GNITED MINDS
Journals

AN ANALYSIS ON EFFICIENT STRATEGIES AND APPLICATIONS OF SPATIAL DATA MINING

An Analysis on Efficient Strategies and Applications of Spatial Data Mining

Sujeet Varshney

Research Scholar, Maharishi Dayanand University, Rajasthan

Abstract – Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. The requirements of mining spatial databases are different from those of mining classical relational databases. The spatial data mining techniques are often derived from spatial statistics, spatial analysis, machine learning, and databases, and are customized to analyze massive data sets. In this report some of the spatial data mining techniques have discussed along with some applications in real world.

Clustering is one of the major tasks in data mining. In the last few years, Clustering of spatial data has received a lot of research attention. Spatial databases are components of many advanced information systems like geographic information systems VLSI design systems. In this paper, we introduce several efficient algorithms for clustering spatial data.

Spatial clustering, which groups similar spatial objects into classes, is an important component of spatial data mining. Spatial clustering can be used in the identification of areas of similar land usage in an earth observation database or in merging regions with similar weather patterns, etc. As a data mining function, spatial clustering can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. It may also serve as a preprocessing step for other algorithms, such as classification and characterization, which will operate on the detected clusters.

Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. In this paper, we explore whether clustering methods have a role to play in spatial data mining.

- - - - - - - - - - - - - X - - - - - - - - - - - - - -

INTRODUCTION

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets, including NASA, the National Imagery and Mapping Agency (NIMA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT). These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology.

General purpose data mining tools such as Clementine and Enterprise Miner, are designed to analyze large commercial databases. Although these tools were primarily designed to identify customer-buying patterns in market basket data, they have also been used in analyzing scientific and engineering data, astronomical data, multi-media data, genomic data, and web data. Extracting interesting and useful patterns from spatial data sets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation.

Specific features of geographical data that preclude the use of general purpose data mining algorithms

are: i) rich data types(e.g., extended spatial objects) ii) implicit spatial relationships among the variables, iii) observations that are not independent, and iv) spatial autocorrelation among the features.

The spatial data mining can be used to understand spatial data, discover the relation between pace 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. 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 (camalig) and other related data and knowledge bases, is original data of the spatial data mining.

In the event that data mining is about concentrating examples from substantial databases, then the biggest databases have an in number spatial part. Case in point, the Earth Observation Satellites, which are systematically mapping the whole surface of the earth, gather in the vicinity of one terabyte of data consistently. Other expansive spatial databases might be the U.s. enumeration, and the climate and atmosphere databases. The necessities of mining spatial databases are not the same as those of mining established social databases. Specifically, the thought of spatial autocorrelation that comparative items have a tendency to bunch in geographic space, is key to spatial data mining.

The complete data-mining process is a blend of numerous sub processes which are worthy of study in their right.. Some significant sub processes are data extraction and data cleaning, characteristic determination, calculation outline and tuning, and the analysis of the yield when the calculation is connected to the data. For spatial data, the issue of scale the level of aggregation at which the data are constantly examined, is additionally exceptionally critical. It. is well known in spatial analysis that indistinguishable investigations at. diverse levels of scale can at times lead to contradictory results. Our center in this part is constrained to the configuration of data-mining algorithms.

Specifically we portray how traditional data-mining algorithms might be stretched out to model the spatial autocorrelation property. Here it. is essential to comprehend the qualification between spatial data mining and spatial data analysis. As the name intimates, spatial data analysis blankets a wide range of strategies that arrangements with both the spatial and nori spatial qualities of the spatial items. Then again spatial data mining procedures are frequently inferred from spatial statistics, spatial analysis, machine taking in and data bases, and are modified to dissect gigantic data sets.

Spatial Database Systems (SDBS) are database systems for the management of spatial data. Both, the number and the extent of spatial databases are quickly developing in provisions for example, geo marketing, movement control and ecological studies. This development by far surpasses human limits to break down the databases with a specific end goal to discover verifiable regularities, manages or bunches covered up in the data. Thusly, mechanized knowledge discovery gets to be more essential in spatial databases. Knowledge discovery in databases (KDD) is the non-minor extraction of implied, at one time obscure, and possibly convenient information from databases.

The meaning of such a set of fundamental operations and their effective backing by an SDBS will speed up both, the improvement of new spatial KDD algorithms and their execution.

