Diabetic Data Warehouses

Joseph L. Breault
Ochsner Clinic Foundation, USA

INTRODUCTION

The National Academy of Sciences convened in 1995 for a conference on massive data sets. The presentation on health care noted that "massive applies in several dimensions . . . the data themselves are massive, both in terms of the number of observations and also in terms of the variables . . . there are tens of thousands of indicator variables coded for each patient" (Goodall, 1995, paragraph 18). We multiply this by the number of patients in the United States, which is hundreds of millions.

Diabetic registries have existed for decades. Data-mining techniques have recently been applied to them in an attempt to predict diabetes development or high-risk cases, to find new ways to improve outcomes, and to detect provider outliers in quality of care or in billing services (Breault, 2001; He, Koesmarno, Van, & Huang, 2000; Hsu, Lee, Liu, & Ling, 2000; Kakarlapudi, Sawyer, & Staecker, 2003; Stepaniuk, 1999; Tafeit, Moller, Sudi, & Reibnegger, 2000).

Diabetes is a major health problem. The long history of diabetic registries makes it a realistic and valuable target for data mining.

BACKGROUND

In-depth examination of one such diabetic data warehouse developed a method of applying data-mining techniques to this type of database (Breault, Goodall, & Fos, 2002). There are unique data issues and analysis problems with medical transactional databases. The lessons learned will be applicable to any diabetic database and perhaps to broader medical databases.

Methods for translating a complex relational medical database with time series and sequencing information to a flat file suitable for data mining are challenging. We used the classification tree approach with a binary target variable. While many data mining methods (neural networks, logistic regression, etc.) could be used, classification trees have been noted to be appealing to physicians because much of medical diagnosis training operates in a fashion similar to classification trees.

MAIN THRUST

Three major challenges are reviewed here: a) understanding and converting the diabetic databases into a data-mining data table, b) the data mining, and c) utilizing results to assist clinicians and managers in improving the health of the population studied.

The Diabetic Database

The diabetic data warehouse we studied included 30,383 diabetic patients during a 42-month period with hundreds of fields per patient.

Understanding the data requires awareness of its limitations. These data were obtained for purposes other than research. Clinicians will be aware that billing codes are not always precise, accurate, and comprehensive. However, the codes are widely used in outcomes modeling. Epidemiologists and clinicians will be aware that important predictors of diabetic outcomes are missing from the database, such as body mass index, family history of diabetes, time since the onset of diabetes, diet, and exercise habits. These variables were not electronically stored and would require going to the paper chart and patient interviews to obtain.

Developing the Data-Mining Data Table

The major challenge is transforming the data from the relational structure of the diabetic data warehouse with its multiple tables to a form suitable for data mining (Nadeau, Sullivan, Teorey, & Feldman, 2003). Data-mining algorithms are most often based on a single table, within which is a record for each individual, and the fields contain variable values specific to the individual. We call this the data-mining data table. The most portable format for the data-mining data table is a flat file, with one line for each individual record.

SQL statements on the data warehouse create the flat file output that the data-mining software then reads. The steps are as follows:

- Review each table of the relational database and select the fields to export.

- Determine the interactions between the tables in the relational database.
- Define the layout of the data-mining data table.
- Specify patient inclusion and exclusion criteria. What is the time interval? What are the minimum and maximum number of records (e.g., clinic visits or outcome measures) each patient must have to be included? What relevant fields can be missing and still include the individual in the data-mining data table?
- Extract data, including the stripping of patient identifiers to protect human subjects.
- Determine how to handle missing values (Duhamel, Nuttens, Devos, Picavet, & Beuscart, 2003)
- Perform sanity checks on the data-mining data table, for example, that the minimum and maximum of each variable make clinical sense.

Handling time series medical data is challenging for data-mining software. One example in our study is the HgbA1c, the key measure of glycemic control. This is closely related to clinical outcomes and complication rates in diabetes. Health care costs increase markedly with each 1% increase in baseline HgbA1c; patients with an HgbA1c of 10% versus 6% had a 36% increase in 3-year medical costs (Blonde, 2001). How should this time series variable be transformed from the relational database to a vector (column) in the data-mining data table? A given diabetic patient may have many of these HgbA1c results. We could pick the last one, the first, a median or mean value. Because the trend over time for this variable is important, we could choose the slope of its regression line over time. However, a linear function may be a good representation for some patients, but a very bad one for others that may be better represented by an upside down *U* curve. This difficulty is a problem for most repeated laboratory tests. Some information will be lost in the creation of the data-mining data table.

We used the average HgbA1c for a given patient and excluded patients who did not have at least two HgbA1c results in the data warehouse. We repartitioned this average HgbA1c into a binary variable based on a meaningful clinical cut-point of 9.5%. Experts agree that an HgbA1c >9.5% is a bad outcome, or a medical quality error, no matter what the circumstances (American Medical Association, Joint Commission on Accreditation of Healthcare Organizations, & National Committee for Quality Assurance, 2001).

Our final data-mining data table had 15,902 patients (rows). Mean HgbA1c > 9.5% was the target variable, and the 10 predictors were age, sex, emergency department visits, office visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, retinopathy, and end stage renal disease. All these patients had at least two HgbA1c tests and at least two office visits, the criteria we used for minimal continuity in this 42-month period.

Data-Mining Technique

We used the classification tree approach as standardized in the CART software by Salford Systems. As detailed in Hand, Mannila, and Smyth (2001), the principle behind all tree models is to recursively partition the input variable space to maximize purity in the terminal tree nodes. The partitioning split in any cell is done by searching each possible threshold for each variable to find the threshold split that leads to the greatest improvement in the purity score of the resultant nodes. Hence, this is a monothetic process, which may be a limitation of this method in some circumstances.

In CART's defaults, the Gini splitting criteria are used, although other methods are options. This could recursively continue to the point of perfect purity, which would sometimes mean only one patient in a terminal node. But overfitting of the data does not help in accurately classifying another data set. Therefore, we divide the data randomly into learning and test sets. The number of trees generated is halted or pruned back by how accurately the classification tree created from the learning set can predict classification in the test set. Cross-validation is another option for doing this, though in the CART software's defaults this is limited to n = 3000. This could be changed higher to use our full data set, but some CART consultants note, "The n-fold cross-validation technique is designed to get the most out of datasets that are too small to accommodate a hold-out or test sample. Once you have 3,000 records or more, we recommend that a separate test set be used" (Timberlake-Consultants, 2001). The original CART creators recommended dividing the data into test and learning samples whenever there were more than 1,000 cases, with cross-validation being preferable in smaller data sets (Breiman, Friedman, Olshen, & Stone, 1984).

The 10 predictor variables were used with the binary target variable of the HgbA1c average (cut-point of 9.5%) in an attempt to find interesting patterns that may have management or clinical importance and are not already known.

The variables that are most important to classification in the optimal CART tree were age (100, where the most important variable is arbitrarily given a relative score of 100), number of office visits (51), comorbidity index (CMI) (44), cardiovascular disease (16), cholesterol problems (17), number of emergency room visits (7), and hypertension (0.6).

CART can be used for multiple purposes. Here we want to find clusters of deviance from glycemic control.