www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-9, Issue-6, December 2021
Research Paper

Int. J. Sc. Res. in Network Security and Communication

E-ISSN:2321-3256

Optimizing and Enhancing Performance Classification Algorithm on Heart Disease through Feature Selection

Vikas Mongia

Dept. of Computer Science, Guru Nanak College, Moga, Panjab- India, email: vikasmongia@gmail.com

Received: 22/Sept/2021, Accepted: 19/Nov2021, Published: 31/Dec/2021

Abstract- The ever-increasing size of datasets in the Big Data era requires effective methods for extracting meaningful information. Data Mining provides a means to analyze large datasets and uncover valuable patterns that can inform future decisions. In this study, we analyze a healthcare dataset of heart diseases to predict the likelihood of a patient having a heart disease based on specific parameters. To accomplish this, we implement decision tree classification algorithms such as ADTree, J48, and RandomForest. Additionally, a feature selection algorithm is applied to remove the least significant three attributes from the dataset, resulting in improved classification performance. Comparing the previous and current results reveals the effectiveness of this approach in enhancing the classification accuracy.

Keywords- Data Mining, classification algorithms, Feature selection

I. INTRODUCTION

Data mining involves the process of discovering patterns in large datasets, utilizing a range of techniques from artificial intelligence, machine learning, statistics, and database systems. It is an interdisciplinary field of computer science with the primary objective of transforming data into a structured and understandable format for further analysis and use [1]. In the era of Big Data, there has been a significant increase in the volume of raw data. According to a survey [2], the size of data is expected to grow from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. Consequently, there is a critical need to analyze this raw data and extract useful patterns that can aid in making informed decisions. Data mining involves performing exploratory analysis on this vast dataset, including database and data management aspects, data preprocessing, model and derivation considerations, interestingness metrics, complexity considerations, post-processing of visualization, and online updating [3].

In this study, we utilize data mining techniques for early detection of heart diseases in patients based on certain attributes. To achieve this, we analyze a dataset of 270 patients with 13 attributes obtained from the UCI repository. We apply Decision Tree Classifiers to the dataset, allowing us to identify key patterns that can aid in early detection of heart diseases.

II. DATA MINING TECHNIQUES

Researchers use various techniques to extract information from datasets, such as clustering, classification [4], association rules, and regression analysis. Clustering is used to group similar objects together and is applied in diverse fields such as market research, image processing, and biology [5]. Classification and Regression Analysis are used to predict the class or value of a data item and generate models based on training data. Decision Tree Classification is a recursive method that creates nodes and splits the data until all records have the same classification. This paper focuses on Decision Tree Classification techniques using J48, ADTree, and Random Forest decision tree classifiers. Association Rules are used to analyze patterns from data based on if-then association rules, and algorithms like AIS, SETM, and Apriori can be used to implement them. The support and confidence relationships between items are used to create association rules.

III. LITERATURE SURVEY:

The use of Data Mining techniques in the field of healthcare has been widespread in recent years. Researchers have applied various classification algorithms to identify patterns and predict disease probabilities. In [6], the focus is on using Artificial Neural Network and Decision Tree approaches to understand why many individuals in the United States lack health coverage. Another researcher in [7] applies Data Mining techniques to predict the likelihood of breast cancer in patients by discovering hidden patterns in the dataset. Similarly, [8] implements an Intelligent Heart Disease Prediction System (IHDPS) using Decision Tree, Naïve Bayes and Neural network techniques to predict the probability of heart disease in patients.

Furthermore, in [9], the accuracy of different classification algorithms is compared and analyzed on a healthcare dataset to identify the best classifier. However, Data Mining classification is not limited to healthcare only, as [10] has

used these techniques to create a vaccination schedule for mothers and provide alerts for the next vaccination.

Additionally, another approach in [11] presents an examination toolbox based on open-source modules that facilitates the analysis of healthcare-related datasets. The toolbox provides detailed analysis of doctor and hospital ratings data and can be useful for software engineers, big data architects, hospital administrators, policy makers, and patients. As an illustration of the toolbox's capabilities, the researcher examines the relationship between the position of medical professionals and clinical outcomes using a freely available dataset of national hospital ratings in the USA. The analysis suggests that there is no significant relationship between the experience of medical professionals and hospital ratings as defined by the US government.

