Volume 8, No. 5, May-June 2017

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Disease Prediction and Risk Analysis using Classification Algorithms

Mixymol V.K.
Assistant Professor,
Department of Computer Science,
Kanchi Mamunivar Centre for Post Graduate Studies,
Puducherry, India

Abstract: Data mining techniques helps to extract important information from a large set of data set which may be widely spread and also helps to find correlation between these data sets. A huge amount of medical data is available at present which are widely spread. Currently a large number of tests are done to reach to a conclusion and this large number of test results may lead to lots of confusion and may lead to complication of the diagnosis process. The major issue with medical diagnosis is to reach to a correct diagnosis to take the correct decision. In this context, machine learning with the help of artificial intelligent algorithms can make use of decision making for rule generation process in healthcare data. Objective of this work is to analyze the performances of disease prediction using various classification techniques available on WEKA data mining tool. Data set for the present work is collected from UCI repository. Data set is collected for Dermatology, Hepatitis and Heart disease. Three classification techniques, Naïve Bayesian, Nearest Neighbor and Reptree, are considered for the performance analysis since it is found that these three algorithms are commonly used in disease prediction in many studies. With the future development of information and communication technologies, data mining will achieve its full potential in the discovery of knowledge hidden in the medical data.

Keywords: Data mining, Machine learning, Medical data, Data mining tool, Disease prediction, Performance Analysis.

1. INTRODUCTION

Data mining is the process of uncovering useful patterns or knowledgeable data that are hidden in a large volume of data. It involves uncovering patterns from large data set; and these uncovered patterns can be used to build new models. In healthcare, data mining has proven effective in areas such as predictive medicine[1], customer relationship management, detection of fraud and abuse[2], management of healthcare[3] and measuring the effectiveness of certain treatments[4].

A huge amount of medical data is available at present which are widely spread. The major issue with medical diagnosis is to reach to a correct diagnosis to take the correct decision [5]. Currently a large number of tests are done to reach to a conclusion and this large number of test results may lead to lots of confusion and may lead to complication of the diagnosis process. In this context, machine learning with the help of artificial intelligent algorithms can make use of decision making for rule generation process in healthcare data [6]. Data mining techniques helps to extract important information from a large set of data set which may be widely spread and also helps to find correlation between these data sets.

Objective of this work is to analyze the performance of different classification methods using data mining tool WEKA, for disease prediction. Three types of diseases are considered to analyze the performance.

2. DATASET DESCRIPTION

Data for this research was collected from UCI repository on health surveys. Three sets of dataset were collected for the study; namely Dermatology, Hepatitis and Heart disease.

Data Preparation

Data set for analyzing dermatology consists of 366 instances and contains 34 attributes. Some of the attributes are , erythema, itching, koebner phenomenon, polygonal papules, follicular papules, oral mucosal involvement, knee and elbow involvement, scalp involvement, family history, (0 or 1), exocytosis, perifollicular parakeratosis. Data set for analyzing Heart disease consists of 258 instances and contains 13 attributes and the attributes include Age , Sex , Chest pain types , Resting blood pressure , Serum cholesterol, Fasting blood sugar. Dataset for Hepatitis consists of 155 instances and 20 attributes which includes Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver.

Data Analysis

The present work performs analysis of the dataset using three algorithms available in WEKA data mining tool, Naïve Bayesian, Nearest Neighbor and Reptree and it predicts Precision, Recall, and Accuracy. Ten fold Cross validation is used in the implementation. In tenfold cross validation, the original sample is randomly partitioned into ten subsamples. Of the ten samples, a single sample is retained as the validation data for testing the model, and the remaining nine sub samples are used as training data. The cross validation process is then repeated ten times, with each of the ten samples used exactly once as the validation data. The ten results from the folds then can be averaged to produce a single estimation. Classification techniques are applied to the training set and classification models are applied on these training sets to find the pattern.

WEKA Data Mining Tool

WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this

functionality. WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of WEKs techniques are predicated on the assumption that the data is available as a single flat file or relation.

3. METHODS/ CLASSIFICATION ALGORITHMS USED

1. Naive Bayesian Algorithm

Naïve Bayesian is the WEKA implementation of the C4.5 algorithm. In this algorithm, the dataset is partitioned recursively to generate decision trees. Depth-first strategy is used to develop the tree. This process is repeated for each new node until a leaf node has been reached. It creates a binary tree to classify the dataset. This decision tree algorithm is used for comparison, it made on rules, accuracy, sensitivity and specificity using true positive and false positive in confusion matrix generated by the respective algorithm. This classifier uses the correct and incorrect instances that give a most efficient method for classification by using the confusion matrix. This method is applied in all the instances of the dataset and then classification of the dataset can be done.

