International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 3 – FEBRUARY 2017 (SPECIAL ISSUE).

AN INTRODUCTION TO GEOGRAPHICAL KNOWLEDGE DISCOVERY IN SPATIAL MINING

B.Geetha[#1],D.B.Shanmugam[*2] and C.Karthi [*3]

[#1]*M.Phil, Research Scholar, Dr.M.G.R.Chockalingam Arts College, Arni.*

[*2] *Associate Professor, Department of MCA, Sri Balaji Chockalingam Engineering College , Arni*

[*3]*Assistant Professor, Department of MCA, Sri Balaji Chockalingam Engineering College , Arni*

***Abstract*— The data mining methods that are combined with Geographic Information Systems (GIS) for carrying out spatial analysis of geographic data. We will first look at data mining functions as applied to such data and then highlight their specificity compared with their application to classical data. Voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. We first briefly review the literature on several common spatial data-mining tasks, including spatial classification and prediction; spatial association rule mining; spatial cluster analysis; and geo visualization. The articles included in this special issue contribute to spatial data mining research by developing new techniques for point pattern analysis, prediction in space–time data, and analysis of moving object data, as well as by demonstrating applications of genetic algorithms for optimization in the context of image classification and spatial interpolation. To address these challenges, spatial data mining and geographic knowledge discovery has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases There is an urgent need for effective and efficient methods to extract unknown and unexpected information from spatial data sets of unprecedentedly large size, high dimensionality, and complexity.**

Index Terms*— *Spatial data mining, spatial database, Clustering, geographic data, geocomputation, geovisualization.

I. INTRODUCTION

Researchers acquire new knowledge by searching for patterns, formulating theories, and testing hypotheses with observations. With the continuing efforts by scientific projects, government agencies, and private sectors, voluminous geographic data have been, and continue to be, collected. We now can obtain much more diverse, dynamic, and detailed data than ever possible before with modern data collection techniques, such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. Generally speaking, geography and related spatial sciences have moved from a data-poor era to a data-rich era. The availability of vast and high-resolution spatial and spatiotemporal data provides opportunities for gaining new knowledge and better understanding of complex geographic phenomena, such as human–environment interaction and social– economic dynamics, and address urgent real-world problems, such as global climate change and pandemic flu spread.

However, traditional spatial analysis methods were developed in an era when data were relatively scarce and computational power was not as powerful as it is today. Facing the massive data that are increasingly available and the complex analysis questions that they may potentially answer, traditional analysis methods often have one or more of the following three limitations. First, most existing methods focus on a limited perspective (such as univariate spatial autocorrelation) or a specific type of relation model (e.g., linear regression). If the chosen perspective or assumed model is inappropriate for the phenomenon being analyzed, the analysis can at best indicate that the data do not show interesting relationships, but cannot suggest any better alternatives. Second, many traditional methods cannot process very large data volume. Third, newly emerged data types (such as trajectories of moving objects, geographic information embedded in web pages, and surveillance videos) and new application needs demand new approaches to analyze such data and discover embedded patterns and information.

Spatial data illustrates information associated with the space engaged by objects. The data consists of geometric information and can be either distinct or continuous. Discrete data possibly will be a single point in multi-dimensional space; on the other hand discrete spatial data is different from non-spatial data in that it has a distance feature that is used to locate the data in space. Continuous data spans a region of space. This data may perhaps include medical images, map regions, or star fields. Spatial databases are database systems that handle spatial data. It is intended to manage both spatial information and the non-spatial characteristics of that data. With the purpose of providing improved and effective access to spatial data it is essential to develop indices. These indices

are most useful when based on multi-dimensional trees. The structures for these indices comprise quad trees, k-d trees, R trees and R* trees. Data mining, or knowledge discovery in databases (KDD), is the method of investigating data to determine previously unidentified potential information. The objective is to show the regularities and relationships that are non-trivial. This is possible through an examination of the patterns that form in the data. Several algorithms have been developed by many researchers to carry out this spatial data mining, but the majority of these approaches are not scalable to very huge databases.

Spatial data mining is the finding of useful associations and characteristics that may well exist implicitly in spatial databases. Spatial data mining concentrates on automating such a knowledge discovery process. It plays an essential role in (i) obtaining interesting spatial patterns and characteristics; (ii) capturing inherent associations among spatial and non-spatial data; (iii) presenting data reliability concisely and at conceptual levels; and (iv) assisting in reorganizing spatial databases to accommodate data semantics, in addition to accomplish enhanced performance. Spatial data clustering is a most important constituent of spatial data mining and is implemented as such to retrieve a pattern from the data objects distribution in a particular data set and as mentioned earlier it has several applications like satellite imagery, geographic information systems, medical image analysis, etc..

Spatial data mining has deep roots in both traditional spatial analysis fields (such as spatial statistics, analytical cartography, exploratory data analysis) and various data mining fields in statistics and computer science (such as clustering, classification, association rule mining, information visualization, and visual analytics). Its goal is to integrate and further develop methods in various fields for the analysis of large and complex spatial data. Not surprisingly, spatial data mining research efforts are often placed under different umbrellas, such as spatial statistics, geocomputation, geovisualization, and spatial data mining, depending on the type of methods that a research focuses on.

