International Journal of Internet of Things and its Applications
Vol. 1, No. 1 (2017), pp. 1–12
http://dx.doi.org/10.21742/ijiota.2017.1.1.01

Analyzing the Usage of Data Mining in Spatial Database

Weijie Kong

Harbin University of Commerce, China
1185596327@qq.com

Abstract

This paper is based on the data mining methods that are combined with Geographic Information Systems (GIS) methods for carrying out spatial analysis of geographic data. Spatial data mining is the application of data mining techniques to spatial data. Large amount (in terabytes) of data that may be obtained from satellite images, medical equipments, video cameras, etc. These are far beyond the human ability to analysis of them, which is urgently in need of spatial data mining. This makes spatial data mining become a promising research field. National Data Centre(NDC) which provides various spatial data types like digital maps, satellite images etc. In this paper we will go on discussing how the spatial data is gathered from different areas and modifying the data. By applying data correction techniques, we will remove errors obtained in transformation of data. A suitable technique like ID3, Claran's, Birch, DBSCAN are used.

Keywords: *Data Mining, Geographic Information System, Spatial Data, ID3, Claran's, Birch, DBScan.*

1. Introduction

Geographical Information System is generally characterized as a methodical integrating of equipment and programming for updating, capturing, storing, displaying, updating and dissecting spatial data. The immense development of spatial information of spatial databases accentuate the requirement for the disclosure of spatial learning. Spatial information mining is the procedure of finding fascinating and interesting obscure, yet conceivably helpful examples from spatial databases. Geographical data requires a merging of various information sources to make a solitary source in which all qualifications are coordinated to shape a solitary database.

1.1. Components of GIS

A working GIS integrates five key components: hardware, software, data, people, and methods.

1.1.1. Data: The most vital segment of a GIS is the information. Data is one of the important and most expensive, components of geographical features and their corresponding attributes information. Geographic information and related plain information can be gathered in-house or bought from a business information supplier. A GIS will combined spatial information with other information assets and can even uses a DBMS, utilized by most associations to sort out and keep up their information, to oversee spatial information. The process involves digital encoding the geographical features such as trees, buildings, rivers, boundaries.

1.1.2. People: The genuine force of GIS originates from the general population we use them.Over the previous decade, PCs have turned out to be much less demanding for individuals to utilize and more reasonable for organizations, schools and associations to buy. GIS innovation is of constrained worth without the general population who deal with framework and create plans for applying it to general issues. GIS clients range from specialized experts who create and keep up the framework to the individuals who use it to offer them some assistance with performing their ordinary work.

1.1.3. Methods: An effective GIS works as indicated by an all around composed arrangement and business principles, which are the models and working practices extraordinary to every association.

1.2. Characteristics of Spatial Data

SDM implies extracting verifiable learning, spatial relations. It need incorporating mining and spatial database innovation, the understanding that can be utilized to spatial data.Spatial information mining(SDM) implies removing certain spatial relations.Spatial articles have spatial area and separation properties. There is a sure connection between adjoining objects, so the relationship between the spatial information is more mind boggling (Not only including topological relations and course relations, additionally measure relationship is identified with spatial area and the separation between spatial items). There are evident contrast between spatial information and other type's information. Spatial information has the accompanying complex attributes.

1) Massive information Massive information regularly make a few calculations can not be executed for trouble or exorbitant figuring sum, and in this way one of the assignments of spatial information mining is to make another processing methodology and grow new productive calculation to beat the specialized challenges created by monstrous information.
2) Non-straight relationship between the spatial trait It is an imperative image of the many-sided quality of space frameworks, mirrors the unpredictable instruments of the framework inner capacity, and is one of the primary assignments of spatial information mining.
3) Scale normal for spatial information The accompanying law and in addition the encapsulating normal for spatial information is not the same at various watching levels. Scale trademark is another indication of the many-sided quality of spatial information, and can be utilized to investigate slow change law of the trademark during the time spent speculation and refinement of data.
4) Spatial measurement expanding The properties of spatial items increments quickly, as in the field of remote detecting, because of the fast improvement of sensor innovation, the quantity of groups expanded from a couple to tens or even hundreds, in this way, how to mine information and revelation learning from tens or even hundreds dimensional space gets to be another problem area study.
5) The uncertainty of spatial data Ambiguity exists in a wide range of spatial data, for example, the vagueness of spatial area, the equivocalness of spatial relationship, and in addition the characteristic estimations of fluffy, and so forth.

