USING DATA MINIG FOR DIAGNOSIS OF DIABETES: AN EXPLORATORY ANALYSIS

Abbas Heiat, College of Business, Montana State University-Billings, 1500 University Drive, Billings Montana 59101, 508-657-1627, aheiat@msubillings.edu

ABSTRACT

Determining ways to see trends in diabetes data would be a big help for health care professionals. Three data mining methods decision trees, neural network, and rule induction were used and the results of these models were compared through a Model Comparison node. The performance results of all algorithms are very close with artificial neural network yielding the best performance.

INTRODUCTION

Health care reform has been the focus of many efforts in the U.S for several years now. We spend a lot of money for health care, but health outcomes are not as good as they should be. One approach to health care reform is the prevention and cost savings that might result from taking preventive measures to improve the population's health. The disease of diabetes is of particular concern due to its chronic nature. Diabetes that normally strike adults is now affecting the younger population. Treatment is expensive and ongoing. CDC estimates the cost of diabetes and its' complications treatment amounts to $ 170 billion dollars per year. It can take a toll on a person's life expectancy and quality of life.

Scientists have discovered a number of genetic risk factors for type 2 diabetes, and tests to detect them are available directly to people via the internet, along with an analysis of that person's lifetime risk. But a new study confirms, what some have said all along, that these analyses add little to our ability to predict who will get the disease. So-called phenotypic factors, such as age, body mass index, waist circumference, and cholesterol levels are much more accurate predictors [15].

Determining ways to see trends in diabetes data would be a big help for health care professionals. It would allow them to focus on preventative treatment in the right places. The data could also show if there is an upward trend in cases of diabetes in society. This would lead to better and larger public health initiatives in communities in the U.S.

The goal of this paper is to develop a model that makes early diagnosis of diabetes possible by applying data mining methods to a database of diabetic and non-diabetic patients.

DATA PROCESSING

The data used in this research provided by SAS Corporation for their 2010 Data Mining Conference. According to SAS a large study was done in the United States to collect information on individuals, their health picture and how much money is spent on their behalf for health care. The sample provided by SAS is representative of the population and represents a snapshot of the country and its health care costs at a point in time.

Data included census region, age, marital status, years educated, highest degree, served armed forces, food stamps purchased, total income, more than one job, wears eyeglasses, blindness, hearing aids, deaf, child BMI, dental checkup, cholesterol check, last checkup, last flu shot, lost all teeth, last PSA, last pap smear, last breast exam, last mammogram, adult BMI, seat belt wearing, asthma diagnosis, high blood pressure diagnosis, heart disease diagnosis, angina diagnosis, heart attack diagnosis, other heart disease, stroke diagnosis, emphysema diagnosis, join pain, currently smoke, amount paid in Medicare, amount paid in Medicaid, total health care expenses, total office visits, weight, and diabetes diagnosis.

The information is organized with much of the data organized by code. For example, if the person was married there would be a “1” under Marital Status, “2” for Widowed, “3” for Divorced, and so on. There is a code book supplied with the data so that the values can be understood [15].

To analyze the data certain variables had to be taken into account and others excluded. The excluded variables did not have any likely impact on the early prevention of diabetes. Diabetes diagnosis and age were the primary variables needed to be taken into account. Previous diagnosis of other medical problems including joint pain, asthma diagnosis, and high blood pressure were weighted against the target variable i.e. diabetes diagnosis. Other less important variables were taken into account to help the overall process of data mining.

The data was prepared and run through exploratory analysis in order to find the most influential variables. The data was doctored to help fill the gaps with the missing data. The large data set was then broken into parts. Any person over the age of 45 was used in the analysis; the other part of the data set that contained subjects under age 45 was dropped. The remaining data was then broken into training and validation sets to be used in analysis of database.

The target or dependent variable was DIABETES_DIAG_BINARY. The following clinical and demographic variables were use in the analysis:

CURRENTLY_SMOKE, HIGH_BLOOD_PRESSURE, JOINT_PAIN, NUM_VISITS, ASTHMA_DIAGNOSIS, ADULT_BMI, TOTAL_EXP, DIABETES_DIAG_BINARY,

SEX, AGE, YEARS OF EDUCATION, TOTAL_INCOME, MORE_THAN_ONE_JOB,

CENSUS_REGION.

METHODOLOGY

Data Mining may be defined as the process of finding potentially useful patterns of information and relationships in data. More and more healthcare organizations are storing large amounts of data about patients and their medical conditions. As the quantity of clinical data has accumulated, domain experts using manual analysis have not kept pace and have lost the ability to become familiar with the data in each case as the number of cases increases. Data visualization techniques can assist in the manual analysis of data, but ultimately the human factor becomes a bottleneck as an organization using a large database can receive hundreds or even thousands of matches to a simple query [1,2,4].

Improved data and information handling capabilities have contributed to the rapid development of new opportunities for knowledge discovery. Interdisciplinary research on knowledge discovery in databases has emerged in this decade. In healthcare, pattern recognition has long been linked with expertise. Data mining, as automated pattern recognition, is a set of methods applied to knowledge discovery that attempts to uncover patterns that are difficult to detect with traditional statistical methods. Patterns are evaluated for how well they hold on unseen cases. Databases, data warehouses, and data repositories are becoming ubiquitous, but the knowledge and skills required to capitalize on these collections of data are not yet widespread. Innovative discovery-based approaches to healthcare data analysis warrant further attention [5,6,7,8]. There are situations where healthcare organizations would like to search for patterns but human abilities are not well suited to search for those patterns. This usually involves the detection of "outliers", pattern recognition over large data sets, classification, or clustering using statistical modeling. Medical data has a lot of information buried within it that will reveal patterns relating to successes and failures in clinical operations. Data mining by discovering these patterns could provide new medical information[9,10, 12].

