International Journal of Future Innovative Science and Engineering Research (IJFISER)
Volume - 2, Issue - II ISSN (Online): 2454- 1966

Research Manuscript Title

Performance Evaluation of Classification Techniques in Diagnosing the Risk Factors of CVD

[1]R. Rajalakshmi, [2]Dr.R.Latha

Research Scholar, Professor & Head,
Department of computer Science,
St. Peter's University, Avadi
Chennai, Tamil Nadu
E-Mail: elangoraji.79@gmail.com

JUNE – 2016

www.istpublications.com

Performance Evaluation of Classification Techniques in Diagnosing the Risk Factors of CVD

[1]R. Rajalakshmi, [2]Dr.R.Latha

Research Scholar, Professor & Head,
Department of computer Science,
St. Peter's University, Avadi
Chennai, Tamil Nadu
E-Mail: elangoraji.79@gmail.com

ABSTRACT

Data mining is a systematic process which conducts research in multidimensional views in order to provide the clear cut prediction measures, classification and decision making ideas for experts in the various disciplines. In the current era, data mining plays a vital role in the part of medical diagnosis. Various tools and software are introduced to reveal the facts and produce the reports in order to diagnose the disease. Now-a-days heart related diseases has become one of the major issue which increases percentage of the death rate in the present generation. This paper deals with the intension of evaluating the performance of various classification algorithms to measure the risk factors which leads to heart related diseases. Further it is determined that one of the classification model named multilayer perceptron model out performs than that of the other classification algorithms. The future work recommended in this paper is to introduce various data mining tools in order to assist the medical practitioners and hence that would be helpful to easily diagnose the risk factor of the diseases.

Keywords: Classification, Multilayer perceptron, Simple Naïve bayes, J48, Decision table.

I.INTRODUCTION

In this decade the food habits of our growing young generations have made ridiculous challenges in maintaining the good health in their livelihood. The intake of unlimited meat, chicken, junk and fast food items gradually leads to the unhealthy conditions where there is increase in the bad cholesterol and decrease in the good cholesterol. Low density lipoprotein(LDL) is meant as bad cholesterol and High density lipoprotein(HDL) is referred to as good cholesterol. Lipoprotein is a biochemical assembly that contains both proteins and lipids, bound to the proteins, which allow the fats to move through water inside and outside the cells. Lipids are group of naturally occurring molecules that include fats, waxes, sterols, fat-soluble vitamins such vitamin A, D, E and K, monoglycerides, diglycerides, triglycerides and others. The nature of LDL is to deposit the cholesterol in the blood streams and the artery walls but HDL is contradictory to LDL, it removes cholesterol from the blood streams and the artery walls. When LDL level is increased and HDL level in decreased in the blood stream, it gets deposited in the artery walls, when its thickness increases it blocks the systolic and diastolic system of the heart. At critical stage it results heart attack. The American Heart Association[1] recommends all adults of age 20 or older should have their cholesterol, and other traditional risk factors, checked every six to twelve months.

Being a well wisher to have a healthy society in the future generations the author tries to analyse the risk factors of the heart related diseases using various measuring parameters. Although there are numerous medical diagnosis procedures it has become a much more complicated job in evaluating the risk factors. Because the quantity of data has been increased, it has become a critical task to maintain, store and evaluate the data in the medical diagnosis. The risk factors may hold numerous measures in the dataset and considering all the measures increases the time of medical practitioners for decision making process. The main aim of this work is to pave the way to maintain the huge datasets with the appropriate measures and choose a right approach which would provide faster and accurate suggestions depending on the accuracy rate of the measures. In order to determine the major risk factors and based on the evaluation, the root cause of the seriousness of the diseases can be evaluated using various data mining algorithms. The outcome of those techniques may reveal different facts and reports. The accuracies of the various algorithms applied can be further estimated by using various validation approaches.

II.LITERATURE REVIEW

Evolving computing algorithms are introduced and discussed in multidimensional views by various experts of their own interest. Now-a-days medical diagnosis and their reports are prepared, maintained and evaluated only with the assistance of computing techniques. Being assisted by the computing techniques it has become an ease task to provide the detailed reports for all types of diseases. Data mining algorithms play a vital role in predicting and comparing the risk factors of various diseases.

Deepali Chandna[5] has discussed how information gain method, feature selection technique, can be used in collaboration with adaptive neuro fuzzy inference systems (ANFIS) in diagnosing new patient cases.

