Quality Assessment and Uncertainty Handling in Uncertainty-Based Spatial Data Mining Framework

Sk. Rafi, Sd. Rizwana

Abstract: **Spatial data mining is to extract the unknown knowledge from a large-amount of existing spatial data repositories areas. The spatial data are to represent the spatial existence of an object in the infinitely complex world. Although uncertainties exist in spatial data mining, they have not been paid much attention to. If the uncertainties are made appropriate use of, then it may be able to avoid the mistaken knowledge discovered from the mistaken spatial data. The uncertainty parameters (i.e. supportable level, interesting level and confidence level) may further decrease the complexity of spatial data mining. Else, it is unable to discover suitable knowledge from spatial databases via taking the place of both certainties and uncertainties with only certainties. The main purpose of this paper is to uncertainty-based spatial data mining to develop a framework for quality assessment and uncertainty handling in spatial data mining. Hence, we have adopted several quality assessment indices for the results of spatial data mining and a methodology for handling uncertainty in spatial data mining.**

I. INTRODUCTION

Now-a-days many advanced technologies have been developed to store and record large quantities of data continuously. The world is an infinitely complex system that is large nonlinear and multi-parameter, which is referred about 80% of the information. The rapid development of the instruments and infrastructures on Geo-Informatics makes spatial data complex that has been beyond the human ability to analyze and interpret. The spatial entity in the world includes historical information, status, and future trend. By using the single conventional technique, we cannot resolve the human facing bottleneck with large amount of the spatial data is still short of knowledge [1] [2]. In uncertain data management, the data records are typically represented by probability distributions rather than deterministic values. The field of uncertain data management poses a number of unique challenges on several fronts. There are uncertainties in spatial data, and they may directly or indirectly affect the quality of spatial data mining [3] [4].

Due to the widespread application of geographic information systems (GIS), GPS technology, and the increasingly mature infrastructure for data collection, sharing, and integration, more and more research domains have gained access to high-quality geographic data and created new ways to incorporate spatial information and analysis in various studies. Spatial data mining and knowledge discovery has emerged as an active research field that focuses on the development of theory and practice for the extraction of useful information and knowledge from massive and complex spatial databases. The data cannot tell stories unless we formulate appropriate questions to ask and use appropriate methods to solicit the answers from the data. New types of data and new application areas have significantly expanded the frontier of spatial data mining research. Handling the very large volume and understanding complex structure in spatial data are another two major challenges for spatial data mining that demand both efficient computational algorithms to process large data sets and effective visualization approaches to present and explore complex patterns.

Spatial data mining is a growing research field that is still at a very early stage. Spatial data of the database is to represent the spatial existence of an object in the infinitely complex world. Although there are virtually uncertainties inherent in most of the spatial data capture and data analyzing due to the limitations or constraints of current instruments, human skills, capital and technologies. The information of different entities may be overlapped,

deformed, or mixed. Two entities of the same classification may eradiate different spectrum information that may belong to different classifications. Finally, it may confuse to correctly classify the pixels with the same gray degrees in the boundary area where two different classifications overlap. Traditionally it was presumed that the spatial world stored in spatial database was crisply defined. Some true spatial values are even inexact or inaccessible, which are actual characteristics of the spatial entity reality. It is impossible to obtain some true spatial values that exist. One is unobservable for they are spatial data with long history and the other is impartial observe because they are too complex. It is a fundamental function to determinate whether or not the spatial element belongs to the predefined entity and the classification determination is performed on the accessible spatial values that are measured by sensors. In the Spatial Data Transfer Standard (SDTS), the data quality is further divided into five fundamental components:

- Positional accuracy
- Attribute accuracy
- Logical consistency
- Lineage
- Completeness

The uncertainty is an essential part of many models of spatial data based decision-making that has become the subject of a growing volume of research and figured prominently in research agendas. Spatial uncertainties include the following uncertainties like positional, attribute, inaccuracy, incompleteness, topological, inconsistency, noisy, omittance, knowledge, and misclassification. If the uncertainties hidden in the database have been taken as the input of spatial data mining, the resulting discovered output might be the wrong knowledge. It is necessary to deal with uncertainty to make the discovered knowledge aware of the level of uncertainty present. The uncertainties in spatial data mining have not been addressed to the same degree to spatial data mining itself [4], [5], [6], and [7]. There have been considerable theories and techniques on either spatial data mining [8]; each effort is focused on its own field. Many efforts on the uncertainties specialize on the general autocorrelation and are not oriented to discover knowledge from spatial data sets in the uncertainty context. Models have been proposed recently allowing enriching database models to manage uncertain spatial data. The major motivation for this is that there exist geographic objects with uncertain boundaries, and fuzzy sets are a natural way to represent this uncertainty [9]. An ontology for spatial data has been developed in which the terms imperfection, vagueness and imprecision are organized into a hierarchy to assist in management of these issues.