For example, if there is a negative or an unacceptable value for income it is set to missing. Later, the Impute node will replace the missing values with appropriate values that had been specified by the analyst. One simple possible approach is the mean of the values.

Next, data was partitioned through Partition node into training and validation sets.

Three data mining methods decision trees, neural network, and rule induction were used and the results of these models were compared through a Model Comparison node.

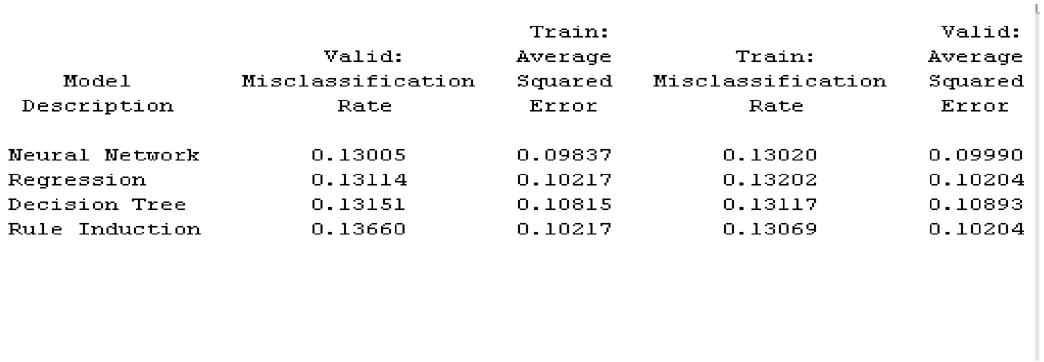
Stepwise regression was use to select the best variables and then its' was fed into neural network and rule induction models.

RESULTS OF ANALYSIS AND CONCLUSION

The decision tree model is using the total medical expenditure, high blood pressure, and body mass index variables for classifying and predicting diabetes versus non-diabetes patients.

Table 1 shows the performance results of data mining techniques used by misclassification rate and average squared error. The performance results of all algorithms are very close with artificial neural network yielding the best performance. This is confirmed by looking at the ROC chart and score rankings of the models.

The misclassification rate for all models is close to 13% which is an acceptable level of tolerance of error for a disease like diabetes. However, assigning costs to misclassification rate for each class i.e. diabetic and non-diabetic and performing a cost benefit analysis should give us a better picture of the consequences of the misclassifications by our models.



Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Neural Network	0.13005	0.09837	0.13020	0.09990
Regression	0.13114	0.10217	0.13202	0.10204
Decision Tree	0.13151	0.10815	0.13117	0.10893
Rule Induction	0.13660	0.10217	0.13069	0.10204

Table 1- Performance Results

REFRENCES

[1] Bentley, T., “Mining for Information,” *Management Accounting-London,* Vol. 75, No. 6, June 1997, pp. 56.

[2] Brachman, R., Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, “Mining Business Databases”, *Communications of the ACM,* Vol. 39, No. 11, November 1996, pp. 42-48.

[3] Brooks, P., “Data Mining Today,” *DBMS*, February 1997, pp. 59-64.

Caudill, M. & Butler, C., Naturally Intelligent Systems, *MIT Press,* Cambridge, 1990.

[4] Fayyad U., and R. Uthurusamy, “Data Mining and Knowledge Discovery in Databases” *Communications of the ACM,* Vol. 39, No. 11, November 1996, pp. 24-26.

[5] Fayyad U., “Data Mining and Knowledge Discovery: Making Sense Out of Data,” *IEEE Expert*, October 1996, pp. 20-25.

[6] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, “The KDD Process for Extracting Useful Knowledge from Volumes of Data,” *Communications of the ACM,* Vol. 39, No. 11, November 1996, pp. 27-34.

[7] Hammergren, Tom, Data Warehousing: building the corporate knowledge base, *International Thomson Computer Press*, Boston, MA, 1996.

[8] Krivda, Cheryl D, Data Mining Dynamite, *Byte Magazine*, October1995.

[9] Mark, B., “Data Mining-Here We Go Again?” *IEEE Expert*, October 1996, pp. 18-19.

[10] Meltzer, Michael, Customer Management Solutions: Using Data Mining Successfully, *NCR Corporation*, 1998.

[11] Parsaye, K.A., “Characterization of Data Mining Technologies and Processes”, *The Journal of Data Warehousing*, January 1998.

[12] Parsaye, K., “The Sandwich Paradigm for Data Warehousing and Mining”, *Database Programming and Design*, April 1995.

[13] Parsaye, K, Measuring the Dollar Value of Mined Information, *Data Management Review*, March1998.

[14] Parsaye, K., Surveying Decision Support: New Realms of Analysis, *Database Programming and Design*, April 1996.

[15] SAS Institute Inc, M2010 Data Mining Shootout, www.sas.com/events /dmconf/contestform.html

[16] Simoudis, E., “Reality Check for Data Mining,” *IEEE Expert*, October 1996,

[17] Watterson, Karen, A Data Miner’s Tools, *Byte Magazine*, October 1995.