Data mining and knowledge discovery is an iterative process that involves multiple steps, including data selection, cleaning, preprocessing, and transformation; incorporation of prior knowledge; analysis with computational algorithms and/or visual approaches, interpretation and evaluation of the results; formulation or modification of hypotheses and theories; adjustment to data and analysis method. In the literature, knowledge discovery refers to the above multistep process while data mining is narrowly defined as the application of computational, statistical or visual methods. In practice, however, the application of any data mining method should be carried out following the above process to ensure meaningful and useful findings. In this paper, ‘‘spatial data mining” and ‘‘geographic knowledge discovery” are used interchangeably, both referring to the overall knowledge discovery process.

Spatial data mining encompasses various tasks and, for each task, a number of different methods are often available, whether computational, statistical, visual, or some combination of them. Here we only briefly introduce a selected set of tasks and related methods, including classification (supervised classification), association rule mining, clustering (unsupervised classification), and multivariate geovisualization.

II. 2. SPATIAL DATA MINING APPLICATIONS

Spatial data mining 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 costly and often unrealistic for users to examine spatial data in detail. Spatial data mining aims to automate such a knowledge discovery process. Thus it plays on 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.

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 databases have many features distinguishing them from relational databases. They 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.

III. 3. SPATIAL DATA MINING STRUCTURE

The spatial data mining can be used to understand spatial data, discover the relation between space and the non space data, set up the spatial knowledge base, excel the query, reorganize spatial database and obtain concise total characteristic etc..

The system structure of the spatial data mining can be divided into three layer structures mostly such as the Figure 1 show The customer interface layer is mainly used for input and output, the miner layer is mainly used to manage data, select algorithm and storage the mined knowledge, the data source layer, which mainly includes the spatial database and other related data and knowledge bases, is original data of the spatial data mining.

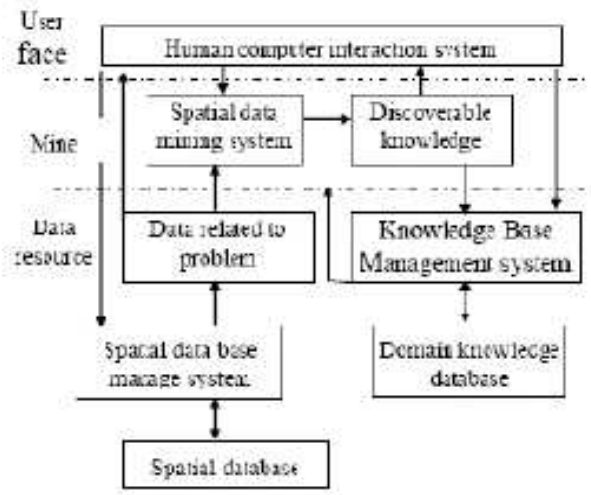

Figure.1 The systematic structure of spatial data mining

IV. 4. SPATIAL DATA MINING TASKS

As shown in the table below, spatial data mining tasks are generally an extension of data mining tasks in which spatial data and criteria are combined. These tasks aim to: (i) summarize data, (ii) find classification rules, (iii) make clusters of similar objects, (iv) find associations and dependencies to characterize data, and (v) detect deviations after looking for general trends. They are carried out using different methods, some of which are derived from statistics and others from the field of machine learning.

SDM Tasks	Statistics	Machine Learning
Summarization	Global autocorrelation Density analysis Smooth and contrast analysis Factorial analysis	Generalization Characteristic rules
Class identification	Spatial classification	Decision trees
Clustering	Point pattern analysis	Geometric clustering
Dependencies	Local autocorrelation Correspondence analysis	Association rules
Trends and deviations	Kriging	Trend rules

Table 1: Comparison between statistical and machine learning approaches to SDM

The rest of this section is devoted to describing data mining tasks that are dedicated to GIS.

V. SPATIAL DATA SUMMARIZATION:

The main goal is to describe data in a global way, which can be done in several ways. One involves extending statistical methods such as variance or factorial analysis to spatial structures. Another entails applying the generalization method to spatial data.

STATISTICAL ANALYSIS OF CONTIGUOUS OBJECTS:

Global autocorrelation: The most common way of summarizing a dataset is to apply elementary statistics, such as the calculation of average, variance, etc., and graphic tools like histograms and pie charts. New methods have been developed for measuring neighborhood dependency at a global level, such as local variance and local covariance, spatial auto-correlation by Geary, and Moran indices. These methods are based on the notion of a contiguity matrix that represents the spatial relationships between objects. It should be noted that this contiguity can correspond to different spatial relationships, such as adjacency, a distance gap, and so on.

Density analysis: This method forms part of Exploratory Spatial Data Analysis (ESDA) which, contrary to the autocorrelation measure, does not require any knowledge about data. The idea is to estimate the density by computing the intensity of each small circle window on the space and then to visualize the point pattern. It could be described as a graphical method.

