SPATIAL DATA MINING USING K MEANS ALGORITHM AND FIND USERS HAVING COMMON INTEREST USING GENETIC ALGORITHM

Er. Yogesh Kumar
Assistant Professor
Taruna
Researcher

Abstract— **Spatial Data Mining (SDM) is an important branch of data mining. With the large amount of spatial data stored in a Geographic Information System (GIS) spatial database, many knowledge and laws need to be mined. This paper presents the understanding of Geographical Information System (GIS) for analysing data using data mining techniques. Spatial Data Mining (SDM) technology has emerged as a new area for spatial data analysis. Geographical Information System (GIS) stores data collected from heterogeneous sources in varied formats in the form of geodatabases representing spatial features, with respect to latitude and longitudinal positions. The intent of this paper is to introduce with GIS, and spatial data mining, GIS and SDM tasks, issues and challenges, and role of spatial association rule mining in big data of GIS.**

Index Terms— **Spatial Data Mining, Geographic Information System, Spatial Association Rule.**

INTRODUCTION

Data mining is a process which finds useful patterns from large amount of data. Spatial data mining is a special kind of data mining. The main difference between data mining and spatial data mining is that in spatial data mining tasks we use not only non-spatial attributes, but also spatial attributes. Spatial data provides the location information of the features whereas non-spatial data describes characteristics of the features.

In this paper we discuss few of the spatial data mining rules, tasks,applications and provide better understanding of Geographical Information System (GIS) for analysing data using data mining techniques. Spatial data are data that have a spatial or location component. Spatial data can be viewed as data about objects that they are located in a physical space. This may be implemented with a specific location attributed such as address or latitude/longitude or may be more implicitly included such as by a portioning of the database based on location.

II DATA MINING

Data Mining or Knowledge Discovery is needed to make sense and use of data. Knowledge Discovery in Data is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [1]. Data mining is the automated process of discovering patterns in data. The purpose is to find correlation among different datasets that are unexpected. There are several applications for Machine Learning (ML), the most significant of which is data mining. Numerous ML applications involve tasks that can be set up as supervised. It is different from other searching and analysis techniques because data mining is highly exploratory, where other analyses are typically problem-driven and confirmatory. Through the combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge, hidden trends and patterns are able to be uncovered. These trends and patterns form the predictive models that enable to assist organizations with uncovering useful information then guide decision-making [4].

The main difference between data mining in relational DBS and in spatial DBS is that attributes of the neighbors of some object of interest may have an influence on the object and therefore have to be considered as well. The explicit location and extension of spatial objects define implicit relations of spatial neighborhood (such as topological, distance and direction relations), which are used by spatial data mining algorithms. Therefore, new techniques are required for effective and efficient data mining. The comprehension of phenomena related to movement not only of people and vehicles but also of animals and other moving objects – has always been a key issue in many areas of scientific investigation or social analysis. Many applications track the movement of mobile objects, using location- acquisition technologies such as Global Positioning System (GPS), Global System for Mobile Communications (GSM) etc., and it can be represented as sequences of time stamped locations. [2-3].

III SPATIAL DATA AND NON-SPATIAL DATA

Spatial Data: Spatial data are data that have a spatial or location component. It includes location, shape, size and orientation information of features or objects. Spatial data can be viewed as data about objects that they are located in a physical space. This may be implemented with a specific location attributed such as address or latitude/longitude or may be more implicitly included such as by a portioning of the database based on location.For example, a particular square in which its center (the intersection of its diagonals) specifies its location; its shape is a square; length of one of its sides specifies its size and angle its diagonals e.g., the x-axis specifies its orientation. Spatial data includes spatial relationships, for example, the arrangement of three stumps in a cricket ground. Spatial data carries topological and/or distance information and it is often organized by spatial indexing structures and accessed by spatial access methods. These distinct features of a spatial database pose challenges and bring opportunities for mining information from spatial data [7]. Spatial data mining, or know ledge discovery in spatial database, refers to the extraction of implicit know ledge, spatial relations, or other patterns not explicitly stored in spatial databases [6].

