Indian Journal of Science and Technology, *Vol 7(10), 1498–1503, October 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Prediction of Chances - Diabetic Retinopathy using Data Mining Classification Techniques

K. R. Ananthapadmanaban* and G. Parthiban

SRM Arts and Science College, Kattankulathur, Tamilnadu, India; toananths@yahoo.com, trgparthi@gmail.com

Abstract

Diabetic retinopathy the most common diabetic eye disease, is caused by complications that occurs when blood vessels in the retina weakens or distracted. It results in loss of vision if early detection is not done. Several data mining technique serves different purposes depending on the modeling objective. The outcome of the various data mining classification techniques was compared using rapid miner tool. We have used Naive bayes and Support Vector Machine to predict the early detection of eye disease diabetic retinopathy and found that Naive bayes method to be 83.37% accurate. The performance was also measured by sensitivity and specificity. The above methodology has also shown that our data mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class which we are trying to predict.

Keywords: Data Mining, Diabetes, Naive Bayes Method, Retinopathy, Support Vector Machine

1. Introduction

The commonest cause of blindness among working class is Diabetic Retinopathy which often leads to the complete loss of vision[1]. The World Health Organization (WHO) has estimated that Diabetic Retinopathy is responsible for 4.8% of the 37 million cases of blindness throughout the world. Therefore a prediction technique is conceived so that early precautions or controls can be implemented. People with diabetes are susceptible to impairment of other vital organs such as heart, kidney and eyes[2]. At the initial stage of Diabetic Retinopathy, there will be some changes in the vision that can be noticed. But over time, Diabetic Retinopathy can get worsen and cause vision loss. Image analysis tools can be used for automated detection of these various features and stages of Diabetes Retinopathy and can be referred to the specialist accordingly for intervention. Thus such tools will be useful for effective screening of Diabetic Retinopathy patients[3]. Prevalence of high rate of retinopathy cases found worldwide is due to delay in diagnosis for retinopathy since it is asymptomatic[4]. Therefore, a prediction technique has been conceived so that early precautions or controls can be implemented.

Lanord Stanley et al.[5] devised a method to diagnose diabetics in the Indian community with the help of four simple questions viz. age, abdominal obesity, physical activity and family history along with one measurement for waist circumference.

'Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner'[6] analyze a Pima Indians diabetes data set containing information about patients with and without diabetes. This work focuses on data pre-processing, including attribute identification and selection, outlier removal, data normalization and numerical discretization, visual data analysis, hidden relationships discovery, and a diabetes prediction model construction.

IHDPS[7] prototype predicts the possibility of patients getting a heart disease from the Cleveland heart disease database using data mining techniques decision trees, naive Bayes and neural network with 9 medical attributes. The results show that the most effective model to predict patients with heart diseases is naive Bayes (86.12%) followed by neural network and decision trees. Furthermore, it can incorporate other data mining techniques such as time series, clustering and association rules.

**Author for correspondence*

'Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets'[8] system has been proposed to improve the diagnostic accuracy of diabetic disease by selecting informative features of Pima Indians Diabetes dataset. The hybrid prediction model proposed combines two different functionalities of data mining clustering and classification with F-score selection approach to identify the optimal feature subset of the Pima Indians Diabetes dataset. The proposed model was validated using four parameters, namely the accuracy of the classifier, area under ROC curve, sensitivity and specificity.

The two traditional classification methods (logistic regression and Fisher linear discriminant analysis) and four machine-learning classifiers (neural networks, support vector machines, fuzzy c-mean, and random forests) were compared[9] to classify persons with and without diabetes.

During the recent years there have been many studies on automatic diagnosis of diabetes, diabetic retinopathy, heart disease etc. In [10] a method has been proposed for automated detection and classification of vascular abnormalities using several techniques such as scale and orientation, selective Gabor filter banks. In [11] Kaplan-Meier method to generate univariate survival curves to identify patients who were at a higher risk for retinopathy, and results showed duration of diabetes, systolic blood pressure, glycosylated haemoglobin, albuminuria, gender and diabetes therapy were significantly associated with the occurrence of retinopathy.

