Maejo Int. J. Sci. Technol. **2012**, *6*(03), 334-343

Maejo International
Journal of Science and Technology

ISSN 1905-7873
Available online at www.mijst.mju.ac.th

Short Communication

Integrating association rules and case-based reasoning to predict retinopathy

Vimala Balakrishnan [1,*], Mohammad R. Shakouri [2] and Hooman Hoodeh [2]

[1] Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

* Corresponding author, e-mail: vim.balakrishnan@gmail.com

Received: 19 August 2011 / Accepted: 11 September 2012 / Published: 11 September 2012

Abstract: This study proposes a retinopathy prediction system based on data mining, particularly association rules using Apriori algorithm, and case-based reasoning. The association rules are used to analyse patterns in the data set and to calculate retinopathy probability whereas case-based reasoning is used to retrieve similar cases. This paper discusses the proposed system. It is believed that great improvements can be provided to medical practitioners and also to diabetics with the implementation of this system.

Keywords: predictive system, data mining, association rules, Apriori algorithm, case-based reasoning

INTRODUCTION

Diabetes is a major chronic metabolic disorder which is characterised by persistent hyperglycaemia (elevated blood glucose). It has been estimated that the total number of people in the world with diabetes will rise from 171 million in 2000 to 366 million in 2030 [1]. With a growing number of people diagnosed with diabetes, Malaysia is also experiencing the same phenomenon, as prevalence of the disease stands at 14.9% of adult population. It has been reported that one in six adult Malaysians above 30 years of age have diabetes. Interestingly, the World Health Organisation (WHO) has projected that Malaysia will have a total of 2.48 million people with diabetes by 2030 [2].

Diabetics are prone to develop various complications, especially when the disease is not well controlled. The major complications are generally categorised as micro-vascular (diabetic nephropathy, retinopathy and neuropathy) and macro-vascular (coronary artery disease, stroke and peripheral arterial disease) [3]. Diabetic retinopathy (DR) or simply retinopathy is one of the most common micro-vascular complications of diabetes. It is a sight threatening complication that affects the retina and is a leading cause of blindness among the working age population [4]. According to WHO, approximately 4.8% of cases of blindness globally are due to retinopathy [5]. A study conducted at the Ophthalmology Clinic, Medical Centre, University of Malaya showed the overall prevalence of retinopathy to be 51.6% [6], which was higher than the previous prevalence rate of 48.6% reported from University Sains Malaysia Hospital in 1996 [7]. The severity of retinopathy can be categorised into five levels: no-DR, mild non-proliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR and proliferative diabetic retinopathy (PDR) [8]. In 2007, the National Eye Database in Malaysia reported that 36.8% of 10,856 registered diabetic patients were inflicted with at least one of these severity levels [9].

Diabetes and its complications are becoming increasingly prevalent worldwide and they impose a heavy burden on the health system in any country. In Malaysia about 14.5 billion MYR (USD 4.75 billion) was estimated for 60,000 diabetic patients per year that were registered with the Ministry of Health [10]. Current methods of detecting, screening and monitoring retinopathy are based on subjective human evaluation, which is slow and time-consuming. The high prevalence and severity of retinopathy suggests the need for a screening programme that can recognise it as early as possible. Early detection becomes more important since retinopathy can be asymptomatic even in its more advanced stages. Medical guidelines suggest that Type 2 diabetics should have a comprehensive eye examination shortly after the diagnosis as retinopathy is often already present then [11]. This paper aims to propose a prediction system for retinopathy among diabetic patients. The system is mainly intended for the physicians as it is expected to simplify their decision-making process.

LITERATURE REVIEW

The majority of work involving retinopathy were related to improving algorithms to perform image processing in order to diagnose retinopathy. The current retinopathy diagnostic method involves the use of seven-field stereo fundus photography reviewed by a trained reader. Research has shown that combining fundus photography and computer algorithms improves diagnostic performance. For example, Sanchez et al. [12] came up with an automatic image processing algorithm to detect hard exudates (bright lipids leaked from a blood vessel) based on Fisher's linear discriminant analysis. Hann et al. [13] developed a computer-vision method of isolating and detecting two of the most common retinopathy dysfunctions, i.e. dot haemorrhages and exudates, using specific colour channels and segmentation methods to separate these retinopathy manifestations from physiological features in the digital fundus images. Other studies related to enhancing algorithms for detecting retinopathy were done using artificial neural network method [14], computer-based classification methods [15], and rule-based classifiers [16], among others.

