World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:6, No:3, 2012

Predictions Using Data Mining and Case-based Reasoning: A Case Study for Retinopathy

Vimala Balakrishnan, Mohammad R. Shakouri, Hooman Hoodeh and Loo, Huck-Soo

***Abstract*—**Diabetes is one of the high prevalence diseases worldwide with increased number of complications, with retinopathy as one of the most common one. This paper describes how data mining and case-based reasoning were integrated to predict retinopathy prevalence among diabetes patients in Malaysia. The knowledge base required was built after literature reviews and interviews with medical experts. A total of 140 diabetes patients' data were used to train the prediction system. A voting mechanism selects the best prediction results from the two techniques used. It has been successfully proven that both data mining and case-based reasoning can be used for retinopathy prediction with an improved accuracy of 85%.

***Keywords*—**Case-Based Reasoning, Data Mining, Prediction, Retinopathy.

I. Introduction

DIABETES is a major chronic disorder which has no cure. It can be categorized into two major types: Type 1 (insulin-dependent) and Type 2 (non-insulin dependent). People with diabetes tend to develop various complications, with retinopathy as one of the most common complications among the adults [1]. The report from the National Eye Database in Malaysia revealed that almost 36.8% of 10, 856 registered diabetes patients are inflicted with at least one severity levels of retinopathy in 2007[2].

The high prevalence and severity of retinopathy suggest the need for a screening program that is able to recognize it as early as possible. Though current clinical treatments for retinopathy slow its progression but they cannot fully reverse vision loss. Studies have confirmed that clinical prognosis is better if patients are screened and treated early. Therefore, the current study aims to design a prediction system integrating two knowledge-based approaches: data mining and case-based reasoning (CBR) to predict retinopathy among diabetes patients. The system was evaluated using 33 diabetes patients' records obtained from the University of Malaya Hospital.

V. Balakrishnan is with Department of Information System, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, MALAYSIA, (phone: +60-3-79676377; e-mail: vimala.balakrishnan@um.edu.my).

M.R. Shakouri is with Department of Computer System and Technology, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, MALAYSIA (e-mail: mimrezash@siswa.um.edu.my).

H. Hoodeh is with Department of Computer System and Technology, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, MALAYSIA (e-mail: hoodeh@siswa.um.edu.my).

H.S. Loo is with Department of Mechanical Engineering, University Technology Mara, Shah Alam, MALAYSIA, (e-mail: jloohs@pd.jaring.my).

II. Related Work

Most of the work related to retinopathy predictions was based on statistical tests. For example, Cho[3] assessed the diagnostic efficacy of macular and peripapillary retinal thickness measurements for the staging of retinopathy and the prediction of retinopathy's progression. Conway et al. [4] investigated the role of hemoglobin level in predicting proliferative retinopathy among 426 Type 1 diabetes patients. They used stereo fundus photography to determine the presence of proliferative retinopathy, followed by Cox proportional hazards modeling with stepwise regression to determine the association of hemoglobin level with proliferative retinopathy. They found that higher hemoglobin level predicts the incidence of proliferative retinopathy, though the association varies by gender, being linear and positive in men and quadratic in women.

Studies were also done using the measurement of retinal vessel diameter from fundus photographs to examine the relationship between the retinal vessel diameter and the risk of retinopathy [5-7]. These studies revealed that a larger retinal vessel diameter can be used to predict the progression of retinopathy, independent of other risk factors such as duration of disease and glycemic levels.

On the other hand, Chan[8] explored the relationship between physiological data and retinopathy, nephropathy and neuropathy in Taiwan using two data mining methods, namely C5.0 and neural network. In the C5.0 method, data with diabetes duration more than seven years were used to generate 22 rules needed for prediction whilst for the neural network method, retinopathy predictions were made based on a hidden layer with 52 neurons. The sensitivity and specificity for retinopathy prediction were found to be 58.62 and 74.73, respectively using C5.0 whereas the values were 59.48 and 99.86 for neural network, indicating the latter method has a higher prediction power.

III. Data Mining and Case-based Reasoning

Current study emphasized on two techniques: data mining and CBR. When a new case arrives, a CBR system retrieves similar cases and adapts the new case according to old cases. Systems developed using CBR are usually able to predict, diagnose and even suggest solutions to a problem [9]. There are many algorithms for calculating the distance or differences between cases in CBR. This study adapted k-nearest neighbor algorithm for this purpose.

On the other hand, data mining uncovers explicit relationships in large databases and provides models to predict a new value of a variable using those relationships. There are many predictive algorithms to create predicting models such as Artificial Neural Network (ANN), Classification and Regression Tree (CART), and C5.0 which was adapted in this study.

