A Survey of Spatial Data Mining Approaches: Algorithms and Architecture

D.Muralir# ,G.Hari Shankar* ,R.Sateesh Kumar$

Professor, CSE Department MRCEW Hyderabad , * Lecturer, CSE Department, Govt Polytechnic, Chitoor,

$ Asst. Professor, CSE Department , VCE, Hyderbad

Abstract:

Voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. There is an urgent need for effective and efficient methods to extract unknown and unexpected information from spatial data sets of unprecedentedly large size, high dimensionality, and complexity. To address these challenges, spatial data mining and geographic knowledge discovery has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases.This paper highlights recent theoretical and applied research in spatial data mining and knowledge discovery. We first briefly review the literature on several common spatial data-mining tasks, including spatial classification and prediction; spatial association rule mining; spatial cluster analysis; and geo visualization. The articles included in this special issue contribute to spatial data mining research by developing new techniques for point pattern analysis, prediction in space–time data, and analysis of moving object data, as well as by demonstrating applications of genetic algorithms for optimization in the context of image classification and spatial interpolation. The papers concludes with some thoughts on the contribution of spatial data mining and geographic knowledge discovery to geographic information sciences.

KEYWORDS : Spatial Data Mining,Classification, Spatial data Bases, GPS

1. Introduction

We are often interested in analyzing complex situations to more precisely predict the effect of some spatial phenomenon. Once its behavior is approximated by a model, the spatial phenomenon can be understood more correctly. However, currently used spatial models are usually created in a very simple way and represent only the general trend. To give the model a more realistic form, advanced methods for spatial data analyzes should be employed. When a more accurate representation of a spatial phenomenon exists, more can be discovered about its possible impact. Recently, the amount of natural and man-made disasters has increased. Therefore, actions concentrating on prediction and assessment of possible consequences for nature as well as human lives are becoming more important. Consequently, principal changes to the existing risk models for rescue operations are essential. Due to the fast development of geo-information technologies, a variety of new opportunities arise. Therefore, more accurate analyzes can be performed on spatial data. In this research the possible use of spatial data mining (SDM) methods is investigated for identifying factors that may influence occurrences of incidents .

Due to advanced data collection techniques such as remote sensing, census data acquiring, weather and climate monitoring etc. contemporary geographical datasets contain an enormous amount of data of various types and attributes. Analyzing this data is challenging for traditional data analysis methods which are mainly based on extensive statistical operations. Since classical data mining methods enable us to detect valuable information from extensive relational databases, SDM can be an appropriate technique for detecting possible interesting

patterns in geographical datasets. Spatial data mining is a knowledge discovery process of extracting implicit interesting knowledge, spatial relations, or other patterns not explicitly stored in databases. Knowledge discovery from database is a complex concept integrating several research fields including machine learning, database systems, statistics, visualization etc. Data mining is a core component of the KDD process. The KDD process assumes that interesting and unexpected patterns in very large databases are deeply hidden and often difficult or impossible to specify *a priori*. Consequently, traditional database queries and statistical methods do not reveal any implicit information from a large database. KDD is a tool for exploring domains that are too difficult to perceive with unaided human abilities

2. Spatial classification and prediction

Classification is about grouping data items into classes (categories)according to their properties (attribute values). Classificationis also called supervised classification, as opposed to the unsupervisedclassification
(clustering).‘‘Supervised” classification needs atraining dataset to train (or configure) the classification model, avalidation dataset to validate (or optimize) the configuration, anda test dataset to evaluate the performance of the trained model.Classification methods include, for example, decision trees, artificialneural networks (ANN), maximum likelihood estimation
(MLE), linear discriminant function (LDF), support vector machines(SVM), nearest neighbor methods and case-based reasoning (CBR).Spatial classification methods extend the general-purpose classificationmethods to consider not only attributes of the object tobe classified but also the attributes of neighboring objects and theirspatial relations (Ester, Kriegel, & Sander, 1997; Koperski, Han, &
Stefanovic, 1998). A visual approach for spatial classification wasintroduced in (Andrienko&Andrienko, 1999), where the decisiontree derived with the traditional algorithm C4.5 (Quinlan, 1993) is combined with map visualization to reveal spatial patterns ofthe classification rules. Decision tree induction has also been usedto analyze and predict spatial choice behaviors (Thill&Wheelerm,
2000). Artificial neural networks (ANN) have been used for a broadvariety of problems in spatial analysis (Fischer, 1998; Fischer, Reismann
andHlavackova-Schindler, 2003; Gopal, Liu and Woodcock,2001; Yao &Thill, 2007). Remote sensing is one of the major areasthat commonly use classification methods to classify image pixels
into labeled categories (for example, Cleve, Kelly, Kearns, & Morltz,2008).Spatial regression or prediction models form a special group of
regression analysis that considers the independent and/or dependentvariable of nearby neighbors in predicting the dependent variableat a specific location, such as the spatial autoregressive
models (SAR) (Anselin, Syabri, &Kho, 2006; Cressie, 1983; Pace,Barry, Clapp, & Rodriquez, 1998). However, spatial regressionmethods such as SAR often involve the manipulation of an n by n
spatial weight matrix, which is computationally intensive if n islarge. Therefore, more recent research efforts have sought to developapproaches to find approximate solutions for SAR so that it can
process very large data sets (Griffith, 2004; Kazar, Shekhar, Lilja,Vatsavai, & Pace, 2004; Smirnov &Anselin, 2001).