Non-spatial Data: It is also known attribute or characteristic data. It consists of the characteristics of spatial features which are independent of all geometric considerations. Let us illustrate this with the help of an example. The non-spatial data of town comprise of name of the town, its population, settlement type, means of transportation and communication, administration set-up, education institutions, occupations and facilities. It is important to note that all the above mentioned data of town are not dependent on their location identities. Hence, non-spatial data is independent from location information. The fundamental difference between spatial and non-spatial data is given in Table 1.

Table 1: Basic difference between spatial and non-spatial data

Spatial data	Non-spatial data
It has multi-dimensional nature and autocorrelated	It has one-dimensional nature and independent

These above distinctions put spatial and non-spatial data into different philosophical camps with far-reaching implications for conceptual, processing and storage issues. For example, sorting is, perhaps, the most common and important non-spatial data processing function that is performed. It is not obvious how to even sort locational data such that all points end up nearby their nearest neighbours. These distinctions justify a separate consideration of spatial and non-spatial data models. In GIS, georelational data model stores spatial and non-spatial (attribute) data separately and also links them on the basis of identity of features. Further, georelational data model arranges these two data-sets in such a way that they can simultaneously be queried, analysed and displayed.

IV SPATIAL DATA MINING

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets [5]. It is the application of data mining techniques to spatial data. Data mining in general is the search for hidden patterns that may exist in large databases. Spatial data mining is the discovery of interesting the relationship and characteristics that may exist implicitly in spatial databases. Because of the huge amounts (usually, terabytes) of spatial data that may be obtained from satellite images, medical equipments, video cameras, etc. 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. This has wide applications in Geographic Information Systems (GIS), remote sensing, image databases exploration, medical imaging, robot navigation, and other areas where spatial data are used. Knowledge discovered from spatial data can be of various forms, like characteristic and discriminant rules, extraction and description of prominent structures or clusters, spatial associations, and others.

The development of information and communication technologies in GIS Domain has generated huge volume of data representing spatial information of water bodies, forest reserves, urbanization, etc., GIS databases stores spatial and non spatial data received from heterogeneous components connected, with each other such as sensors, laptop, mobile etc. Analysis of data deposited in GIS has gained importance in domains related to knowledge management and data mining. Recent widespread use of spatial databases has lead to the studies of Spatial Data Mining (SDM), Spatial Knowledge Discovery (SKD), and the development of SDM techniques. GIS can be viewed as collection of components such as Data, Software, Hardware, Procedures and methods used by people for analysis and decision making with respect to location. Fig 1, represents the components of GIS. The focus of this section is to provide with an overview of GIS data source, data formats, trends and Data Mining applications in GIS.

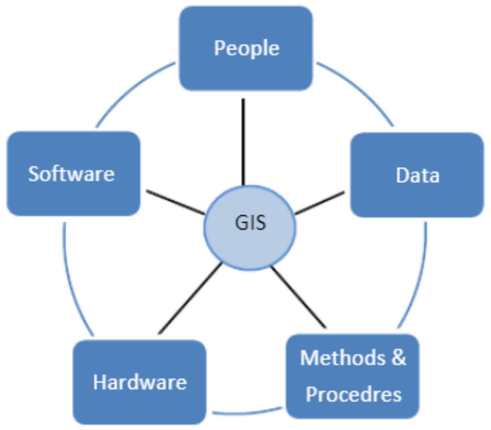

Fig. 1. Components of GIS

Spatial Data Mining techniques collectively used with GIS and satellite imagery in various studies to mine interesting facts associated in diverse domains' applications such as traffic risk analysis, fire accident analysis, analysis of forest extent changes, grading of agriculture land, analysis of railways, farming and forestry, warehouse, transport, tourism, military, geology, soil quality monitoring, water resource monitoring, and deforestation, land allocation, meteorology [12-13]. This section throws the light on GIS data source, data formats, trends and applications in GIS.

