Chapter 19
A Framework on Data Mining on Uncertain Data with Related Research Issues in Service Industry

Edward Hung
Hong Kong Polytechnic University, Hong Kong

ABSTRACT

There has been a large amount of research work done on mining on relational databases that store data in exact values. However, in many real-life applications such as those commonly used in service industry, the raw data are usually uncertain when they are collected or produced. Sources of uncertain data include readings from sensors (such as RFID tagged in products in retail stores), classification results (e.g., identities of products or customers) of image processing using statistical classifiers, results from predictive programs used for stock market or targeted marketing as well as predictive churn model in customer relationship management. However, since traditional databases only store exact values, uncertain data are usually transformed into exact data by, for example, taking the mean value (for quantitative attributes) or by taking the value with the highest frequency or possibility. The shortcomings are obvious: (1) by approximating the uncertain source data values, the results from the mining tasks will also be approximate and may be wrong; (2) useful probabilistic information may be omitted from the results. Research on probabilistic databases began in 1980s. While there has been a great deal of work on supporting uncertainty in databases, there is increasing work on mining on such uncertain data. By classifying uncertain data into different categories, a framework is proposed to develop different probabilistic data mining techniques that can be applied directly on uncertain data in order to produce results that preserve the accuracy. In this chapter, we introduce the framework with a scheme to categorize uncertain data with different properties. We also propose a variety of definitions and approaches for different mining tasks on uncertain data with different properties. The advances in data mining application in this aspect are expected to improve the quality of services provided in various service industries.

DOI: 10.4018/978-1-4666-2625-6.ch019

INTRODUCTION

Data mining has been widely used as an important process to improve the quality of services in service industry, e.g., targeted marketing and churn reduction using classification, store-layout design and promotion design using association rule mining, and customer segmentation using clustering. In fact, nowadays, service systems (such as customer service systems) depend heavily on data mining (business intelligence) to analyze collected user data in order to learn from them and improve the next interactions or services provided to the users again.

There has been a large amount of research work done on mining on relational databases, which are often used in service industry. Commercial vendors such as IBM, Microsoft, Oracle, SPSS Inc., DBMiner Technology Inc. as well as research institutions have been producing commercial products or research prototypes to accomplish these mining tasks. Classical sub-areas of data mining include association rules (patterns that associate features in data to discover, for example, interesting spending patterns among customers), clustering (grouping of similar data such as customers records) and classification (assigning data into predefined classes, e.g., identifying profitable customers that are likely to churn).

These works were done on databases that store data in exact values. However, in many real-life applications such as those commonly used in service industry, the raw data are usually uncertain when they are collected or produced. Sources of uncertain data include readings from sensors (such as RFID tagged in products in retail stores), information extracted using probabilistic parsing of input sources, classification results (e.g., identities of products or customers) of image processing using statistical classifiers, results from predictive programs used for stock market or targeted marketing as well as predictive churn model in customer relationship management. These uncertain data may be in the form of an exact value with margins of error, sometimes with or without a probability distribution (or density) function. The result may also be represented as an interval or a set of values, one of which may be the real value.

In this chapter, we will mainly use an example of a banking system for illustration purpose. Readers are reminded that similar techniques could be applied to other application systems in other application domains. Consider a bank receives a loan application. It may consider the applicant's credit worthiness, assets, and liabilities. The credit worthiness may be uncertain because different agencies may collect different credit histories (possibly with errors) and generate different credit scores. Assets and liabilities are in fact uncertain values due to ever-changing stock prices and interest rates, hard-to-evaluate intangible assets (e.g., patents, copyrights), and the rapid trading in global market. Clustering of past applications into different clusters (groups of similar applications) may provide the manager or decision maker a clear picture of possible main categories of applications. By comparing the new loan application with those clusters (main categories of similar applications) may help the user to make a decision of approving or rejecting the application. Careful examination of some clusters of customers may also reveal valuable customers who could be sold high-value products through up-selling or cross-selling, which is an important element of customer relationship management.

However, since traditional databases only store exact values, uncertain data are usually transformed into exact data by, for example, taking the mean value (for quantitative attributes) or by taking the value with the highest frequency or possibility. This makes the storage, query and mining much simpler by using existing commercial database systems and mining techniques, but the shortcomings are obvious:

1. By approximating the uncertain source data values, the results from the mining tasks