VOL. 10, NO. 17, SEPTEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ACCURACIES OF J48 WEKA CLASSIFIER WITH DIFFERENT SUPERVISED WEKA FILTERS FOR PREDICTING HEART DISEASES

Jothikumar R[1], Sivabalan R. V[2]. and Sivarajan E[1]
[1]Department of Computer Science and Engineering, Adhiparasakthi College of Engineering, G.B. Nagar, Kalavai, Tamil Nadu, India
[2]Department of Master of Computer Application, Noorul Islam University, Kumaracoil, Tuckalay, Kanyakumari, Tamil Nadu, India
E-Mail: rinfotech.jothi@gmail.com

ABSTRACT

Heart disease is one of the life threatening disease overall the globe. As per the survey of world health organization 17 million deaths are due to heart attacks and strokes which cause maximum casualties. So Heart disease diagnosis and prediction is the essential and monotonous task in healthcare industry. The healthcare industry is information rich and knowledge poor. Useful knowledge can be exposed from health care systems using data mining techniques and can be used to predict the likelihood of patients getting heart disease. Number of researchers used many data mining techniques to diagnose and predict the heart disease and many of them were in practice. But most of the techniques outputs approximately and not accurately, because the datasets used by the researcher is impure and subjected to errors, missing values, irrelevant data and outdated data. The process of cleaning the data that is removing the impure data like errors, missing values, irrelevant data and outdated data is called data preprocessing. By applying data preprocessing prior to the actual technique the accuracies of the prediction classifier can be improved. Here, I am applying supervised weka filters Add classification, attribute Selection, Class Order, Discretize and Nominal to Binary filters for preprocessing the data on the Switzerland heart disease dataset. The cleaned datasets obtained as output from those filters is fed as input to the J48 Classifier and the prediction accuracy of each is measured and Tabulated for comparative analysis. It is found that the J48 Pruned tree with Add classification Filter with J48 classifier gives the improved accuracy of 80.9524 % than others. The performance analysis and different measures considered were tabulated and discussed below.

Keywords: add classification, discretize and nominalToBinary filters, J48 classification, data cleaning, data mining.

INTRODUCTION

Heart disease diagnosis and prediction is the essential and monotonous task in healthcare industry. The healthcare industry is information rich and knowledge poor. Useful knowledge can be exposed from health care systems using data mining techniques and can be used to predict the likelihood of patients getting heart disease. These kinds of systems can serve as a training tool to train nurses and medical students to diagnose patients with heart disease [1]. The importance of heart disease prediction system can be visualized from the fact that heart disease is one of the diseases that causes highest mortality rate [2]. The heart disease was thought to be the problem of developed countries but now it is problem for developing countries too. [5]. Predicting the outcome of disease is one of the most interesting and challenging task in data mining. The knowledge discovery process includes data mining techniques has become a popular research tool for medical researchers and it is able to predict the outcome of a disease using historical data records of patients. Number of tests must be requisite from the patient for detecting a disease. However using data mining technique can reduce the number of test that is required. In order to reduce numbers of deaths from heart diseases there have to be a quick and efficient detection technique [7]. Computer Aided Decision Support System plays a major role in medical field [8]. Heart Diseases remain the biggest cause of deaths for the last two decades. These kinds of systems can be used to assist doctors and assist medical professionals in making decision of heart disease in the early stage based on the clinical data of patients [8]. In biomedical diagnosis, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patients suffer from more than one type of disease of the same category. The physicians may not able to diagnose it correctly. Data mining with intelligent algorithms can be used to tackle the said problem of prediction in medical dataset involving multiple inputs [8].Patient's disease is increasing, day by day. So there is a need of an application which can provide information for decision makers, on patient diseases collected data. Computational intelligence methods open up new prospects for diseases diagnostic criteria. Data mining is an approach which can help in decision making [9]. According to world health organization about more than 12 million deaths occurs worldwide, every year due to heart problems. It is also one of the fatal diseases in India which causes maximum casualties [10]. Heart disease should be diagnosed accurately and correctly. Due to limitation of the potential of the medical experts and their unavailability at certain places put their patients at high risk. It would be highly advantageous if the techniques will be integrated with the medical information system. Computer based information or decision support systems can facilitate accurate diagnosis that's too at reduced