As for prediction, most work in the literature focused on determining significant predictors of retinopathy in a diabetic patient. Statistical tests were found to be commonly used for this purpose. For instance, Semeraro et al. [17] used various statistical measures such as Kaplan-Meier method to generate univariate survival curves, which were then compared among themselves using the logrank test, U-statistics, regression analysis and Cox analysis, among others, to identify patients who are at a higher risk for retinopathy. Their results showed that the duration of diabetes, glycosylated hemoglobin, systolic blood pressure, gender (male), albuminuria and diabetes therapy were significantly associated with the occurrence of retinopathy. Similar work was carried out by Cho et al. [18] who assessed the diagnostic efficacy of macular and peripapillary retinal thickness measurements for the staging of retinopathy and the prediction of its progression.

Another study that focused on determining retinopathy predictors was conducted in Taiwan. Chan et al. [19] compared the performance of two data-mining methods, namely C5.0 and neural network, in identifying key predictors for diabetic retinopathy, nephropathy and neuropathy. In the C5.0 method, data with diabetes duration of more than seven years were used to generate 22 rules needed for prediction. On the other hand, with the neural network method, retinopathy predictions were made based on a hidden layer with 52 neurons. The sensitivity and specificity for retinopathy prediction was found to be 58.62 and 74.73 respectively using C5.0, whereas the values were 59.48 and 99.86 respectively for neural network, indicating that the latter method was better in determining retinopathy predictors. Their results revealed ten key predictors based on the number of occurrences in the rules, with creatinine emerging as the most important predictor, followed by diabetes duration and family history.

To the best of our knowledge, only the work of Skevofilakas et al. [20] focused on developing a decision support system to predict the risk of retinopathy occurrences among Type-1 diabetic patients. The system was built by combining a feed-forward neural network, a classification and regression tree and a rule induction C5.0 classifier with an improved hybrid wavelet neural network. The data from 55 Type-1 diabetic patients were used to test the system, which resulted in a performance with an accuracy of 98%. It is to be noted that in determining retinopathy occurrences the authors only used seven risk factors, i.e. age, diabetes duration, glycated hemoglobin, cholesterol, triglycerides, hypertension incidence rate and their treatment duration. Various studies have revealed that factors such as gender, smoking habit and even race or ethnicity play major roles in retinopathy occurrences [6, 17-19]. In addition, the authors only focused on Type-1 patients.

The literature also revealed some other methods of identifying diabetic patients at risk of developing retinopathy, e.g. electroretinogram (ERG) [21], multifocal ERG [22] and visual evoked potentials [23]. However, none of these methods have fully entered the clinical practice due to the high cost of instrumentation and the difficulty of carrying out an examination by both the examiners and the patients [17].

In summary, the literature review has revealed that most previous studies related to retinopathy predictions were more inclined in determining the significant risk factors/predictors associated with retinopathy occurrences [17-19]. Study related to retinopathy prediction and machine learning techniques was limited to that of Skevofilakas and coworkers [20], although it is not without its shortcomings as indicated above. The rest of the techniques were reported to be either costly or difficult for both examiners and patients [21-23]. Therefore, the present study proposes to develop a system for prediction of retinopathy resulting from both Type-1 and Type-2 diabetics by integrating data mining, particularly association rules generated using Apriopri algorithm, and case-based reasoning (CBR). In addition, more important risk factors are identified and included so as to ensure an accurate prediction. We have conducted a similar study in predicting retinopathy by focusing on the use of C5.0, K-nearest neighbour and Hamming algorithms to make the prediction [24]. In this communication, we integrate association rules generated using Apriopri algorithm instead of decision trees generated by C5.0.