IV. Methodology

The current study involves three main phases in developing the prediction system: learning, experience and operation phases. In the learning phase, the knowledge base was created based on literature reviews and interviews with medical experts. This is necessary to determine the main variables required for retinopathy prediction. A total of 16 variables were selected, namely, Gender, Age, Race, Smoking, Alcohol Consumption, Body Mass Index (BMI), Glycated Hemoglobin (HbA1c), High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Triglyceride, Cholesterol, Alanine Aminotranferease (ALT), Aspartate Aminotransferase (AST), Diabetes Duration, Hypertension, and Cardiac Complication. Medical records of 185 diabetes patients were obtained from University of Malaya Hospital, and these were then cleansed using duplication elimination and statistical methods (removing outliers and extremes) resulting in a total of 140 records.

C5.0 algorithm and CBR were used in the experience phase to develop the inference engine. The C5.0 algorithm results in a decision tree as shown in Fig. 1.

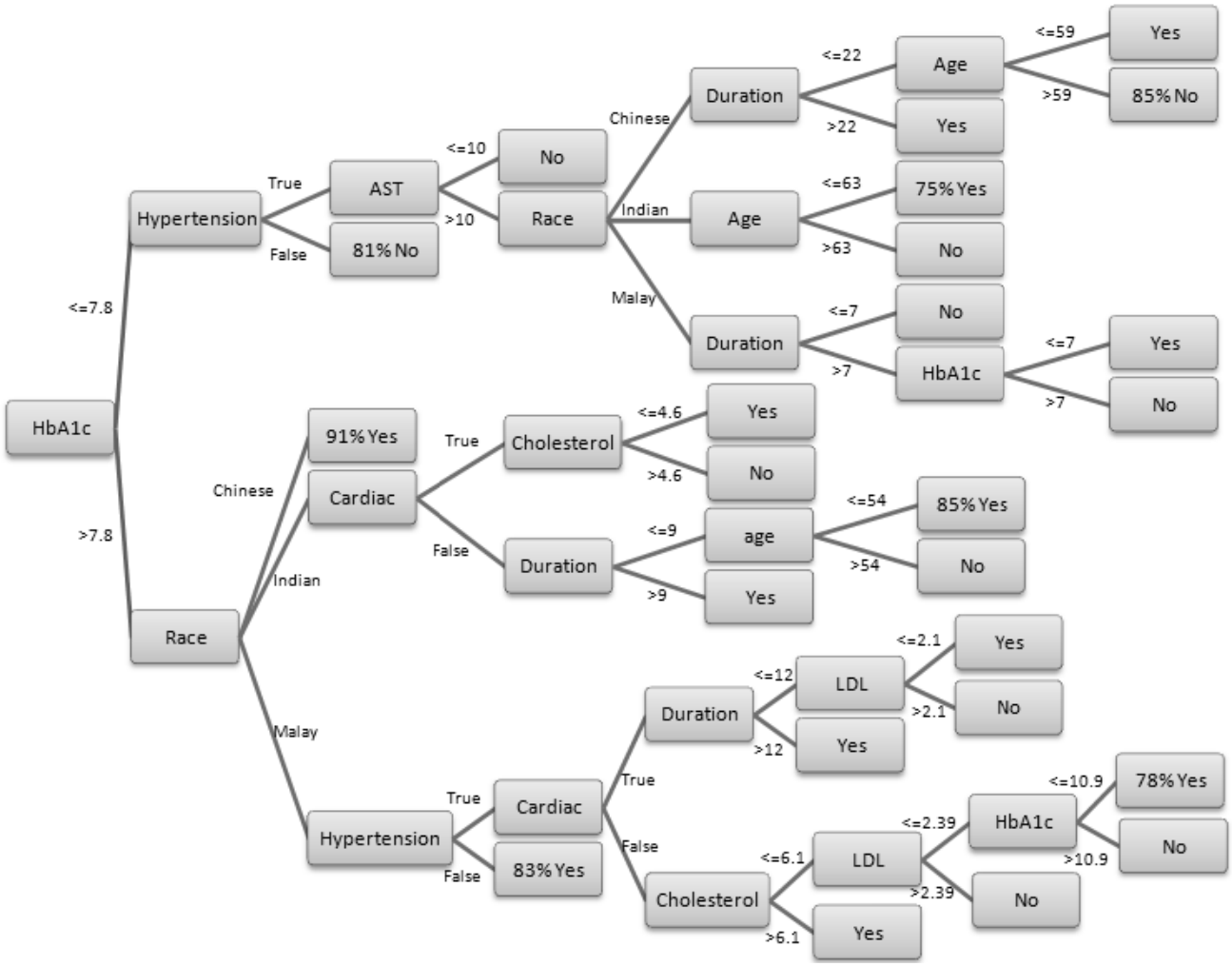

Fig. 1 The Decision Tree Produced by C5.0

Similar cases were retrieved using CBR, particularly the k-nearest neighbor (KNN) algorithm which calculates the similarity rate between two cases, as shown in Formula (1).

