International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 3 Issue 11 November 2016

A Decision Tree Based Classification Technique for Accurate Heart Disease Classification & Prediction

Apurva Joshi[1], Er. Jitendra Dangra[2] and Dr. M. K. Rawat[3]
Master of Engineering (C.S.E.), Lakshmi Narain College of Technology, Indore Madhya Pradesh, India[1]
Asst. Prof, Lakshmi Narain College of Technology, Indore Madhya Pradesh, India[2]
Prof, Lakshmi Narain College of Technology, Indore Madhya Pradesh, India[3]
apurvajoshi2012@gmail.com[1], jitendra.it@lnctindore.com[2], drmkrawat@lnctindore.com[3]

Abstract: *The data mining and their different applications have became more popular now a days. A number of large and small scale applications are now being developed with the help of data mining techniques such as predictors, weather forecasting systems, and business intelligence etc. There are two kinds of model available for these techniques namely supervised and unsupervised learning. The accuracy and performance of the supervised data mining techniques are higher as compared to unsupervised data mining techniques therefore in sensitive applications the supervised data mining techniques are used for prediction and classification. In this presented work the supervised learning based application is proposed and demonstrated. The proposed work is intended to demonstrate the data mining technique in disease prediction systems in medical domain. In order to perform this task the heart disease based data is selected for analysis and prediction.*

Keywords: *Data Mining, Genetic algorithm, Neural Network, Decision tree, Heart disease prediction, Supervised learning.*

1. INTRODUCTION

The data mining [1] is a process of analysis of the data and extraction of the essential patterns from the data. These patterns are used with the different applications for making decision making and prediction related task. The decision making and prediction is performed on the basis of the learning of algorithms. The data mining algorithms supports both kinds of learning supervised and unsupervised. In unsupervised learning only the data is used for performing the learning and in supervised technique the data and the class labels both are required to perform the accurate training. In supervised learning the accuracy is maintained by creating the feedbacks form the class labels and enhance the classification performance by reducing the error factors from the learning model.

The proposed work is intended to investigate these techniques in the application of the predictions. Data mining techniques have been widely used in clinical decision support systems for prediction and diagnosis of many diseases with good accuracy. These techniques have been very effective in designing clinical support systems due to their ability to discover hidden patterns and relationships in medical data. One of the most important use of such systems is in diagnosis of heart diseases because it is one of the leading causes of deaths all over because it is one of the leading causes of deaths all over the world. Almost every system that predicts heart diseases uses clinical dataset having parameter sends inputs from complex tests conducted in laboratory. None of the system predicts heart diseases based on risk factors such as age, family history, diabetes, hypertension, low cholesterol, smoking, alcohol intake, obesity or physical inactivity, etc. Heart disease patients have many of these visible risk factors which are common and can be used very effectively for

diagnosis. System based on such risk factors will not only help medical professionals but it would also give patients a warning about the probable presence of heart disease even before he visit a hospital or tends towards costly medical checkups. This technique has two most successful data mining tools, neural networks and genetic algorithms. The hybrid system implementation uses the global optimization benefit of genetic algorithm for initialization of neural network weights. The learning is fast, more stable and accurate in comparison to back propagation. The system was implemented in Matlab and predicts the risk of heart disease with an accuracy of 89%.

2. LITERATURE REVIEW

Chaitrali S. Dangare et al [2] has analysed prediction systems for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease. Until now, 13 attributes are used for prediction. Jyoti Soni et al [3] intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well.

The main objective of Shadab Adam Pattekari and Asma Parveen [4] is to develop an Intelligent System using data mining modeling technique, namely, Naive Bayes. It is implemented as web based application in this user answers the predefined questions.

N. Aditya Sundar et al [5] describes about a prototype using data mining techniques, namely Naïve Bayes and WAC (weighted associative classifier).This system can answer complex "what if" queries which traditional decision support systems cannot. Using medical profile0073 such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease.

In this paper R. Thanigaivel et al [6] survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. Result from using neural networks is nearly 100%. So that the prediction by using data mining algorithm given efficient results. Applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.