PROPOSED METHODOLOGY

Figure 1 illustrates the proposed methodology for this study, which can be segregated into three phases. Phase one involves the data collection, followed by data analysis and data pre-processing in phase two. In addition, rules required to predict retinopathy are also generated in this phase. Finally, phase 3 involves the use of the system to predict retinopathy. The case database is accessed to retrieve similar cases for the retinopathy predictions. The following sections elaborate these phases.

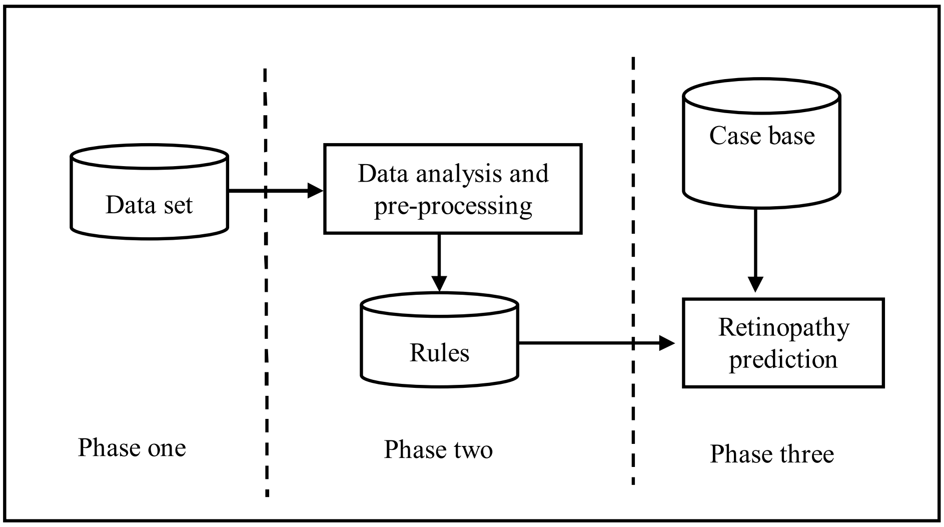

Figure 1. Proposed research methodology

Data Set

The data collection for the study has been accomplished. These data were obtained from the University of Malaya Hospital, comprising Type-1 and Type-2 diabetic patients. The data are needed to generate the necessary rules for retinopathy prediction. Medical experts were interviewed to help assist in understanding the nature of the data collected and also to determine the important risk factors for predicting retinopathy. The literature shows many risk factors that can be used to predict retinopathy, such as body mass index (BMI), smoking history, age and diabetes duration [6, 17-20]. These variables were then reviewed and validated by three medical experts who confirmed that 16 risk factors are crucial for retinopathy prediction. These are BMI, high-density lipoprotein, triglyceride, diabetes duration, glycated hemoglobin, hypertension, age, cholesterol, low-density lipoprotein, alanine aminotranferease, aspartate aminotranferease, cardiac complication, gender, race, smoking and alcohol consumption.

Data Pre-processing

The next step in the study is data pre-processing. Figure 2 shows the overall process involved in preparing the data for rule generation. Before using the data mining algorithm, the data collected may need to be cleaned and filtered to avoid the creation of inappropriate or inaccurate rules [25]. Some of the actions in the data pre-processing are removing duplicate records, normalising the values used to represent information in the database, accounting for missing data points and removing unneeded data fields [25]. The literature revealed four different approaches to clean a data set [26]: parsing, data transformation, duplication elimination and statistical methods. The current study intends to use data transformation, duplication elimination and statistical methods for data cleaning, if necessary. In data transformation the subjected variables in a database are transformed into a right format to fit in the data mining process. For example, in this study continuous attributes such as hypertension are categorised into classes for the purpose of association mining and building decision trees. Duplication elimination deletes duplicate records from the data sets whereas statistical methods analyse the data sets to identify false data. Once the data set is cleaned, the system specifies the explicit relationship between the variables in the data set using association rule mining algorithm.