The computerization of numerous business and government transactions and the developments in experimental data gathering devices give us an enormous and constantly expanding measure of data. This hazardous development of databases has far outpaced the human capability to decipher this data, making an dire necessity for new strategies and instruments that backing the human in converting the data into advantageous information and knowledge. Knowledge discovery in databases (KDD) has been characterized as the non-insignificant process of identifying substantial, novel, and possibly functional, and at last justifiable designs from data. The process of KDD is intelligent and iterative, including some steps, for example, the accompanying ones:

- Selection: selecting a subset of all qualities and a subset of all data from which the knowledge ought to be identified.

- Data lessening: utilizing dimensionality diminishment or conversion systems to lessen the viable number of ascribes to be acknowledged.

- Data mining: the provision of suitable algorithms that, under satisfactory computational proficiency confinements, prepare a specific list of examples over the data.

- Evaluation: translating and assessing the ran across examples regarding their functionality in the given provision.

Spatial Database Systems (SDBS) are database systems for the management of spatial data. To discover verifiable regularities, administers or examples covered up in expansive spatial databases, e.g. for geo-promoting, activity control or ecological studies, spatial data mining algorithms are

exceptionally significant for a diagram of spatial data mining).

Most existing data mining algorithms run on divide and uniquely ready records, yet mixing them with a database management system (DBMS) has the accompanying preferences. Excess space and potential inconsistencies might be kept away from. Besides, business database systems offer different list structures to help distinctive sorts of database questions. This practicality can be utilized without additional execution exertion to speed-up the execution of data mining algorithms (which, as a rule, need to perform numerous database inquiries). Like the social standard dialect SQL, the utilization of standard primitives will speed-up the advancement of new data mining algorithms also will likewise make them more portable.

SPATIAL DATA MINING TASKS

Basic tasks of spatial data mining are:

A. Classification

An object can be classified using its attributes. Each classified object is assigned a class. Classification is the process of finding a set of rules to determine the class of an object.

B. Association Rules

Find (spatially related) rules from the database. An 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%)."

C. Characteristic Rules

The characterization of a selected part of the database has been defined in as the description of properties that are typical for the part in question but not for the whole database. In the case of a spatial database, it takes account not only of the properties of objects, but also of the properties of their neighborhood up to a given level**.**

D. Discriminant Rules

Describe differences between two parts of database e. g. find differences between cities with high and low unemployment rate.

E. Clustering

Clustering means it is the process of grouping the database items in to clusters. All the members of the cluster has similar features. Members belong to different clusters has dissimilar features.

F. Trend Detection

Finds trends in database. A trend is a temporal pattern in some time series data. Spatial trend is defined as follows: consider a non-spatial attribute which is the neighbor of a spatial data object. The pattern of changes in this attribute is called spatial trend.

MOTIVATING SPATIAL DATA ALINING

We now present an illustration which will be utilized all around this section to outline the distinctive ideas in spatial data mining. We are given data in the vicinity of two wetlands on the shores of Lake Erie in Ohio, USA, to anticipate the spatial appropriation of a bog reproducing winged animal, the red-winged blackbird (Agelaius phoeniceus). The names of the wetlands are Darr and Stubble, and the data was gathered from April to June in two progressive years, 1995 and 1996.

An uniform lattice was forced on the two wetlands, and diverse sorts of estimations were recorded at each one unit or pixel. The span of every pixel was five meters. The qualities of seven qualities were recorded at each one unit, obviously area knowledge is pivotal in choosing which traits are imperative and which are definitely not.

For instance, Vegetation Durability was picked over Vegetation Species in light of the fact that specific knowledge about the settling propensities of the red-winged blackbird prescribed that the decision of home area is more reliant on the plant structure and its imperviousness to wind furthermore wave activity than on the plant species.

Measures of Spatial Form : Mean focus is the normal area, figured as the mean of X and mean of Y directions. The mean focus is otherwise called the inside of gravity of a spatial dispersion. Regularly the weighted mean focus is proper measure for some spatial provisions, for e.g., focal point of populace. The weighted mean focus is registered as the degree between the aggregate of the directions of each one focus multiplied by its weight (e.g., number of individuals in piece) furthermore the total of the weights. The measure focus is utilized as a part of some structures. It might be utilized to streamline complex items (e.g., to stay away from space prerequisites and unpredictability of digitations of verges, a geographic item could be spoken to by its focus), or for distinguishing the best area for an arranged movement (e.g. a dissemination focus

ought to be spotted a main issue with the goal that make a trip to it is minimized).