2.1 Spatial data characteristics

Extracting implicit information from geographical databases appears, in comparison to traditional non-spatial databases, to be more challenging. Together with non-spatial attributes, spatially referenced objects also carry information concerning their representation in space by geometrical and topological properties. Topology covers the geographical properties which are not closely connected to the actual position of objects, i.e. it represents the spatial relationships among objects. The topology is a branch of geometry that deals with those properties of a figure (object) that remain unchanged even when

the figure is transformed. On the other side, geometric characteristics of data concerns information related to the actual location of the object in space. The location is usually described by Euclidian coordinates or Latitude and Longitude. Besides the core spatial characteristics dealing with geometry and topology, geographical data also contains information about the behavior of a phenomenon the data represents. In particular, the notion of spatial autocorrelation is fundamental to any spatial related operations. Omitting the fact that nearby items tend to be more similar than items situated more apart, causes inconsistent results in the spatial data analysis. Another important characteristic of geographic data is spatial heterogeneity. Spatial data is not identically distributed in space, therefore data properties are location dependent. It is possible that local trends can sometimes contradict the global trends. In other words, global parameters estimated from a geographic database do not sufficiently describe the geographic phenomenon at any particular location. Due to the spatial data diversity, a composition of geographical databases is crucial. Moreover, the data integration process has to deal with very complicated data transformations, because the collected data are often from different sources. Therefore good database design provides the possibility of analyzing geographical data with maximum efficiency on data processing and in the same time considers their spatial characteristics.

2.2 Spatial data mining techniques

There is no unique way of classifying SDM techniques. Various kinds of patterns can be discovered from databases and can be presented in different forms. The categorization often depends on the background field of a particular researcher. If we assume a person to be interested in data visualization, the criteria for classification will probably be dependent on various visualization techniques, whereas a computer science researcher might see the main variance in the utilization of different algorithms. General data mining tasks can be classified into two main categories: descriptive data mining and predictive data mining. The former concisely describes the behavior of datasets and presents interesting general properties of the data. Whereas the latter attempts to construct models that tend to help predicting the behavior of the new datasets. Forecasting an employee's potential salary based on the salary distribution of similar employees can be seen as an example of a predictive data mining task. While descriptive methods may be used for comparison of sales between a European
and an Asian branch of a certain company. spatial data mining techniques divided into four general groups: spatial association rules, spatial clustering, spatial trend detection and spatial classification. The categorization is based on the KDD algorithms. The three most non-controversial techniques would be classification, clustering and association rules. However, some of those algorithms can be accompanied by supporting methods. For example for identification of so called *Hot Spots* which are areas of a high value of certain activity within a large area of low activity, clustering technique is performed together with outlier detection. Consequently, the basic idea of co-location is derived from a spatial association technique.

The spatial data mining techniques are categorized as

. **clustering and outlier detection**
. association and co-location
. classification
. trend detection

2.3 Clustering and outlier detection

Spatial clustering is a process of grouping a set of spatial objects into groups called clusters. Objects within a cluster show a high degree of similarity, whereas the clusters are as much dissimilar as possible. Unlike classification, clustering is an unsupervised process. This means that clustering does not rely on predefined labels of classes or *a priori* given number of classes. Clustering is a very well known technique in statistics and the data mining role is to scale a clustering algorithm to deal with the large geographical datasets. Clustering algorithms can be separated into four general categories: partitioning method, hierarchical method, density-based method and grid-based method. The categorization is based on different cluster definition techniques.

- **Partitioning method**
- **Density-based method**
- **Grid-based method**

2.4 Association and co-location

When performing clustering methods on the data, we can find only characteristic rules, describing spatial objects according to their non-spatial attributes. In many situations we want to discover spatial rules that associate one or more spatial objects with others.A spatial association rule is of the form

$X \rightarrow Y\ (c\ \%\ )$, where X and Y are sets of spatial or non-spatial predicates and c % is the *confidence* of the rule.