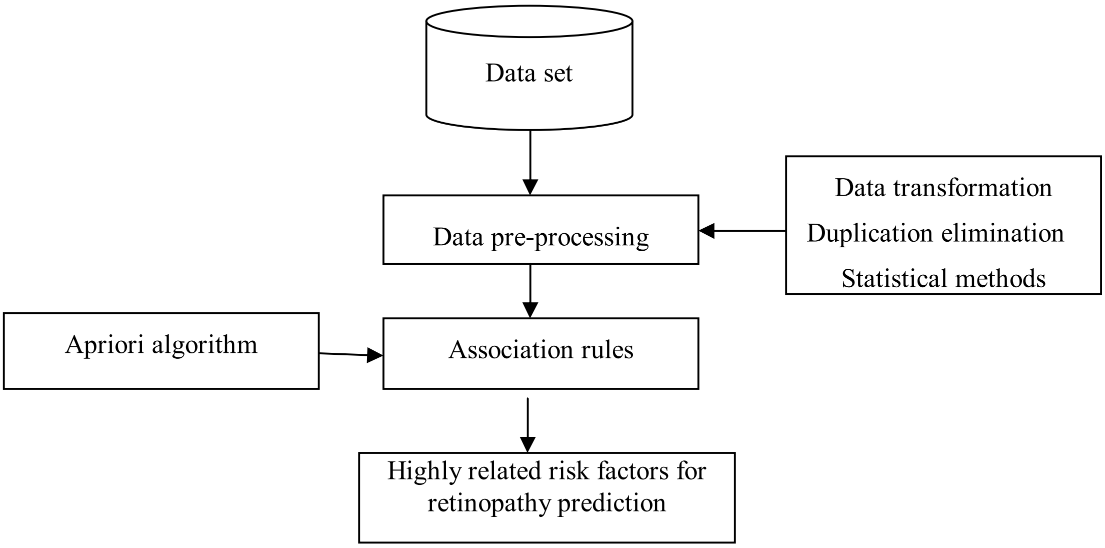

Figure 2. Overall flow of association rules generation

Association Rules

Association rules find interesting associations and/or relationships among a large set of data items. They capture all possible rules that explain the presence of some attributes in relation to the presence of other attributes. An association rule can be expressed in the form of X → Y, whereby X and Y are disjointed item sets. An association rule must satisfy two important measures, namely support and confidence. Support indicates how frequently the items in the rule occur together whereas confidence determines how frequent items in Y appear in transactions containing X [27].

It is important to analyse the relationships among the risk factors and Apriori algorithm is used for this purpose. It is an powerful algorithm for mining frequent item sets for association rules.

It involves two stages. The first stage identifies frequent item sets and the second derives the association rules from the identified frequent item sets [28]. Frequent item sets identification in the first stage is accomplished by scanning over the data multiple times and counting the support of the individual item. Support is calculated by dividing the number of rules in which the item set is found by the total number of transactions. In each subsequent pass a set of item sets found to be frequent in the previous pass are used to generate new frequent item sets, and their support is calculated again. At the end of the pass, item sets satisfying minimum support thresholds are collected and they become the seed for the next pass. This process is repeated until no new frequent item sets are found.

The second stage is to generate the desired association rules from the frequent item sets. This is accomplished by calculating the confidence for each frequent item set. The confidence is the number of times a condition in the most frequent item sets is true over the whole number of frequent item sets [28]. The result of these stages will be a set of highly related risk factors that can be used to predict retinopathy. These association rules are to be first selected based on their support and confidence values (e.g. above 0.9) and will then be verified by the medical experts to ensure an accurate retinopathy prediction.

Case-based Reasoning

Case-based reasoning (CBR) is to be used once the system has calculated the risk of retinopathy occurrence in a patient. CBR is a technique which operates based on older cases. When a new case arrives, a CBR system retrieves similar cases and adapts or justifies the new case according to old cases. Systems developed using CBR can usually predict, diagnose and even suggest solutions for a problem [29]. In CBR systems, retrieving valuable cases from the database is very important. In this study the cases obtained from the University of Malaya Hospital are to be stored in a case database. It is important for these cases to contain at least all the 16 variables needed for the prediction. The indexing function helps to index cases according to their critical risk factors (e.g. demographic data, duration of diabetic and HbA1c) and the risk factors' weights. The better the indexing is, the more accurate the results will be.