R. Chitra and V. Seenivasagam [6] examined that introduction of the Hybrid Intelligent Algorithm improves the accuracy of the heart disease prediction system.

Hlaudi Daniel Masethe et al.[7] applied data mining algorithms like J48, naïve bayes, REPTREE, CART, and bayes net to predict the heart attacks. The result of the research concluded that there is no much difference in the accuracy of above used algorithms in predicting the occurrence of heart diseases.

Jyothi Soni et al.[8] has applied predictive data mining techniques like KNN, Neural Networks, Bayesian classification, Classification based on clustering, and Decision Tree for classification and evaluation of dataset related to heart diseases. The performance evaluation of the above techniques concluded that decision tree and Bayesian classification outperforms than the others.

Jyothi Rohilla et al.[9] has discussed the advantages of applying data mining techniques to maintain huge volumes of data. It is also discussed that the proportion of deaths caused by heart diseases is greater than other diseases. Performance analysis of various classification algorithms are also conducted and it is concluded that ID3 and decision tree outperforms than other techniques.

Mai shouman et al.[10] has evaluated the performance of a number of decision trees with the bench mark datasets and proposed that J4.8 decision tree and bagging algorithm outperforms among all in the diagnosis of heart diseases.

Parvathi et al. [11] discusses how to enable the diseases diagnosis and discover the health care patterns from related database and also to discover the relationship between the health condition using the data mining techniques.

Rajalakshmi et al.[12] applied the k-means algorithm with the combination of various classification approaches like naïve bayes, decision trees and others for diagnosing the heart diseases. The comparative analysis of those applied algorithm concluded that the usage of data mining algorithms helps to reduce the human task and it is cost effective.

Dr. K. Usha Rani[13] has applied the neural network approach for classification of medical datasets using single and multilayer neural network nodes. The work concluded that neural networks provides satisfactory results for classification.

B.Venkatalakshmi et al.[14] applied decision tree and naïve bayes to predict the heart diseases and finally concluded that the accuracy of naïve bayes outperforms when compared to decision tree.

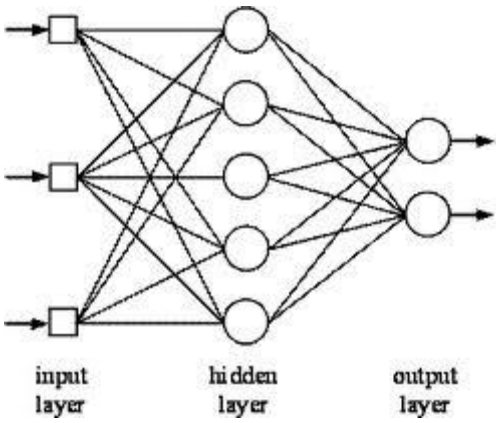
Vikas Chaurasia et al. [15] applied three classifiers naïve bayes, J48, decision tree and bagging algorithm for diagnosing heart diseases.

III. RESEARCH METHODS

At the outset of the literature review conducted in the previous section, it is evidentially understood that there is a pure intervention of computing algorithms in predicting and evaluating the risk factors of death causing diseases in the medical scenario. Enormous tools and approaches are left out for the diagnosis of the root cause of disease indexes. The authors concentration is to diagnose the level of cholesterol components like LDL, HDL, triglycerides which are the risk factors of obesity, diabetes and further leads to CVD(Cardio Vascular Diseases). The classification algorithms used to diagnose the risk factors are Multilayer Perceptron, Simple Naïve Bayes Classifier, J48 Decision Tree classifier, Decision Table classifier.

A.Multilayer Perceptron

Multilayer perceptron model plays a vital role in prediction and classification problems. Here the network is provided with desired input and outputs. Hence, it is easier to model the input-output relationship. Here, the input layer carries out the linear transform function and does not perform a weighted sum or threshold. The hidden layer are forced to perform the sigmoidal function and further alter the learning rules. The input-output mapping process is conducted by the perceptrons in the hidden layer. The output layer carries out the sigmoidal function and produces the network output. Consequently the error and difference between the network output and the desired output is calculated and the weight values are calculated and updated for the proceeding iteration. This routines are carried out until the expected output is received.