IV. PROPOSED METHODOLOGY

In this paper, a dataset of Heart diseases is collected from the UCI repository. This data set has 13 attributes and 270 instances. The various attributes considered in the study are:

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. Thalassemia: 3 = normal; 6 = fixed defect; 7 = reversable defect

With the help of WEKA machine learning tool, a classification model using J48, RandomForest and ADTree is built to clasify the test data. Output of classifiers is as shown below:

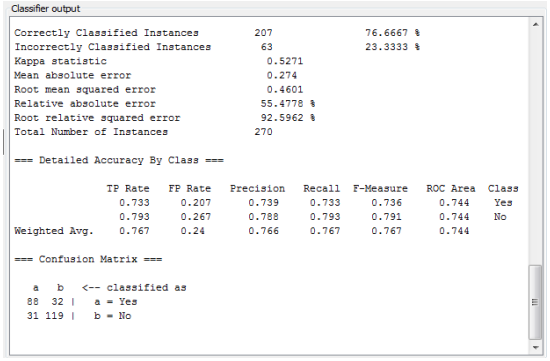

Fig 1: representing output of ADTree classifier

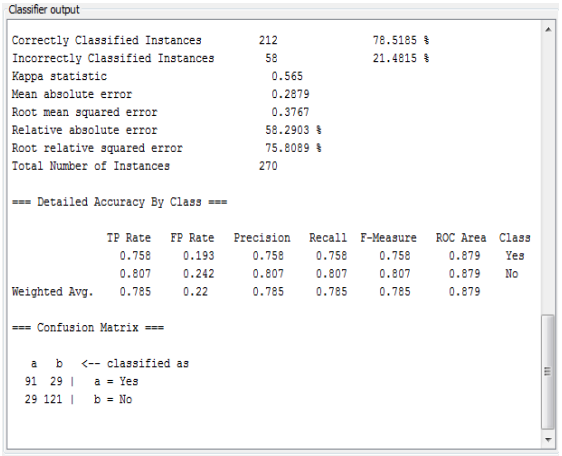

Fig 2: representing output of J48 classifier

From the experiment it can be concluded that the classification accuracy of RandomForest classifier is highest which is 79.6296% ass compared to J48 and ADTree classifier which is 76.6667% and 78.5185% respectively.

Attribute Selection: To determine the least contributing attributes in the dataset,feature selection techniques are applied. Objective of these techniques is to identify those attributes that do not or least contributes to the classification accuracy and removal of these attributes may increase the accuracy rate. In this paper,Information Gain, Gain Ratio and Chisquare attribute selection algorithms are applied and those attributes are removed that are common to the result of all feature selection techniques to ensure right selection of the attributes.Along with these attributes, Rankers algorithm is applied for the ranking of attributes

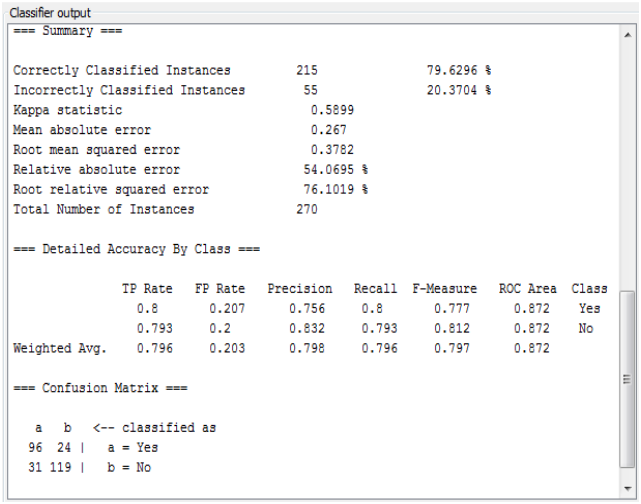

Fig 3: representing output of RandomForest

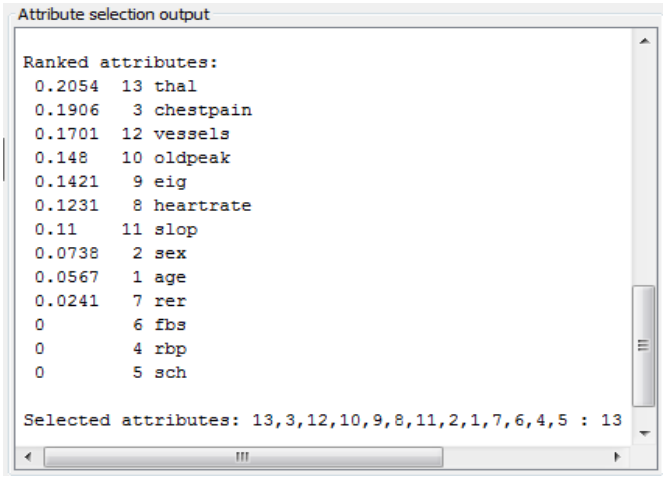

Fig 4: representing result of GainRatioAttributeEval

From the experiment it can be concluded that the classification accuracy of RandomForest classifier is highest which is 79.6296% ass compared to J48 and ADTree classifier which is 76.6667% and 78.5185% respectively.

