MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 6, Issue 2, July-December 2021

The Analysis and Survey of Spatial Data Mining Methods Recent Challenges and Trends

M. Raghavendra Rao
Assistant Professor
Department of Computer Science & Engineering
St. Martin's Engineering College, Dhulapalli, Secunderabad, JNTUH, Hyderabad, TS, INDIA
Corresponding Author's E-mail: *raghavendramarpudi14@gmail.com*

Abstract

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. This paper reviews 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. The collected data far exceeded human's ability to analyze. Recent studies on data mining have extended the scope of data mining from relational and transactional databases to spatial databases. We will conclude by discussing the main differences between these two approaches and the elements they have in common. We will further explain the research conducted in this sector and point out that two separate methods exist: one is focused on space database learning while the other is based on space statistics. Finally, we will address the key distinctions between these two methods and their similar features.

Keywords: *Spatial Data Mining, Spatial Databases, Rules Induction, Spatial Statistics, Spatial Neighborhood, knowledge discovery*

INTRODUCTION

The growing production of maps is generating huge volumes of data that exceed people's capacity to analyze them. It thus seems appropriate to apply knowledge discovery methods like data

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

mining to spatial data. This recent technology is an extension of the data mining applied to alphanumerical data on spatial data. The main difference is that spatial analysis must take into account spatial relations between objects. This recent technology is an extension of the data mining applied to alphanumerical data on spatial data. The main difference is that spatial analysis must take into account spatial relations between objects.

The applications covered by spatial data mining are decisional ones, such as geomarketing, environmental studies, risk analysis, and so on.

Spatial data mining applications are decisive, for example, marketing, environmental analyses, risk mitigation, etc. For instance, a store may set its trading location, i.e., its clients' area, marketing, and evaluate their profile based on both its assets and the properties belonging to the area in which they reside.

Spatial data mining aims to automate such a knowledge discovery process.

Thus it plays on important role in

1. Extracting interesting spatial patterns and features
2. Capturing intrinsic relationships between spatial and non spatial data
3. Presenting data regularity concisely and at higher conceptual levels and
4. Helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance.

Spatial knowledge mining is used for traffic risk analysis in our project. The probability calculation is focused on details on historical incidents like injury accidents, combined to thematic data relating to the road network, population, buildings, and so on. The project aims at identifying regions with a high level of risk and analyzing and explaining those risks with respect to the geographic neighborhood. Spatial data mining technology specifically allows for those neighborhood relationships.

Nowadays, data analysis in geography is essentially based on traditional statistics and multidimensional data analysis and does not take account of spatial data. Yet the main specificity of geographic data is that observations located near to one another in space tend to share similar (or correlated) attribute values. This constitutes the fundamental of a distinct scientific area called “spatial statistics” which, unlike traditional statistics,

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

supposes inter-dependence of nearby observations. There is abundant bibliography in this field, including popular geostatistics, recent advances in Anselin's Spatial Data Analysis (ESDA) and Openshaw's geographic analyzer (GAM). As evidence-based assessments are presented, we maintain that spatial statistics is a part of spatial data mining. Any of them are now deployed in GIS or computational software.

In the field of databases, two main teams have contributed to developing data mining for spatial data analysis. The first, a GeoMiner [8] extension, was established by the DB Research Lab (DB Research Lab) of the Simon Fraser University in Vancouver. The second (University of Munich) developed a neighborhood structure graph [9] based on certain algorithms.

In this paper, we will describe data mining methods for Geographic Information Systems and highlight their value in performing spatial data analysis. It will survey both statistical approaches and those involving inference from databases.

It is structured as follows. In section 2 we define spatial data mining and subdivide it into generic tasks. Then in section 3 we classify spatial data mining methods, whether drawn from the realm of databases, statistics or artificial intelligence, in terms of these different tasks. We go on to compare the statistical analysis approach with the spatial database approach, with the aim of emphasizing their similarities and complementarily. Lastly, we conclude and discuss research issues.

SPATIAL DATA MINING

Spatial data mining (SDM) consists of extracting knowledge, spatial relationships and any other properties which are not explicitly stored in the database. SDM is used to find implicit regularities, relations between spatial data and/or non-spatial data. The feature of SDM is its spatial interaction.

In effect, a geographical database constitutes a spatio-temporal continuum in which properties concerning a particular place are generally linked and explained in terms of the properties of its neighborhood. Data mining are not spatial data suitable since they do not accept position data or tacit connections between objects. New approaches, including spatial relationships and spatial data care, must, therefore, be established. It takes time to approximate these spatial interactions, and the encoding

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

of geometric position produces an immense amount of data.