cost.[11,12]. By the survey of WHO, 17 million deaths are due to heart attacks and strokes. Most of the deaths due to heart disease in many countries occur due to work overload, mental stress and many other problems. On the whole it is found as primary reason behind death in adults. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor's experience and knowledge. This leads to unwanted results and excessive medical costs of treatments provided to patients. Therefore an automatic medical diagnosis system is designed that take advantage of collected data base and decision support system. This system can help in diagnosing disease with less medical tests and effective treatments [13]. Although such diseases are controllable, their early prognosis and a patient's evaluated risk are necessary to curb the high mortality rates it presents. Common cardiovascular diseases include coronary heart disease, cardiomyopathy, hypertensive heard disease, heart failure, etc. Common causes of heart diseases include smoking, diabetes, lack of physical activity, hypertension, high cholesterol diet, etc [14].

MATERIALS AND METHODS

UCI SWITZERLAND DATABASE

The heart disease database from the Hungarian database, Irvine, Uci archive is used. This database contains four data sets from the Cleveland clinic foundation, Hungarian institute of cardiology, v.a. medical center and university hospital of Switzerland. It provides 920 records in total. Originally, the database had 76 raw attributes. However, all of the published experiments use only 13 of these. The Hungarian_csv database with 294 instances and 14 attributes age, sex, cp, trestbps, chol, fbs, restecg, talach, exang, oldpeak, slope, ca, thal and num were used here for the analysis.

The detailed information about the 14 attributes has been given below:

1. (age) age in years
2. (sex) sex (1 = male; 0 = female)
3. (chest_pain) chest_pain: chest pain type
Value 1: typical angina
Value 2: atypical angina
Value 3: non-anginal pain
Value 4: asymptomatic
4. (trestbps) resting blood pressure (in mm Hg on admission to the hospital)
5. (chol) serum cholestoral in mg/dl
6. (fbs) (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. (restecg) restecg: resting electrocardiographic results
Value 0: normal
Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. (thalach) maximum heart rate achieved
9. (exang) exercise induced angina (1 = yes; 0 = no)
10. (oldpeak) ST depression induced by exercise relative to rest
11. (slope) the slope of the peak exercise ST segment
Value 1: upsloping
Value 2: flat
Value 3: downsloping
12. (ca) number of major vessels (0-3) colored by flourosopy
13. (thal) 3 = normal; 6 = fixed defect; 7 = reversable defect
14. (num) (the predicted attribute, diagnosis of heart disease (angiographic disease status)
Value 0: < 50% diameter narrowing.
Value 1: > 50% diameter narrowing.

WEKA

Weka is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. Weka is a state-of-the-art facility for developing machine learning techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. Weka is a data mining tools. It contains many machine leaning algorithms. It provides the facility to classify our data through various algorithms (Pankaj saxena and Sushma lehri, 2013). The algorithms are applied directly to a dataset. Weka implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools. The new machine learning schemes can also be developed with this package. Weka is open source software issued under the general public license.

RESULTS AND DISCUSSIONS

J48 Pruned tree with Add classification Filter in weka. The supervised weka filter Add Classification is applied to the data set for preprocessing. Then the J48 Pruned Tree Weka classifier is applied on the Switzerland heart disease dataset with 294 instances and 14 attributes of the format (arff) Attribute relationship file format which supports weka tool. The result obtained is given below. This implementation uses 10-fold cross-validation test mode. The time taken to build the model is 0.02 seconds. The accuracy obtained by J48 Pruned tree with Add classification Filter in weka is 80.9524 % and the other measures were given in Table-1.

Table-1. J48 pruned tree with add classification filter in WEKA.

Correctly Classified Instances	**238 (80.9524 %)**
Incorrectly Classified Instances	56 (19.0476%)
Kappa statistic	0.5729
Mean absolute error	0.2976
Root mean squared error	0.4027
Relative absolute error	64.4938%
Root relative squared error	83.8622%
Coverage of cases (0.95 level)	98.9796%
Mean rel. region size (0.95 level)	98.9796%
Total Number of Instances	294

The other measures TP Rate, FP Rate, Precision, Recall, F-measure, and ROC Area with respective weighted average are given in the Table-2.

Table-2. Detailed accuracy by class.