Study[12] was made to evaluate the efficiency of three plant components viz, cinnamaldehyde, cinnamic acid and cinnamyl alcohol in inhibiting Aldose Reductase (AR), an enzyme associated with retinopathy of both type 1 and type 2 diabetic patients.

A product[13] made from whole leaf concentrate of Stevia, found to reduce hyper glycaemia in type 2 diabetic women.

In[14], it was suggested that increased awareness and treatment of diabetes should begin with prevention.

According to [15] data mining applications can be developed to evaluate the effectiveness of medical treatments.

2. Methods

Data mining technique was used to predict the chances of diabetic retinopathy. Under the data exploration mode, almost all attribute selection modules applicable for the data to collect optimal subset of attributes were explored. Rapid Miner was chosen as the data mining tool due to its learning operators and operator framework, which allows forming nearly arbitrary processes.

Though there is availability of Cleveland Clinic Foundation Heart Disease dataset, for the sake of determining the accuracy rate in Indian region, we have collected 300 clinical records from Dr. Seshaiah Diabetes Centre, Chennai, Tamil Nadu. The clinical data set specification provides concise, unambiguous definition for items related to diabetes.

Typically, cross-validation is used to generate a set of training, validation folds, and we compared the expected error on the validation folds after training on the training folds. Cross validation works were carried out by using part of the data to train the model, and the rest of the dataset to test the accuracy of the trained model. In this case, we have divided the dataset into 10 parts with training and testing data for each part. The proposed architecture is given in Figure 1. The attributes data view of each records are shown in Table 1.

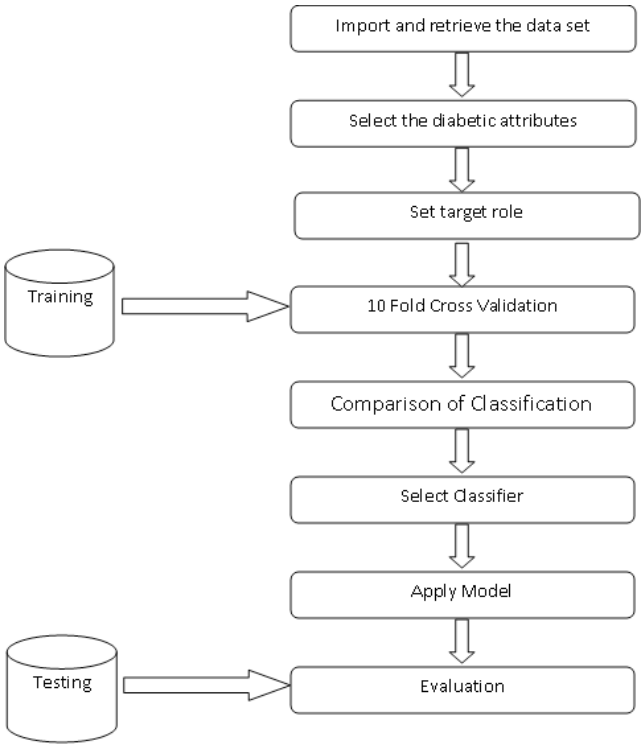

Figure 1. Proposed architecture.