Attribute Selection: To determine the least contributing attributes in the dataset,feature selection techniques are applied. Objective of these techniques is to identify those attributes that do not or least contributes to the classification accuracy and removal of these attributes may increase the accuracy rate. In this paper,Information Gain, Gain Ratio and Chisquare attribute selection algorithms are applied and those attributes are removed that are common to the result of all feature selection techniques to ensure right selection of the attributes.Along with these attributes, Rankers algorithm is applied for the ranking of attributes

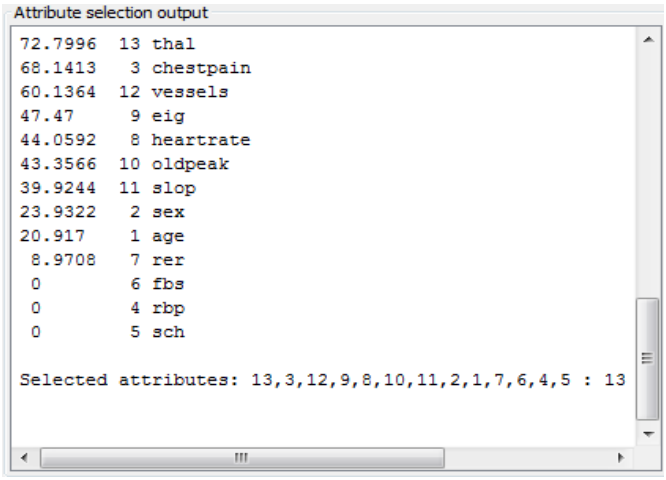

Fig 5: representing result of InfoGainAttributeEval

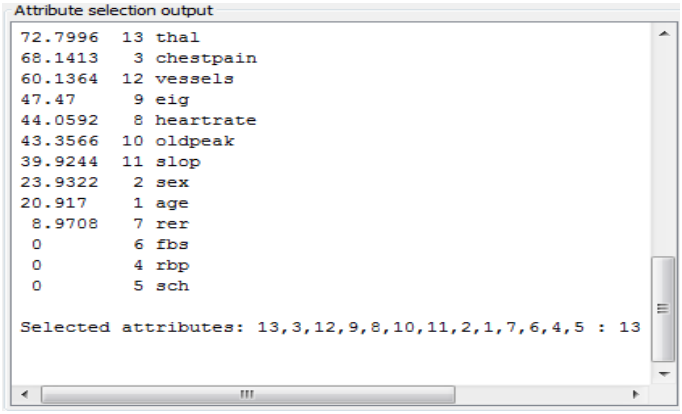

Fig 6: representing results of ChiSquaredAttributeEval

From the analysis it has been observed that the attributes fasting blood sugar (fbs), resting blood pressure (rbp) and serum cholesterol (sch) have no contribution to the classification hence these can be removed from the dataset.

Removing the Attributes: From the above experiment it has been observed that the attributes fbs, rbp and sch have least contribution to the classification decision. Thus, these attributes are removed and the Decision Tree Classification algorithms are again implemented to check the performance difference. The results of these classifiers after the attribute removal are as below:

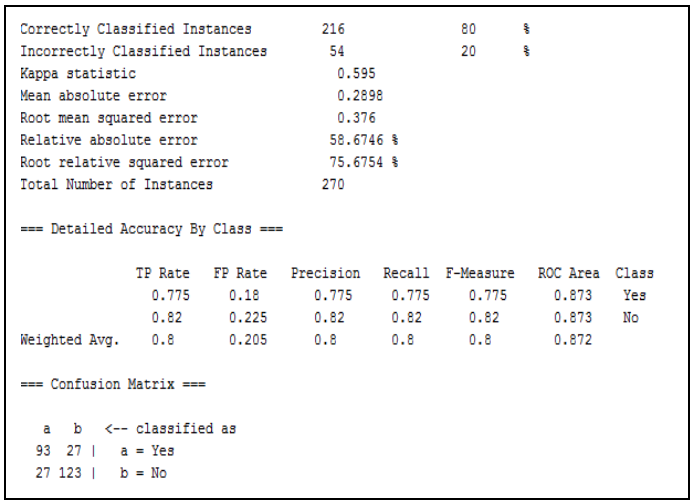

Fig 7: representing results of ADTree classifier after feature extraction


```
Correctly Classified Instances         213               78.8889 %
Incorrectly Classified Instances        57               21.1111 %
Kappa statistic                          0.5714
Mean absolute error                      0.2508
Root mean squared error                  0.4337
Relative absolute error                 50.7902 %
Root relative squared error             87.2721 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.75      0.18       0.769     0.75      0.759      0.779    Yes
                 0.82      0.25       0.804     0.82      0.812      0.779    No
Weighted Avg.    0.789     0.219      0.789     0.789     0.789      0.779

=== Confusion Matrix ===

   a   b   <-- classified as
  90  30 |   a = Yes
  27 123 |   b = No
```