Scattering is a measure of the spread of an appropriation around its focus. Regularly utilized measures of scattering and variability are reach, standard deviation, change and coefficient, of difference. Scattering measures for geographical disseminations are regularly figured as the summation over the proportion of the weight of geographic articles and the closeness between them. Shape is multi-dimensional, and there is no single measure to catch the sum of the measurements of the shape. A hefty portion of shape measures are dependent upon correlation of the shape's border with that of a ring of the same area.

The Data-Mining Trinity : Data mining is a without a doubt multidisciplinary area, and there are numerous novel methods for concentrating designs from data. Still, if one were to name data-mining systems, then the three most non-controversial marks might be arrangement, grouping, and cooperation standards. When we depict each of these classes in portion, we exhibit some illustrative illustrations where these systems might be connected.

The objective of characterization is to gauge the worth of a trait of a connection dependent upon the worth of the connection's different characteristics. Numerous issues could be communicated as characterization issues. Case in point, determining the areas of homes in a wetland based upon the worth of different characteristics (vegetation toughness, water profundity) is a grouping issue frequently additionally called the area expectation issue. Also, foreseeing where to need problem areas in wrongdoing action might be given a part as an area forecast issue. Retailers basically settle a area expectation issue when they choose an area for another store. The well-known representation in land, "Location is everything," is a prevalent indication of this issue.

EXPLORATORY DATA ANALYSIS TECHNIQUES

Spatial data analysis comprises a broad spectrum of techniques that deals with both spatial and non-spatial characteristics of spatial objects. Exploratory techniques allow to investigate first and second order effects of the data. First order variation informs about large scale global trend of phenomena which is spatially distributed through the area. The second order variation defines dependence between observations.

- Global Autocorrelation (Moran's I): Moran's I is a measure of global spatial autocorrelation. Global or local autocorrelation reveal feature similarity based location and attribute values to explore the pattern whether it is clustered, dispersed, or random.

- Hot Spots (Getis-Ord): The G-statistic is often used to identify whether hot spots or cold spots exist based on so-called distance statistics. Hot spots are regions that stand out compared with the overall behavior prevalent in the space. Hot spots can be detected by visualizing the distribution in format of choropleth or isarithmic maps.

- Local Autocorrelation (Auseliu's Local Moran I): Hie local Moran statistic is used as a local indicator of spatial association which is calculated for individual zones around each observation within a defined neighborhood to identify similar or different pattern in nearby. Because the distribution of the statistic is not known,
high positive or high negative standardized scores of Ii are taken as indicators of similarity or dissimilarity respectively.

- Density (Kernel): Kernel density estimation is a nonparametric unsupervised learning procedure (classifier). Kernel, k is bivarate probability density function which is symmetric around the origin.

SPATIAL DATA MINING: AN ALGORITHMS

To help our claim that the expressivity of our spatial data mining primitives is satisfactory, we exhibit how ordinary spatial data mining algorithms could be combined with a spatial DBMS by utilizing the database primitives presented as a part of area.

Spatial Clustering : Clustering is the errand of assembling the objects of a database into genuine subclasses (that is, bunches) with the goal that the parts of a bunch are as comparable as could be expected under the circumstances though the parts of distinctive bunches contrast however much as could be expected from one another. Provisions of bunching in spatial databases are, e.g., the discovery of seismic blames by assembling the passages of a tremor inventory or the formation of topical maps in geographic information systems by grouping characteristic spaces.

Distinctive sorts of spatial grouping algorithms have been proposed, e.g. k-medoid grouping algorithms for example, CLARANS. This is a case of a worldwide clustering calculation (where a change of a solitary database item may impact all groups) which can't make utilization of our database primitives in a common manner. Then again, the essential thought of a solitary output calculation is to assembly neighboring objects of the database into groups dependent upon a nearby group condition performing stand out output through the database. Single output grouping algorithms are productive if the recovery of the neighborhood of an item might be effectively performed by the SDBS. Note that nearby group conditions are generally underpinned by our

database primitives, specifically by the neighbors operation on a suitable neighborhood chart. The algorithmic composition of single output grouping is delineated in figure 1. Diverse bunch conditions yield distinctive thoughts of a group and diverse grouping algorithms.