II. RELATED WORK

Spatial data point to the data that are able to represent the spatial existence of an entity and there are various kinds like attributes, images, temporal data and graphics. It is difficult to define an uncertainty-based spatial data mining completely. The uncertainty-based spatial data mining is to extract knowledge from the vast repositories of practical spatial data under the umbrella of uncertainties with the given perspectives and parameters. We can control and reduce the uncertainty in the acceptable manner in the three ways:

a. Data acquisition
   It highlights the information acquired from the process of data collection and data amalgamation
b. Data cognition
   It emphasizes the knowledge discovered from data extraction process and information generalization
c. The usable techniques and methods that may possibly cope with the uncertainties in spatial data mining are briefly overviewed.

A) UNCERTAIN DATA REPRESENTATION AND MODELING

A database that provides incomplete information consists of a set of possible instances of the database. Since the latter is a more specific definition which creates database models with crisp probabilistic quantification. The direct specification of incomplete databases and probabilistic data is unrealistic and since an exponential number of instances would be needed to represent the table. In many practical applications one may often work with simplifying assumptions on the underlying database.

The assumption is that the presence and absence of different tuple's is probabilistically independent. All possible probability distributions on possible worlds are not captured with the use of independent tuple's.

B) UNCERTAINTY ANALYSIS IN SPATIAL DATA MINING

The uncertainties in spatial data mining may exist in the process of spatial data selection, spatial data preprocessing, data mining and knowledge representing. The original data of spatial data mining stem from uncertain spatial database or uncertain spatial data sets being analyzed. Uncertainties in spatial data may directly or indirectly affect the quality of spatial data mining [10]. As shown in the fig.1 uncertainties will be propagated and accumulated in spatial data mining process.

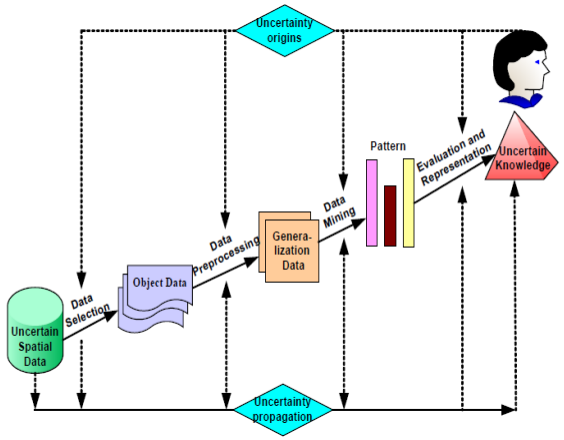

Figure 1: Uncertainties and its propagation in the process of spatial data mining.

Uncertainties mainly stem from a subjectivity of selecting object data according to the task of spatial data mining including what data should be collected and how much data is enough. Spatial data preprocessing mainly include data cleaning, data transformation and data discretization, where many uncertainties will be produced if we do not adopt appropriate uncertainty handling methods. Given continuous attribute data is divide into discrete values by the data discretization and this operation may be a main origin of uncertainties in the whole process of data mining. Most of spatial knowledge discovered by spatial data mining is qualitative knowledge and the best way to represent them is the natural language. It is derived from spatial data mining that is a branch of data mining, and the discovered knowledge is diversity as shown in the table 1.

It is an uncertain process for spatial data mining to discover the little-amount refined knowledge from the large-amount coarse data. The uncertainties in spatial data mining may exist in mining process, theories and techniques, knowledge characteristics, spatial data, etc.

Knowledge	Data mining	Spatial data mining	Uncertainty-based spatial data mining
Association rule	Yes	Yes	Yes
Clustering rule	Yes	Yes	Yes
Classification rule	Yes	Yes	Yes
Characteristics rule	Yes	Yes	Yes
Serial rule	Yes	Yes	Yes
Regression rule	Yes	Yes	Yes
Dependent rule	Yes	Yes	Yes
Spatial topological rule	Yes	Yes	Yes
Spatial distribution rule	Yes	Yes	Yes
Outlier	Yes	Yes	Yes

Table 1: The discovered knowledge of uncertainty-based spatial data mining

The manipulations of spatial data mining are more abundant than common data mining on transaction data. There may be uncertainty in the understanding of entities or in the quality or meaning of the data. As the spatial data are the objectives of spatial data mining, the uncertainties are brought to spatial data mining along with spatial data at the beginning.