Fig 8: representing results of J48 claclassifier after feature extraction

Classifier output

```
Correctly Classified Instances         222               82.2222 %
Incorrectly Classified Instances        48               17.7778 %
Kappa statistic                          0.6406
Mean absolute error                      0.237
Root mean squared error                  0.3602
Relative absolute error                 47.9951 %
Root relative squared error             72.4982 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.808     0.167      0.795     0.808     0.802      0.885    Yes
                 0.833     0.192      0.845     0.833     0.839      0.885    No
Weighted Avg.    0.822     0.181      0.823     0.822     0.822      0.885

=== Confusion Matrix ===

   a   b   <-- classified as
  97  23 |   a = Yes
  25 125 |   b = No
```

Fig9: representing results of RandomForest classifier after feature extraction

Algorithm	**Accuracy before feature extraction**	**Accuracy after feature extraction**	**Mean Absolute error before Feature Extraction**	**Mean Absolute Error After Feature Extraction**
J48	76.6667%	78.8889%	0.274	0.2508
RandomForest	79.6296%	82.2222%	0.267	0.237
ADTree	76.6667%	78.8889%	0.274	0.2508

Table: Representing comparison among algorithms

V. RESULTS AND DISCUSSION

From the above experiment we can conclude that feature extraction technique has impact on accuracy of decision tree classifier. The results are improved up to 2.89 %. This would lead to better prediction of a patient suspectable to heart disease.

VI. CONCLUSION

Machine Learning calculations can be applied to datasets to extract some useful patterns from it which may support future directions. This paper has implemented Decision Tree Classifiers for the early detection of heart diseases in a patient based upon some attributes. A feature extraction approach is also applied to increase the accuracy of prediction and it has been observed that the accuracy has been improved to 2.89%. Further, from the study it can be concluded that the algorithms behave differently for different datasets in terms of accuracy in prediction, execution time and mean square error.

REFERENCES

[1]. F. Chu and C. Zaniolo, “Fast and light boosting for adaptive mining of data streams,” Adv. Knowl. Discov. Data Min.,vol. 3056, pp. 282–292, 2004.

[2]. Sh. Hajirahimova, et. al., Azerbaijan; Aliyeva, Aybeniz S., "About Big Data Measurement Methodologies and Indicators". International Journal of Modern Education and Computer Science. Vol.9, Issue 10, pp.1–9, 2017. doi:10.5815/ijmecs.2017.10.01

[3]. S. Sumathi and S. N. Sivanandam, “Data mining tasks, techniques, and applications,” Stud. Comput. |Intell., Vol. 29, pp. 195–216, 2007.

[4]. S. a. Mingoti and J. O. Lima, “Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms,” Eur. J. Oper. Res., Vol. 174, pp. 1742–1759, 2006.

[5]. GeletawSahle, “Ethiopic maternal care data mining:discovering the factors that affect postnatal care visit in Ethiopia” Sahle Health Inf Sci Syst (2016) 4:4 DOI 10.1186/s13755-016-0017-2

[6]. DursunDelen*, Christie Fuller, Charles McCann, Deepa Ray “Analysis of healthcare coverage: A data mining approach” Available online at www.sciencedirect.com Expert Systems with Applications, Vol.36, pp.995–1003, 2009.

[7]. Shelly Gupta, DharminderKumar,Anand Sharma, “Data Mining Classification Techniques Applied for Breast Cancer Diagnosis And Prognosis “ Vol. 2 No. 2 Apr-May 2011

[8]. SellappanPalaniappan, Rafiah Awang “Intelligent Heart Disease Prediction System Using Data Mining Techniques”AICCA ’08 Proceeding of the 2008 IEEE/ACS International conference on computer Systems and Application pages 108-115, 2008.

[9] Suresh, Annamalai, Rajagopal Kumar, and R. Varatharajan. "Health care data analysis using evolutionary algorithm." The Journal of Supercomputing, Vol.76, Issue 6, pp.4262-4271, 2020.

[10] Skoff, Tami H., et al. "Impact of the US maternal tetanus, diphtheria, and acellular pertussis vaccination program on preventing pertussis in infants< 2 months of age: a case-control evaluation." Clinical Infectious Diseases, Vol.65, Issue.12, pp.1977-1983, 2017.

[11] Lo’ai, A. Tawalbeh, et al. "Mobile cloud computing model and big data analysis for healthcare applications." IEEE Access, Vol.4, pp. 6171-6180, 2016.

