INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND CREATIVE ENGINEERING (ISSN:2045-8711)
VOL.7 NO.03 MARCH 2017, IMPACT FACTOR: 1.04

Diabetic Disease Identifications Using Classification Technique in Orange Tool

Dr.R.Shanmugasundaram
Associate Professor, Department of Computer Science,
Erode Arts and Science College (Autonomous),
Erode, Tamil Nadu, India.
Email: rshanmuga.lec@gmail.com
Dr.S.Prasath
Assistant Professor, Department of Computer Science,
Nandha Arts and Science College
Erode, Tamil Nadu, India.
Email: softprasaths@gmail.com

Abstract-Data mining is a process of extracting information from a dataset and transforms it into understandable structure to discover patterns in large data sets. Data mining for healthcare is useful in evaluating the effectiveness of clinical treatments to its roots in databases records system getting to know and facts visualization. Diabetic ailment refers back to the heart disorder that develops in persons with diabetes. The term diabetes is a continual ailment that occurs both when the pancreas does now not produce sufficient insulin. The blood vessels despite the fact that many data mining type techniques exist for the prediction of heart disorder there is inadequate records for the prediction of heart illnesses in a diabetic character. A number of experiments had been conducted the use of orange tools for contrast of the performance of predictive facts mining techniques on the diabetic dataset with attributes. The SVM classifier method has been carried out in orange tool prediction model using minimal training set to diagnose vulnerability of diabetic sufferers. All the above experiments find the probabilities of risk in diabetic patients for coronary heart sickness.

Keywords— Data Mining, Diabetic, Heart, Orange, SVM.

1. INTRODUCTION

Data mining for healthcare is useful in evaluating the effectiveness of medical treatments and it is interdisciplinary field of study in databases statistics machine learning and data visualization. Diabetic disease refers to the heart disease that develops in persons with diabetes. The term diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin. The cardiovascular disease is class of diseases that involves the heart. Even though many data mining classification techniques exist for the prediction of heart disease there is insufficient data for the prediction of heart diseases in a diabetic individual. The main objective focus on this research is to find an optimal model and test the ability of classification algorithms with state of the art parties in global health care domain. A number of experiments have been conducted using weka and orange tools for comparison of the performance of predictive data mining techniques on the diabetic dataset with 1000 records using different attributes. In this work naive Bayes data mining classifier technique has been applied in weka and orange tools produces an optimal prediction to diagnose of diabetic patients.

2. RELATED WORKS

Anuja Kumari et al. [1] described the Support vector machine, a supervised machine learning method as the classifier for diagnosis of diabetes using Pima Indian diabetic database in Classification of Diabetes Disease Using Support Vector Machine.

Asha Gowda Karegowda et al. [2] describes diabetes can occur in anyone. However, people who have close relatives with the disease are somewhat more likely to develop it. Other risk factors include obesity, high cholesterol, high blood pressure and physical inactivity. The risk of developing diabetes also increases, as people grow older. People who are over 40 and overweight are more likely to develop diabetes, although the incidence of type-2 diabetes in adolescents is growing.

Jayshri Sonawane et al. [3] presented the heart is the organ that pumps blood, with its life giving oxygen and nutrients, to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer and if the heart stops working altogether, death occurs within minutes. The term heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels.

Jianchao Han et al. [4] analyzed 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.

Karthikeyani et al. [5] presented the classification of supervised data mining algorithms based on diabetes disease dataset in Comparative of Data mining classification algorithm in Diabetes disease Prediction.

Sarojini Balakrishnan et al. [6] proposed a system to improve the diagnostic accuracy of diabetic disease by selecting informative features of Pima Indians Diabetes dataset in Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets. They propose a hybrid prediction model that 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.

Selvakuberan et al. [7] presented the diabetes is one of the major causes of premature illness and death worldwide. In developing countries, less than half of people with diabetes are diagnosed. There is no time for diagnoses and adequate treatment, complications and morbidity from diabetes rise exponentially.

Vahid Rafe et al. [8] developed the medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently.

Vijayarani et al. [9] discussed the heart disease plays an important role in data mining due to occurrence of death in heart diseases. To reduce cost for achieving clinical tests an appropriate computer based information and decision support should be provided.

3. METHODOLOGY

A major problem of traditional strategy of encoding is the high dimensionality of the feature vector. The feature vector with a large number of key terms is not only unsuitable for neural networks but also easily to cause the over fitting problem.