$$\mathcal{D}(\mathcal{C}_{new}, \mathcal{C}_{old}) = \sqrt{\sum_i^m \mathcal{D}\left(\mathcal{C}_{new,x_i}, \mathcal{C}_{old,x_i}\right)^2} \quad (1)$$

If there are n cases in a set $\mathcal{C} = \{c_1, c_2, c_3, c_4, \ldots\ldots\ldots, c_n\}$ and each case has m variables $c_i = \{x_1, x_2, x_3, x_4, \ldots\ldots\ldots, x_m\}$, then in KNN algorithm, for a variable x_i with a weight of w_i, the distance between a new case ($\mathcal{C}_{new}$) and an old case($\mathcal{C}_{old}$)is calculated using Formula (1). The current system selects the three most similar cases to be displayed (i.e. k = 3).

Fig.2 shows the main components of the system. For classification purposes, both C5.0 and CBR accesses the database containing the patients' records. C5.0 produces the decision tree (Fig. 1) whereas CBR uses the KNN algorithm (Formula 1) to provide the three most similar cases. A voting mechanism then makes the final prediction.

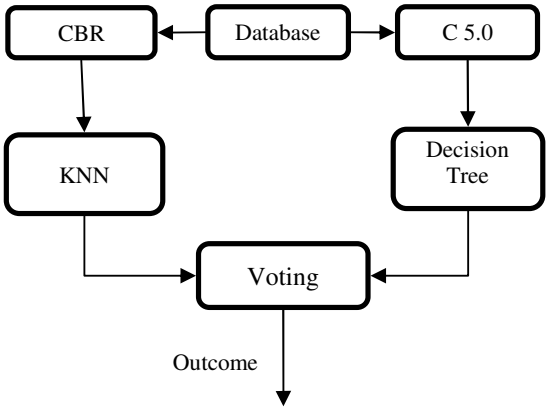

Fig. 2 System Components

When a prediction result from the decision tree is between 0 – 25%, then the decision tree is deemed to be sure that the prediction is negative but when the result is between 25 – 50%, the decision tree is not very sure about the negative prediction. Similarly, when the similarity rate is over 95% for CBR, the prediction has a high assurance but when it is between 90 – 95%, the prediction has lesser assurance. This results in one of two of the following messages:

- Immediate check - happens when both techniques vote on positive predictions or one technique votes with a high assurance on a positive prediction
- Monitoring - happens when both techniques vote on negative predictions or one technique votes with a high assurance on a negative prediction

The final phase which is called operation involves the development of the user interface using Java programming language.

V. Results and Discussion

Fig.3 shows a sample output for a negative retinopathy prediction. In this example, the input for 15 variables was entered for a 62 years old male. Although C5.0 votes on the positive prediction (50%) but CBR predicted negative with a similarity rate of 93%. Because the similarity of cases is relatively higher than the prediction of C5.0, the system votes on a final negative prediction (i.e. monitoring status). The reasoning behind the final prediction is also provided together with the three most similar cases.

On the other hand, Fig.4 shows a positive prediction for a 49 years old male. Contrary to the previous example, the C5.0 outcome is 100% and despite the negative vote of CBR the overall vote was positive (i.e. immediate check).