An association rule is characterized by two parameters:

*support*and *confidence*. The former expresses a ratio of transactions that satisfies both X and Y, to the number of transactions in a dataset. Whereas the latter one presents a conditional probability that Y is true under the condition that X is true.

A large number of associations may be extracted from an extended geographical database. However, a majority of those rules are applicable to only a small number of objects and the extraction of all rules is very computationally expensive. Often the confidence of rules is low. Therefore, the concepts of *minimum* support and *minimum* confidence are used to guarantee that only important transactions are discovered. We state that a rule is *strong* when the support is *large*, i.e., no less than the minimum support threshold, and the confidence is *large*, i.e., no less than the minimum confidence threshold.

However, one of the biggest research challenges in mining association rules is the development of methods for selecting potentially interesting rules from among the mass of all discovered rules.

Co-location patterns represents subsets of Boolean spatial features whose instances are often located close proximity. Co-location rules are models to infer the presence of Boolean spatial features in the neighborhood of instances of other Boolean features. Co-location rule discovery is the process of identifying co-location patterns from the large spatial data sets with large number of Boolean fetures.

2.4 classification

Every data object stored in a database is characterized by its attributes. Classification is a technique, which aim is to find rules that describe the partition of the database into an explicitly given set of classes. Objects with similar attribute values are integrated into the same class. In spatial classification the attribute values of neighboring objects may also be relevant for the membership of objects in a certain group. Therefore, we have to include the neighborhood factor in the calculation. A classification method consists of two parts. First the user defines the number of classes. To test, whether the number of classes was chosen correctly, a set of training data is selected and the classification is performed on it. Consequently, classification rules are derived from the training dataset. Next, those rules are applied to the test dataset. Classification is considered as predictive spatial data mining, because we first create a model according to which the whole dataset is analyzed. A classification process can be performed in many different ways. A method is based on the Linear Regression (LR). To guarantee the spatial dependencies of objects, a Spatial Autoregressive Regression (SAR) technique has been proposed by spatial statisticians.

2.5. Spatial association rule mining

Association rule mining was originally intended to discover regularitiesbetween items in large transaction databases (Agrawal,Imielinski, & Swami, 1993). Let I = {i1, i2, . . .,im} be a set of items(i.e., items purchased in transactions such as computer, milk, bike,etc.). Let D be a set of transactions, where each transaction T is a set of items such that T #I. Let X be a set of items and a transaction T issaid to contain X if and only if X #T. An association rule is in theform: X) Y, where X _ I; Y _ I and X \ Y ¼ £. The rule X) Yholds in the transaction set D with confidence c if c% of all transactions in D that contain X also contain Y. The rule X) Y has supports in the transaction set D if s% of

transactions in D contain X [Y.Confidence denotes the strength and support indicates the fre-404 D. Guo, J. Mennis / Computers, Environment and Urban Systems 33 (2009) 403–408quencies of the rule. It is often desirable to pay attention to thoserules that have reasonably large support (Agrawal et al., 1993).

Similar to the mining of association rules in transactional orrelational databases, spatial association rules can be mined in spatialdatabases by considering spatial properties and predicates (Appice,Ceci, Lanza, Lisi, &Malerba, 2003; Han &Kamber, 2001;Koperski & Han, 1995; Mennis& Liu, 2005). A spatial associationrule is expressed in the form A) B [s%, c%], where A and B are setsof spatial or non-spatial predicates, s% is the support of the rule,

and c% is the confidence of the rule.Obviously, many possible spatial predicates (e.g., close_to, far_-

away, intersect, overlap, etc.) can be used in spatial associationrules. It is computationally expensive to consider various spatialpredicates in deriving association rules from a large spatial datasets.

Another potential problem with spatial association rule miningis that a large number of rules may be generated and many ofthem are obvious or common knowledge. Domain knowledge isneeded to filter out trivial rules and focus only on new and interesting

findings.Spatial co-location pattern mining is spiritually similar to, buttechnically very different from, association rule mining (Shekhar& Huang, 2001). Given a dataset of spatial features and their locations,a co-location pattern represents subsets of features frequentlylocated together, such as a certain species of bird tend tohabitat with a certain type of trees. Of course a location is not atransaction and two features rarely exist at exactly the same location.

Therefore, a user-specified neighborhood is needed as a containerto check which features co-locate in the same neighborhood.Measures and algorithms for mining spatial co-location patternshave been proposed

.