Class	**Values**		**Weighted Avg.**
	<50	**>50_1**	
TP Rate	0.894	0.66	0.81
FP Rate	0.34	0.106	0.256
Precision	0.824	0.778	0.807
Recall	0.894	0.66	0.81
F-Measure	0.857	0.714	0.806
ROC Area	0.752	0.752	0.752

The confusion Matrix obtained for Attribute selected classifier is given below in the Table-3.

Table-3. Confusion matrix.

A	**B**	
168	20	a=<50
36	70	B=>50_1

J48 Pruned tree with attribute Selection Filter in weka. The supervised weka filter attribute Selection is applied to the data set for preprocessing. Then the J48 Pruned Tree Weka classifier is applied on the Switzerland heart disease dataset with 294 instances and 14 attributes of the format (arff) Attribute relationship file format which supports weka tool. The result obtained is given below. This implementation uses 10-fold cross-validation test mode. The time taken to build the model is 0 seconds. The accuracy obtained by J48 Pruned tree with attribute Selection Filter in weka is 77.8912 % and the other measures were given in Table-4.

Table-4. J48 pruned tree with attribute selection filter in WEKA.

Correctly Classified Instances	**229 (77.8912%)**
Incorrectly Classified Instances	65 (22.1088%)
Kappa statistic	0.501
Mean absolute error	0.3084
Root mean squared error	0.4162
Relative absolute error	66.8347%
Root relative squared error	86.663%
Coverage of cases (0.95 level)	98.2993%
Mean rel. region size (0.95 level)	93.8776%
Total Number of Instances	294

The other measures TP Rate, FP Rate, Precision, Recall, F-measure, and ROC Area with respective weighted average is given in the Table-5.

Table-5. Detailed accuracy by class.

Class	**Values**		**Weighted Avg.**
	<50	**>50_1**	
TP Rate	0.878	0.604	0.779
FP Rate	0.396	0.122	0.297
Precision	0.797	0.736	0.775
Recall	0.878	0.604	0.779
F-Measure	0.835	0.663	0.773
ROC Area	0.758	0.758	0.758

The confusion Matrix obtained for Attribute selected classifier is given below in the Table-6.

Table-6. Confusion matrix.

A	**B**	
70	36	a=<50
21	167	b=>50_1

J48 Pruned tree with Class Order Filter in weka. The supervised weka filter Class Order is applied to the data set for preprocessing. Then the J48 Pruned Tree Weka classifier is applied on the Switzerland heart disease dataset with 294 instances and 14 attributes of the format (arff) Attribute relationship file format which supports weka tool. The result obtained is given below. This

implementation uses 10-fold cross-validation test mode. The time taken to build the model is 0 seconds. The accuracy obtained by J48 Pruned tree with Class Order Filter in weka is 80.6122 % and the other measures were given in Table-7.

Table-7. J48 pruned tree with class order filter in WEKA.

Correctly Classified Instances	**237 (80.6122%)**
Incorrectly Classified Instances	57(19.3878%)
Kappa statistic	0.5661
Mean absolute error	0.2976
Root mean squared error	0.4027
Relative absolute error	64.4938%
Root relative squared error	83.8622%
Coverage of cases (0.95 level)	98.9796%
Mean rel. region size (0.95 level)	98.9796%
Total Number of Instances	294

The other measures TP Rate, FP Rate, Precision, Recall, F-measure, and ROC Area with respective weighted average are given in the Table-8.

Table-8. Detailed accuracy by class.

Class	**Values**		**Weighted Avg.**
	<50	**>50_1**	
TP Rate	0.66	0.888	0.806
FP Rate	0.112	0.34	0.257
Precision	0.769	0.823	0.803
Recall	0.66	0.888	0.806
F-Measure	0.711	0.854	0.802
ROC Area	0.752	0.752	0.752

The confusion Matrix obtained for Attribute selected classifier is given below in the Table-9.

Table-9. Confusion matrix.

A	B	
160	28	a=<50
31	75	b=>50_1

J48 Pruned tree with Discretize Filter in weka. The supervised weka filter Discretize is applied to the data set for preprocessing. Then the J48 Pruned Tree Weka classifier is applied on the Switzerland heart disease dataset with 294 instances and 14 attributes of the format (arff) Attribute relationship file format which supports weka tool. The result obtained is given below. This implementation uses 10-fold cross-validation test mode. The time taken to build the model is 0.02 seconds. The accuracy obtained by J48 Pruned tree with Discretize Filter in weka is 79.932% and the other measures were given in Table-10.

