Predictive Factors of Glycemic Control: A Comparison of Decision Tree and Neural Networks

CHUN-LIANG LAI[1], PI-SHAN HSU[2],
SHOW-WEI CHIEN[3], CHUNG-LIANG LAI[4], KWOTING FANG[5]
[135]Department of Information Management,
National Yunlin University of Science & Technology
123, Section 3, University Road, Douliou, Yunlin 64002, Taiwan

[2]Department of Family Medicine, [4]Department of Physical Medicine and Rehabilitation,
Taichung Hospital, Department of Health, Executive Yuan, Taiwan
199, Section 1, San Min Road, Taichung 40343, Taiwan

Abstract: - Way back in the past decade, we are confronted with the sharply increasing diabetes patients who have become one of the most important issues of public health. Accompanied with different complications, there is overwhelming evidence that threatens the patient's life. The complications of diabetes can be slow or even prevented by glycemic control in advance. The purpose of this study is twofold. First, from the comparison standpoint, decision tree and back-propagation neural networks were adopted to pursue the underlying characteristics of the glycemic control of the achieving target, or in risk control level, so as to provide guidelines for physicians and diabetes educators. Second, for the cross validity purpose, 512 patients, enrolled in Diabetes Healthcare Quality Improvement Program, were divided into estimated data and holdout data in a teaching hospital in Taiwan. Armed with the comparison, the finding revealed that back-propagation neural networks is more precision than C5.0 and C4.5 with the classification rate 74.5%, 68.6% and 60.8%, respectively. We concluded that age, years of diabetes onset, patient-physician relationship, medical department to visit, family history of diabetes, body mass index, during of enrollment in Diabetes Healthcare Quality Improvement Program, and education status are most important attributes influence the glycemic control.

Key-Words: - Neural networks, Decision Tree, Diabetes.

1 Introduction

During past decade, diabetes incidence and prevalence are rapidly increasing in the developing countries and newly industrialized countries. The prevalence of diabetes for all age-groups worldwide was estimated to be 2.8% in 2000 and 4.4% in 2030. The total number of people with diabetes is projected to rise from 171 million in 2000 to 366 million in 2030 [29]. Total prevalence of diabetes in the United States was 7% in 2005 with 20.8 million diabetes patients. Estimated total cost for diabetes in the United States in 2002 was US$132 billion [8].

The mortality of diabetes is still increasing year by year, besides causing loss of life; it also causes the acute and chronic complication of diabetes such as diabetic ketoacidosis, hyperglycemia, hypoglycemia, cardiovascular disease, retinopathy, nephropathy, neuropathy and foot amputation. The complications can be slow or even prevented by glycemic control in advance. In general, the diabetes patients were evaluated health care quality by using A1C (Hemoglobin A1C) [2, 19].

From the standpoint of practice, classification is one of the most useful techniques for extracting meaningful knowledge from databases. This study used decision tree algorithm C4.5, C5.0 [22, 23] and Back-Propagation Neural Networks (BPNN) [25] in terms of A1C as a decision attribute, to analyze 512 patients in a teaching hospital in Taiwan. The discovered rules may assist the physicians and diabetes educators in precisely determining the behavior characteristics of patients, in order to provide guidelines to improve glycemic control in diabetic patients.

The paper organized as follows. The next section presents the background of diabetes and thoroughly reviews the previous research in diabetes. Section 3 describes the data and prediction models, C4.5, C5.0, and BPNN. Section 4 discusses the classification results and evaluates the important attributes to classification. Finally, Section 5 provides the conclusions.

2 Background

2.1 Diabetes

The vast majority of cases of diabetes fall into two broad categories [2]. In one category, type 1 diabetes, the cause is an absolute deficiency of insulin secretion. In the other, type 2 diabetes, the cause is a combination of resistance to insulin action and an inadequate compensatory insulin secretory response. In this paper, we focus on type 2 diabetic patients.

Second we state criteria of the diagnosis of diabetes. Three ways to diagnose diabetes are available, and each must be confirmed on a subsequent day unless unequivocal symptoms of hyperglycemia are present. Criteria for the diagnosis of diabetes in nonpregnant adults were shown as follows [2]: (1). Symptoms of diabetes and a casual plasma glucose $\geqq$ 200 mg/dl. Casual is defined as any time of day without regard to time since last meal. The classic symptoms of diabetes include polyuria, polydipsia, and unexplained weight loss. OR (2). FPG $\geqq$126 mg/dl. Fasting is defined as no caloric intake for at least 8 h. OR (3). 2-h plasma glucose $\geqq$ 200 mg/dl during an OGTT (Oral Glucose Tolerance Test). The test should be performed as described by the World Health Organization, using a glucose load containing the equivalent of 75-g anhydrous glucose dissolved in water.