The overall flow of prediction is depicted in Figure 3. The retinopathy risk evaluator receives its input when the medical expert enters the patient's data (age, duration of diabetics, HbA1c, etc.). These inputs will be analysed by the risk evaluator and the probability of retinopathy occurrence will be displayed if the risk is deemed to be low (e.g. < 25%). On the other hand, if the risk is high, then the CBR will retrieve similar cases from the existing case database. Once the most similar case is identified from the database, the system will modify (if necessary) the retrieved cases to solve the new case, a process known as adaptation. The outputs of this phase are the probability of the incidence of retinopathy with the proximate time of occurrence (in years).

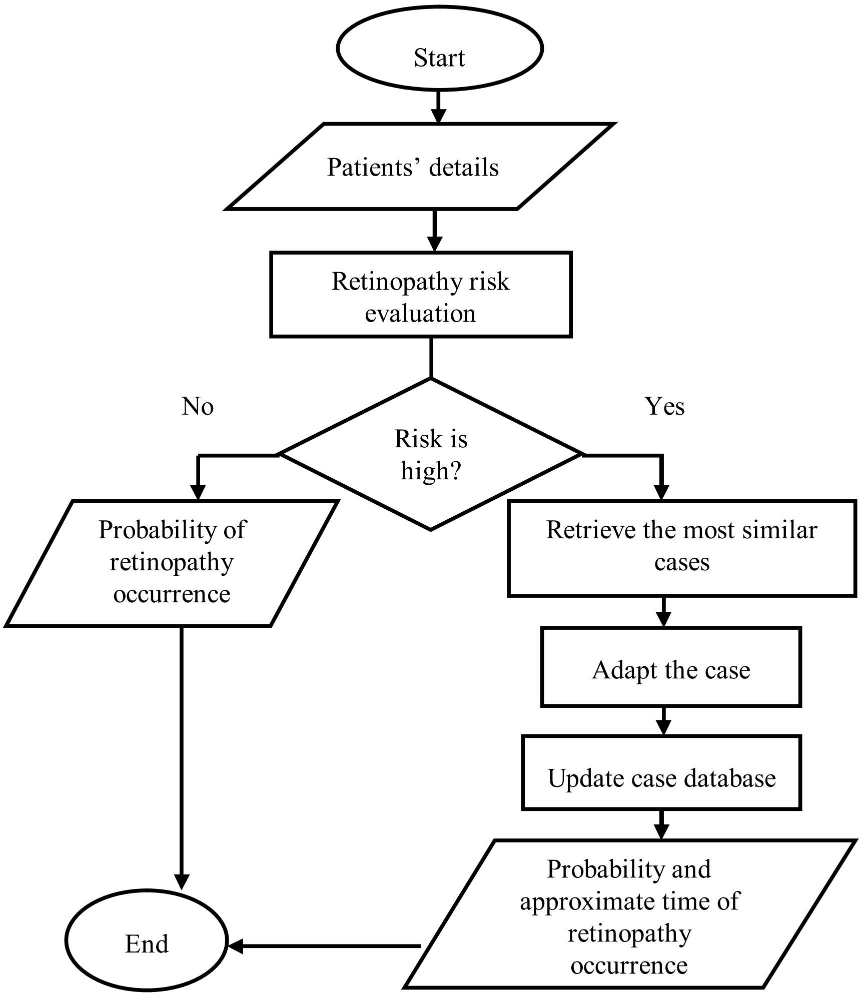
Start
Patients' details
Retinopathy risk evaluation
No
Yes
Risk is high?
Probability of retinopathy occurrence
Retrieve the most similar cases
Adapt the case
Update case database
Probability and approximate time of retinopathy occurrence
End

Figure 3. Retinopathy prediction flowchart

CONCLUSIONS AND FUTURE WORK

This paper has proposed to develop a prediction system for retinopathy using two popular approaches: data mining (association rules from Apriori algorithm) and CBR. The data set in the study goes through the data mining process first before being analysed by the CBR technique. The data set and cases can be obtained by interviewing some domain experts. Apriopri algorithms are used to generate the required association rules for the risk factors. These rules cam be used to make the prediction of retinopathy. Similar cases are then retrieved using CBR technique. It is believed that the implementation of the system will enable medical practitioners to predict the chances of retinopathy occurrences among their diabetic patients, and hence to be able to start the treatment or control measures early.