V SPATIAL DATA MINING TASKS

Brief overview of Spatial Data Mining tasks are discussed in the section below. Mining of Spatial Data using data mining techniques such as association, classification, clustering, and trend detection generates interesting facts associated in various domains. Spatial Data Mining tasks are generally an extension of data mining tasks in which spatial data and criteria are combined [8],[9],[10] to form various tasks to find class identification, to find association and colocation of Spatial and Non-Spatial data, make the clustering rules to detect the outliers and to detect the deviations of trends. Basic tasks of spatial data mining are:

Classification – finds a set of rules which determine the class of the classified object according to its attributes. In spatial classification the attribute values of neighboring objects are also considered. e. g. “IF population of city = high AND economic power of city = high THEN unemployment of city = low” or classification of a pixel into one of classes, e. g. water, field, forest.

Association rules – find (spatially related) rules from the database. Spatial association rule is a rule indicating certain association relationship among a set of spatial and possibly some non-spatial predicates. The association rule has the following form: A → B(s%,c%), where s is the support of the rule (the probability, that A and B hold together in all the possible cases) and c is the confidence (the conditional probability that B is true under the condition of A e. g. “if the city is large, it is near the river (with probability 80%)” or “if the neighboring pixels are classified as water, then central pixel is water (probability 80%).”

A multilevel Association Rule has been generated to find association between the data in a large database [11] and suggested a method by applying different minimum confidence thresholds for mining associations at different levels of abstraction.

Characteristic rules – A spatial characteristic rule is a general description of a set of spatial-related data. e. g. “bridge is an object in the place where a road crosses a river.”

Discriminant rules –A spatial discriminant rule is the general description of the contrasting or discriminating features of a class of spatial-related data from other class(es). e. g. find differences between cities with high and low unemployment rate.

Characterization – It is the process of finding a description for a dataset or some subset thereof.

Clustering rules –Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. e. g. we can find clusters of cities with similar level of unemployment or we can cluster pixels into similarity classes based on spectral characteristics.

Trend detection –A spatial trend is a regular change of one or more non-spatial attributes when spatially moving away from a start object. Spatial trend detection is a technique for finding patterns of the attribute changes with respect to the neighbourhood of some spatial object. One of the trend detection techniques is kriging to predict the location from outside the sample. e. g. “when moving away from Brno, the unemployment rate increases” or we can find changes of pixel classification of a given area in the last five years.

Table 2 summarizes SDM techniques used in various domain as listed from various literature.

Table 2 Spatial Data Mining techniques used

SDM Domain	Usage	Technique/Method	Reference
Forest	Identify false alarms in forest fire hotspots	Region growing method, Hough transform	Satellite images are used for experiments [12]
	GIS based fireproof system	Frequency theory based method	Sample spatial dataset [14]
	to evaluate forest extent changes	Spatial Data Mining, Back propagation algorithm	Satellite images are taken as dataset [15]
Transport	To increase the effectiveness of railway MIS such as monitoring, railway tracks, and geographical spread	Association rule mining, classification, forecast, trend analysis and planning	[16]

	Represent link between the GIS street data and roadway connections	Object oriented modeling transportation	[17]
Warehouse	Improving spatial data mining effectiveness	Spatial data cube	Spatial data from warehouse [18]
Agriculture	Precision agriculture	Cross-validation technique	Spatial dataset [19]
	Crop yield prediction for wheat	Neural Network	Experiments are made on satellite images [20]
	accessing the quality of soil, management of water resources.	GIS and Spatial Data mining	[21]
Tourism Management	integrating the ICT techniques with tourism	association technique	[22]
Land allocation	provide the selection a model for land use, illegal land fills	Location prediction	[23]
Meteorology	Estimation of rainfall for homogenous monsoon region	genetic approach and correlation	[24]
	forecast the rainfall of Rajasthan state	Multiple liner regression	[25]
	Regionwise rainfall fluctuation	classification	[26]
	Estimation of rainfall	Spatial interpolation and association rule	[13]

VI SPATIAL DATA MINING APPLICATIONS

Spatial Trend Detections in GIS – Spatialtrends describe a regular change of non-spatial attributes when moving away from certain start objects. Global and local trends can be distinguished. To detect and explain such spatial trends, e.g. with respect to the economic power, is an important issue in economic geography.

