DATA MINING: CONCEPTS, BACKGROUND AND METHODS OF INTEGRATING UNCERTAINTY IN DATA MINING

Yihao Li, Southeastern Louisiana University
Faculty Advisor: Dr. Theresa Beaubouef, Southeastern Louisiana University

ABSTRACT

The world is deluged with various kinds of data-scientific data, environmental data, financial data and mathematical data. Manually analyzing, classifying, and summarizing the data is impossible because of the incredible increase in data in this age of net work and information sharing. This research investigates the fundamentals of data mining and current research on integrating uncertainty into data mining in an effort to develop new techniques for incorporating uncertainty management in data mining.

INTRODUCTION

What is data mining?

Briefly speaking, data mining refers to extracting useful information from vast amounts of data. Many other terms are being used to interpret data mining, such as knowledge mining from databases, knowledge extraction, data analysis, and data archaeology. Nowadays, it is commonly agreed that data mining is an essential step in the process of knowledge discovery in databases, or KDD. In this paper, based on a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories [2].

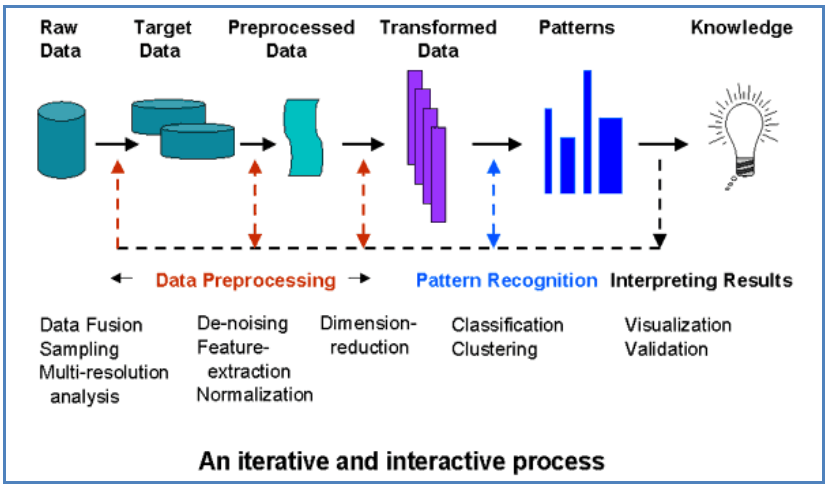

Figure 1. Knowledge Discovery in Databases [8]

Background

Necessity is the mother of invention. Since ancient times, our ancestors have been searching for useful information from data by hand. However, with the rapidly increasing

volume of data in modern times, more automatic and effective mining approaches are required. Early methods such as Bayes' theorem in the 1700s and regression analysis in the 1800s were some of the first techniques used to identify patterns in data. After the 1900s, with the proliferation, ubiquity, and continuously developing power of computer technology, data collection and data storage were remarkably enlarged. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms in the 1950s, Decision trees in the 1960s and support vector machines in the 1980s.

Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns [3]. Data mining or data mining technology has been used for many years by many fields such as businesses, scientists and governments. It is used to sift through volumes of data such as airline passenger trip information, population data and marketing data to generate market research reports, although that reporting is sometimes not considered to be data mining.

Data mining commonly involves four classes of tasks [1]: (1) classification, arranges the data into predefined groups; (2) clustering, is like classification but the groups are not predefined, so the algorithm will try to group similar items together; (3) regression, attempting to find a function which models the data with the least error; and (4) association rule learning, searching for relationships between variables.

According to Han and Kamber [2], data mining functionalities include data characterization, data discrimination, association analysis, classification, clustering, outlier analysis, and data evolution analysis. Data characterization is a summarization of the general characteristics or features of a target class of data. Data discrimination is a comparison of the general features of target class objects with the general features of objects from one or a set of contrasting classes. Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Classification is the process of finding a set of models or functions that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering analyzes data objects without consulting a known class model. Outlier and data evolution analysis describe and model regularities or trends for objects whose behavior changes over time.