Data mining is not spatial data suitable since they do not accept position data or tacit connections between objects. New approaches, including spatial relationships and spatial data care, must, therefore, be established. It takes time to approximate these spatial interactions, and the encoding of geometric position produces an immense amount of data. This uncertainty would impact global efficiency

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. Spatial data access methods and often require spatial reasoning, geometric Computation, and spatial knowledge representation techniques.

Primitives of Spatial Data Mining:

Rules*:* There are several kinds of rules can be discovered from databases in general. For example characteristic rules, discriminate rules, association rules, or deviation and evaluation rules can be mined [1]. A Spatial characteristic rule is a general description of the spatial data. For example, a rule describing the general price range of houses in various geographic regions in a city is a spatial characteristic rule.

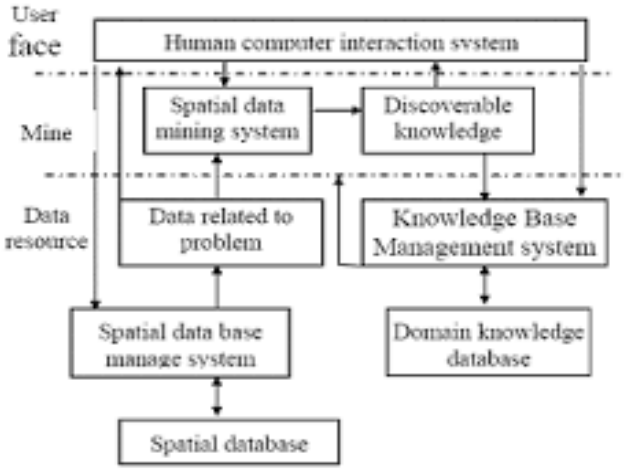


Figure 1 The systematic structure of spatial data mining

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

A discriminate rule is general description of the features discriminating or contrasting a class of spatial data from other class(es) like the comparison of price ranges of houses in different geographical regions. A spatial association rule is a rule which describes the implication of one a set of features by another set of features in spatial databases. For example, a rule associating the price range of the houses with nearby spatial features, like beaches, is a spatial association rule.

Thematic Maps: Thematic map is map primarily design to show a theme, a single spatial distribution or a pattern, using a specific map type. These maps show the distribution of features over limited geography areas [1]. Each map defines a partitioning of the area into a set of closed and disjoint regions; each includes all the points with the same feature value. Thematic maps present the spatial distribution of a single or a few attributes. This differs from general or reference maps where the main objective is to present the position of the object in relation to other spatial objects. Thematic maps may be used for discovering different rules. For example, we may want to look at temperature thematic map while analyzing the general weather pattern of a geographic region. There are two ways to represent thematic maps: Raster, and Vector.

In the raster image form thematic maps have pixels associated with the attribute values. For example, a map may have the altitude of the spatial objects coded as the intensity of the pixel (or the color). In the vector representation, a spatial object is represented by its geometry, most commonly being the boundary representation along with the thematic attributes. For example, a park may be represented by the boundary points and corresponding elevation values.

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.

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

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

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

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

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 [12, 13]. 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

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

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 [6]. 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 need to look at spatial organization spawned several research studies [6, 14]. 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.

Generalization:

This method consists of raising the abstract level of non- spatial attributes and reducing the detail of geometric description by merging adjacent objects. It is derived from the concept of attribute-oriented induction as described in [7]. Here, a concept hierarchy can be spatial (like the hierarchy of administrative boundaries) or non-spatial (thematic) [15]. An example of thematic hierarchy in agriculture can be represented as follows: “cultivation type (food (cereals (maize, wheat, rice), vegetable, fruit, other)”. That kind of hierarchy can be directly introduced by an expert in the field or generated by an inference process related to the attribute. A spatial hierarchy may preexist, like the administrative boundaries one, or it may be based on an artificial geometric splitting like a quad-tree [16], or it may result from a spatial clustering (see below). There are two kinds of generalization: non-spatial dominant generalization, where we first use a thematic hierarchy and then merge

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

adjacent objects; and spatial dominant generalization, which is based on a spatial hierarchy to begin with, followed by the aggregation or generalization of non-spatial values for each generalized spatial value. The complexity of the corresponding algorithms is O(NlogN), where N is the number of actual objects. This approach could be treated as a first step towards a method of inferring rules, such as association rules or comparison rules.