2.2 Knowledge discovery in database and data mining

Data mining is the process of discovering interesting knowledge from large amount data stored in databases, or other information repositories [16]. Dunham [13] define data mining as the use of algorithms to extract the information and patterns derived by the Knowledge Discovery in Data process. KDD is the process of finding useful information and pattern in data. KDD consists of following steps: Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation and Knowledge presentation [13]. Data mining is a step in the KDD process. It discovered the hidden patterns, therefore it is the most important step.

2.3 Previous research

Demographics such as age, sex, and ethnicity, affect the development of diabetes [10]. Obesity, physical inactivity, smoking, family history of diabetes have been described as risk factors for diabetes [1]. Socioeconomic status, such as education level, occupational status, and income, are implicated in the development of diabetes.

The Diabetes Control and Complications Trial (DCCT) [12] in 1993, indicated that the intensive control of blood sugar could reduce the risk of complications and diabetes-related death. Consequently, in essential, the complications of diabetes could be slow or prevented by better control on blood sugar. These researches proposed to use A1C as the criteria of glycemic control.

Quinlan [22] applied C4.5 on PIDD and it was 71.1% accurate. Breault et al. [5] applied CART [6] algorithm to classify the glycemic control, and reported the accuracy was 59.5%. The important attributes are age, Number of office visits in the given time period, Number of major complications, lipid disorder, coronary artery disease (CAD)/ Peripheral Vascular Disease (PVD) disorder, and Number of ER (emergency room) visits in the given time period. Stepaniuk [26] applied rough set to identify the most important attributes from the medical data set. The important attributes are Age of disease diagnosis, and disease duration.

3 Method

3.1 Data Source

This study used the clinical database of a teaching hospital in central Taiwan, where there are 110 physicians, six hundred hospital beds, 60 thousand outpatient services, and ten thousand inpatient services annually. The diabetes patients are collected from one-year outpatient and inpatient services and enrolled in Diabetes Healthcare Quality Improvement Program (DHQIP) [7] from Jan. 1, 2005 to Dec. 31, 2005 to serve as an analysis dataset.

3.2 Data preparation

Armed with the above-mentioned points, the data has collected includes: medical record code, name, date of birth, gender, address, postcode, medical department to visit, date of clinic visit, prescription, date of enrolled in DHQIP, year of diabetes onset, possess blood glucose meter, body mass index (BMI), education status, tobacco and alcohol use, regular exercise, family history of diabetes, ICD-9 code of diagnosis, and A1C test data. We transformed the data collected by the database into suitable format for data mining. Date of birth is transformed into age, based on the date Jan.1, 2005. The average of age is 61.83. We discretized the age into three categories: 18-40, 41-60, 61 and above. Address and postcode are transformed into the area of residence, grouped into residing in the metropolitan area (population excess one million) and sub-metropolitan area (population excess 300 thousand).

We calculated times of visit from the clinic visit time records, and the medical department is selected from the highest ratio among the clinic visits. The patient-physician relationship (PPR) is determined by the proportion of visiting same physician, if the proportion greater than or equal to 70%, we defined the PPR is stable; if the proportion less than 70%, the PPR is unstable. We classified BMI into two classes according the standard weight status categories of CDC [9]: first class is overweight and obese, BMI greater than or equal to 25.0 kg/m^2; second class is normal and underweight, BMI less than 25.0 kg/m^2. The complications, have divided into 6 categories are judged from ICD-9 diagnosis codes which listed in Table 1. Decision attribute is A1C in terms of glycemic control which is classified into two levels: A1C less than 7.0% [2], reach target level; A1C greater than 9.5% [19], fall into risk level. If A1C has more than one data, we adopted the average value. The attribute information are listed in Table 2.

Table 1 Categories of complications

Complications	ICD No.
Diabetes with Acute complication	250.1-250.3
Diabetes with Renal complication	250.4, 583.81, 581.81
Diabetes with Ophthalmic complication	250.5, 369.00-369.9, 366.41, 365.44, 362.83, 362.01-362.02
Diabetes with Neurological complication	250.6, 358.1, 354.0-355.9, 713.5, 337.1, 357.2
Diabetes with Vascular complication	250.7, 785.4, 443.81, 410-414, 430-438
Diabetes with Foot complication	707.15, 250.8