ACKNOWLEDGEMENTS

This study is funded and supported by the University of Malaya Research Grant (Grant no. RG052/11ICT). The authors also wish to extend their gratitude to Professor Umah Kuppusamy, Associate Professor Chua Kek Heng and Associate Professor Tajunisah Begam from the Faculty of Medicine, University of Malaya for their valuable comments and assistance.

REFERENCES

1. S. H. Wild, G. Roglic, A. Green, R. Sicree and H. King, "Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030", *Diabetes Care*, **2004**, *27*, 1047-1053.
2. M. Nor, N. Safiza, G. K. Lin, S. Suzana, C. C. Kee, H. Jamaiyah, A. Geeta, R. Rahmah, N. F. Wong, A. A. Zainuddin, A. R. Jamaluddin, A. T. Ruzita and A. F. Yusoff, "The third national health and morbidity survey (NHMS III) 2006: Nutritional status of adults aged 18 years and above", *Malaysian J. Nutr.*, **2008**, *14*, 1-87.
3. M. J. Fowler, "Microvascular and macrovascular complications of diabetes", *Clin. Diabetes*, **2008**, *26*, 77-82.
4. P. Zimmet, K. G. M. M. Alberti and J. Shaw, "Global and societal implications of the diabetes epidemic", *Nature*, **2001**, *414*, 782-787.
5. S. Resnikoff, D. Pascolini, D. Etya'ale, I. Kocur, R. Pararajasegaram, G. P. Pokharel and S. P. Mariotti, "Global data on visual impairment in the year 2002", *Bull. World Health Organ.*, **2004**, *82*, 844-851.
6. I. Tajunisah, H. Nabilah and S. C. Reddy, "Prevalence and risk factors for diabetic retinopathy-A study of 217 patients from University of Malaya Medical Centre", *Med. J. Malaysia*, **2006**, *61*,451-456.
7. S. R. Shriwas, A. B. R. Isa, S. C. Reddy, M. Mohammad, W. B. Mohammad and M. Mazlan, "Risk factors for retinopathy in diabetes mellitus in Kelantan, Malaysia", *Med. J. Malaysia*, **1996**, *51*, 447-452.
8. American Academy of Ophthalmology Retina/Vitreous Panel, "Preferred practice pattern: Diabetic retinopathy 2003", Technical Report of American Academy of Ophthalmollogy, San Francisco, **2003**, www.aao.org/ppp (Accessed: June 2011).
9. P. P. Goh, "Status of diabetic retinopathy among diabetics registered to the diabetic eye registry, National Eye Database, 2007", *Med. J. Malaysia*, **2008**, *63*, 24-28.
10. M. Y. Tan, "Self-care practices of adults with poorly controlled diabetes mellitus in Malaysia", *PhD Thesis,* **2009**, University of Adelaide, Australia.
11. Amercian Diabetes Association, "Executive summary: Standards of medical care in diabetes—2008", *Diabetes Care*, **2008**, *31*, S5-S11.
12. C. I. Sanchez, A. Mayo, M. Garcia, M. I. Lopez and R. Hornero, "Automatic image processing algorithm to detect hard exudates based on mixture models", Proceedings of 28th Annual International Conference of the IEEE, **2006**, New York, USA, pp.4453-4456.
13. C. E. Hann, J. A. Revie, D. Hewett, J. G. Chase and G. M. Shaw, "Screening for diabetic retinopathy using computer vision and physiological markers", *J. Diabetes Sci. Technol.*, **2009**, *3,* 819-834.
14. G. G. Gardner, D. Keating, T. H. Williamson and A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: A screening tool", *Br. J. Ophthalmol.*, **1996**, *80*, 940-944.