Characteristic rules:

The characterization of a selected part of the database has been defined in [17] 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.

Consider a subset S of objects to analyze. This method uses the following parameters: 1) significance (relative frequency to the database in S); 2) confidence (ratio of objects in S which satisfy the significance threshold in the neighborhood) ; and 3) the maximum extension max- neighbors to the neighbors. This method throws up the properties pi = (attribute, value), the relative frequency factors freq-fac i (higher than the significance parameter) and the number ni of neighbors on which the frequency of the property is extended. The characterization can be expressed by the following rule

S (n1, freq-fac 1)Λ□...Λ □p k (nk ,freq- fac k).

Class identification:

This task, also called supervised classification, provides a logical description that yields the best partitioning of the database. Classification rules constitute a decision tree where each node contains a criterion on an attribute. The difference in spatial databases is that this criterion could be a spatial predicate and, because spatial objects are dependent on neighborhood, a rule involving the non-spatial properties of an object should be extended to neighborhood properties. In spatial statistics, classification has essentially served to analyze remotely-sensed data, and aims to identify each pixel with a particular category. Homogeneous pixels are then aggregated in order to form a geographic entity [4]. In the spatial database approach [18], classification is seen as an arrangement of objects using both their properties (non-spatial values) and their neighbors' properties, not only for direct neighbors

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

but also for the neighbors of neighbors and so on, up to degree N. Let us take as an example the classification of areas by their economic power. Classification rules are described as follows:
High population Λ neighbor = road Λ neighbor of neighbor = airport => high economic power (95%).

In GeoMiner, a classification criterion can also be related to a spatial attribute, in which case it reflects its inclusion in a wider zone. These zones could be determined by the algorithm, whether by clustering or by merging adjacent objects, or it could arise from a predefined spatial hierarchy. A new algorithm [19] extends this classification method in GeoMiner to spatial predicates. For example, to determine high level wholesale profits, a decision factor can be the proximity to densely populated districts.

Clustering:

This task is an automatic or unsupervised classification that yields a partition of a given dataset depending on a similarity function.

Database approach:

Paradoxically, clustering methods for spatial databases do not appear to be very revolutionary compared with those applied to relational databases (automatic classification). The clustering is performed using a similarity function which was already classed as a semantic distance. Hence, in spatial databases it appears natural to use the Euclidean distance in order to group neighboring objects. Research studies have focused on the optimization of algorithms. Geometric clustering generates new classes, such as the location of houses in terms of residential areas. This stage is often performed before other data mining tasks, such as association detection between groups or other geographic entities, or characterization of a group.

GeoMiner combines geometric clustering applied to a point set distribution with generalization based on non-spatial attributes. For example, we may want to characterize groups of major cities in the United States and see how they are grouped. Cluster results will be represented by new areas, which correspond to the convex hull of a group of towns. A few points could stay outside clusters and represent noise. A description of each group may be generated for each attribute specified.

Many algorithms have been proposed for performing clustering, such as CLARANS

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

[20], DBSCAN [8] or STING [9]. They usually focus on cost optimization. Recently, a method that is more specifically applicable to spatial data, GDBSCAN, was outlined in [21]. It applies to any spatial shape, not only to points data, and incorporates attributes data.

Statistic approach:

Clustering arises from point pattern analysis [22, 23] and was mainly applied to epidemiological research. This is implemented in Opens haw's well-known Geographical Analysis Machine (GAM) and could be tested by using the K-function [24]. The clusters could also be detected by the ratio of two density estimates: one of the studied subset and the other of the whole reference dataset.

Trend and Deviation Analysis:

In relational databases, this analysis is applied to temporal sequences. In spatial databases, we want to find and characterize spatial trends.

Database approach:

Using the process described in [18], which is based on the central places theory, the analysis is performed in four stages. The first one involves discovering centers by computing local maxima of particular attributes; in the second, the theoretical trend of these attributes is determined by moving away from the centers; the third stage determines the deviations in relation to these trends; and finally, we explain these trends by analyzing the properties of these zones. One example is the trend analysis of the unemployment rate in comparison with the distance to a metropolis like Munich. Another example is the trend analysis of the development of house construction.