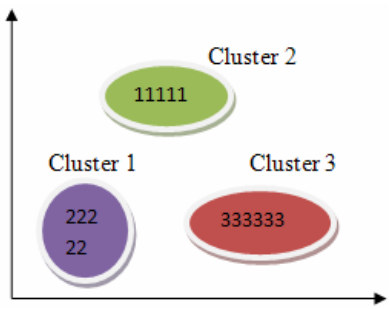

Figure 2. Cluster Analysis for numbers

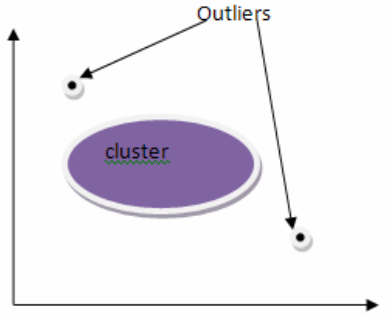

Figure 3. Outlier Analysis

DATA MINING APPLICATIONS [2]

There are approximately 100,000 genes in a human body and each gene is composed of hundreds of individual nucleotides which are arranged in a particular order. Ways of these nucleotides being ordered and sequenced are infinite to form distinct genes. Data mining technology can be used to analyze sequential pattern, to search similarity and to identify particular gene sequences that are related to various diseases. In the future, data mining technology will play a vital role in the development of new pharmaceuticals and advances in cancer therapies.

Financial data collected in the banking and financial industry is often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. Typical cases include classification and clustering of customers for targeted marketing, detection of money laundering and other financial crimes as well as design and construction of data warehouses for multidimensional data analysis.

The retail industry is a major application area for data mining since it collects huge amounts of data on customer shopping history, consumption, and sales and service records. Data mining on retail is able to identify customer buying habits, to discover customer purchasing pattern and to predict customer consuming trends. Data mining technology helps design effective goods transportation, distribution polices and less business cost.

Data mining in telecommunication industry can help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources and improve service quality. Typical cases include multidimensional analysis of telecommunication data, fraudulent pattern analysis and the identification of unusual patterns as well as multidimensional association and sequential pattern analysis.

UNCERTAINTY IN DATA MINING INTEGRATION

There are many factors causing data uncertainty in real-world applications. These factors include outdated resources, sampling errors, imprecise calculation and other errors, and so on. This is especially true for applications that require interaction with the physical world, such as location-based services [6] and sensor monitoring [7]. Recently, research has been done in the area of data uncertainty management in databases. It is proposed that when data mining is performed on uncertain data, data uncertainty has to be considered in order to obtain high quality data mining results [4]. This is called "Uncertain Data Mining".

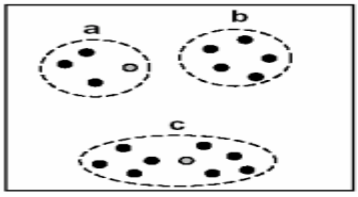

Figure 4. Real-world data

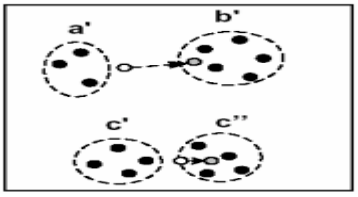

Figure 5. Recorded data

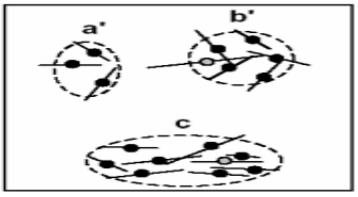

Figure 6. Uncertain data

According to Chau, et.al [4], Figure 4 shows that the real-world data are partitioned into three clusters (a, b, c). Figure 5 is the recorded locations of some objects (shaded) that

are not the same as their true location, thus creating clusters a', b', c' and c''. Note that a' has one fewer object than a, and b' has one more object than b. Also, c is mistakenly split into c' and c''. In Figure 6, line uncertainty is considered to produce clusters a', b' and c. The clustering result is closer to that of Figure 4 than Figure 5.