Table-10. J48 pruned tree with discretize in WEKA.

Correctly Classified Instances	**235 (79.932%)**
Incorrectly Classified Instances	59 (20.068%)
Kappa statistic	0.5621
Mean absolute error	0.2829
Root mean squared error	0.3987
Relative absolute error	61.3096%
Root relative squared error	83.0261%
Coverage of cases (0.95 level)	99.3197%
Mean rel. region size (0.95 level)	89.966 %
Total Number of Instances	294

The other measures TP Rate, FP Rate, Precision, Recall, F-measure, and ROC Area with respective weighted average are given in the Table-11.

Table-11. Detailed accuracy by class.

Class	**Values**		**Weighted Avg.**
	<50	**>50_1**	
TP Rate	0.851	0.708	0.799
FP Rate	0.292	0.149	0.241
Precision	0.838	0.728	0.798
Recall	0.851	0.708	0.799
F-Measure	0.844	0.718	0.799
ROC Area	0.801	0.801	0.801

The confusion Matrix obtained for Attribute selected classifier is given below in the Table-12.

Table-12. Confusion Matrix.

A	B	
165	23	a=<50
42	64	b=>50_1

J48 Pruned tree with NominalToBinary Filter in weka. The supervised weka filter NominalToBinary is applied to the data set for preprocessing. Then the J48 Pruned Tree Weka classifier is applied on the Switzerland heart disease dataset with 294 instances and 14 attributes of the format (arff) Attribute relationship file format which supports weka tool. The result obtained is given below. This implementation uses 10-fold cross-validation test mode. The time taken to build the model is 0.02 seconds. The accuracy obtained by J48 Pruned tree with NominalToBinary Filter in weka is 76.5306 % and the other measures were given in Table-13.

Table-13. J48 pruned tree with NominalToBinary in WEKA.

Correctly Classified Instances	**225 (76.5306 %)**
Incorrectly Classified Instances	69(23.4694 %)
Kappa statistic	0.4748
Mean absolute error	0.3037
Root mean squared error	0.4276
Relative absolute error	65.8173 %
Root relative squared error	89.045 %
Coverage of cases (0.95 level)	97.619 %
Mean rel. region size (0.95 level)	95.7483 %
Total Number of Instances	294

The other measures TP Rate, FP Rate, Precision, Recall, F-measure,and ROC Area with respective weighted average is given in the Table-14.

Table-14. Detailed accuracy by class.

Class	**Values**		**Weighted Avg.**
	<50	**>50_1**	
TP Rate	0.856	0.604	0.765
FP Rate	0.396	0.144	0.305
Precision	0.793	0.703	0.761
Recall	0.856	0.604	0.765
F-Measure	0.824	0.65	0.761
ROC Area	0.772	0.772	0.772

The confusion Matrix obtained for Attribute selected classifier is given below in the Table-15.

Table-15. Confusion matrix.

A	**B**	
161	27	a=<50
42	64	b=>50_1

Performance Analysis

The accuracies obtained by J48 Classification Technique with different supervised weka filter are tabulated in the Table-16. It is found that the J48 Pruned tree with Add classification Filter gives the higher accuracy of 80.9524 % and the next nearby accuracy by J48 Pruned tree with Class Order Filter is 80.6122 %. The accuracies obtained were charted in Figure-1 for analysis.

Table-16. J48 pruned tree with different supervised filter.

S. No.	**J48 Pruned tree with different supervised filters**	**Accuracy**
1	J48 Pruned tree with Add classification Filter in weka	80.9524 %
2	J48 Pruned tree with attribute Selection Filter in weka.	77.8912 %
3	J48 Pruned tree with Class Order Filter in weka.	80.6122 %
4	J48 Pruned tree with Discretize Filter in weka	79.932 %
5	J48 Pruned tree with NominalToBinary Filter in weka	76.5306 %