III. QUALITY ASSESSMENT OF SPATIAL DATA MINING

The traditional data-mining algorithm and the pattern/rule recognition were considered as useful, certainty and correct. In fact many number of uncertainties hid in spatial data and the operations of spatial data mining.

A) SPATIAL DATA CLEANING

Spatial data cleaning is an essential in uncertainty-based spatial data mining to improve spatial data quality. Spatial data cleaning includes understanding the semantic fields and their relationships in databases, checking and affirming the completeness and consistence of acquired data.

Spatial data cleaning is not a simple processing of turning the records into the right records and it analyzes and recombines spatial data. The methods on spatial data cleaning are tightly related to the exact task of spatial data mining and its basic methods are data merging or data purging. Under the umbrella of the techniques, spatial data mining can be classified into three types:

- *Data migration*
  It gives simple migration regulations
- *Data scrubbing*
  It makes use of specific field knowledge
- *Data auditing*
  It makes data clean with statistical analysis

It may be able to avoid the mistaken knowledge discovered from the mistaken spatial data, if the uncertainties are made good and right use of.

B) Quality assessment of spatial data clustering

Spatial data clustering points to partition the spatial data to different clusters according to the similar degree of spatial objects. The objective of the spatial clustering methods is to provide optimal partitions of a spatial data set. The number of clusters and the evaluation of clustering results have been subject of several research efforts [11], [12], [13]. Several assessment indices have been introduced, in practice they not be used by most of the clustering methods. One of quality measures that can be used in clustering was describe as:

Variance of spatial data set: The variance of spatial data set X, called σ (X), the value of the p-th dimension is defined as follows:

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}(x_k^p - x^{-p})^2$$

Where x^{-p} the p-th dimension of

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_k \,, \forall x_k \in X$$

The total variance of spatial data set with respect to c clusters:

$$\sigma = \sum_{i=1}^{c} \sigma(v_i)$$

The *distance between clusters* is defined by the average distance between the centers of specified clusters is

$$d = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c}||v_i - v_j||}{c(c-1)}$$

c) Quality assessment of spatial data classification

Spatial data classification points to assign a spatial data object to a predefined class set. The classification can be described as a function that maps (classifies) a spatial data item into one of the several predefined classes. The problems of these traditional methods may be included as follows:

- The clusters are not overlapping
- The data values are treated equally in the classification process

A successful classification scheme should contain a significant amount of information. Another requirement is the minimization of the entropy in the defined classes. A quality assessment index of classification based on the information theory will be introduced. Consider C= {c_1, c2, c3, c4… cnc} to be a classification scheme for a data set *S* into n_c clusters.

Uncertainty of a class: It evaluates the uncertainty within a class based on the memberships (degrees of belief) of the data into the specific class. It is suppressed as

$$Unc\_Cic_j = -\sum_{i=1}^{N} log_2^{(\mu_j)} \Big/ N$$

where *N* = number of tuple's in the dataset

Information coefficient of a class: It is an index of the quality of the class under consideration defined as

$$Info\_Coef(C) = \frac{1}{n}\sum_{i=1}^{n_c} InfoCl_i$$

where $InfoCl_i = DoBc_j(log_2(n_c) - Unc\_Cl_i$

IV. UNCERTAINTY HANDLING IN SPATIAL DATA MINING

Another issue will be how to handle the uncertainty in spatial data preprocessing phase. We apply the ways of Maximum Variance to discrete the continuous attribute, and fuzzy logic to "soft" the "hard" discretization interval.

a) Maximum variance data discretization

Suppose a numerical data have *N* attributes that are arranged from little too big, and partition it into *K* groups. Every group has n_i data respectively

$$x_{ij}, i = 1,2,3 \dots \dots, k; j = 1,2,3 \dots \dots, n_i$$

Variance between groups:

$$S_A = \sum_{i=1}^{k} n_i(\bar{x_i} - \bar{x})^2 = \sum_{i=1}^{k} n_i \bar{x_i}^2 - N\bar{x}^2$$

Where

$$\bar{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\ , \bar{x} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$$

We select the largest variance between groups as the optimal partition. The larger S_A is the best the partition.

b) Fuzzy belief function based on fuzzy logic

The Maximum Variance of data discretization is "hard" discretization interval. Based on Maximum Variance of data discretization, we adopt fuzzy belief function to "soft" discretization intervals as shown in the fig.2.