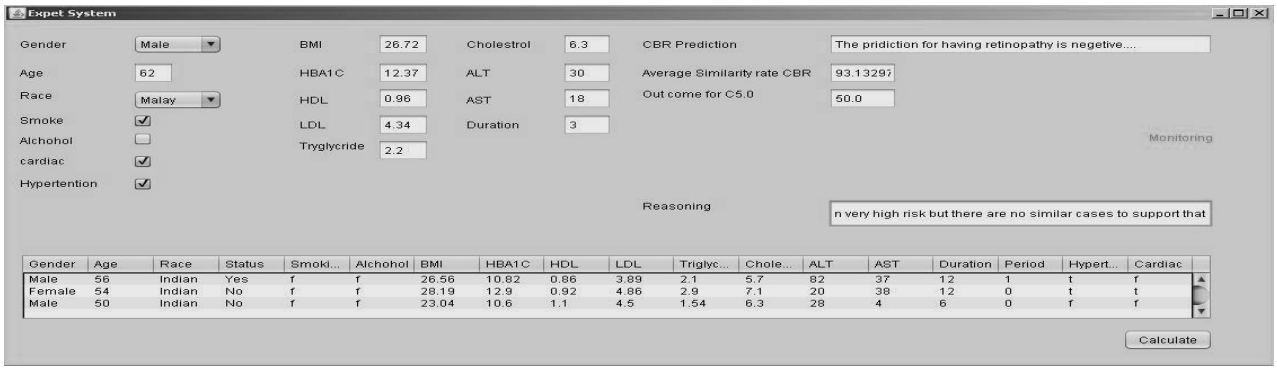

Fig. 3 Application Screen Shot for a Negative Prediction

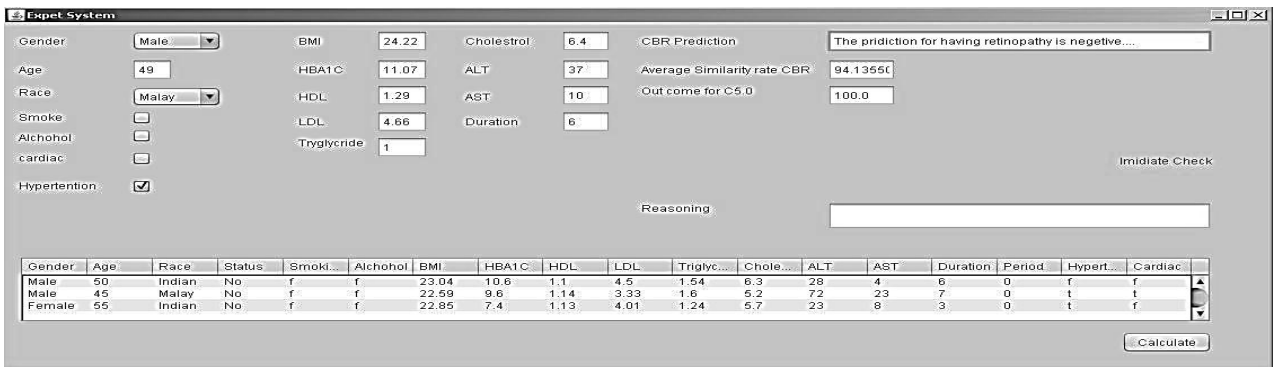

Fig. 4 Application Screen Shot for a Positive Prediction

A total of 33 patients' data were used to evaluate the prediction system. These 33 data refers to those patients whom the authors managed to trace back based on the 140 records. It was found that C5.0 resulted in an accuracy of 76% whereas CBR's accuracy was 73%. The integrated techniques resulted in a better prediction, that is, 85%. It is believed higher accuracy can be achieved with the availability of more patients' data.

ACKNOWLEDGMENT

The authors wish to thank University of Malaya for assisting us with the study.

REFERENCES

[1] P. Zimmet, K.G.M. Alberti, and J. Shaw, "Global and societal implications of the diabetes epidemic," *Nature*, vol. 414, pp. 782-787, 2001.

[2] P. P. Goh, "Status of Diabetic Retinopathy Among Diabetics Registered to the Diabetic Eye Registry, National Eye Database, 2007," *Med. J. Malaysia,* vol. 63, pp. 24-28, 2008.

[3] H.Y. Cho, D.H. Lee, S.E. Chung, and S.W. Kang, "Diabetic Retinopathy and Peripapillary Retinal Thickness," *Korean J Ophthalmol*, vol. 24, pp. 16-22, 2010.

[4] B.N. Conway, R.G. Miller, and R. Klein, T.J. Orchard, "Prediction of Proliferative Diabetic Retinopathy With Hemoglobin Level," *Arch Ophthalmol,* vol. 127, pp. 1494-1499, 2009.

[5] R. Klein, B.E.K. Klein, S.E. Moss, T.Y. Wong, L. Hubbard, K.J. Cruickshanks, and M. Palta, "The Relation of Retinal Vessel Caliber to the Incidence and Progression of Diabetic Retinopathy: XIX: The Wisconsin Epidemiologic Study of Diabetic Retinopathy," *Archives of Ophthalmology*, vol. 122, pp. 76-83, 2004.

[6] N. Cheung, S.L. Rogers, K.C. Donaghue, A.J. Jenkins, G. Tikellis, and T.Y. Wong, "Retinal Arteriolar Dilation Predicts Retinopathy in Adolescents with Type 1 Diabetes," *Diabetes Care*, vol. 31, pp. 1842-1846, 2008.

[7] T.T. Nguyen, J. Wang, A. Sharrett, F. Islam, R. Klein, K. Klein, M. Cotch, and T. Wong, "Relationship of Retinal Vascular Caliber With Diabetes and Retinopathy," *Diabetes Care,* vol. 31, pp. 544-549, 2008.

[8] C.L. Chan, Y.C. Liu, and S.H. Luo, "Investigation of diabetic microvascular complications using data mining techniques," in Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). *IEEE International Joint Conference on Neural Networks*, 2008, pp. 830-834.

[9] J. L. Kolodner, "An introduction to case-based reasoning," *Artificial Intelligence Review*, vol. 6, pp. 3-34, 1992.