2. Literature Survey

- Jiayupan have introduced a new tool „geoplot" which is used to find spatial patterns within a single group of video clips and across two groups of video clips, where video clips are grouped around the globe to patterns across the spatial distributions of points.
- Xiaofangzhou proposed a database-oriented processing framework that optimizes the performance of spatial data generalization from inside a spatial database, rather than treating such operations as post-database processing.
- .sumathi and geetha they presented the techniques of spatial data mining in the following four categories Clustering and Outlier Detection, Association and Co-Location, Classification and Trend-Detection.They also discussed some trends and applications of spatial data mining.
- Raymond T. Ng and Jiawei Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", IEEE Computer Society.
- Leo at hangoro invented the GKD and is a dynamic field that is only just beginning at the time of this writing. GKD will continue to grow as the scope and volume of digital geo-references.

3. Spatial Clustering Methods

Group examination is an essential exploration theme at the field information mining. The supposed bunching [2], is gathering information objects in light of the likeness, fmding the conveyance attributes of spatial information, making each the information of each grouping has high likeness, and the information of various grouping distinctive however much as could be expected. Spatial bunching investigation is to isolate the objects of spatial database into diverse important sub-classes as indicated by a few qualities; objects of the same sub-class have certain qualities with high closeness, which has self-evident contrasts to qualities of various sub-classes.

The upsides of utilizing bunch investigation are: the structure then again bunches which are needed to acquire can be discovered straightforwardly from the information, does not require any foundation learning. As such, individuals have as of now proposed four sorts of spatial grouping techniques:

3.1. Segmentation-based method

Division based technique incorporates the K-normal technique , K-focus strategy and CLARANS grouping technique. They all receive an iterative repositioning innovation, and attempt to utilizing development of items among division to enhance grouping impact. Since this kind strategy is suitable for finding comparative size globular groups, so is regularly utilized as a part of the applications, for example, the office area.

3.2. Hierarchical-based technique

This kind technique is just breaking down the accumulation of objects. As indicated by disintegration method of various leveled, these strategies can be isolated into two sorts: union and division. The surely understood progressive based strategies are BIRCH calculation and CURE calculation , and so on.

3.3. Density-based strategy

For every information purpose of a given class, in a zone of given scope must contain information focuses that surpass a certain limit, and after that proceed with the group. It can

be utilized to find groups of subjective shape, and channel "commotion". Representation calculations are DBSCAN calculation , OPTICS calculation , GDBSCAN calculations , DBRS calculation and DENCLUE calculations , and so on.

3.4. Grid-based strategy

This sort of strategy uses a multi-determination framework information structure to isolate the space into set number units. Bunching operations carry on inside of the unit. The preparing time of this sort of strategy is autonomous of the number of articles, handling speed. The surely understood techniques are STING calculations ,WaveCluster calculation and CLIQUE calculation , and so forth.

4. K-means clustering algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed appriori. The main idea is to define k centers, one for each cluster. These centers should be placed in cunning way because of different location cause different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate is to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of clusters resulting from previous step. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

Where

'$||x_i - v_j||$' is the Euclidean distance between xi and vj.
Cj is the number of data points in ith cluster.
C is the number of cluster centers.

4.1. Algorithm steps for k-means clustering

Let X={x1,x2,x3,……xn} is the set of data points and V={v1,v2,v3…..,vc} be the set of centers

1) Randomly select „C" cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster centeris minimum of all cluster centers.
4) Recalculate the new cluster center using:
   where, „ci" represents the number of data points in ith cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

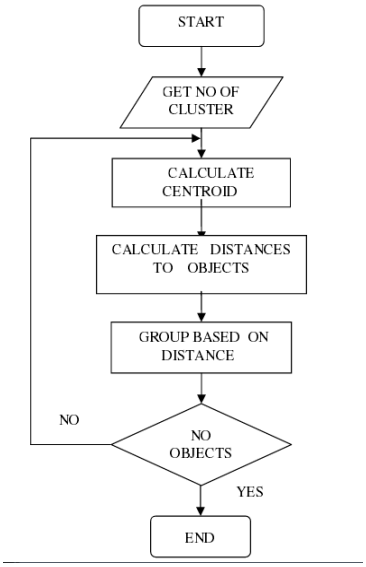

Figure 1. K-means flowchart

4.1.1. Advantages

1) Fast, vigorous and simpler to get it.
2) Relatively proficient: O(tknd), where n is #objects , k is # groups, d is # measurement of every article, and t is # cycles. Ordinarily, k, t, d<<n

4.1.2. Disadvantages

1) The learning calculation requires apriori determination of the quantity of group focuses.
2) The utilization of Exclusive Assignment - If there are two exceptionally covering information then k-means won't have the capacity to determine that there are two bunches.