Table 2 Attribute information

Attribute		Total	%
Age	Average (S.D.)	61.83 (12.77)	
Gender	Male	244	47.7%
	Female	268	52.3%
Times of diabetic clinic visit	Average (S.D.)	5.15 (3.86)	
Patient-Physician relationship	Stable	366	71.5%
	Unstable	98	19.1%
	Only visit once	48	9.4%
Medical department to visit	Family department	190	37.4%
	Internal department	318	62.6%
Area of residence	Metropolitan area	402	79.3%
	Submetropolitan area	105	20.7%
Months of enrolled in DHQIP	Average (S.D.)	23.28 (11.41)	
Possess Blood Glucose Meter	Yes	100	19.5%
	No	412	80.5%
Years of diabetes onset	Average (S.D.)	9.93 (6.87)	
Body Mass Index	Obese & Overweight	224	45.3%
	Normal & Underweight	271	54.7%
Education status	Primary school	259	51.0%
	High school	150	29.5%
	College and above	99	19.5%
Regular smoking everyday	Yes	72	14.1%
	No	438	85.9%
Regular drinking everyday	Yes	47	9.2%
	No	463	90.8%
Regular exercising everyday	Yes	216	42.6%
	No	291	57.4%
Family history of diabetes	Yes	224	43.8%
	No	288	56.2%
With complication	Yes	363	70.9%
	No	149	29.1%
Hemoglobin A1C	Target: ＜7.0%	296	57.8%
	Risk ：＞9.5%	216	42.2%

3.3 Development of Prediction models

The existence of multicollinearity among independent variables can affect the parameters of the model. The common measures for assessing multiple variables collinearity are the *Tolerance* value and its inverse, the *Variance Inflation Factor (VIF)*. A common cutoff threshold is a *Tolerance* value below 0.10, which corresponds to a *VIF* value above 10 [15]. The *Tolerance* value of our training set all above 0.10, which indicated there is no multicollinearity between the independent variables.

3.3.1 Decision tree

From the standpoint of practice, classification is one of the most useful techniques for extracting meaningful knowledge from databases. The decision tree approach is currently one of the useful techniques in place. Its main advantage is to generate the decision rules which can be easily understood and applied. Decision tree can perform feature selection and complexity reduction. Popular decision tree algorithms include ID3, C4.5, C5 and CART.

The decision tree algorithm C4.5 has been used in a variety of field, including the diagnosis of carapal tunnel syndrome [24], and web user behavior analysis [20]. The default splitting criterion used by C4.5 is gain ratio [22]. The attribute with the highest gain ratio is chosen as the test attribute for the current node. C5.0 is a commercial version of C4.5, the major improvement of C5.0 is boosting, a technique for constructing multiple classifiers to improve predictive accuracy.

3.3.2 Neural networks

The attractiveness of neural networks comes from the remarkable characteristics such as nonlinearity, high parallelism, and robustness. Neural network are very flexible with respect to incomplete, missing and noisy data. The learning and adaptivity of neural networks allow the system to update its internal structure in response to changing environment [4, 17]. Multi-layer perceptron (MLP) are feed-forward neural networks trained with the standard back-propagation algorithm. They are supervised networks so they require a desired response to be trained [11].

The neural networks with back-propagation algorithm has been used in a variety of field, including the diagnosis of breast and ovarian cancer [28], and diabetes prediction [21]. We used a popular ANN architecture called MLP with back-propagation. In most function approximation problems, one hidden layer is sufficient to approximate continuous functions [4].

3.4 Performance evaluation

Regarding performance evaluation, we used three performance measures: accuracy, sensitivity and specificity. The classification accuracy measures the proportion of correctly classified cases. Sensitivity measure the fraction of positive cases that are classified as positive. Specificity measure the fraction of negative cases that are classified as negative [16, 18].

$$\text{Accuracy}=TP+TN \,/\, TP+TN+FP+FN. \quad (1)$$

$$\text{Sensitivity} = TP \,/\, TP+FN. \quad (2)$$

$$\text{Specificity} = TN/TN+FP. \quad (3)$$

where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively.

4 Result

4.1 Classification results

In this study, we used decision tree algorithm C4.5, C5.0 and BPNN for classification. We split the data into a training set and a test set, 80% for training and 20% for testing, i.e. 410 randomly selected cases for training and remain 102 cases for testing [14]. We applied WEKA [27] for C4.5, and Clementine for C5.0 and BPNN to classify the dataset. The models were evaluated based on classification accuracy, sensitivity and specificity.