Based on whether data imprecision is considered, Chau, et.al [4] propose that data mining methods can be classified through a taxonomy. Common data mining techniques such as association rule mining, data classification and data clustering need to be modified in order to handle uncertain data. Moreover, there are two types of data clustering: hard clustering and fuzzy clustering. Hard clustering aims at improving the accuracy of clustering by considering expected data values after data uncertainty is considered. On the other hand, fuzzy clustering presents the clustering result in a "fuzzy" form [5].

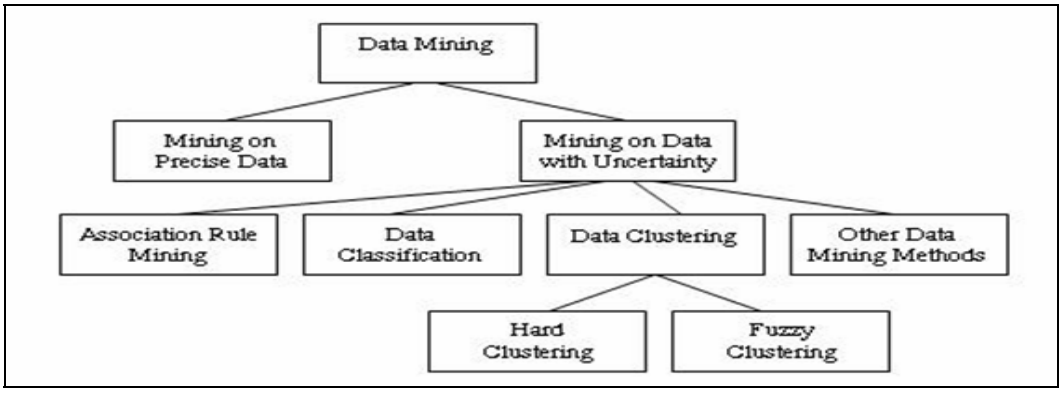

Figure 7. A taxonomy of data mining on data with uncertainty

K-means clustering for precise data [4]

The classical K-means clustering algorithm which aims at finding a set C of K clusters C_j with cluster mean $\mathbf{c}_j$ to minimize the sum of squared errors (SSE). The SSE is usually calculated as follows:

$$\sum_{j=1}^{K} \sum_{x_i \in C_j} \left\| \mathbf{c}_j - \mathbf{x}_i \right\|^2 \qquad (1)$$

Where || . || is a distance metric between a data point $\mathbf{x}_i$ and a cluster means $\mathbf{c}_j$. For example, the Euclidean distance is defined as:

$$\left\| \mathbf{x} - \mathbf{y} \right\| = \sqrt{\sum_{i=1}^{V} \left| x_i - y_i \right|^2} \qquad (2)$$

The mean (centroid) of a cluster C_i is defined by the following vector:

$$\mathbf{c}_j = \frac{1}{\left| C_j \right|} \sum_{i \in C_j} \mathbf{x}_i \qquad (3)$$

The K-means algorithm is as follows:

Assign initial values for cluster means c_1 to c_k

1 **repeat**
2 **for i**=1 to n **do**
3 Assign each data point x_i to cluster C_j where $\| c_j - x_i \|$ is the minimum
4 **end for**
5 **for** j=1 to K **do**
6 Recalculate cluster mean c_j of cluster C_j
7 **end for**
8 **until** convergence
9 **return** C

K-means clustering for uncertain data [4]

In order to take into account data uncertainty in the clustering process, Chau, et.al [11] propose a clustering algorithm with the goal of minimizing the expected sum of squared errors E(SSE). Notice that a data object $\mathbf{x}_i$ is specified by an uncertainty region with an uncertainty $f(\mathbf{x}_i)$. Given a set of clusters, C_j's the expected SSE can be calculated as:

$$\begin{aligned} & E\left(\sum_{j=1}^{k}\sum_{i\in C_j}\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2\right) \\ &=\sum_{j=1}^{k}\sum_{i\in C_j}E\left(\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2\right) \\ &=\sum_{j=1}^{k}\sum_{i\in C_j}\int\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2 f(\mathbf{x}_i)d\mathbf{x}_i \end{aligned} \qquad (4)$$