Table 1. Diabetic Attributes used in our Experimentation

Attribute Role	Attribute Name	Attribute Type	Description
Regular	Sex	Binomial	Sex of the patient. Values: Male, Female
Regular	Age	Integer	Age of the patient
Regular	Family / Heredity	Polynomial	Indicates whether the patient's parents were affected by diabetes. Values: Father, Mother, Both
Regular	Weight	Numeric	Weight of the patient
Regular	BP	Polynomial	Blood pressure of the patient
Regular	Fasting	Integer	Fasting Blood Sugar
Regular	PP	Integer	Post prondial Blood Glucose
Regular	A1C	Numeric	Glycosylated Hemoglobin Test
Regular	LDL	Integer	Low Density Lipoprotein
Regular	VLDL	Integer	Very Low Density Lipoprotein
Label	Vulnerability	Binomial	Indicates the Vulnerability of the patients to Retinopathy. Values : High, Low

3. Data Mining Classification Techniques for Predicting Diseases

By using data mining technique method we have found a model which described and differentiated data classes, which in turn helped to predict accurately the class label which is unknown. We have also used regression method which helped to analyze the current and past states of the attributes and prediction of the future. Many researchers in the past used data mining techniques in diagnosis of various diseases.

3.1 Naives Bayes Method

Naives Bayes method is based on probabilities which are conditional and given the probability of another event that has already occurred, the probability of an event occurring is found using Bayes theorem[16]. If 'A' is referred as prior event and 'B' as dependent event, Bayes' theorem can be given as

$$\text{Prob (B given A)} = \text{Prob(A and B)/Prob(A)}$$

The Naive bayes performance screen and plots for each attributes are shown in Figure 2 to 5.

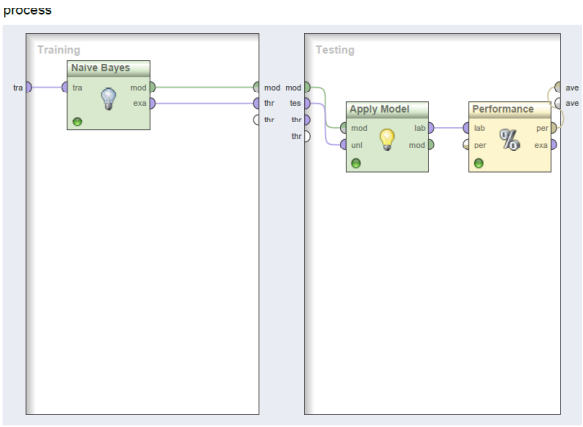

Figure 2. Naive Bayes performance screen

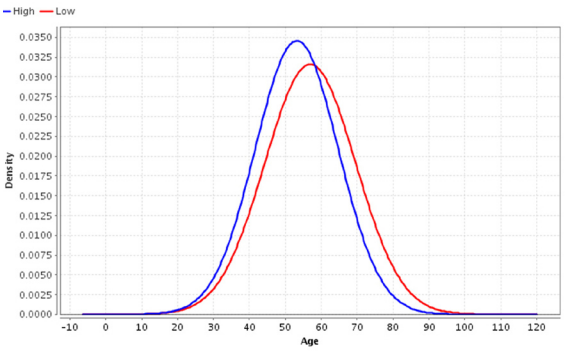

Figure 3. Bayes age attribute plot.

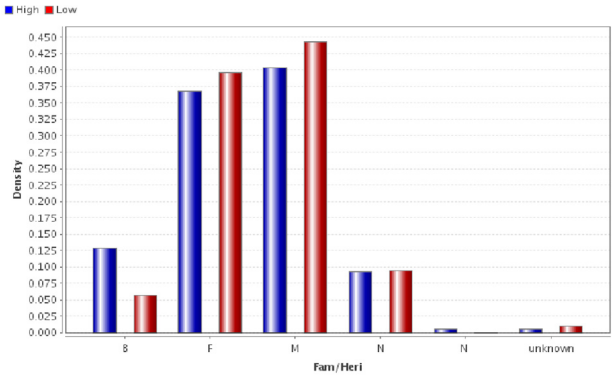

Figure 4. Bayes heredity attribute plot.

