Advanced Data Mining Approaches of Knowledge Data Discovery for Geospatial Data

Aamir khan[1] , Dr. Umesh Chandra [2] and Dr. Parag Jain[3]

[1]Uttarakhand Technical University, Dehradun

[2]Glocal University, Saharanpur (U.P.)

[3]RIT Roorkee Uttarakhand

Abstract: This paper discusses the concept of data mining that helps in knowledge data discovery of spatial database. It represents approaches to extract data to gain data discovery of spatial data. All the phases of data mining are discussed with algorithm that helps in extraction of spatial database.This study includes the various steps to extract knowledge discovery from the geospatial data. Various techniques of spatial data mining are also discussed that eventually help in interpreting patterns.Knowledge based data discovery has now globally magnetized the attention of academic, government agencies, industry and many other administrations. Data mining and knowledge based data discovery can be categorized in the disciplinary area of previous geospatial data handling theory and methodology. The maximization of volume and varying structures of collected geospatial data offer challenges in storing, arranging, negotiate, processing, visualizing and analysing various kind of data.This paper requisites the existing geospatial data mining patterns, techniques and theories to understand and determine the capability to handle emerging geospatial data that focus on knowledge based data.

Keywords: Spatial data, Data mining, Knowledge data discovery

1. Introduction

Data analyst, Business Intelligence and people from the field of statistics usually deals with the term of data mining whereas “Data discovery” word was first used by KDD workshop in 1989 to represent the knowledge result in the end of the data-driven process[1]. Knowledge data discovery is the process of determining patterns that are understandable, valid, novel [2]. Geospatial data stands for the data which is related to geographical location. It may be vector or raster format of data. Geospatial data is very important for the government as well as private agencies. About 90% of government policies depend upon the geographic information. Unstructured data which is captured by remote sensing device is to be store in database, for unknown information which can be extract with the help of patterns matching process /algorithm, an implicit process which is define as knowledge data discovery process.GIS is usedto represent all position from earth’s surface. Position means a mapping system which is used to projected object that are different and also include different attributes.

Relational databases are used to store data (facts, figures) in tables. Similarly spatial database store and query data in that represent objects defined in space. Spatial database deals with geometric objects (line, point’s anal 3D objects).In **1998** Open Geospatial Consortium added a new functionality to database and extend the database into spatial database[3]. In the world of Big Data, variety, volume, velocity kinds of terms are used to describe the production of data. Big Data consist of a mixture of different kinds of data i.e. structure and as well as unstructured. This data has images, text, videos, emails and spatial data too. A large portion of Big Data is related to geographic data. This geographic data helps in discovering new weather forecast and earthquake zones etc.

According to upcoming next 15-20 years trends and co-relation in present geospatial data that must be help to change overall structure of geo-location view and process of decision making for government agencies and as well private sector[13].

1.1 **Features of Spatial Database**[4]

	Features
1.	Spatial Index Used for fast Access
2.	Spatial Measurement
3.	Spatial Predicates
4.	Geometric constructors
5.	Observer Function

The above features help researchers to predict pattern from geospatial database. This prediction of pattern is termed as Knowledge data discovery [5]. Patterns in large spatial database results in traffic controlling, environmental controlling, satellite, remote sensing, navigation etc. Geospatial is actually a combination of geography world with spatial term that tells about an object with the help of co-ordinate (x,y) anywhere in the place present on earth.

2. Objective: Discover Patterns

2.1. Data Selection:

During data selection select the relevant set of spatial features from spatial database management system. GIS provide operations to analyse and to work with spatial data. Like in above figure there are geospatial data represented through different colours and dimensions.

2.2. Data Reduction:

Data reduction is a technique to reduce volume of data without compromising its originality.Some of the famous techniques for data reduction is data compression, data duplication etc. Data reduction is reducing geographic data with the help of filters. These filters remove non-spatial attribute of the selected Data. Low pass filter and high pass filter are used for removal for spurious data.

2.3. Data Transformation:

Data transformation is converting the data values from geospatial database to another format of a new destination system. This conversion is done by various algorithm and programming languages. Smoothing, Aggregation, Normalization, Discretization are some of the techniques for data transformation.