B.Simple Naïve Bayes

Naïve Bayes Classifier classifies the dataset based on conditional probability of the occurrence of attribute values in a relation. When a sample data is taken for classification, the prior probabilities of the present class is determined. Based on the prior probability the probability of the classification of the test data is evaluated and its appropriate class is allocated. This approach can be applied when the number of prediction is very large.

C.J48 Decision Tree

J48 is a software version of C4.5 algorithm. It is a classification algorithm which builds the decision tree. The initial process of decision tree is to generate a model that predicts the value of class attribute. The rule set is framed from the model. Decision tree classifies the objects based on rule set framed by the training set. Every node in the decision tree is the test over the attribute values and branches are the outcome of the test. It further uses pruning technique to prune the decision tree. Pruning is an approach which removes the over fitting nodes and branches. Decision trees are commonly used in decision analysis.

D.Decision table classifier

Decision table is a predictive modeling tool for classification. A decision table is a hierarchical classification of data which contains two attributes at each level of hierarchy. The inducers of the decision table identifies the precise attribute to classify the data. The attributes or columns are classified based on its domain values as the conditional attributes and decision attribute. Based on its attribute classification the hierarchical structure is generated. It has four quadrants condition, condition alternatives, actions, action entries.

Condition	Condition alternative
Action	Action entries

Action is a procedure or operation to perform and entries specifies that whether the action is to be performed for the condition alternatives.

IV.DATA SET DESCRIPTION AND PREPROCESSING

The dataset used in this work is the Pima-Indian data set. This dataset contains 10 attributes (which have been extracted from a larger set of 76). For about 768 instances are taken for the evaluation. Among the 768 instances 60% of the instances are allocated for the trainset and the remaining are equally shared for the test dataset and for cross validation.The type of cross-validation applied over the data set is 10-fold cross validation.

Attribute Information:

-- 1. age

-- 2. sex

-- 3. serum cholestoral in mg/dl
-- 4. LDL in mg/dl
-- 5. HDL in mg/dl
-- 6. Triglycerides in mg/dl
-- 7. resting blood pressure
-- 8. Obesity
-- 9. Family history
-- 10. normclass - value a / b

Attributes types

Real, String, Nominal

Variable to be predicted

a = tested_negative

b = tested_positive

V.METHODOLOGY AND EXPERIMENTAL RESULTS

The dataset extracted for the evaluation numbers for about 154.

A.Classifiers performance

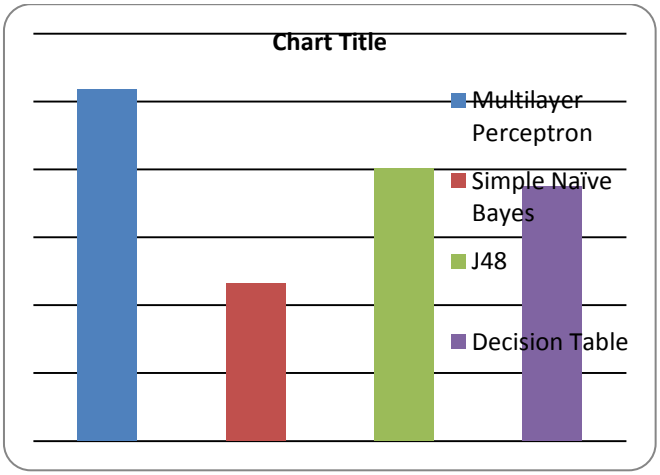
As mentioned in the previous section the performance evaluation of the classifiers is represented using the parameter accuracy in the Table I

Table I

Classifiers	Accuracy	Correctly classified	Incorrectly classified	Total instances
Multilayer Perceptron	91%	140	14	154
Simple Naïve Bayes	77%	118	36	154
J48	85%	131	23	154
Decision Table	84%	129	25	154

The following graph explains the efficiency of the above said algorithm in predicting the risk factors of heart diseases.

Figure I – Graphical representation of accuracy of classifiers



B.Comparative Analysis

The table II reveals the results of the evaluation for the data set

Table II

Classifiers	Evaluation on data set					
	TT (Sec)	**KS**	**MAE**	**RMSE**	**RAE**	**RRSE**
Multilayer	0.64	0.8021	0.1456	0.2686	31.9406%	56.2892%
Simple Naïve	0.01	0.4686	0.2692	0.4059	59.033%	85.055%
J48	0.01	0.662	0.2535	0.356	55.5917%	74.606%
Decision Table	0.03	0.6293	0.2806	0.3609	61.5458%	75.6273%

TT – Time Taken

KS – Kappa Statistics

MAE – Mean Absolute Error

RMSE – Root Mean Square Error

RAE – Relative Absolute Error

RRSE – Root Relative Squared Error

Based on the comparative analysis conducted in the Table II, it clearly lists out the error rate at various statistics.