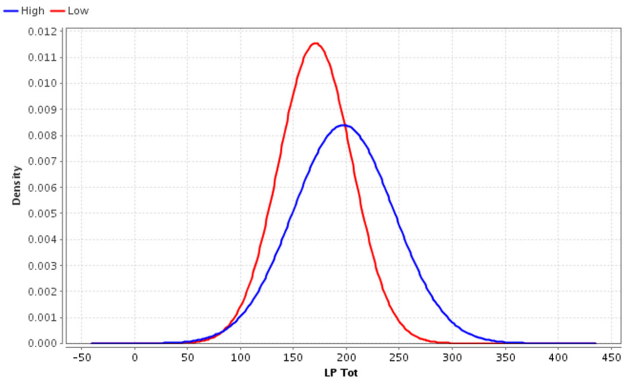

Figure 5. Bayes LP total cholesterol attribute Plot.

3.2. Support Vector Machine (SVM)

SVM was introduced by Corinna Cortes and Vladimir Vapnik[17] used for classification and regression analysis. It constructs a hyper plane or set of hyper planes in a high- or infinite dimensional space. SVMs can efficiently perform non-linear classification using kernel functions which implicitly map their inputs into high-dimensional feature spaces. It also helps to plot the training vectors.

SVM uses kernel functions to map the data set to a high dimensional data space for performing classification. The different types of kernel functions are as follows:

Linear:

$$K(x_i, x_j) = x_i^T x_j.$$

Polynomial:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)_d,\ \gamma > 0$$

Radial basis function:

$$K(x_i, x_j) = \exp\left(r\left\|x_i - x_j\right\|^2\right),\ \gamma > 0$$

$$\text{Sigmoid}: K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j + r\right)$$

The SVM performance screen is shown in Figure 6.

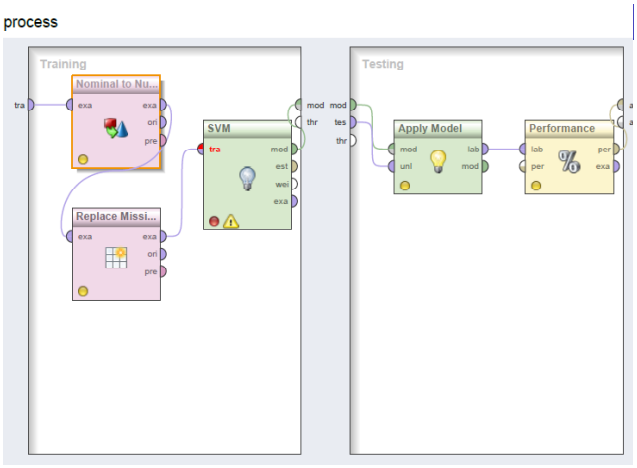

Figure 6. SVM performance screen.

3.3 Statistica Tool

The procedure for data analysis, data management, statistics, data mining, and data visualization procedures have been provided using Statistica tool developed by StatSoft[18]. The Statistica pie chart of age attribute screen is shown in Figure 7. It shows the high risk factor of getting retinopathy of diabetic patients at the age between 40 to 70 years.

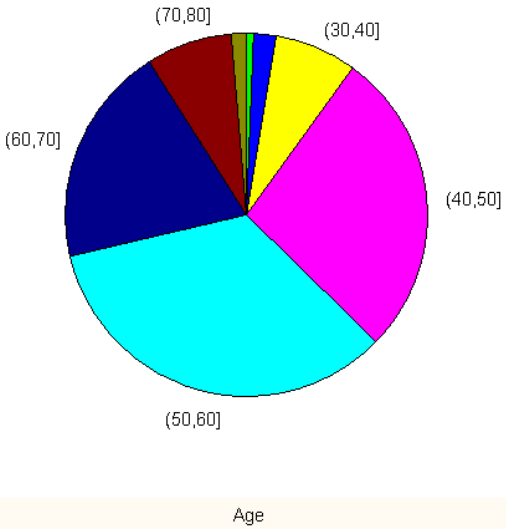

Figure 7. Pie chart of age using statistica tool

3.4. Rapid Miner Tool

Rapid miner tool is used in our experimentation. The application of Rapid miner software is being used for business and industries besides research, education, training, rapid prototyping. It helps in coordinated activities in machine learning, data mining, text mining, predictive analytics and business analytics. Thus it supports all stages of the data mining process to get valid and optimized results with clear visualization[19,20].