2.4. Transformation of Geospatial data into useful information

Transformation stands change the objects attributes such as angle, direction ,scalability ,measuring points etc. transformation is very important step to develops a knowledge domain project . In each and every iteration of project result may be effected and work as an input for next phase.

Basically transformation is done for better data, for data mining is proposed and developed.

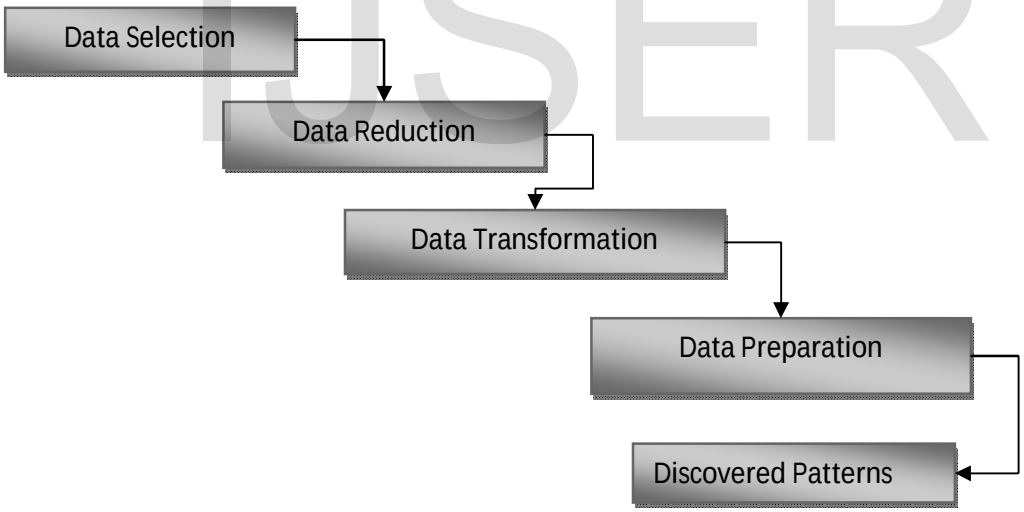
Various methods are used in this phase:

3.1-dimension reduction
3.2 – best features/attributes selection
3.3- extraction
3.4- Attributes transformation
3.1.1- numerical attributes
3.1.2 – functional attributes

This step can be crucial for the success of the entire KDD project and it also focus to project specific .The main point should be remember it that right transformation at the beginning we may obtain a surprising effect that give a hint about the transformation needed in the next phase. KDD process reflect upon itself[14].

3. Data Preparation:

Data preparation involves maximum process of entire KDD. It interprets patterns on the basis of input data sets [2]. After data reduction data preparation times get reduced. Data preparation is an important phase for data mining. Data improves the quality of data and that's ultimately improves the data mining result.

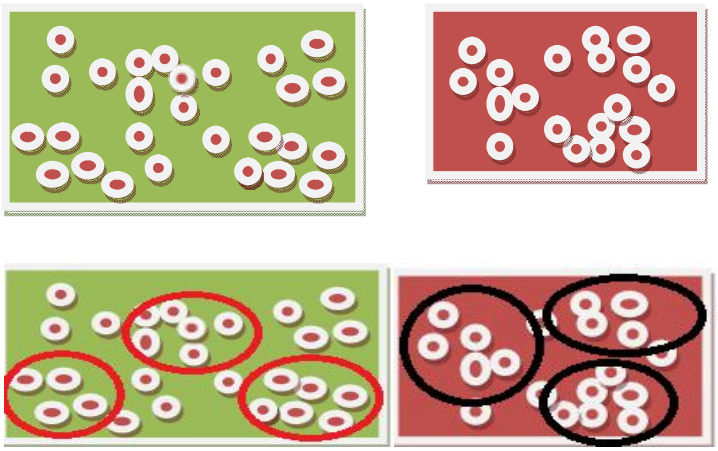


Above figure is in the respect of geospatial data

4. Various techniques on Spatial Data Mining

4.1. Spatial Clustering

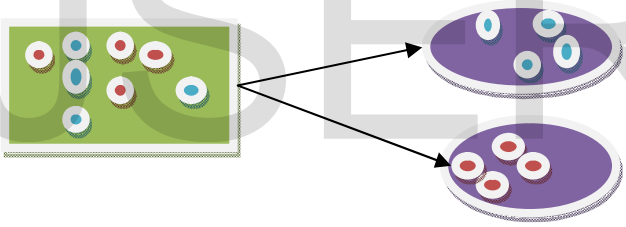
Spatial Clustering is grouping of data of similar spatial objects. It is done by identifying land same land area and earth data and grouping this area into one cluster [6]. After grouping different clusters various a particular patterns can be achieved for knowledge data discovery. Some of the famous clustering algorithms used for spatial data are: DBSCAN [7], OPTICS [8], Chameleon [9], and K-means [10].