Cluster means are then given by:

$$\begin{aligned} \mathbf{c}_j &= E\left(\frac{1}{|C_j|}\sum_{i\in C_j}\mathbf{x}_i\right) \\ &=\frac{1}{|C_j|}\sum_{i\in C_j}E(\mathbf{x}_i) \\ &=\frac{1}{|C_j|}\sum_{i\in C_j}\int\mathbf{x}_i f(\mathbf{x}_i)d\mathbf{x}_i \end{aligned} \qquad (5)$$

They also propose a new K-means algorithm for clustering uncertain data.

Assign initial values for cluster means c_1 to c_k

1 **repeat**
2 **for i**=1 to n **do**
3 Assign each data point x_i to cluster C_j where $\| c_j - x_i \|$ is the minimum
4 **end for**
5 **for** j=1 to K **do**
6 Recalculate cluster mean c_j of cluster C_j
7 **end for**
8 **until** convergence
9 **return** C

Based on [4], the main difference between UK-mean clustering and K-means clustering

lies in the computation of distance and clusters. In particular, UK-means compute the *expected* distance and cluster centroids based on the data uncertainty model. Again, convergence can be defined based on different criteria. Note that if the convergence is based on squared error, E(SSE) as in Equation (4) should be used instead of SSE. In Step 4, they point out that it is often difficult to determine $E(\| \mathbf{c}_j - \mathbf{x}_i \|)$ algebraically. In particular, the variety of geometric shapes of uncertainty regions (e.g., line, circle) and different uncertainty pdf imply that numerical integration methods are necessary. In view of this, $E(\| \mathbf{c}_j - \mathbf{x}_i \|^2)$, which is easier to obtain, is used instead. This allows us to determine the cluster assignment (i.e., Step 4) using a simple algebraic expression.

CONCLUSIONS

This paper gives a general introduction of data mining, the process of discovering interesting knowledge from large amounts of data stored in information repositories. It also discusses background on data mining and methods to integrate uncertainty in data mining such as K-means algorithm. It is also shown that data mining technology can be used in many areas in real life including biomedical and DNA data analysis, financial data analysis, the retail industry and also in the telecommunication industry. One of the biggest challenges for data mining technology is managing the uncertain data which may be caused by outdated resources, sampling errors, or imprecise calculation. Future research will involve the development of new techniques for incorporating uncertainty management in data mining.

REFERENCES

1. Fayyad, Usama, Gregory Piatetsky-Shapiro, Padhraic Smyth, *From Data Mining to Knowledge Discovery in Databases*, 1996.
2. Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*, London: Academic Press, 5, 2001.
3. Kantardzic, Mehmed, *Data Mining: Concepts, Models, Methods, and Algorithms*, New York: John Wiley & Sons Inc publishes, 2003.
4. Michael Chau, Reynold Cheng, Ben Pao, *Uncertain Data Mining: A New Research Direction, Introduction*, 2005.
5. Mika Sato, Yoshiharu Sato, L.C.Jain, J.Kacprzyk, *Fuzzy Clustering Models and Applications*. UK: Physica-Verlag Heidelberg, 1999.
6. Ouri Wolfson, A. Prasad Sistla, Sam Chamberlain, and Yelena Yesha, *Updating and Querying Databases that Track Mobile Units*, MA: Kluwer Academic Publishers, 1999.
7. Reynold Cheng, Dmirti V.Kalashnikov, Sunil Prabhakar, *Evaluating Probabilistic Queries over Imprecise Data*, UK: Elsevier Science Ltd, 2007.
8. Sapphire, *Large Scale Data Mining and Pattern Recognition*, https://computation.llnl.gov/casc/sapphire/overview/data_mining_steps.gif, 1999.