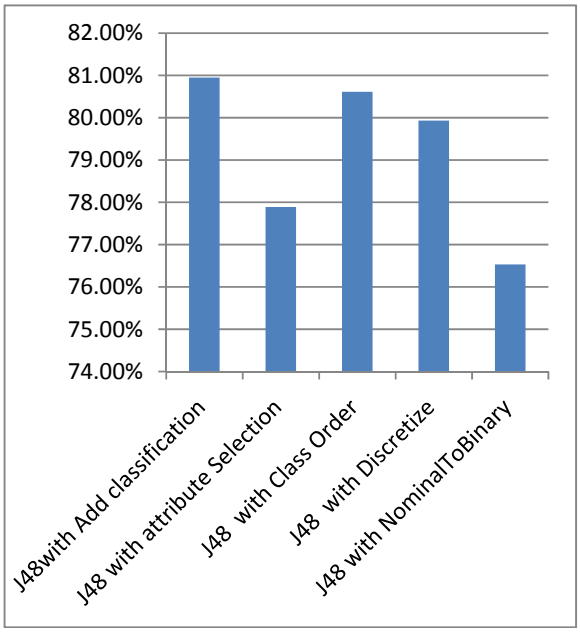

Figure-1. Accuracies of J48 with different filters.

CONCLUSIONS

Many researchers implemented heart disease prediction systems using classification techniques. But, the accuracies obtained by those are poor because none of them concentrated on the data cleaning. In this paper, we have applied supervised weka filters Add classification, attribute Selection, Class Order, Discretize and NominalToBinary filters for preprocessing the data on the Switzerland heart disease dataset. The preprocessed data is given as input to the J48 classifier and the result obtained is analyzed for measuring performance. It is found that the J48 Pruned tree with Add classification Filter gives the higher accuracy of 80.9524 % and the next nearby accuracy by J48 Pruned tree with Class Order Filter is 80.6122 %. Hybrid data preprocessing techniques can be applied in future to improve the accuracies still in heart disease prediction systems for further research.

REFERENCES

[1] Lakshmi K. R., Veera Krishna M. and Prem Kumar S.2013.Performance Comparison of Data Mining Techniques forof Data Mining Techniques for Predicting of Heart Disease Survivability, International Journal of Scientific and Research Publications. 3(6): 1-10.

[2] Juneja U. and Dhingra D. 2014. Multi Parametric Approach Using Fuzzification on Heart Disease Analysis. International Journal of Engineering Sciences and Research Technology. 3(5):492-497.

[3] Sathish Kumar L. and Padmapriya A. 2013. Prediction for Common Disease using ID3 Algorithm in Mobile Phone and Television. International Journal of Computer Science and Network Security. 13(6): 83-86.

[4] Desikan P., Kuo-Wei Hsu. and Srivastava J. 2011. Data Mining for Healthcare Management. SIAM International conference on data mining.

[5] Lovepreet Kaur. 2014. Predicting Heart Disease Symptoms using Fuzzy C-Means Clustering. International Journal of Advanced Research in Computer Engineering and Technology. 3(12): 4232-4235.

[6] Sudha A., Gayathri P. and Jaisankar N. 2012. Utilization of data mining approaches for prediction of life threatening diseases survivability. International Journal of computer applications. 41(17):51-55.

[7] Deepali Chandna. 2014. Diagnosis of Heart Disease Using Data Mining Algorithm. International Journal of Computer Science and Information Technologies. 5(2):1678-1680.

[8] Chitra R. and Seenivasagam V. 2013.Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques. ICTACT Journal on Soft Computing. 3(4): 605-609.

[9] Dhiraj Pandey and Santosh kumar. 2012. Prediction System to Support Medical Information System Using Data Mining Approach. International Journal of Engineering Research and Applications. 2(3): 1988-1996.

[10] Taneja A. 2013.Heart Disease Prediction System Using Data Mining Techniques.Oriental Journal of Computer Science and Technology. 6 (4):457- 466.

[11] Olson D.L. and Dursun D. 2008. Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg.

[12] Fayyad U.M., Piatetsky-Shapiro G.,and Smyth P. 2008. Disease Prediction System Using Data Mining Techniques. International Journal of Computer Science and Network Security. 8(8): 343-350.

[13] Dangare C.S. and Apte S.S. 2012. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications. 47(10): 44-48.

[14] Mythili T., Mukherji D., Padalia N. and Naidu A. 2013. A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications. 68(16): 11-15.

[15] Sivagowry S. and Durairaj M. 2014. PSO - An Intellectual Technique for Feature Reduction on Heart Malady Anticipation Data. International Journal of Advanced Research in Computer Science and Software Engineering. 4(9): 610-621.