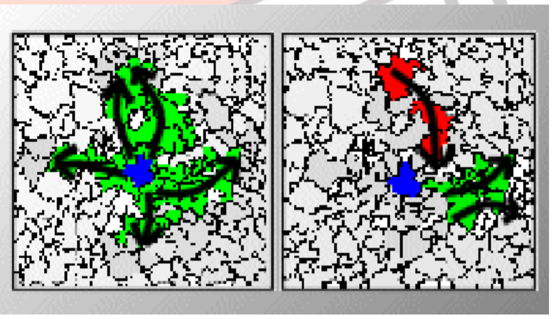

Fig.2 Spatial Trend Detections in GIS

Spatial Characterization of Interesting Regions – Anotherimportant task of economic geography is to characterize certain target regions such as areas with a high percentage of retirees. Spatial characterization does not only consider the attributes of the target regions but also neighboring regions and their properties.

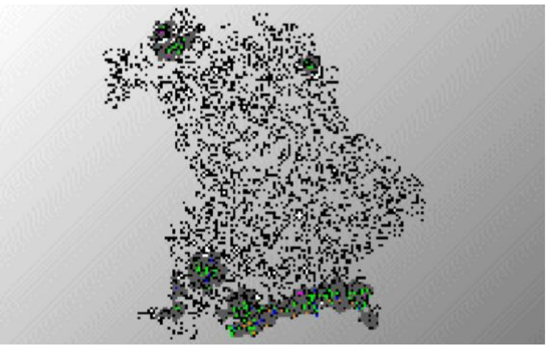

Fig.3 Spatial Characterization of Interesting Regions

VII ISSUES AND CHALLENGES

The issues and challenges in applying Spatial Data Mining in GIS can be viewed in terms of Integration of data and Mining huge volume of data. Architecture is proposed to address the issues of data integration and volume of data based on the analysis of data of GIS from literature is given in Fig 4.Data warehousing technology is used as a tool for data integration and stores summarized data. Currently a Bigdata approach has gained attention for mining data parallel using architectures like Hadoop and Mapreduce. In this paper we propose an architecture which can have a Bigdata platform modeled for representing a data warehouse. The other challenge in integration of data is semantic representation of information organized from various data sources. An ontology layer is proposed for semantic representation of data [27]. The data mining techniques can be applied above these layers by modifying the algorithmic approaches suitable for BigData architecture.

VIII CONCLUSION

This paper provides with a detailed survey on spatial data bases and provide better understanding of Geographical Information System (GIS) for analysing data using data mining techniques. The spatial data mining is newly arisen area when computer technique, database applied technique and management decision support techniques etc. have been developed at certain stage. The spatial data mining gathered productions that come from machine learning, pattern recognition, database, statistics, artificial intelligence and management information system etc. Different theories, put forward the different methods of spatial data mining, such as methods in statistics, proof theories, rule inductive, association rules, cluster analysis, spatial analysis, fuzzy sets, cloud theories, rough sets, neural network, decision tree and spatial data mining technique based on information entropy etc. Spatial data mining, has established itself as a complete and potential area of research. This article tries to explain spatial data mining as well the its different tasks. It also explains how we explain a particular task in relation to the spatial data. The challenge is to take up spatial data mining as an important area and work on it using the various domains such as statistics, artificial intelligence, machine learning and geography etc.

REFRENCES

[1] Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy: “Advances in Knowledge Discovery and Data Mining”, AAAI/MIT Press 1996

[2] Ajaya Kumar Akasapu, Lokesh Kumar Sharma, G. Ramakrishna in Volume 9– No.5, November 2010, “Efficient Trajectory Pattern Mining for both Sparse and Dense Dataset”, International Journal of Computer Applications (0975 – 8887).

[3] Applications Volume-9 –No. 5. Jae-Gil Lee, Jiawei Han &Kyu-Young Whang” Trajectory Clustering: A Partition-and-Group Framework in University of Illinois at Urbana-Champaign KAIST.

[4] Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2012). Analytics: The Widening Divide. MIT Sloan Management Review, 53(2), 1-22.

[5] N. Sumathi, R. Geetha and Dr. S. SathiyaBama, “spatial data mining-techniques trends and its applications”, Journal of Computer Applications, Vol.1, No.4, Oct – Dec 2008.