The Table III illustrates the detailed accuracy by class. Various parameters like Total positive, False positive, Precision, Recall, F-Measure, ROC Area are estimated for the various types of classifiers. The accuracy rate of various classifiers are comparatively not related based on different views of parameters considered in this evaluation. Every classifier shows significant difference in their evaluation rate and among those classifiers every one vary in their performance level for selected parameter and the outperforming classifier is evaluated based on their performance rate.

Table III

Classifiers	Detailed accuracy by class						
	class	**TP Rate**	**FP Rate**	**Pre**	**Re**	**F-M**	**R-A**
Multilayer Perceptron	a	0.92	0.11	0.93	0.92	0.92	0.94
	b	0.88	0.08	0.85	0.88	0.87	0.94
Simple Naïve Bayes	a	0.86	0.40	0.79	0.86	0.82	0.83
	b	0.59	0.14	0.69	0.59	0.64	0.83
J48	a	0.92	0.27	0.86	0.92	0.88	0.82
	b	0.72	0.08	0.83	0.72	0.77	0.82
Decision Table	a	0.92	0.31	0.84	0.92	0.88	0.85
	b	0.68	0.08	0.82	0.68	0.74	0.85

a = tested_negative

b = tested_positive

Pre – Precision

Re – Recall

F-M – F-Measure

R-A -ROC Area

VI.CONCLUSION AND FUTURE ENHANCEMENT

In this paper, the performance analysis of various data mining classifiers have been conducted and while comparing the efficiency of accuracy of the discussed algorithms the author concludes that the multilayer perceptron model outperforms than that of the other algorithms in diagnosing the risk factors of the heart diseases. Further this evaluation can be continued by using various combinations of data mining tools and methodologies which would be helpful for the medical practitioners for easily diagnosing the major risk factors of forth coming diseases.

REFERENCES

[1] Han, J., Kamber, M.: ―Data Mining Concepts and Techniques‖, Morgan Kaufmann Publishers, 2006

[2] Jiawei Han and Micheline kamber,‖Data Mining Concepts and Techniques‖, Second Edition, Elsevier Inc, San Francisco, 2006.

[3] www.watchlearnlive.heart.org

[4] www.cholesterollevels.net

[5] Chandna Deepali, "Diagnosis of Heart Disease Using Data Mining Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 - pp. 1678-1680, Issue 2, 2014 - ISSN:0975-9646.

[6] R. Chitra and V. Seenivasagam, " Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques", ICTACT Journal on Soft Computing, Vol. 03, Issue 4, July 2013, ISSN: 2229-6956.

[7] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction Of Heart Disease Using Classification Algorithms", Proceedings of the World Congree on Engineering and Computer Science 2014, vol. II, San Francisco, USA.

[8] Jyothi Soni,Ujma Ansari, Dipesh Sharma, Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, Vol. 17- No.8, March 2011.

[9] Jyothi Rohilla, Preeti Gulia, "Analysis of Data Mining Techniques for Diagnosing Heart Disease", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 7, July 2015.

[10] Mai Shouman, Tim Turner, Rob Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients", Proceedings of the 9th Australasian Data Mining Conference, Ballarat, Australia.

[11] Parvathi I and Siddharth Rautaray, " Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5. - pp. 838-846, Issue 1, 2014, ISSN:0975-9646.

[12] K.Rajalakshmi, Dr.S.S. Dhenakaran, N. Roobin, "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research, Vol. 4, Issue 7, July 2015.

[13] Dr. K. Usha Rani, "Analysis Of Heart Diseases Dataset Using Neural Network Approach", International Journal Of Data Mining And Knowledge Management Process, Vol. 1, No.5, September 2011.

[14] B.Venkatalakshmi, M.V. Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining", 2014 International Conference on Innovations in Engineering and Technology, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Special Issue 3, March 2014.

[15] Vikas Chaurasia and Saurabh Pal, "Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2. - pp. 56-66., Issue 4, 2013, ISSN: 2296-1739.