude that optics is the beat out of all above methods because it is scalable to large data sets, it has the ability to handle noise and it supports arbitrary shape cluster.

IV. FUTURE SCOPE

1. Multimedia Data Mining: Finds huge application in predicting diverse types of multimedia formats, techniques etc. suitable for different formats.

2. Distributed Data Mining: Can be used in distributed time-based measurements for various companies and organizations. The data from these various locations are extracted using highly sophisticated algorithms and reports are generated using them.

3. Super Market Analysis: Can be used in Market Basket analysis for determining the interests of the customers and this can help in improving the business.

4. Sequence Data mining: This branch of data mining finds huge application in the study of cyclical and seasonal trends. This technique eases out the process of examining random events (differing from normal trends of events). Retail companies use this technique to observe the customers' behavior and buying patterns.

REFERENCES

[1] Ranga Raju Vatsavai, Varun Chandola, Auroop Ganguly, Scott Klasky, Shashi Shekhar, Anthony Stefanidis : Spatiotemporal Data Mining in the Era of Big Spatial Data: Algorithms and Applications
[2] Michael Goebel, Le Gruenwald : A Survey Of Data Mining And Knowledge Discovery Software Tools.
[3] Oracle database online documentation 12c Release1 (12.1)
[4] Krzysztof Koperski, Junas Adhikary, Jiawei Han : Spatial Data Mining:Progress and Challenges Survey paper
[5] Smark Wallson, Jake Harry: Spatial Big Data Challenges
[6] Pavel Berkhlin : Survey of clustering of data mining techniques.
[7] Martin Ester, Xiaowei Xu, Michael Wimmer: Incremental clustering for mining in a data warehousing environment.
[8] Alan T. Murray,Tung-Kai Shy : Integrating attribute and space characteristics in Chloropleth display and spatial data mining.
[9] Daniel A.Keim *Christian Panse Mike Sips,Stephen C.Korth:Pixel based visual mining of geo spatial data.
[10] Derya Birant*,Alp kut : ST-DBSCAN An algorithm for clustering spatio temporal data.
[11] Martin Ester ,Hans Peter Kriegal,Jorg Sander:Algorithms and Applications of Spatial data mining.
[12] Raymond T.NG*,Jiawei Han*CLARANS:A method for clustering objects for spatial data mining.
[13] Diansheng Guo a,1, Jeremy Mennis b,* a Department of Geography, University of South Carolina, 709 Bull Street, Room 127, Columbia, SC 29208, United States. b Department of Geography and Urban Studies, Temple University, 1115 W. Berks Street, 309 Gladfelter Hall, Philadelphia, PA 19122, United Statesa r:Spatial data mining and geographic knowledge discovery—An introduction.
[14] Jiawei Han Krzysztof Koperski Nebojsa Stefanovic GeoMiner Research Group, Database Systems Research Laboratory School of Computing Science Simon Fraser University Burnaby, BC, Canada V5A 1S6 Geo Miner: A System Prototype for Spatial Data Mining.
[15] Shashi Shekhar, Michael R. Evans, James Kang: Spatial Data Mining
[16] Xiaobai Yao:Research Issues in Spatio-temporal Data Mining.
[17] Wei Wang, Jiong Yang, and Richard Muntz: STING : A Statistical Information Grid Approach to Spatial Data Mining.
[18] Martin Ester, Hans-Peter Kriegel, Jörg Sander: Spatial Data Mining: A Database Approach.
[19] Dawei Wang · Wei Ding · Henry Lo · Tomasz Stepinski · Josue Salazar · Melissa Morabito:Crime hotspot mapping using the crime related factors—a spatial data mining approach
[20] Shuliang Wang and Hanning Yuan:Spatial Data Mining: A Perspective of Big Data.
[21] Bradley P. Carlin, Sudipto Banerjee , and Alan E. Gelfand:Hierarchical Modeling and Analysis for Spatial Data.
[22] Enrico Feoli , Rufino Pérez-Gómez , Cecilio Oyonarte , Juan J. Ibáñez :Using spatial data mining to analyze area-diversity patterns among soil, vegetation, and climate: A case study from Almería, Spain.