For example, *GDBSCAN* (*Generalized Density Based Spatial Clustering of Applications with Noise*) relies on a density-based notion of clusters. The key idea of a density based cluster is that for each point of a cluster its *Eps*-neighborhood for some given *Eps* > 0 has to contain at least a minimum number of points, i.e. the "density" in the *Eps*-neighborhood of points has to exceed some threshold. This idea of "density-based clusters" can be generalized in two important ways. First, any notion of a neighborhood can be used instead of an *Eps*-neighborhood if the definition of the neighborhood is based on a binary predicate which is symmetric and reflexive.

Second, instead of simply counting the objects in a neighborhood of an object other measures to define the "cardinality" of that neighborhood can be used as well. Whereas a distance-based neighborhood is a natural notion of a neighborhood for point objects, it may be more appropriate to use topological relations such as *intersects* or *meets* to cluster spatially extended objects such as a set of polygons of largely differing sizes. for a detailed discussion of suitable neighborhood relations for different applications.

Figure 1. Schema of single scan clustering algorithms

Spatial Characterization : The task of *characterization* is to find a compact description for a selected subset of the database.

In this section, we discuss the task of characterization in the context of spatial databases and review two relevant methods.

The algorithm to find spatial association rules consists of 5 steps. Step 2 (coarse spatial computation) and step 4 (refined spatial computation) involve spatial aspects of the objects and are briefly examined in the following. Step 2 computes spatial joins of the object type to be characterized (such as town) with each of the other specified object types (such as water, road, boundary or mine) using a neighborhood relation (such as close-to). For each of the candidates obtained from step 2 (and which passed an additional filter-step 3), the exact spatial relation, for example *overlap*, is determined in step 4. Finally, a relation such as the one depicted in figure 2 results which is the input for the final step of rule generation. It is easy to see that the spatial steps 2 and 4 of this algorithm can be well supported by the neighbors operation on a suitable neighborhood graph.

Town	Water	Road	Boundary
Saanich	<meet, J.FucaStrait>	<overlap,highway1>, <close-to,highway17>	<close-to,US>
PrinceGeorge		<overlap, highway97>	
Petincton	<meet,OkanaganLake>	<overlap, highway97>	<close-to,US>
. . .	. . .	. . .	. . .

Figure 2. Input for the step of rule generation.

This paper introduces the following definition of spatial characterization with respect to a database and a set of target objects which is a subset of the given database. A *spatial characterization* is a description of the spatial and non-spatial properties which are typical for the target objects but not for the whole database. The relative frequencies of the non-spatial attribute values and the relative frequencies of the different object types are used as the interesting properties. For instance, different object types in a geographic database are communities, mountains, lakes, highways, railroads etc. To obtain a *spatial* characterization, not only the properties of the target objects, but also the properties of their neighbors (up to a given maximum number of edges in the relevant neighborhood graph) are considered.

CLUSTERING METHODS FOR SPATIAL DATA MINING

A. Partitioning Around Medoids(PAM)

PAM is similar to K- means algorithm. Like k- means algorithm, PAM divides data sets in to groups but based on mediods. Whereas k- means is based on centroids. By using mediods we can reduce the dissimilarity of objects within a cluster. In PAM, first calculate the mediod, then assigned the object to the nearest mediod, which forms a cluster.

B. Clustering LARge Applications(CLARA)

Compared to PAM, CLARA can deal with much larger data sets. Like PAM CLARA also finds objects that are centrally located in the clusters. The main problem with PAM is that it finds the entire dissimilarity matrix at a time. So for n objects, the space complexity of PAM becomes O($n2$). But CLARA avoid this problem. CLARA accepts only the actual measurements (i.e,. *n ´ p* data matrix).

CLARA assigns objects to clusters in the following way:

- BUILD-step*:* Select *k* "centrally located" objects, to be used as initial medoids. Now the smallest possible average distance between the objects to their mediods are selected, that forms clusters.

- SWAP-step: Try to decerase the average distance between the objects and the mediods.

This is done by replacing representative objects. Now an object that does not belong to the sample is assigned to the nearest mediod.