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 [25]. 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

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

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

CONCLUSION

We have shown that spatial data mining is a promising field of research with wide applications in GIS, medical imaging, remote sensing, robot motion planning, and so on. We also outlined and categorized this analysis and contrasted the two methods, stressing each method's specific effectiveness and future benefits to be integrated. We have summarized and classified this research and compared the two approaches, emphasizing the particular utility of each method and the possible advantages of combining them. 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 the case of graphical methods, these relationships could be optimized using spatial indexes. As regards the other methods that use neighborhood structures, instantiation of the structure is costly and should be pre-computed as far as possible.

REFERENCES

1. Bédard, y., lam, s., proulx, m.j., caron, p.y., létourneau, f.: data warehousing for spatial data: research issues, proceedings of the international symposium geomatics in the era of radarsat (ger'97), ottawa, may 1997, pp. 25-30
2. Benali, h., escofier, b.: analyse factorielle lissée et analyse factorielle des différences locales, revue statistique appliquée, 1990, xxxviii (2), pp 55-76
3. Cliff a.d., ord j.k., 1973 : "spatial utocorrelation", pion, london.
4. Diggle p.j., 1993, point process modeling in environmental epidemiology. In barnett v., turkman k. (eds) statistics for the environment, chichester, john wiley & sons, pp 89-110 The contract prism -inrets, december 1998, 33 p.
5. Ester, m., kriegel ,h.-p., sander, j., xu, x.: density-connected sets and their application for trend detection in spatial databases, proc. 3rd int. Conf. On knowledge discovery and data mining, newport beach, ca, 1997, pp. 10-15.
6. Ester, m., kriegel, h.-p., sander, j.: spatial data mining: a database approach, proc. 5th symp. On spatial databases, berlin, germany, 1997

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

7. Moran p.a.p., the interpretation of statistical maps, journal of the royal statistical society, b: 10, pp 234-251., 1948.
8. Fayyad et al., "advances in knowledge discovery and data mining", aaai press / mit press, 1996
9. Fotheringham s., zhan b., 1996 : "a comparison of three exploratory methods for cluster detection in spatial point patterns", geographical analysis, vol. 28, n° 3, pp. 200-218
10. Gatrell a., bailey t., diggle p., rowlingson b., 1996 : "spatial point pattern analysis and its application in geographical epidemiology", transactions of the institute of british geographers, n° 21, pp. 256-274
11. Geary r.c.: the contiguity ratio and statistical mapping, the incorporated statistician, 5 (3), pp 115-145.
12. Gunther o., "efficient computation of spatial joins", proc of data engineering, vienna, austria, april 1990, pp. 50- 59.
13. Han j., cai y. & cerone n., "knowledge discovery in databases; an attribute-oriented approach." proceedings of the 18th vldb conference. Vancouver, b.c., august 1992. Pp. 547-559
14. Sanders, l.: l'analyse statistique des données en géographie, gip reclus, 1989
15. Longley p. A., goodchild m. F., maguire d. J., rhind d. W., geographical information systems - principles and technical issues, john wiley & sons, inc., second edition, 1999.
16. Lebart l. Et al., "statistique exploratoire multidimensionnelle", editions dunod, paris, 439 p., 1997.
17. Lebart, l. (1984) correspondence analysis of graph structure. Bulletin technique du cesia, paris:2, 1-2, pp 5- 19.
18. Koperski, k., han, j., and stefanovi,c n.: an efficient two-step method for classification of spatial Data, in proc. International symposium on spatial data handling (sdh'98) , pp. 45-54, vancouver, canada, july 1998
19. Lebart l. Et al., "statistique exploratoire multidimensionnelle" , editions dunod, paris, 439 p., 1997.
20. Lebart, l. (1984) correspondence analysis of graph structure. Bulletin technique du cesia, paris:2, 1-2, pp 5-19.

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved

21. Longley p. A., goodchild m. F., maguire d. J., rhind d. W., geographical information systems - principles and technical issues, john wiley & sons, inc., second edition, 1999.
22. Lu, w., han, j. And ooi, b.: discovery of general knowledge in large spatial databases, in proc. Of 1993 far east workshop on geographic information systems (fegis'93), singapore, june 1993, pp. 275-289
23. Mathsoft inc., "s-plus for arcview gis - users guide version 1.0" and "s-plus spatial stat.", data analysis products division, seattle, washington, april 1998.
24. Moran p.a.p., the interpretation of statistical maps, journal of the royal statistical society, b: 10, pp 234-251., 1948.

Page 69-81 © MANTECH PUBLIATIONS 2021. All Rights Reserved