Each algorithm requires submission of data in a specified format. The conversion of raw data into machine understandable format is called preprocessing. The data preparation phase covers all activities to construct the final dataset from the initial raw data. These raw data can be stored in several formats including text, excel or other database types of files. Then the raw data is changed into data sets with a few appropriate characteristics.

3.1 NAIVE BAYES APPROACH

Naive Bayes classifier as a term dealing with a simple probabilistic classifier based on application of Bayes theorem with strong independence assumptions. Since independent variables are assumed, only the variances of the variables for each class need to be determined. It can be used for both binary and multi class classification problems. Naive Bayes data mining classifier technique has been applied which produces an optimal prediction model using minimum training set to predict the chances of diabetic patient getting heart disease. The diagnosis of diseases plays vital role in medical field.

Algorithm

Step 1: Load the dataset and divides the data into training set.
Step 2: Generate random weights for each input data point.
Step 3: Determine the value of the bias term b and initialize the error for each point randomly.
Step 4: Initialize random values.
Step 5: Apply SVM classifier algorithm to train the data are identified.
Step 6: Calculate number of support vectors.
Step 7: Loop until stopping criteria is met, usually until reach maximum number of iterations.
Step 8: Identify the class label for new test data.
Step9: Performance is evaluated for SVM classifier using predicted class label for test data and expected class labels using confusion matrix.

4. EXPERIMENTATION AND RESULTS

The predictive facts mining techniques on the diabetic dataset with attributes are collected from the Pima diabetic database. The experimentation is carried out by Orange. Some of the sample data are experimented and is presented in the Fig.4.1 to Fig. 4.3.

4.1 ORANGE

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front- end for explorative data analysis and visualization and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation and exploration techniques. Its graphical user interface builds upon the cross-platform framework

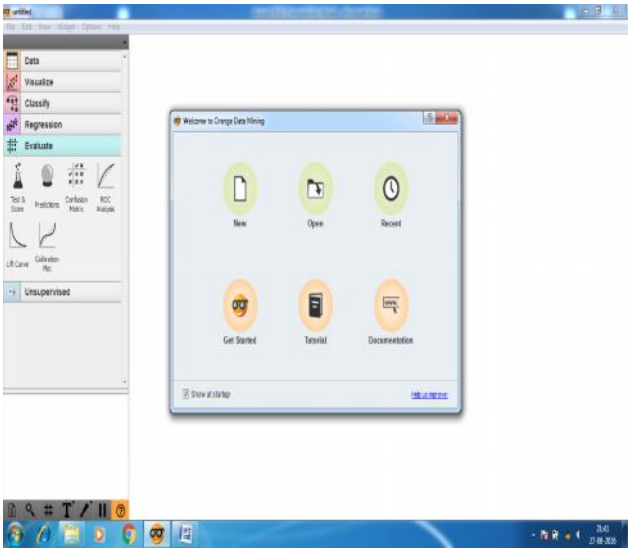

Fig 4.1 Explorer in ORANGE

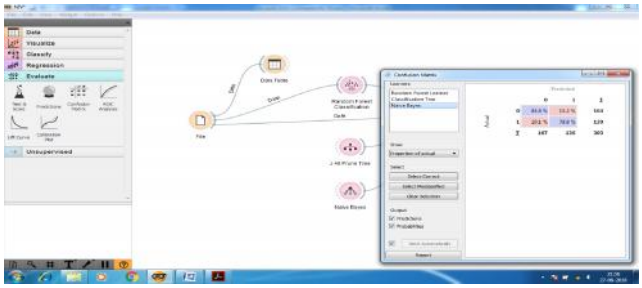

Fig 4.2 Naive Bayes Diabetic Dataset Classifier Predicted Output

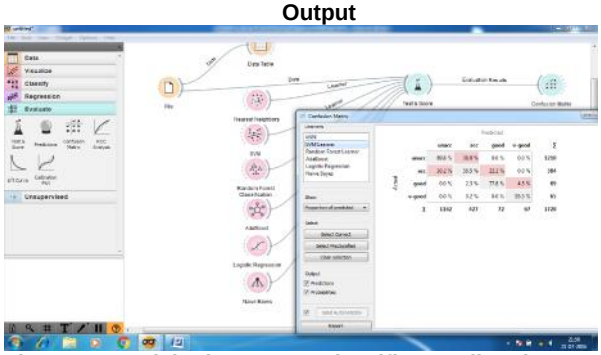

Fig 4.4 SVM Diabetic Dataset Classifier Predicted Output

5. PERFORMANCE EVALUATION

To measure the performance sensitivity, accuracy and specificity are used. TP is true positive, FP is false positive, TN is true negative and FN is false negative. TPR is true positive rate, which is equivalent to Recall.