4.2. STING Algorithm

Wang et al. proposed a STatisticalINformation Grid-based clustering technique (STING) to bunch spatial databases and to encourage district situated inquiries. STING isolates the spatial region into rectangular cells and stores the cells in a progressive lattice structuretree. Every cell (with the exception of leaves in the tree) is apportioned into 4 type cells at the following level with every kid relating to a quadrant of the guardian cell. A guardian cell is the union of its kids; the root cell at level 1 compares to the entire spatial region. The leaf levelcells are of uniform size, decided comprehensively from the normal thickness of articles. Forevery cell, both quality ward and trait free parameters of the factual data are kept up. These parameters are characterized in. STING keeps up synopsis insights for every cell in its various leveled tree. Therefore,factual parameters of guardian cells can without much of a stretch be registered from the parameters of youngster cells. Note that the circulation sorts might be typical, uniform, exponential and none. Esteem ofdist may either be doled out by the client or got by speculation tests, for example, the χ 2test. The algorithm is summarized in Algorithm

4.2.1. STING Algorithm

1) Determine a level in the first place.
2) For every cell of this level, we compute the certainty interim (or evaluated extent) of likelihood that this cell is pertinent to the inquiry.
3) From the interim computed above, we name the cell as important or not applicable.
4) If this level is the leaf level, go to Step 6; generally, go to Step 5.
5) We go down the progressive system structure by one level. Go to Step 2 for those cells that shape the significant cells of the more elevated amount.
6) If the determination of the question is met, go to Step 8; generally, go to Step 7.
7) Retrieve those information fall into the significant cells and do assist preparing. Return the result that meet the prerequisite of the inquiry. Go to Step 9.
8) Find the locales of important cells. Return those districts that meet the prerequisite of the question. Go to Step 9.
9) Stop.

4.2.2. Advantages

1) Query independent, simple to parallelize, incremental update
2) O (K), where K is the quantity of framework cells at the most minimal level

4.2.3. Disadvantages

1) All the bunch limits are either flat or vertical, and no slanting limit is identified

5. Implementation

5.1. ID3 Algorithm

ID3(Iterative Dichotomiser 3) is used to generate a decision tree.ID3 algorithm is invented by Ross Quinlan.Decision tree classifies data using the attributes,Tree consist of decision nodes and decision leafs.Nodes can have two or more branches which represents the value for attributes tested.

Algorithm:

1) Create a root node for the tree.
2) If all examples are positive, Return the single-node tree root with label = +.
3) If all examples are negative, Return the single-node tree root with label = -.
4) If number of predicting attributes is empty, then return the single node tree root with label = most common value of the target attribute in the examples.
5) Else
   A = The Attribute that best classifies examples.
   Decision Tree attribute for Root = A.
   For each possible value, *vi*, of A,
6) Add a new tree branch below Root, corresponding to the test A = *vi*.
7) Let Examples(*vi*), be the subset of examples that have the value *vi* for A.
8) If Examples(*vi*) is empty then below this new branch add a leaf node with label = most common target value in the examples
10) Else below this new branch add the subtree ID3 (Examples(*vi*), Target_Attribute, Attributes – {A})
11) End

12) Return Root

Formula to calculate Entropy is

$$Entropy(S) = \sum_{i=1}^{c} p_i \log_2 p_i$$

Table 1. The weather data (Witten and Frank; 1999, p.9)

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

In this dataset, there are five straight out properties viewpoint, temperature, humidity, bwindy, and play. We are occupied with building a framework which will empower us to choose whether or not to play the amusement on the premise of the climate conditions, i.e. we wish to foresee the estimation of play utilizing standpoint, temperature,stickiness, and breezy. We can think about the credit we wish to foresee, i.e. play, as the yield property,The main aim of the id3 algorithm is building a decision tree ,it consists of nodes and arcs

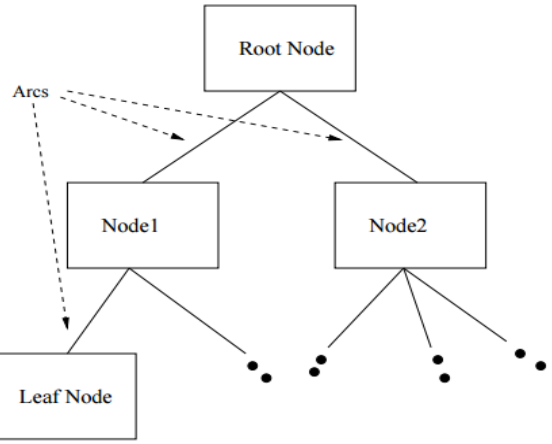

Figure 2. Basic Decision tree structure

Idea of ID3 algorithm is

1) Each non-leaf hub of a choice tree compares to a data property, and every circular segment to a conceivable estimation of that trait. A leaf hub relates to the normal estimation of the yield characteristic when the data properties are portrayed by the way from the root hub to that leaf hub.
2) In a "decent" choice tree, each non-leaf hub ought to compare to the info trait which is the most educational about the yield trait amongst all the information properties not yet considered in the way from the root hub to that hub. This is on account of we might want to anticipate the yield quality utilizing the littlest conceivable number of inquiries all things considered.