2.6. Spatial clustering, regionalization and point pattern analysis

Cluster analysis is widely used for data analysis, which organizesa set of data items into groups (or clusters) so that items inthe same group are similar to each other and different from thosein other groups Many different clustering methods have been developedin various research fields such as statistics, pattern recognition,data mining, machine learning, and spatial analysis.Clustering methods can be broadly classified into two groups:partitioning clustering and hierarchical clustering. Partitioning

clustering methods, such as K-means and self-organizing map(SOM) (Kohonen, 2001), divide a set of data items into a numberof non-overlapping clusters. A data item is assigned to the ‘‘closest”

cluster based on a proximity or dissimilarity measure. Hierarchicalclustering, on the other hand, organizes data items into a hierarchywith a sequence of nested partitions or groupings. Commonly-used hierarchical clustering methods includethe Ward’s method (Ward, 1963), single-linkage clustering, average-

linkage clustering, and complete-linkage clustering (Gordon,1996; Jain &Dubes, 1988).

To consider spatial information in clustering, three types ofclustering analysis have been studied, including spatial clustering(i.e., clustering of spatial points), regionalization (i.e., clusteringwith geographic contiguity constraints), and point pattern analysis(i.e., hot spot detection with spatial scan statistics). For the firsttype, spatial clustering, the similarity between data points or clustersis defined with spatial properties (such as locations and distances).Spatial clustering methods can be partitioning orhierarchical, density-based, or grid-based. Readers are referred to(Han, Kamber, & Tung, 2001) for a comprehensive review of various

spatial clustering methods.Regionalization is a special form of clustering that seeks togroup spatial objects into spatially contiguous clusters (i.e., regions)while optimizing an objective function. Many geographicapplications, such as climate zoning, landscape analysis, remotesensing image segmentation, often require that clusters are geographicallycontiguous. Existing regionalization methods that arebased on a clustering concept can be classified into three groups:

(1) multivariate (non-spatial) clustering followed by spatial processingto rearrange clusters into regions (Fovell&Fovell, 1993);

(2) clustering with a spatially weighted dissimilarity measure,which considers spatial properties as a factor in forming clusters(Wise, Haining, & Ma, 1997) and

(3) contiguity constrained clusteringthat enforces spatial contiguity during the clustering process(Guo, 2008).Point pattern analysis, which is also known as ‘‘hot spot” analysis(Brimicombe, 2007), focuses on the detection of unusual concentrationsof events in

space, such as geographic clusters ofdisease, crime, or traffic accidents. The general research problem is to determine whether there is an excess of observed event points(e.g., disease incidents) for an area (e.g., within a certain distance toa location). Several scan statistics have been developed to find suchspatial clusters .For each replication, the test statistic value is calculatedagain (i.e., the maximum likelihood ratio is found over all enumeratedlocal areas). Then the actual test statistic value is compared tothe test values of all replications to derive the significance level forthe most likely cluster (and the secondary clusters).

3. Overview of the articles

Here we provide an overview of the articles in the special issue.These articles make contributions to the spatial data mining literaturein a variety of ways. Some of the articles extend establishedtechniques, such as artificial neural networks (ANN) and spatial clustering, to account for issues of spatial dependency and spatialscale. Others develop new techniques for types of spatial data thatare only recently becoming widely available, such as path and trajectorydata describing moving objects. The contribution of otherarticles concern new applications of data mining techniques.As noted earlier, classification and prediction is a fundamental data-mining task and ANNs are among the commonly used classificationmethods. However, conventional ANN does not considerthe spatial dependence and associations between neighboring objects.Cheng and Wang (2009) seek to address this issue in developingan ANN for space–time prediction. Their approachincorporates spatial associations among observations into dynamic recursive neural networks (DRNN), an ANN approach that incorporatesfeedback from previous iterations of the model inputs andoutputs. Such feedbacks make DRNN a good candidate for modelingtime-series data. In the present article, the authors proposethat a target prediction can be improved by not only incorporatingthe value of the target at the previous time interval but the values of nearby observations. Three case studies serve to demonstratethis approach using a variety of types of data with varying spatialand temporal records – the prediction of forest fires, economicgross domestic product, and temperature. Results indicate that including spatial association information can improve the computationalperformance and accuracy of DRNN for space–timeprediction.One of the major challenges to spatial data mining arises from handling new kinds of data. Recent advances in embedding GPS to create location-aware devices have generated a massive volumeof data about moving objects. Detecting patterns in these data are challenging due to both the massive volume and temporal natureof the data. Dodge, Weibel, and Forootan (2009) address this challengein their article, which focuses on the classification of moving trajectories. The authors present a way to characterize moving objecttrajectories from both a global perspective, i.e. those propertiesthat characterize the entire trajectory of the object, and a local perspective, i.e. those properties that characterize portions of the object'strajectory. Properties include characteristics such as pathlength and straightness as well as velocity and acceleration. Withthese extracted characteristics, a SVM is applied to classify trajectoriesinto categories. Two types of data, transportation data ofmoving vehicles as well as eye-tracking data, are used to demonstratethe proposed approach.