[6] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In Proc. 4th Int'l Symp. on Large Spatial Databases (SSD'95), pp. 47{66, Portland, Maine, August 1995

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

[8] Guarino, A. Jarvis, R.J. Hijmans and N. Maxted, “Geographic Information Systems (GIS) and the Conservation and Use of Plant Genetic Resources”, IPGRI 2002.

[9] “Luc Anselin,”“Exploring Spatial Data with GeoDaTM : A Workbook”“,Revised Version, March 6, 2005 Copyright 2004-2005 Luc Anselin, All Rights Reserved”.

[10] S. Schockaert, P.D. Smart, F.A. Twaroch, Generating approximate region boundaries from heterogeneous spatial information: an evolutionary approach, Inform. Sci, 2011, 181, 257–283.

[11] A. Pope III, R. T. Burnett, G. D. Thurston, M. J. Thun, E. E. Calle, D. Krewski, J. J. Godleski, “Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution, Circulation”,2004,109,pp. 71-77.

[12] Sengchuan, T. 2003. Spatial Data Mining: Clustering of Hot Spots and Pattern Recognition. IEEE. pp.3685-3687.

[13] Teegavarapu, R. S. V., 2009. Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. Journal of Hydroinformatics, vol. 11, no. 2, pp.133–146.

[14] Liang,Y., and Fuling, B. 2007. An Incremental Data Mining Method for Spatial Association Rule in GIS Based Fireproof System. IEEE. pp.5983-5986.

[15] Jayasinghe, P.K.S.C., and Masao, Y. 2013. Spatial data mining technique to evaluate forest extent changes using GIS and Remote Sensing.

[16] Wei, X., Yong, Q., Houkuan, and H. 2003. The Application of Spatial Data Mining in Railway Geographic Information Systems. IEEE. pp.1467-1471

[17] Marzolf, F., Trépanier, M., and Langevin. 2006. A Road network monitoring algorithms and a case study. Journal of Computer and Operation Research, pp.3494–3507.

[18] Yuanzhi, Z., XieKunqing, M., Xiujun, X., Dan, C., and Tang S. 2005. Spatial Data Cube: Provides Better Support for Spatial Data Mining. IEEE. pp.795-798

[19] Rub, G., and Brenning, A. 2010. Data Mining in Precision Agriculture: Management of Spatial Information, Computational Intelligence for Knowledge Based System Design, Volume 6178, pp. 350-359.

[20] Stathakis, D., Savin, I., and Nègre T. Neuro-Fuzzy Modeling for Crop Yield Prediction, The International Archives of the Photogrammetry, Remote Sensing and Spatial Info. Sc., Vol. 34, pp.1-4.

[21] Vaagh, Y. 2012. The application of a visual data mining framework to determine soil, climate and land use relationships. Journal of Procedia Eng. 32, pp.299–306 .

[22] Buhalis, D., and Law, R. 2008. Progress in information technology and tourism management,The state of eTourism research. Journal of Tourism Mgmt.

[23] Chakraaborty, A., Mandal, J.K., Chandrabanshi, S.B., and Sarkaar, S. 2013. A GIS Anchored system for selection of utility service stations through Hierarchical Clustering. International Conference on Computational Intelligence: Modeling techniques and Application, CIMTA

[24] Kashid, S.S., and Maity, R., 2012. Prediction of monthly rainfall on homogenous monsoon regions of India based on large scale circulation patterns using Genetic Programming. Journal of Hydrology, pp.26-41.

[25] Vyas, P., 2015. To predict rainfall in desert area of Rajasthan using data mining techniques. vol.3, no.5.

[26] Priya, R.L., and Manimannan, G., 2014. Rainfall fluctuation and regionwiseclassificatrion in Tamilnadu using geographical information system. IOSR Journal of Mathematics (IOSR-JM), vol. 10, pp.5-12.

[27] Boomashanthini.S, “Gene ontology similarity metric based on DAG using diabetic gene”compusoft on International Journal of Advances Computer Technology, ISSN: 2320 0790.