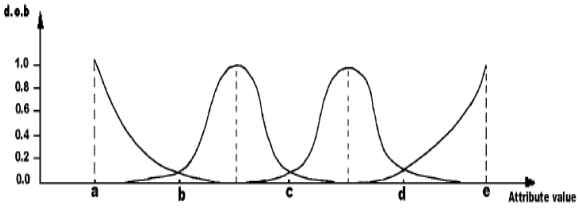

Figure 2: The fuzzy belief function for spatial data.

Where the interval [a, e] is the scope of a attribute.

V. MANAGEMENT & CONTROL

The strategies for managing uncertainties in data mining may develop formal & rigorous models of uncertainty. Through spatial processing and decision-making:

a. For the communication different levels of users in more ways that are meaningful, this refers as Communicate uncertainty.
b. To assess the fitness for use of geographic information and reducing uncertainty, design techniques are use to manageable levels for any given application
c. when uncertainty is present in geographic information leads to how take decision

There are two main technical directions to control and reduce uncertainty in an acceptable degree.

- Data acquisition that highlights the information acquired from the process of data collection and data amalgamation
- Data cognition that emphasizes the knowledge discovered from data extraction process and information generalization

The direction of data acquisition has achieved many results data amalgamation system, new sensors, computerized network, and database technology. The new algorithms on object track and object capture have further ameliorated the quality of produced information and knowledge.

VI. CONCLUSION

This paper proposed the uncertainty-based spatial data mining to achieve of both objectives, quality of spatial data mining and uncertainty handling in spatial data mining. The uncertainty-based spatial data mining is to extract knowledge from the vast repositories of practical spatial data under the umbrella of uncertainties with the given perspectives and parameters. The quality of spatial data mining can be improved by analyzing the uncertainties and its characteristics in each phase of spatial data mining and finding efficient method to reduce its uncertainties. Although the uncertainties of spatial data mining cannot be eliminated, the quality of data mining results can be evaluated in order to make use of the knowledge discovered in spatial data mining. A series of assessment indices are adopted regarding data clustering, classification and association rule mining. New techniques should be developed to handle the cases when there is more than one uncertainty in spatial data mining at the same time. Further work aims at experimental study based upper theories and methods and uncertainty propagation in spatial data mining is our interesting.

VII. REFERENCES

[1] Du Y., Li D.Y., 2001, Cloud-based concept division and its application in associated rule mining. Journal of Software, 12(2): 196-203

[2] LI D.Y., 1997, Knowledge representation in KDD ased on linguistic atoms. Journal of Computer Science and Technology, 12(6), 481-496

[3]HAN J., KAMBER M., 2001, Data Mining: Concepts and Techniques (San Francisco: Academic Press)

[4] SHI W.Z., WANG S.L., 2002**,** Further Development of Theories and Methods on Attribute Uncertainty in GIS, Journal of Remote Sensing, 6(4): 282-289

[5] DI K.C., 2001, Spatial Data Mining and Knowledge Discovery (Wuhan: Wuhan University Press)

[6]ESTER M. et al., 2000, Spatial data mining: databases primitives, algorithms and efficient DBMS support. Data Mining and Knowledge Discovery, 4, 193-216

[7]WANG S.L. et al., 2003, A method of spatial data mining dealing with randomness and fuzziness. Proceedings of the 2nd International Symposium on Spatial Data Quality, edited by Wenzhong Shi, Michael F Goodchild, Peter F Fisher, Hong Kong, March 19th – 20th, pp.370-383.

[8]LI D.R., et al., 2002, Theories and technologies of spatial data mining and knowledge discovery. Geomatics and Information Science of Wuhan University, 27(3): 221-233

[9] Burrough, P.: Natural objects with indeterminate boundaries. In: Burrough, P.A., Frank, A. (eds.) Geographic Objects with Indeterminate Boundaries, pp. 3–28. Taylor & Francis, London (1996)

[10]Miller, H.J. and Han, J., 2001: *Geographic Data Mining and Knowledge Discovery*. London, Taylor & Francis.

[11]Gath, I. and Geva., B., 1989:Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(7).

[12]Dave, R.N., 1996: Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters* 17.

[13]Rezaee, R., Lelieveldt, B.P.F. and Reiber, J.H.C, 1998: A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters* 19, 237-246.

About Authors:

Sk. Rafi, working as assistant Professor in the Department of CSE in Tirumala Engineering College, Jonnalagadda, Narasaraopet.

Sd. Rizwana working as assistant Professor in the Department of CSE in Narasaraopeta Engineering College, Narasaraopet.