4. Result Analysis

The basic approach taken with this Rapid miner tool is to prepare a process model which uses 10-Cross validation along with the machine learning algorithm to increase the accuracy of the model. Further out of 10 subsets, which are classifier trained, one subset is tested. By this the whole training set is provided once in each instance and cross validation arrived. It shows the correct classification data in percentage. To distinguish between the actual class and the predicted class we use the labels High/Low for the class predictions produced by a model.

With the given classifier and instances, four possible outcomes can be arrived as given in the Table 2.

The performance of the proposed model is evaluated by accurately calculating the correctly predicted True Positive and True Negative classifications, arrived from out of proportion of instances.

The Sensitivity, Specificity and Accuracy values are calculated using the formulas.

Sensitivity = True Positive / (True Positive + False Negative).

Specificity = True Negative / (True Negative + False Positive).

Accuracy = (True Positive + True Negative) / True Positive + False Positive + True Negative + False Negative.

Table 2. Classifier table

Sl.No	Instance	Classification	Counted as
1.	Positive	Positive	True Positive
2.	Positive	Negative	False Negative
3.	Negative	Negative	True Negative
4.	Negative	Positive	False Positive

The confusion matrix indicating the accuracy of the Naive bayes classifier for the given data sets is shown in Table 3.

The proposed Naive bayes model was able to classify 83% of the input instances correctly. It exhibited a precision of 91% in average, recall of 82% in average. The results show clearly that the proposed method performs well compared to other similar methods in the literature.

The confusion matrix indicating the accuracy of the SVM classifier for the given data sets is shown in Table 4.

The cross validation tests prevent over fitting problem. Based on the exhaustive trials conducted, we found that for C = 5.0 and γ = 1.0 the classifier exhibited the best accuracy of 64.91%. From the results obtained, it can be seen that the classifier exhibits a very high classification accuracy i.e 64.91% overall. It also shows a very high precision for the positive class (64.90%) and also the recall of the positive class is quite good.

As in table 5, a total of 300 records with 16 medical attributes, having the results of two models, Naive bayes appears to be most effective as it has the highest percentage of correct predictions (83.37%) for patients with retinopathy, followed by SVM.

Table 3. Bayes distribution: accuracy: 83.37%

	true High	true Low	class precision
pred. High	162	16	91.01%
pred. Low	34	90	72.58%
class recall	82.65%	84.91%	

Table 4. SVM accuracy: 64.91%

	true High	true Low	class precision
pred. High	196	106	64.90%
pred. Low	0	0	0.00%
class recall	100.00%	0.00%	

Table 5. Accuracy of various classification techniques

Technique	Accuracy
Naive Bayes	83.37%
SVM	64.91%

5. Discussion and Conclusion

Naives Bayes is more efficient than SVM. Thus this work presents a successful Diabetic Retinopathy Diagnosing method which helps to predict the disease in early stage that can eventually reduce the manual work.

We started with a preprocessing operation to improve image quality by eliminating defects caused by lighting and acquisition processes. In the second step the optic disc has disrupted the automatic detection. In the third step, the segmentation of graph cuts is used in order to detect exudates regions. Finally, the neural network gave better results with a feature extraction of images by descriptors and Hu moment of GIST. The final results were compared quantitatively with a manual exudates segmentation produced by an expert in ophthalmology. Performances of our method were also measured by Specificity 95% and Sensitivity 96.65%. Our future work for this paper is to implement other algorithms like neural network and clustering with use of medical datasets in Weka tool.