Entropy is nothing but measure of uncertainity in communication systems. The relations used for entropy and information gain are

For entropy

$$H(P) = -\sum_{i=1}^{n} p_i \log(p_i).$$

For information gain

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i).$$

Total information gain is

$$\text{Gain}(X,T) = H(T) - H(X,T).$$

For above data the information gain is

H(T)=H(Pr) = -(5/14 log(5/14) + 9/14 log(9/14))

= 0.94

$$\begin{aligned} H(\text{temperature},T) &= \frac{4}{14}H(T_{cool}) + \frac{6}{14}H(T_{mild}) + \frac{4}{14}H(T_{hot}) \\ &= \frac{4}{14}\left(-\left(\frac{1}{4}\log(\frac{1}{4}) + \frac{3}{4}\log(\frac{3}{4})\right)\right) + \frac{6}{14}\left(-\left(\frac{2}{6}\log(\frac{2}{6}) + \frac{4}{6}\log(\frac{4}{6})\right)\right) + \\ &\quad \frac{4}{14}\left(-\left(\frac{2}{4}\log(\frac{2}{4}) + \frac{2}{4}\log(\frac{2}{4})\right)\right) \\ &= \frac{4}{14} \times 0.811 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 1.00 \\ &= 0.911 \end{aligned}$$

The information gain for attribute *temperature* for set T is thus

$$\begin{aligned} \text{Gain}(\text{temperature},T) &= 0.940 - 0.911 \\ &= 0.029 \text{ bits.} \end{aligned}$$

5.1.1. Advantages

1) It builds a fastest tree.
2) Builds a shortest tree.
3) Finding leaf nodes enables test data to be pured, reducing number of tests.

5.1.2. Disadvantages

1) Information might be over-fitted or over-grouped, if a little example is tried.
2) Only one attribute at a time is tested.

6. Spatial Trend Analysis

Spatial trend patterns refers to the progressions of non-spatial objects when away given space object. For instance, the progressions pattern of monetary circumstance while getting further and more distant far from the downtown area. Its examination results might be a positive pattern, reverse pattern or no pattern. For the most part, dissecting spatial pattern on spatial information structure and spatial access techniques need to utilize relapse investigation and related examination strategies. Because of the disposition of space protests possess, the conventional relapse model may not be suitable. The above strategies frequently utilized artificially as a part of down to earth applications. Moreover, the information mining strategy need to coordinate completely with the traditional database innovation, the more the innovation utilized by information mining, the higher the precision of the outcomes.

7. Issues Faced by Spatial Data Mining

(1) The larger part of spatial information mining calculations is transplanted movement from general information mining calculations also, did not consider capacity, handling of the spatial information what's more, the spatial information itself qualities. Spatial information is unique in relation to the information of social database, and has their own spatial information access strategies, so customary information mining procedures are frequently not able to appropriately investigate complex spatial marvels and spatial items.

(2) The proficiency of spatial information mining calculations is definitely not high; the disclosure examples of spatial information mining calculations are not refined. The vulnerability showed up in procedure of spatial information mining, the likelihood of mistake examples and the measurement of the issue to be illuminated are all extensive, not just builds the calculation's inquiry space additionally expands the probability of visually impaired inquiry. In this way, we should use area learning to discover and evacuate information having nothing to do with the errand, successfully lessen the measurement of the issue, and outline a more compelling learning revelation calculation.

(3) There is no by and large acknowledged institutionalized spatial information mining question dialect. One reason of fast improvement of database innovation is nonstop change and improvement t of database question dialect. Along these lines, to always enhance and create spatial information mining must create spatial information mining question dialect to set up the establishment for effective spatial information mining and premise.

(4) Interaction of learning revelation arrangement of spatial information mining association is not solid. It is hard to completely furthermore, adequately use area

specialists' information in the learning revelation process. The client can't well control the spatial information mining process.

These issues portrayed above make it more hard to concentrate information from spatial database than customary social database, which have conveyed difficulties to spatial information mining research.

