CHALLENGES IN HANDLING UNCERTAINTY IN DATA MINING RESEARCH

Abhishek Kajal, Isha Kajal
Deptt. of Computer Science, VDIET, Julana, Jind, India.
Deptt. of Computer Science, CDLU, Sirsa, India.
abhishekkajal@gmail.com, dearishakajal@gmail.com

Abstract— The digital information is growing so rapidly in organizations, due to which it's very tedious process to examine some useful information. For this, firstly we classify the explosive growth of information in some pattern, rules etc. which represent useful data. The main requirement in classifying is that the pattern which we have discovered must be comprehensible and valid. The useful data is intelligently transformed in Knowledge using new analysis techniques. This paper gives a technical overview on the problem faced during knowledge discovery process like uncertainty in rules extraction, classification, data mining process of clustering. Finally the suggestions for handling uncertainty in proper way to achieve finest results.

Keywords: Data Mining, Knowledge Discovery, uncertainty.

1. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Uncertainty in data is often associated due to outdated data source, inaccuracy in measurement, or some other errors. Like an example of a scenario in which an object (such as traffic or any person) in a moving position. Then it is impossible to create a database to track exact position of all objects at all time instants. So there is uncertainty between updates occurred in the location of every objects. To get accurate query, we consider the sources of uncertainty.

2. TOXONOMY OF UNCERTAIN DATA MINING

There are number of techniques like association rule, data clustering, and classification of data that can handle uncertainty in data mining by some modification in these techniques.

Data clustering technique is further classified in two types:

a. Fuzzy clustering:- In which each result is presented in fuzzy form. [2]e.g. each data item is given a probability of being assigned to each member in a set of clusters.
b. Hard clustering:-When data uncertainty is considered then it is used to improve the accuracy of clustering by taking expected values of data.

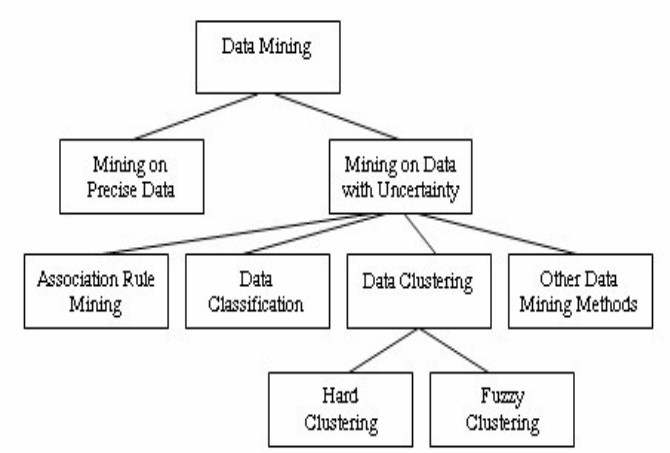

Figure 1. A taxonomy of data mining on data with uncertainty

Example of uncertainty:-

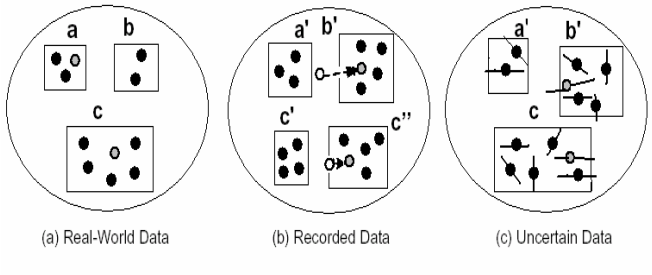

Fig2. (A) a,b,c are the partition of real world data (B) The recorded locations of some objects (shaded) are not the same as their true location, thus creating clusters a', b', c' and c'' (c) When line uncertainty is considered, clusters a', b' and c are produced The clustering result is closer to that of (a) than (b) is.

3. Challenges in Research

In KDD process Data Mining is a step which is concerned with techniques used for the extraction of the knowledge from large data. The main purpose of the data mining methods are : classification of values in database according to categories , definition of association rules, clusters that provide classification ideas. Clustering is a unsupervised process. This means there is no already defined classes and some algorithms work on assumptions. It is assumed that number of clusters extracted by an algorithm is best fitting of data set.
As a consequence, in most application there is requirement of some evaluation in the final cluster.

There are some issues which we have addressed in the above described approaches, as following:-

(i) No overlapping of clusters: - The limits of clusters are crusty and each value of database is classified in to at most one cluster. In some cases "interesting" tuples cross the limits of a cluster so they can not classified. A value can be classified in to more categories as like a person **A** with a height of 110cm may be consider as "SMALL" in India and as "Medium" in any other country.

(ii) The data values are preserved uniformly in classification process:- In data mining process the classification of database values is done in the predefined and categories in disciplined manner. As the person stated in above example, a person **A** with height 110 cm is considered in small category but if there is another person **B** which satisfied the higher degree than **A**. So it is difficult to find out the difference between **A** and **B** person which is small with the help of this existing classification method.

(iii) Knowledge may be put out of sight due to some subsequent rules:- A rule is an association of procedure X --> Y where X,Y represent collections of categories. Then X,Y categories denoted by all sets of values equally donate to the forte of rules. Each value set carries a different classification degree of belief in which every tuple sponsor its own sponsorship to the rules. Detected rule cannot internment the difference in forte of union in tuple basis. It is clear that there is no interesting knowledge extracted during the data mining process. The fact behind this is that no consideration of uncertainty.

(iv) Minor concentration on cluster excellence issues:- Classification of data set in clusters by many clustering algorithm is based on some stricture such as number of objects in cluster should be minimum, small size of cluster, less number of cluster etc. They hunt for finest clusters on basis of well sharp standards with expectation that the out coming arrangement for extracted clusters is ideal.[4] As a significance, If the algorithm stricture have not been allotted right values, the clustering way may result in a classification which is not ideal for particular data set. The main problem of choosing total number of clusters as well as calculation of the clustering out put that subject to many research efforts.

4. CURRENT SOLUTIONS

There are so many methods offered in handling uncertainty in data mining process like fuzzy decision tree, fuzzy c means. Every value of data is allotted one or more categories with attachment of d.o.b. (Degree of Belief). So there is no right way offered to handle the partition of information and feat it for decision making. Quality is also interesting point on which mostly research clustering algorithms have been distillated. The evaluation of indices also failed due to some reason like that it based on some stricture that may impact values and take it to variable output. Evaluation method will affect the data mining process.

5. PROPOSED METHOD TO HANDLE UNCERTAINTY

We can generate the clusters by performing some experiments. We suggest to follow the Uncertainty K – mean methods. In this method the actual location of any object based on Recorded and Uncertainty. When location data is stored in a set is called Recorded. So the original location is registered in Recorded.

We compute and compare the clusters output by following output of set:-

a. Recorded (using K-mean)
b. Recorded + Uncertainty (using UK- mean)

c. Actual (using classical K-mean)

After verify these results we found that the cluster generate by the UK-mean is close to the cluster generated by K-mean.

6. REFERENCES

[1] Michael Chau, Reynold Cheng, and Ben Kao "Uncertain Data Mining: A New Research Direction".

[2] Sato, M., Sato, Y., and Jain, L. "Fuzzy Clustering Models and Applications" Physica-Verlag, Heidelberg
(1997).

[3] Maria Halkidi "Quality assessment and Uncertainty Handling in Data Mining Process".

[4] R. Dave. "Validating fuzzy partitions obtained through c-shells clustering", Pattern Recognition
Letters, Vol .17, pp613-623, 1996.