2. Reptree Algorithm

Decision trees can also built using Reptree algorithm. At runtime, this decision tree is used to classify new unseen test cases by working down the decision tree using the values of this test case to arrive at a terminal node that tells you what class this test case belongs to. In this method greedy approach by selecting the best attribute is used to split the dataset on each iteration. One improvement that can be made on the algorithm is to use backtracking during the search for the optimal decision tree.

Reptree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That wills be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree. Reptree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once.

3. Nearest Neighbor Algorithm

An RBF (Radial Basis Function) network is a type of ANN (Artificial Neutral Network), which is simpler network structure with better approximation capabilities. It is an artificial neural network that uses the radial basis function as the artificial network.

4. RESULT ANALYSIS

The following are the summary of the findings using Naive Bayesian classification (NBTree).

Table 1 shows the result analysis for the prediction of dermatology. It can be observed that using Nearest Neighbor algorithm, 96.17 % of the data set are correctly classified and 3.82% of the data set is wrongly classified. In NBTree algorithm, 95.62 % of the data set are correctly classified and 4.37% of the data set is wrongly classified. Using Reptree algorithm, only 91.53 % the data set are correctly classified and 8.46% of the data set is wrongly classified. Among the three algorithms Nearest neighbor gives the highest True Positive rate (96.2%) even though NBTree algorithm gives the minimum value root mean square error. Figure 1 shows the graphical representation of the performance analysis of these three algorithms.

Table 1 : Performance analysis for Dermatology

Dermatology	Correctly classified	Incorrectly classified	TP-Rate	Root Mean Square Error
	(in percentage)			
Nearest Neighbor	96.17	3.82	96.2	11.29
NBTree	95.62	4.37	95.6	10.02
Reptree	91.53	8.46	91.5	15.78

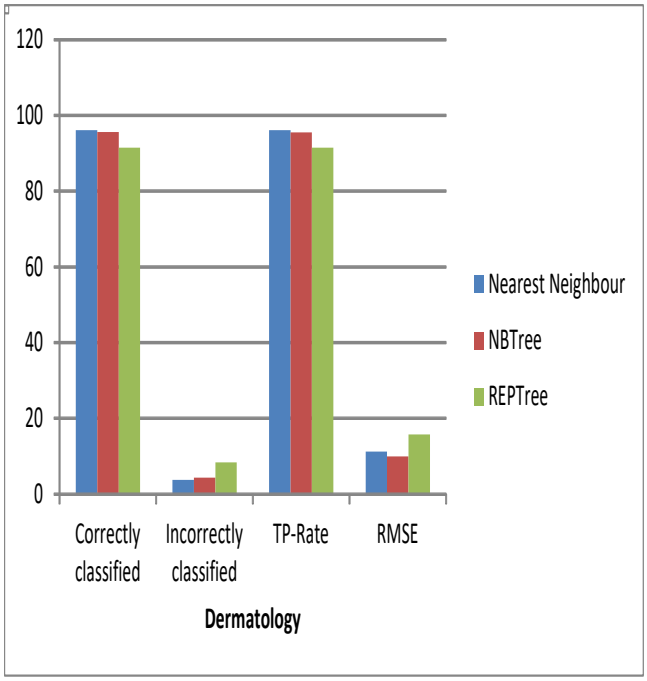

Figure 1: Graphical representation of performance analysis for Dermatology

Table 2 shows the result analysis for the prediction of hepatitis. It can be observed that using Nearest Neighbor algorithm, 84.51 % of the data set are correctly classified and 15.49% of the data set is wrongly classified. In NBTree algorithm, 82.58 % of the data set are correctly classified and 17.42% of the data set is wrongly classified. Using Reptree algorithm only 78.71% the data set are correctly classified and 21.29% of the data set is wrongly classified. Among the three algorithms Nearest neighbor gives the highest True Positive rate (84.51%) even though NBTree algorithm gives the minimum value root mean square error. Figure 2 shows the graphical representation of the performance analysis of these three algorithms.

Table 2: Performance analysis for Hepatitis

Hepatitis	Correctly classified	Incorrectly classified	TP-Rate	Root Mean Square Error
	(in percentage)			
Nearest Neighbor	84.51	15.49	84.5	39.35
NBTree	82.58	17.42	82.6	37.7
Reptree	78.71	21.29	78.7	40.67