C. Spatial dominant approach SD (CLARANS)

In SDCLARANS, all the data containing spatial components are collected. After that clustering is used based on CLARANS. It should be mentioned that CLARANS is used to find the most natural number, knat, of clusters. One may ask, how is *nat k* determined in the first place. It is indeed a very difficult and open question. The authors however, adopt a heuristic of detemining *nat k* , which uses silhoutee coefflients , introduced by Kaufman and Rousseeuw. Each of the clusters thus obtained is processed by generalizing its nonspatial components using DBLEARN. Note that this algorithm differs from the spatial dominant generalization algorithm (without clustering), in that the latter requires the user to provide the spatial concept hierarchies. however, in this case, it can be said that SD(CLARANS) computes spatial hierarchy dynamically. The hierarchy thus found is more "data oriented" rather than "human oriented".

D. Non- spatial dominant approach NSD (CLARANS)

This nonspatial dominant approach first applies nonspatial generalizations and spatial clustering afterwards. DBLEARN is used to perform attribute-oriented generalizations of the nonspatial attributes and produce a number of generalized tuples. Then, for each such generallized tuple, all the spatial components are collected and clustered using CLARANS to find *nat k* clusters. In the final step, the clusters thus obtained are checked to see if they overlap with eachother. If so, then the clusters are merged, and the corresponding generalized tuples are merged as well.

If the rules to find are nonspatial characterizations of spatial attributes, then SD(CLARANS) has an edge. This is because NSD(CLARANS) separates the objects into different groups before clustering which may weaken the inter object similarity, or cluster tightness. On the other hand, NSD(CLARANS) is suitable if the spatial clusters within groups of data that has been generalized nonspatially is sought. However, both algorithms arrive at the same result (or rules).

CONCLUSION

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. Hie 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

The main objective of the spatial data mining is to discover hidden complex knowledge from spatial and not spatial data despite of their huge amount and the complexity of spatial relationships computing. However, the spatial data mining methods are still an extension of those used in conventional data mining. Spatial data is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. Spatial Data Mining extends relational data mining with respect to special features of spatial data, like mutual influence of neighboring objects by certain factors (topology, distance, direction).

It is based on techniques like generalization, clustering and mining association rules. Some algorithms require further expert knowledge that cannot be mined from the data, like concept hierarchies. Spatial data mining is a niche area within data mining for the rapid analysis of spatial data. Spatial data can potentially influence major scientific challenges, including the study of global climate change and genomics.

REFERENCES

- Agrawal R., Imielinski T., Swami A.: *"Database Mining: A Performance Perspective"*, IEEE Transactions on Knowledge and Data Engineering, Vol.5, No.6, 2003, pp. 914-925.

- Brinkhoff T., Kriegel H.-P., Schneider R., Seeger B.: *'Efficient Multi-Step Processing of Spatial Joins'*, Proc. ACM SIGMOD Int. Conf. on Management of Data, Minneapolis, MN, 2004, pp. 197-208.

- Erwig M., Gueting R.H.: *"Explicit Graphs in a Functional Model for Spatial Databases"*, IEEE Transactions on Knowledge and Data Engineering, Vol.6, No.5, 2004, pp. 787-803.

- Ester M., Kriegel H.-P., Sander J.: *"Spatial Data Mining: A Database Approach"*, Proc. 5th Int. Symp. on Large Spatial Databases, Berlin, Germany, 2007, pp. 47-66.

- Koperski K., Adhikary J., Han J.: *"Knowledge Discovery in Spatial Databases: Progress and Challenges"*, Proc. SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Technical Report 96-08, University of British Columbia, Vancouver, Canada, 1996.

- Koperski K., Han J.: *"Discovery of Spatial Association Rules in Geographic Information Databases"*, Proc. 4th Int. Symp. on Large Spatial Databases, Portland, ME, 2005, pp.47-66.

- Lu W., Han J.: *"Distance-Associated Join Indices for Spatial Range Search"*, Proc. 8th Int. Conf. on Data Engineering, Phoenix, Arizona, 1992, pp. 284-292

- Ng R. T., Han J.:*"Efficient and Effective Clustering Methods for Spatial Data Mining"*, Proc. 20th Int. Conf. on Very Large Data Bases, Santiago, Chile, 1994, pp. 144-155.

- Sashi Sekhar ,Chang-Tien Lu and Pusheng Zhang. A Unified Approach To Detecting Spatial Outlier. *Published in:Journal Geoinformatica* Volume 7 Issue 2, June 2003

- 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 2003.

- Zhong yong a, Zhang jixian a, Yan qin. Research on spatial data mining technique applied in land use dynamic monitoring. *Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion. Volume XXXVI 2/W25, 2005*