8. Development Trend of Spatial Data Mining

Spatial information mining is an extremely youthful and promising field, there are numerous issues require inside and out study. Because of the spatial information have qualities of huge, non-straight, multi-scale and uncertainty, so extricating learning from spatial database is more troublesome than removing information from customary social database, conveyed difficulties to spatial information mining research, which is examining and creating bearing of this field.

(1) Research of spatial information mining calculations and specialized. Spatial affiliation guideline mining calculation, time arrangement mining innovation, spatial assembled calculation, spatial characterization strategies, spatial anomaly calculation are examination problem area of spatial information mining, while enhancing the effectiveness of spatial information mining calculations are likewise essential.

(2) Pre-handling of multi-source spatial information. Spatial information incorporate computerized line information, picture information, and advanced height models and trait information of surface components. As its very own result many-sided quality and the challenges of information accumulation, it is definitely that opportunity esteem, commotion information and irregularities information exist in spatial information, so multi-source spatial information pre-handling turns out to be more vital.

9. Proposed System

The GIS field has developed rapidly especially in the last ten years.Developments in database technologies and the growing interest in GIS by many new disciplines have developed new questions and problems.

- Mobile GIS:Carrying mobile devices that hold geographic information is becoming very common.Receving online updates and querying a database while on road will be an major application area.Mobile GIS will be an important topic in major payoffs in research.
- Temporal GIS:The geographic landscape surrounding us changes will all the time.Man-made features are constantly built and other landscapes are disappear.Many of these changes are must be recorded.All man made changes such as dams,reservoirs,roads,buildings are updated for each period of time.
- Data Models:The two commonly used data models are raster and vector.Merging the raster and vector data is limited and fully incorporating the model remains a challenge.
- Data Source:The methods of collecting data are introduced to the GIS community.Newsatellites,aerial cameras and GPS must be incorporated into the system promptly and efficiently.The large amount of information introduced by some of the newest resources is a major database obstacle.

- GPS:GPS is one of the major application in GIS.GPS can be used for navigation purpose.Other uses are military(tracing enemies missiles),transportation(gps vehicles),location etc.

10. Expected Outcomes

GIS(geographical information system) basically provides output in the form of maps.After applying the clustering and outlier techniques and some other preprocess techniques we can get the output in the form of other streams also:

- **Maps:** Everyone recognizes knows that this is the most common output from a GIS.The maps which are sub divided to show level by level wise i.elocal,national,international and global
- **Cartograms:** These are the special type of maps that distort geographic features based on their output values rather than their size.
- **Charts:** GIS can also produce pie charts, histograms (bar charts), line charts, and even pictures in addition to maps to show variation in different levels.
- **Directions:** Another common output, directions show you how to get from one place to another. This is also known as application of GIS.
- **Image:**The output also be displayed in the form of images.The images are of satellite images which are modified to give correct output.

11. Conclusion

In recent years, spatial data mining techniques developed and get some initial results. Spatial data mining is nothing but extracting interested results from large databases, and can be used to understand spatial relationships between spatial and non-spatial data.

There are more number of algorithms for spatial data mining each one has its own advantages and disadvantages. One algorithm is suited for large data sets but it takes more search space, one can suitable for small data set but the efficiency is less. So no algorithm is perfectly suited for spatial data mining . Spatial data mining is now used in many fields, and also achieved great results. It can be predicted spatial data mining will not only promote the development of spatial science and computer science, but also will enhance human ability to understand and transform the world, thus serve the human society better. Besides to develop and update their own algorithms and methods, spatial data mining need to fully achieve its heights and draw professional theoretical method from data discovery, databases, robotics, artificial intelligence, mathematical statistics, visualization, remote sensing, land use land cover ,graphic image, science and other fields.

References

[1] M. Anjireddy, “Remote sensing and GIS”, third edition, Professor and Head, Center of Environment.

[2] R.T. Ng and J. Han, “Efficient and Effective clustering methods for spatial data mining”, IEEE, Computer Society, **(1994)**.

[3] D.H. Bae, J.H. Baek, H.K. Oh, J.W. Song and S.W. Kim, “SD-Miner: A Spatial Data Mining System”, **(2009)**.

[4] L. Yang, Zhanmingjin School of Economics and Management, Tsinghua University, Beijing, 100084, China.

[5] R. Elmasri and S.B. Navathe, “Fundamentals of Database Systems”, 5th Edition Pearson, **(2007)**.

[6] J.Y. Pan and C. Faloutsos, “GeoPlot: Spatial Data Mining on Video Libraries”, Computer Science Department, Carnegie Mellon University Pittsburgh, PA, **(2002)**.

