Data Mining in Diabetes Diagnosis and Detection

Indranil Bose
The University of Hong Kong, Hong Kong

INTRODUCTION

Diabetes is a disease worrying hundreds of millions of people around the world. In the USA, the population of diabetic patients is about 15.7 million (Breault et al., 2002). It is reported that the direct and indirect cost of diabetes in the USA is $132 billion (Diabetes Facts, 2004). Since there is no method that is able to eradicate diabetes, doctors are striving for ways to fight this doom. Researchers are trying to link the cause of diabetes with patients' lifestyles, inheritance information, age, and so forth in order to get to the root of the problem. Due to the prevalence of a large number of responsible factors and the availability of historical data, data mining tools have been used to generate inference rules on the cause and effect of diabetes as well as to help in knowledge discovery in this area. The goal of this chapter is to explain the different steps involved in mining diabetes data and to show, using case studies, how data mining has been carried out for detection and diagnosis of diabetes in Hong Kong, USA, Poland, and Singapore.

BACKGROUND

Diabetes is a severe metabolic disorder marked by high blood glucose level, excessive urination, and persistent thirst, caused by lack of insulin actions. There are usually three forms of diabetes—Type 1, Type 2, and gestational. It is believed that diabetes is a particularly opportune disease for data mining technology for a number of reasons (Breault, 2001):

- There are many diabetic databases with historic patient information.
- New knowledge about treatment of diabetes can help save money.
- Diabetes can produce terrible complications like blindness, kidney failure, and so forth, so physicians need to know how to identify potential cases quickly.

The availability of historical data on diabetes naturally leads to the application of data mining techniques to discover interesting patterns (Apte et al., 2002; Hsu et al., 2000). The objective is to find rules that help to understand diabetes, facilitate early detection of diabetes, and discover how diabetes may be associated with different segments of the population.

The data mining process for diagnosis of diabetes can be divided into five steps, though the underlying principles and techniques used for data mining diabetic databases may differ for different projects in different countries. Following is a brief description of the five steps.

Step 1: Data Cleaning

Before carrying out data mining on the diabetic patient database, the data should be cleaned. Errors such as missing values, typographical errors, or wrong information are contained in the patient records, and, worse still, many records are duplicate records. Two approaches can be used to clean the data, namely standardized format schema and sorted neighborhood method (Hsu et al., 2000). By generating the standardized format schema, a user defines mappings among attributes in different formats, and each of the database files are modified into this standardized format. The next task is to look for duplicate records that have to be removed from the standardized format using the sorted neighborhood method. Under this scheme, the database is sorted on one or more fields that uniquely identify each record, and the chosen fields of the records are compared within a sliding window. When duplicates are detected, the user is called in to verify it, and the duplicates are removed from the database.

Step 2: Data Preparation

The importance of the data preparation step cannot be overstated, since the success of data analysis depends on it (Breault et al., 2002). A fundamental issue is whether to use a relational database comprising multiple tables or a flat file that is best suited for data mining (Breault, 2002).

Generally, either of the following two methods is adopted. In the first method, a set of flat files is used to represent all the data, then data mining tools are used separately on each file, and the results obtained from the flat files are linked together in some fashion. An alternative method is to find a data reduction technique that allows fields in a complex database to be transformed into vector scores instead of considering them separately. For example, for an Australian study, the 26 items recorded for a patient from a diabetes database was converted to a four-component vector <HgbA1c, Eye, Lipids, Microalbumin> by assigning them to each component with an associated weight (determined by domain experts) that indicated how strongly the item related to the vector, and significant data reduction was obtained.

Step 3: Data Analysis

Data mining tools are used to convert the data stored in databases into useful knowledge. Different software programs using different models are run on the same data to provide different results. Sometimes, the result may be unexpected, but many of the rules and causal relationships discovered can conform to the trends. A software that is commonly used for data mining is Classification and Regression Tree (CART). CART recursively partitions the input variable space to maximize purity in the terminal tree nodes (Breault et al., 2002).

Step 4: Knowledge Evaluation

The rules obtained from data mining may not be meaningful or true. The experts have to evaluate the knowledge before using it. For example, multiple random samples should be used to evaluate the data mining tools used to insure that results are not just by chance (Breault, 2001). For diabetes, similar data mining studies in other geographic and cultural locations are needed to prove any results that are suspected to be specific to one region or segment of population.

Step 5: Knowledge Usage

After knowledge is extracted as a result of data mining, the developers have to address the issue of cost effectiveness of applying that knowledge for detection of diabetes. As a result of Step 4, the data miners may obtain a costly solution that can help a small group of people. However, this may not prove helpful for diagnosing diabetes for the entire population. It is the effective usage of the knowledge gathered from data mining for real-life application that will mark the success of the endeavor.

MAIN THRUST

Hong Kong Case—Diabetes Registry and Statistical Analysis

In Hong Kong, around 10% of the population is suffering from diabetes, of which 5% belong to Type 1 and 95% belong to the Type 2 form of diabetes. The prevalence of Type 2 diabetes is increasing at an alarming rate among Chinese, and its development is believed to involve the interplay between genetic and environmental factors. In view of this, the major hospitals have established their own diabetes registry for knowledge discovery. For instance, the diabetes clinic at the Prince of Wales Hospital of Hong Kong has adopted a structured diabetes care protocol and has created a diabetes registry based on a modified Europe DIAB-CARE format (Apte et al., 2002) since 1995.

In September 2000, a diabetes data mining study was conducted to investigate the patterns of diabetes and their relationships with clinical characteristics in Hong Kong Chinese patients with late-onset (over age 34) Type 2 diabetes (Lee et al., 2000). This study involved 2,310 patients selected from a hospital clinic-based diabetes registry. A statistical analysis tool—Statistical Package for Social Sciences (SPSS)—was used for conducting t-test, Mann-Whitney U test, analysis of variance, and χ^2 test. Many useful results were generated, which can be found on the two tables on page 1366 of the paper by Lee, et al. (2000). For example, it was found that the patients, irrespective of their sex, were more likely to have a diabetic mother than a diabetic father. Also, female patients with a diabetic mother were found to have higher levels of plasma total cholesterol compared to those having a diabetic father. In two-group comparisons, there was also evidence that the male patients with a diabetic father had higher body mass index (BMI) values than the male patients with a diabetic mother. It also was shown that both maternal and paternal factors may be responsible for the development of Type 2 diabetes in the Chinese population. All these rules greatly improved the physicians' understanding of diabetes.

United States Case—Classification and Regression Trees (CART)

Diabetes is a major health problem in the United States. According to current statistics, about 5.9% of the population, or 16 million people in the USA, have diabetes. It is estimated that the percentage will rise to 8.9% by 2025. In the United States, there is a long history of diabetic