Spatial Clustering

4.2. Spatial Classification

Classification of spatial data is according to the analysis of spatial data. The features of the spatial objects (river,glacier, and plateau) are characterized on the basis of some predefined characters. KNN algorithm [11] is an example of spatial classification



Data classification

Red Dots: Area with less rain falls, Green Land: area with high vegetation
Blue dots: Area with high rain fall, above figure helps in classifying various dots of land area with some features.

4.3. Spatial Association

Spatial Association consist of some association rules that define the relation between spatial data and event occurring. Some of the commands used for spatial association rules are **MAPSPEC** subcommand that specifies the map specification file, **DATASET** subcommand that specifies the data source that contains event data [10].

Example of Spatial Association is:

Isa(X,city)^within(X,DB)^adjacent to(X, Water)→close to(X,SA) is 92%

A city in Dubai which is adjacent to water and close to Saudi Arabia is 92%. Above statement helps to determine it with the help of predicates like X→Y where some of the predicates are spatial [10].

4.4. Learning of Environmental Data

The objective of analysis of Environmental data is the study of the ecology, conservation and preservation of the environment.There are various methods to analyse environmental data for example collection of data, data can be the count of population, temperature, Distance. Methods like regression and Analysis of variance are some of the methods to extract the statistics of the environment data. Environmental data are of different types, continuous data, Count Data, proportion data, Binary data, circular data,and time series data. Different data are deals with different methods to gain the measure of the respective variable[11].

4.5. Data mining of geospatial data

Data mining means that fetch useful knowledge or pattern contents which is extracted from large amount of data set .After that first of all we choose data mining task then we come to the position for ready to decide which type of data mining is suitable for our project and useful for further expectation.

- Classification
- Regression analysis and trends
- Clustering /segmentation
- Association
- Deviation
- Generalization

Data mining and KDD process mostly depends upon the behavioural structure of our task and one major factor which is depend upon the previous step.

There are two major goals in data mining:-

Prediction: prediction describes the supervised data mining.

Description: while description data mining combined the unsupervised data and it also focus visualization aspect of data mining.

Conclusion

This paper concludes that there are different techniques to interpret patterns from spatial data mining. Different phases from data selection to data preparation are discussed that conclude pattern interpretation and knowledge data discovery. Geospatial data is one of the finest fields where data can be used to conclude knowledge data discovery.

References

1. Piatetsky-Shapiro, G. 1991. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Work- shop. AI Magazine 11(5): 68–70
2. From data mining to data discovery
3. https://www.geospatialworld.net/article/spatial-database-services-for-location-aware-applications/
4. Han j. and Kamber , M. 2000, Data Mining : concepts and Techniques, Morgan Kaunfman
5. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei X,1996

6. Chameleon: hierarchical clustering using dynamic modeling - IEEE ...
7. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
8. https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/
9. https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/syn_spatial_association_rules_overview.html].
10. Discovery of association rule in geographic Information Database by Krzysztof of koperski and Jiawei Han by School of computing Science, Simon Fraser University
11. Analysis of Environmental data Conceptual Foundations: Environmental data
12. https://books.google.co.in/books?hl=en&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=data+selection+in+data+mining+for+geospatial+data&ots=tyMv1ZpC1X&sig=RG77c5Ij7nruK_YZoVf2YUeIZLI#v=onepage&q&f=false
13. "Application of geographical concepts and spatial technology to the Internet of Things Research Memorandum 2013-33 Erik van der Zee Henk Scholten Application of geographical concepts and spatial technology to the Internet of Things The role of location in real-time smart environments," 2013.
14. https://blog.udemy.com/knowledge-discovery-in-databases