The accuracy of training set produced by C4.5 is 82.7%, C5.0 is 89.3%, while produced by BPNN is 76.6%. The results obtained from test set are the C4.5 model achieved a classification accuracy of 60.8% with a sensitivity of 79.7% and a specificity of 34.9%. The C5.0 model achieved a classification accuracy of 68.6% with a sensitivity of 83.1% and a specificity of 48.8%. The BPNN model achieved a classification accuracy of 74.5% with a sensitivity of 84.8% and a specificity of 60.5%. Table 3 shows the complete set of results and the confusion matrix of classification.

4.2 Important Attribute to Classification

Our analysis of BPNN used Clementine with sensitivity analysis can produce the relative importance of attributes. The sensitivity of an input field is calculated by varying the value of that input field for each record in the test set. As the value is varied, the maximum and minimum outputs are stored and the maximum difference in the outputs is calculated. This maximum difference is calculated for every record, and then averaged. The input units (features) are subjected to sensitivity analysis to rank their importance. The results of important attributes produced from C4.5 and C5.0 are listed in Table 4.

As shown in the Table 4, the first important attribute for C4.5 and C5.0 is Years of diabetes onset, while for BPNN is Age. We generated ten predictive rules from C5.0. From the rule, we have found that the Years of diabetes onset above eight, the glyermic control most belong to risk level. From the rule generated from C4.5, we have found that the patients who have stable physician-patient relationship and the age strata in 41 to 60 have better glycemic control.

Department of Health of Taiwan encourages diabetic patients to enroll in DHQIP, so as to control blood sugar effectively and slow or prevent the complications by integrating healthcare of physicians, dietarians, and case managers. The patients in the program examined by the physician who were specialist and all were certificated by DHQIP, and were also mutually taken care by case managers and dietarians. During of enrollment in DHQIP is affected the glycemic control, this may verify the clinical condition.

Family medical department varies from the other department because of it is the gatekeeper of the medical treatment. So in the composed characteristics of the patients, some undifferentiated diseases or the patient's initial diseases are more than the other departments thus the ratio of target control is higher. That is, in the initial period of diabetes, only dining control or simple medicine control can lead to a higher ratio of target control than the other

departments. This major point also conforms to the clinical phenomenon.

5 Conclusion

In this paper we used C4.5, C5.0, and BPNN algorithm with Hemoglobin A1C as a decision attribute to classify the glycemic control status and to discover the behavior characteristics of patients. We acquired a very large dataset from HIS (Hospital Information System) from a teaching hospital. The results indicated that the BPNN method performed the best with a classification accuracy of 74.5%. C5.0 model was the second best with a classification accuracy of 68.6%, and the C4.5 model was the worst with a classification accuracy of 60.8%. Medical databases may consist of a large volume of heterogeneous data. The heterogeneity of the data may decrease the classification accuracy rate.

In addition to the prediction model, we also generated important attributes from classification.

The glycemic control could be predicated by some characters of diabetic patients. From the three algorithms, we concluded that age, years of diabetes onset, patient-physician relationship, medical department to visit, family history of diabetes, BMI, during of enrollment in Diabetes Healthcare Quality Improvement Program (DHQIP), and education status are most important attributes affecting the glycemic control.

The complications of diabetes have great influence on the individual quality of life and the use of medical resources. Ideally, the value of this finding may assist the physicians and diabetes educators to understand the behavior characteristics of patients, in order to provide guidelines to improve glycemic control in diabetic patients. The future work could focus on different areas of residence, the insurance conditions for medical treatment, and socioeconomic status to find more interesting knowledge.

Table 3 Classification results of test sample

Algorithm	Confusion Matrix			Accuracy %	Sensitivity %	Specificity %
	Actual class	Predicted class				
		Target	Risk			
C4.5	Target	47	12	60.8	79.7	34.9
	Risk	28	15			
C5.0	Target	49	10	68.6	83.1	48.8
	Risk	22	21			
BPNN	Target	50	9	74.5	84.8	60.5
	Risk	17	26			

Table 4 The Relative Importance of Attributes

C4.5	C5.0	BPNN
Years of diabetes onset	Years of diabetes onset	Age
Age	Age	Patient-Physician relationship
BMI	Regular smoking everyday	During of enrollment in DHQIP
Medical department to visit	Family history of diabetes	Medical department to visit
Patient-Physician relationship	Education status	Area of residence
Family history of diabetes	Medical department to visit	Regular drinking everyday
Education status	BMI	Regular exercising everyday
Times of diabetic clinic visit	Times of diabetic clinic visit	Family history of diabetes

References:

[1] Agardh, E.E. and colleagues, Explanations of socioeconomic differences in excess risk of type 2 diabetes in Swedish men and women, *Diabetes Care*. Vol.27, 2004, pp. 716-721.