M.I. López et al [7] proposes a classification via clustering approach to predict the final marks in a university course on the basis of forum data. The objective is twofold: to determine if student participation in the course forum can be a good predictor of the final marks for the course and to examine whether the proposed classification via clustering approach can obtain similar accuracy to traditional classification algorithms. Experiments were carried out using real data from first-year university students. Several clustering algorithms using the proposed approach were compared with traditional classification algorithms in predicting whether students pass or fail the course on the basis of their Moodle forum usage data.

3. PROPOSED WORK

The existing version of decision tree construction algorithm do not consider the elimination of the outliers or the noisy data. Due to this reason the accuracy of the existing version reduces. Our proposed algorithm takes into consideration the elimination of noisy data or outlier data.

Input: D – Data Partition A – Attribute List GR – Gain Ratio

Output: A Decision Tree

1. Create a node N

2. If samples in N are of same class, C then

3. Return N as a leaf node and mark class C;

4. If A is empty then

5. Return N as a leaf node and mark with majority class;

6. else

7. Apply Gain Ratio(Dw , Aw)

8. Label root node N as f(A)

9. for each outcome j of f(A)do

10. subtree j =New Decision Tree(Dj,A)

11. connect the root node N to subtree j

12. endfor

13. endif

14. endif

15. Return N

16. For each instance I in D,

If there is an instance I which is not classified then remove the instance from the data set.

4. RESULT ANALYSIS

The given section provides the study about the proposed classification algorithm and the comparative performance study among the implemented classifiers in different performance factors. The performance outcomes and the estimated analysis are provided in this chapter.

5. Accuracy

The accuracy is a measurement of the data model for finding the amount of correctly classified data using the input samples. The performance of the algorithm in terms of accuracy can be evaluated using the following formula.

$$accuracy\ \% = \frac{total\ correctly\ classified\ data}{total\ input\ datasets} X100$$

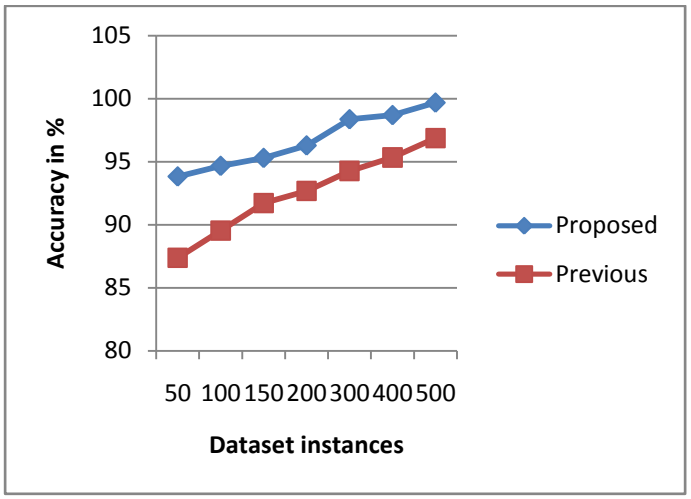

Figure 1: figure accuracy

The performance of the proposed hybrid classifier and the traditional genetic back propagation neural network is compared.

Table 1: Table accuracy

Dataset size	Proposed	Previous
50	93.82	87.38
100	94.66	89.53
150	95.29	91.71
200	96.28	92.68
300	98.36	94.26
400	98.69	95.33
500	99.68	96.86

6. CONCLUSION

The data mining is helpful for analysis the data, when the manually analysis of the data is not feasible in the data mining techniques are applied for analysis. The data mining

techniques are computer based algorithms which identify the relationship among the data and extraction of the similar pattern data on which they are trained. This paper presented a decision tree based classification technique for accurate heart disease classification & prediction. The accuracy of heart disease classification using this method is better than other methods.

REFERENCES

[1] Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm
[2] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888)Volume 47– No.10, June 2012
[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887)Volume 17– No.8, March 2011
[4] Shadab Adam Pattekari and AsmaParveen, "Prediction System for Heart Disease Using Naive Bayes", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294
[5] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base", International Journal of Engineering Science & Advanced Technology, Volume-2, Issue-3, 470 – 478.
[6] R. Thanigaivel, Dr. K. Ramesh Kumar, "Review on Heart Disease Prediction System using Data Mining Techniques", Asian Journal of Computer Science and Technology (AJCST)Vol.3.No.1 2015 pp 68-74.
[7] M.I. López, J.M Luna, C. Romero, S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums", Proceedings of the 5th International Conference on Educational Data Mining.