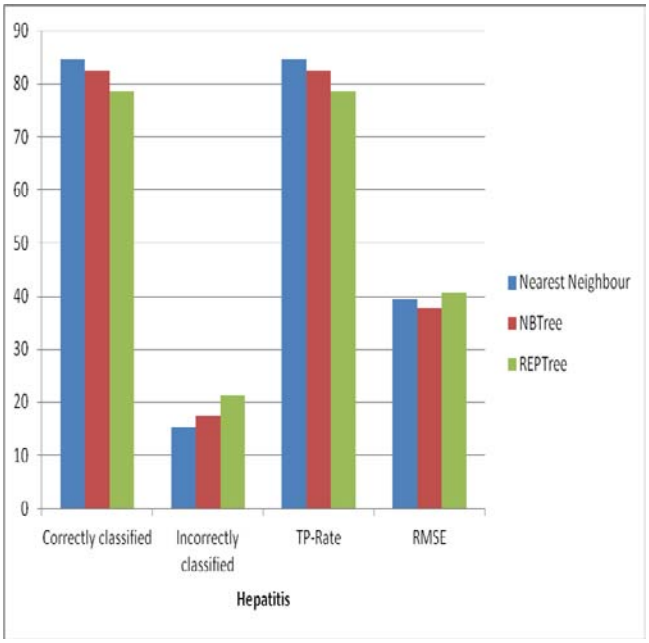

Figure 2: Graphical representation of performance analysis for Hepatitis

Table 3 shows the result analysis for the prediction of heart disease. It can be observed that using Nearest Neighbor algorithm, 78.14 % of the data set are correctly classified and 21.86% of the data set is wrongly classified. In NBTree algorithm, 80.37 % of the data set are correctly classified and 19.63% of the data set is wrongly classified. Using Reptree algorithm only 76.67% the data set are correctly classified and 23.33% of the data set is wrongly classified. Among the three algorithms NBTree gives the highest True Positive rate (80.37%). Also NBTree algorithm gives the minimum value root mean square error. Figure 3 shows the graphical representation of the performance analysis of these three algorithms.

Table 3: Performance analysis for Heart Disease

Heart Disease	Correctly classified	Incorrectly classified	TP-Rate	RootMean Square Error
Nearest Neighbor	78.14	21.86	78.1	46.75
NBTree	80.37	19.63	80.4	39.6
Reptree	76.67	23.33	76.7	42.63

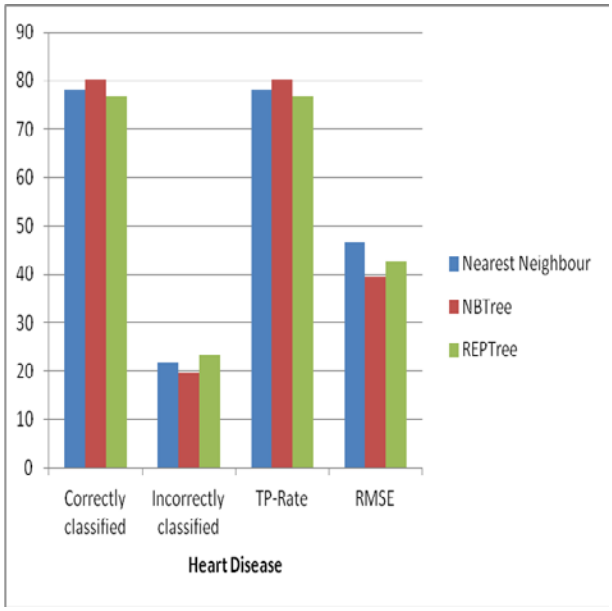

Figure 3 : Graphical representation of performance analysis for Heart Disease

5. CONCLUSION

The major issue with medical diagnosis is to reach to a correct diagnosis to take the correct decision from a widely spread huge amount of data. The present study focused to analyze the performance of disease prediction and risk analysis using various classification tools available in WEKA data mining tool. Performance of the systems had been measured using precision, recall, and Accuracy. It is found that the commonly used classification methods are Nearest Neighbor, Reptree and Naïve Bayesian algorithms. It can be observed for the results that Nearest Neighbor outperforms for dermatology and hepatitis while Naive Bayesian (NBTree) algorithm outperforms for heart disease compared to Reptree.

REFERENCES

[1] Neesha Jothi, Nur Aini Abdul Rashid, Wahidah Husain, “Data Mining in Healthcare – A Review”, Procedia Computer Science, 306 – 313, 2015

[2] L. Jiang, H. Zhang, and Z. Cai. Dynamic K-Nearest-Neighbor Naive Bayes with attribute weighted. In proceedings of the 3rd International Conference on Fuzzy Systems and Knowledge Discovery, 365–368. Springer, 2006.

[3] Neesha Jothi, Nur Aini Abdul Rashid, Wahidah Husain, “Data Mining in Healthcare – A Review”, The Third Information System and International Conference, Procedia Computer Science72(2015),pp 306 – 313.

[4] S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita “Prediction of Heart Disease using Data Mining Techniques”, Indian Journal of Science and Technology, Vol 9(39), October 2016,

[5] F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.),” Comparison of Different Classification Techniques Using WEKA for Breast Cancer” IFMBE Proceedings 15(2007), pp. 520-523.

[6] Krishnaiah V., Narsimha G., Chandra N.S.,”Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach” , Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India (CSI), (2015), Volume 1. Advances in Intelligent Systems and Computing, 371-384.