Smooth, contrast and factorial analysis: In density analysis, non-spatial properties are ignored. Geographic data analysis is usually concerned with both alphanumerical properties (called attributes) and spatial data. This requires two things: integrating spatial data with attributes in the analysis process, and using multidimensional data to analyze multiple attributes. To integrate the spatial neighborhood into attributes, two techniques exist that modify attribute values using the contiguity matrix.

The first technique performs a smoothing by replacing each attribute value by the average value of its neighbors. This highlights the general characteristics of the data. The other contrasts data by subtracting this average from each value. Each attribute (called variable) in statistics can then be analyzed using conventional methods. However, when multiple attributes (above tree) have to be analyzed together, multidimensional data analysis methods (i.e. factorial analysis) become necessary. Their principle is to reduce the number of variables by looking for the factorial axes where there is maximum spreading of data values. By projecting and visualizing the initial dataset on those axes, the correlation or dependencies between properties can be deduced. In statistics and especially in the above methods, the analyzed objects were originally considered to be independent. The extension of factorial analysis methods to contiguous objects entails applying common Principal Component Analysis or Correspondence Analysis methods once the original table is transformed using smoothing or contrasting techniques.

VI. GEOSTATISTICAL APPROACH

Geostatistics is a tool used for spatial analysis and for the prediction of spatio-temporal phenomena. It was first used for geological applications (the geo prefix comes from geology). Nowadays, geostatistics encompasses a class of techniques used to analyze and predict the unknown values of variables distributed in space and/or time. These values are supposed to be connected to the environment. The study of such a correlation is called structural analysis. The prediction of location values outside the sample is then performed by the “kriging” technique. It is important to remember that geostastics is limited to point set analysis or polygonal subdivisions and deals with a unique variable or attributes.

Under those conditions, it constitutes a good tool for spatial and spatio-temporal trend analysis.

VII. GEOVISUALIZATION

Geovisualization concerns the development of theory and method to facilitate knowledge construction through visual exploration and analysis of geospatial data and the implementation of visual tools for subsequent knowledge retrieval, synthesis, communication and use. As an emerging domain, geovisualization has drawn interests from various cognate fields and evolved along a diverse set of research directions, as seen in a recently edited volume on geovisualization. The main difference between traditional cartography and geovisualization is that, the former focuses on the design and use of maps for information communication and public consumption while the latter emphasizes the development of highly interactive maps and associated tools for data exploration, hypothesis generation and knowledge construction.

Geovisualization also has close relations with exploratory data analysis (EDA) and exploratory spatial data analysis (ESDA). Statistical graphics and maps and relies on the human expert to interact with data, visually identify patterns, and formulate hypotheses/models. However, to cope with today's large and diverse spatial data sets and facilitate the discovery and understanding of complex information, geovisualization needs to address several major challenges, including

(1) processing very large datasets efficiently and effectively;
(2) handling multiple perspectives and many variables simultaneously to discover complex patterns and
(3) the design of effective user interface and interactive strategy to facilitate the discovery process.

To process large data sets and visualize general patterns, visual approaches are often combined with computational methods (such as clustering, classification, and association rule mining) to summarize data, accentuate structures and help users explore and understand patterns

VIII. CONCLUSION

The abundance of spatial data provides exciting opportunities for new research directions but also demands caution in using these data. The data are often from different sources and collected for different purposes under various conditions, such as measurement uncertainty, biased sampling, varying area unit, and confidentiality constraint. It is important to understand the quality and characteristics of the chosen data.

Among the other issues in the area of spatial data mining, one approach is to consider the temporality of spatial data, while another is to see how linear or network shape (like roads) can have a particular influence on graphical methods. In any event, it remains essential to continue enhancing the performance of these techniques. One reason is the enormous volumes of data involved, another is the intensive use of spatial proximity relationships. In the case of graphical methods, these relationships could be optimized using spatial indexes.

New types of data and new application areas (such as the analysis of moving objects and trajectories, spatially embedded social networks, spatial information in web-based documents, geocoded multimedia, etc.) have significantly expanded the frontier of spatial data mining research. Handling the very large volume and understanding complex structure in spatial data are another two major challenges for spatial data mining, which demand both efficient computational algorithms to process large data sets and effective visualization approaches to present and explore complex patterns.

REFERENCES

[1] M.Hemalatha.M; Naga Saranya.N. A Recent Survey on Knowledge Discovery in Spatial Data Mining, IJCI International Journal of Computer Science, Vol 8, Issue 3, No.2, may, 2011.
[2] Shiode, S., & Shiode, N. (2009). Detection of multi-scale clusters in network space. International Journal of Geographical Information Science, 23, 75–92.
[3] Pei, T., Zhu, A. X., Zhou, C., Li, B., & Qin, C. (2009). Detecting feature from spatial point processes using collective nearest-neighbor. Computers, Environment and Urban Systems, 33(6), 435–447.
[4] Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann. Rogerson, P., & Yamada, I. (2009). Statistical detection and surveillance of geographic clusters. Taylor and Francis Group.
[5] 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.
[6] Raymond T. Ng and Jiawei Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", IEEE Computer Society.