The proposed approach has shown that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict. Moreover these data analysis results can be used for further research in enhancing the accuracy of the prediction system in future.

6. Acknowledgement

We are grateful to Dr. V.Shesiah, Chairman and Managing director of Dr. V.Shesiah Diabetes Centre, Chennai for providing an access to medical diabetic data and for his involvement in this domain.

7. References

1. Sopharak A, Uyyanonvara B, Barman S. Automatic microaneurysm detection from non-dilated diabetic retinopathy retinal images using mathematical morphology methods. IAENG, IJCS. 2011 Aug 24; 38(3):38_3_15
2. Ashwinkumar UM, Dr. Anandakumar KR. (2012), 'Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques', ICCDE, IACSIT Press, Singapore, 106–115.
3. Singh N, Tripathi RC. Automated early detection of diabetic retinopathy using image analysis techniques. Int J Comput Appl. 2010; 8(2):18–23.
4. Balakrishnan V, Kuppusamy U, Heng CKAP. An Intelligent Predictive System for Diabetic Retinopathy among Diabetes Patients in Malasia. 2012; 1–4.
5. Stanley JML, Elantamilan D, Mohanasundaram K, Kumaravel TS. Evaluation of Indian diabetic risk score for screening undiagnosed diabetes subjects in the community. Indian Journal of Science and Technology. 2012; 5(6):2798–9.
6. Han J, Rodriguze J, Beheshti M. Diabetes data analysis and prediction model discovery using rapid miner. 2nd International Conference on Future Generation Communication and Networking; 2008 Dec 13–15; Hainan island. IEEE; 2008. p. 96–9.
7. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. International Journal of Computer Science and Network Security. 2008; 8(8):343–50
8. Balakrishnan S, Narayanaswamy R. An empirical study on the performance of integrated hybrid prediction model on the medical datasets. Int J Comput Appl. 2011 Sep; 29(5):1–6.
9. Tapak L, Mahjub H, Hamidi O, Poorolajal J. Real-data comparison of data mining methods in prediction of diabetes in Iran. Healthc Inform Res. 2013; 19:177–85
10. Vallabha D, Dorairaj R, Namuduri K, Thompson H. Automated detection and classification of vascular abnormalities in diabetic retinopathy. Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers. 2004; 2:1625–9.
11. Cho HY, Lee DH, Chung SE, Kang SW. Diabetic retinopathy and peripapillary retinal Thickness. *Korean J Opthalmol*. 2010; 24(1):16–22
12. Sivakumari K, Rathinabai FMCA, Kaleena PK, Jayaprakash P, Srikanth R. Molecular docking study of bark-derived components of *Cinnamomum cassia* on aldose reductase. Indian Journal of Science and Technology. 2010; 3(8):1081–8.
13. Parimalavalli R, Radhaisri S. Glycaemic index of stevia product and its efficacy on blood glucose level in type 2 diabetes. Indian Journal of Science and Technology. 2011; 4(3):318–21.
14. Sa-ngasoongsong A, Chongwatpol J. An Analysis of Diabetes Risk Factors Using Data Mining Approach. 2012. Available from: e http://www.mwsug.org/proceedings/2012/PH/MWSUG-2012-PH10.pdf
15. Diwani S. Overview applications of data mining in health care: the case study of arusha region. IJCER. 2013 Aug 8; 3(8):73–7.
16. Naive Bayes Classifier, '*Bayes theorem*' <http://en.wikipedia.org/wiki/Naive bayes classifier>
17. Support Vector Machine, Machine Learning and Data Mining. Available from: http://en. wikipedia. org/wiki/Support_vector_machine
18. Statistica tool. Available from: <http://en.wikipedia.org/wiki/STATISTICA>
19. Rapid Miner, Machine learning software getting started. Available from: <http:// rapidminer.com/learning/getting-started/>
20. Rapid miner data mining tool. Available from: http://en.wikipedia.org/wiki/RapidMiner