[2] American Diabetes Association, Clinical Practice Recommendations 2006. *Diabetes care*. Vol.29, S1, 2006.

[3] Barriga, K.J., Hamman, R.F., Hoag, S., Marshall, J.A., and Shetterly, S.M., Population screening for glucose intolerant subjects using decision tree

analyses. *Diabetes Research and Clinical Practice*, Vol.34, Suppl. 1996, pp. S17-S29.

[4] Basheer, I.A., Hajmeer, M., Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*. Vol.43, No.1, 2000, pp. 3-31.

[5] Breault, J.L., Goodall, C.R., Fos, P.J., Data mining a diabetic data warehouse. *Artif Intell in Med.* Vol. 26, 2002, pp. 37-54.

[6] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software, 1984.

[7] Bureau of National Health Insurance, Department of Health, Executive Yuan, R.O.C., Diabetes Healthcare Quality Improvement Program, 2005. Retrieved March 8, 2006, from http://www.nhi.gov.tw/webdata/AttachFiles/Attach_3078_2_w0950059032-a1.pdf

[8] Centers for Disease Control and Prevention, U.S. Department of Health and Human Services., National diabetes fact sheet: general information and national estimates on diabetes in the United States, 2005. Retrieved March 8, 2006, from http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2005.pdf

[9] Centers for Disease Control and Prevention (CDC), U.S., BMI — Body Mass Index: About BMI for Adults. Retrieved Oct 5, 2006, from http://www.cdc.gov/nccdphp/dnpa/bmi/adult_BMI/about_adult_BMI.htm

[10] Congdon, P., Estimating diabetes prevalence by small area in England. *J. Public Health Med*. Vol.28, 2006, pp. 71-81.

[11] Delen, D., Walker, G., Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*. Vol.34, No.2, 2005, pp. 113-127.

[12] Diabetes Control and Complications Trial research group., The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine*. Vol.329, 1993, pp. 977-986.

[13] Dunham, M. H., *Data mining: introductory and advanced topics*. Prentice Hall, Upper Saddle River NJ, 2003.

[14] Eklund, P. W. and Hoang, A., Classifier Selection and Training Set Features: LMDT. 1998, Retrieved March 8, 2006, from citeseer.nj.nec.com/309003.html.

[15] Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C., *Multivariate Data Analysis, 5th edn*., Prentice Hall, Upper Saddle River, NJ, 1998.

[16] Han, J. and. Kamber, M., *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco CA, 2001.

[17] Jain, A.K., Mao, J., Mohiuddin, K.M., Artificial neural networks: a tutorial. *Comput. IEEE*, March 1996, pp. 31–44.

[18] Lavrac, N., Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*. Vol.16, 1999, pp. 3-23.

[19] National Committee for Quality Assurance (NCQA)., Diabetes Quality Improvement Project Initial Measure Set, 2003. http://www.ncqa.org/dprp/dqip2.htm

[20] Pabarskaite, Z., Decision trees for web log mining. *Intelligent Data Analysis*. Vol.7, 2003, pp. 141-154.

[21] Park, J., Edington, D.W., A sequential neural network model for diabetes prediction, *Artificial Intelligence in Medicine*, Vol.23, No.3, 2001, pp. 277-293.

[22] Quinlan, J.R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Diego, 1993.

[23] Quinlan, J.R.: Improved use of continuous attributes in C4.5. J. *Artificial Intelligence Research*. Vol.4, 1996, pp. 77-90.

[24] Rudolfer S. M., Paliouras G., Peers I. S., A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome, *Computers and biomedical research*. Vol.32, No.5, 1999, pp. 391-414.

[25] Rumelhart, D., Hinton, G. and Williams, R., *Learning internal representations by error propagation*. In : Anderson, J. and Rosenfeld, E. (eds.): Neurocomputing. MIT Press, Camb ridge, MA, 1988, pp. 675-695.

[26] Stepaniuk J, *Rough set data mining of diabetes data*. In: Ras, Z., Skowron, A. (eds.): Proceedings of the 11th International Symposium on Foundations of Intelligent Systems, 1999 June 8–11 Warsaw, Poland. ISMIS 1999. Springer, Berlin, 1999, pp. 457-465.

[27] Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

[28] Wilding, p., Morgan, M.A., Grygotis, A.E., Shoffner, M.A., and Rosato, E.F., Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. *Cancer Letter*, Vol.77, 1994, pp. 145-153.

[29] Wild, S., Roglic, G., Green, A., Sicree, R. and King, H., Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care*. Vol.27, 2004, pp. 1047-1053.