The third paper by Pei, Zhu, Zhou, Li, and Qin (2009) focuses onthe development of a new method for point pattern analysis. Theauthors note that established spatial clustering methods are often sensitive to the parameterization of the clustering algorithm, particularlyto the scale at which one theorizes clustering occurs, assuch an assumption often must be made a priori to the application of the clustering technique. Consequently, the results of clusteringmay be highly subjective. To address this issue the authors presenta new method of clustering they call the collective nearest neighbor (CLNN) method. The basis for CLNN is the distinction betweenpoints whose distribution may be explained by a causal mechanismversus those whose distribution may be explained by random'noise,' where the distinguishing characteristics between the twoprocesses is intensity of clustering. CLNN extends previous researchby developing a procedure for iterating over various scalesof measurement to assess intensity. The authors demonstrate CLNN using both synthetic data as well as a case study focusingon identifying clusters of earthquakes in China from seismic data.

4. Conclusion

Due to the widespread application of geographic informationsystems (GIS) and GPS technology and the increasingly matureinfrastructure for data collection, sharing, and integration, moreand more research domains have gained access to high-quality geographic data and created new ways to incorporate spatial informationand analysis in various studies.

Private industries and thegeneral public also have more and more interest in both contributing and using geographic data. These data have become more diverse,complex, dynamic, and much larger than ever before andtherefore are more difficult to analyze and understand. Spatial datamining and knowledge discovery has emerged as an active researchfield that focuses on the development of theory, methodology,and practice for the extraction of useful information andknowledge from massive and complex spatial databases. The articlesin this special issue highlight a selected set of approaches andapplication in spatial data mining. As noted earlier, spatial datamining is still at a very early stage and its bounds and potentialsare yet to be defined. There are both opportunities and challengesfacing spatial data mining research.

Spatial data mining is not a push button task. We often claim to‘‘let the data speak for themselves”. However, the data cannot tellstories unless we formulate appropriate questions to ask and use appropriate methods to solicit the answers from the data. Datamining is data-driven but also, more importantly, human-centered,with the user controlling the selection and integration of data, cleaning and transformation of the data, choice of analysis methods,and the interpretation of results. It is an iterative and inductivelearning process that is embedded in an overall deductiveframework.

5.References

1. Agarwal, P., &Skupin, A. (2008). Self-organising maps: Applications in geographic information science. Chichester: Wiley.
2. Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association rules between sets of items in large databases. In ACM SIGMOD international conference onmanagement of data (pp. 207–216).
3. Andrienko, G., &Andrienko, N. (1999). Data mining with C4.5 and interactivecartographic visualization. In N. W. G. T. Paton (Ed.), User interfaces to dataintensive systems (pp. 162–165). Los Alamitos, CA: IEEE Computer Society.
4. Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis.
5. P.A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), Geographical information systems–principles and technical issues (pp. 253–266). New York, NY:
6. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining toknowledge discovery—an review.
7. Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. Journal of Geography, 69(4), 211–221.
8. Gordon, A. D. (1996). Hierarchical classification.In P. Arabie, L. J. Hubert, & G. D. Soete (Eds.), Clustering and classification (pp. 65–122). River Edge, NJ, USA:
World Scientific Publisher.

Griffith, D. (2004). Faster maximum likelihood estimation of very large spatial autoregressive models: An extension of the Smirnov–Anselin result. Journal of Statistical Computation and Simulation, 74(12), 855–866.

Guo, D. (2008). Regionalization with dynamically constrained agglomerativeclustering and partitioning (REDCAP). International Journal of Geographical Information Science, 22(7), 801–823.

9. Guo, D., Gahegan, M., MacEachren, A. M., & Zhou, B. (2005). Multivariate analysisand geovisualization with an integrated geographic knowledge discoveryapproach. Cartography and Geographic Information Science, 32(2), 113–132.
10. Guo, D., Peuquet, D., &Gahegan, M. (2003). ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. Geoinformatica, 7(3),229–253.
11. Han, J., &Kamber, M. (2001). Data mining: Concepts and techniques. MorganKaufmann Publishers.