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.

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

The 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.