$$\text{Sensitivity} = \frac{\text{True Positive Rate}}{(\text{True Positive} + \text{False Negative})} \quad \text{..... equ.1}$$

$$\text{Specificity} = \frac{\text{True Negative Rate}}{(\text{False Positive} + \text{True Negative})} \quad \text{..... equ.2}$$

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \quad \text{..... equ.3}$$

Table.5.1 Comparison Results

Methods / Parameters	Random Forest	Naive Bayes Classifier	SVM Classifier
Number of Instances	788	788	788
Accuracy	80.14%	82.74%	85.86%

From the above table 5.1 shows the performance of SVM classifier. The fig.5.1 shows comparison graphical representation of methods. The method can over perform the traditional method with classify recall rate.

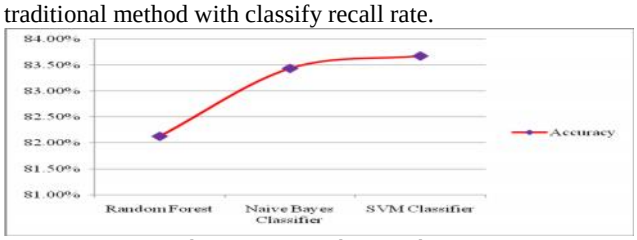

Fig 5.1 Comparison Values

6. CONCLUSION

Data mining for healthcare is useful in evaluating the effectiveness of medical treatments and ensures detection of fraud and abuse. The data mining techniques give the necessary standard in prediction. The performance in prediction depends on the various attributes which are helpful in predicting disease efficiently and patients receive better and more affordable healthcare services. The SVM data mining classifier technique has been applied which produces an optimal prediction model using minimum training set to predict the chances of diabetic patient. Orange tool is considered being a successful tool for classification purpose and evidence is the proposed system is quite good, since it has proved and shown good accuracy on the prediction of diabetic. To determine the most accurate technique to predict the risk in diabetic patients. The diabetic patients based on their predictive accuracy. In overall accuracy, in terms of precision and recall exhibited a very consistent performance.

REFERENCES

[1] Anuja Kumari, V and Chitra, R 2013, 'Classification of Diabetes Disease Using Support Vector Machine', International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797-1801.

[2] Asha Gowda Karegowda, Manjunath AS and Jayaram MA 2011, 'Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes', International Journal on Soft Computing, vol. 2, no. 2, pp.15-23.

[3] Jayshri Sonawane, S, Dharmaraj Patil, R & Vishal Thakare, S 2013, 'Survey on Decision Support System For Heart Disease, International Journal of Advancements in Technology, vol.4, no.1, pp. 89-96.

[4] Jianchao Han, Juan Rodriguze & Mohsen Beheshti 2008, 'Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner', In Proceedings of the 2nd International Conference on Future Generation Communication and Networking, vol.3, pp. 96-99.

[5] Karthikeyani, V & Parvin Begum 2012, 'Comparative of Data mining classification algorithm in Diabetes disease Prediction', International Journal of Computer Applications, vol. 60, no. 12, pp. 26-31.

[6] Sarojini Balakrishnan, Ramaraj Narayanaswamy & Ilango Paramasivam 2011,'An Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets', International Journal of Computer Applications, vol.29, no.5, pp. 1-6.

[7] Selvakuberan, K, Kayathiri, D, Harini, B & Indra Devi, M 2011, 'An Efficient Feature Selection Method for Classification in Health care Systems using Machine Learning Techniques', In Proceedings of the 3rd International Conference on Electronics Computer Technology, Kanyakumari, India vol. 4, pp. 223-226.

[8] Vahid Rafe & Roghayeh Hashemi Farhoud 2013, 'A Survey on Data Mining Approaches in Medicine', International Research Journal of Applied and Basic Sciences, vol.4, no.1, pp. 196-202.

[9] Vijayarani, S & Sudha, S 2013, 'An Effective Classification Rule Technique for Heart Disease Prediction',International Journal of Engineering Associates, vol. 1, no.4, pp. 81-85.