15. A. J. Frame, P. E. Undrill, M. J. Cree, J. A. Olson, K. C. McHardy, P. F. Sharp and J. V. Forrester, "A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein angiograms", *Comput. Biol. Med.*, **1998**, *28*, 225-238.
16. A. W. Reza and C. Eswaran, "A decision support system for automatic screening of non-proliferative diabetic retinopathy", *J. Med. Syst.*, **2011**, *35*, 17-24.
17. F. Semeraro, G. Parrinello, A. Cancarini, L. Pasquini, E. Zarra, A. Cimino, G. Cancarini, U. Valentini and C. Costagliola, "Predicting the risk of diabetic retinopathy in type 2 diabetic patients", *J. Diabetes Complicat.*, **2011**, *25*, 292-297.
18. H. Y. Cho, D. H. Lee, S. E. Chung and S. W. Kang, "Diabetic retinopathy and peripapillary retinal thickness", *Korean J. Ophthalmol.*, **2010**, *24*, 16-22.
19. C. L. Chan, Y. C. Liu and S. H. Luo, "Investigation of diabetic microvascular complications using data mining techniques", Proceedings of International Joint Conference on Neural Networks, **2008**, Hong Kong, China, pp.830-834.
20. M. Skevofilakas, K. Zarkogianni, B. G. Karamanos and K. S. Nikita, "A hybrid decision support system for the risk assessment of retinopathy development as a long term complication of type 1 diabetes mellitus", Proceedings of 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, **2010**, Buenos Aires, Argentina, pp.6713-6716.
21. C. D. Luu, J. A. Szental, S. Y. Lee, R. Lavanya and T. Y. Wong, "Correlation between retinal oscillatory potentials and retinal vascular caliber in type 2 diabetes", *Invest. Opththalmol. Vis. Sci.*, **2010**, *51*, 482-486.
22. Y. Han, M. E. Schneck, M. A. Jr Bearse, S. Barez, C. H. Jacobsen, N. P. Jewell and A. J. Adams, "Formulation and evaluation of a predictive model to identify the sites of future diabetic retinopathy", *Invest. Opththalmol. Vis. Sci.*, **2004**, *45*, 4106-4112.
23. A. Verrotti, L. Lobefalo, D. Trotta, G. D. Loggia, F. Chiarelli, C. Luigi, G. Morgese and P. E. Gallenga, "Visual evoked potentials in young persons with newly diagnosed diabetes: A long-term follow-up", *Dev. Med. Child Neurol.*, **2000**, *42*, 240-244.
24. V. Balakrishnan, M. R. Shakouri, H. Hoodeh and H. S. Loo, “Predictions using data mining and case-based reasoning: A case study for retinopathy”, *World Acad. Sci. Eng. Technol.*, **2012** ,*63,* 573-576.
25. W. H. Mong, W. Hsu, M. L. Lee, B. Liu and L. W. Tok, "Exploration mining in diabetic patients databases: Findings and conclusions", Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **2000**, Boston, USA, pp. 430-436.
26. H. Müller and J. C. Freytag, "Problems, methods, and challenges in comprehensive data cleansing", Technical Report HUB-IB-164, **2003**, Institut für Informatik, Humboldt University, Berlin, Germany.
27. P. N. Tan, M. Steinbach and V. Kumar, "Association analysis: Basic concepts and algorithms", in "Introduction to Data Mining", Addison Wesley, **2006,** Ch.6, www.users.cs.umn.edu/ ~kumar/dmbook/ch6.pdf (Accessed: August 2012).

28. M. J. Zaki, "Mining non-redundant association rules", *Data Mining Knowl. Disc.*, **2004**, *9*, 223-248
29. J. L. Kolodner, "An introduction to case-based reasoning", *Artif. Intell. Rev.*, **1992**, *6,* 3-34.