Comparative Analysis of Classification Approaches for Heart Disease Prediction

Conference Paper · February 2018

CITATIONS
11

READS
556

4 authors:

S M Mahedy Hasan
Rajshahi University of Engineering & Technology
53 PUBLICATIONS **228** CITATIONS
SEE PROFILE

Md. Al Mamun
Rajshahi University of Engineering & Technology
103 PUBLICATIONS **871** CITATIONS
SEE PROFILE

Md Palash Uddin
Deakin University
86 PUBLICATIONS **1,253** CITATIONS
SEE PROFILE

Dr Md Ali Hossain
UNSW Sydney
90 PUBLICATIONS **1,085** CITATIONS
SEE PROFILE

Comparative Analysis of Classification Approaches for Heart Disease Prediction

S. M. M. Hasan[1], M. A. Mamun[2], M. P. Uddin[3] and M. A. Hossain[4]
Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh
[1]mahedycseruet@gmail.com, [2]a.mamun@ruet.ac.bd, [3]palash_cse@hstu.ac.bd and [4]ali.hossain@ruet.ac.bd

***Abstract*—Heart disease is one of the most common causes of death around the world nowadays. Often, the enormous amount of information is gathered to detect diseases in medical science. All of the information is not useful but vital in taking the correct decision. Thus, it is not always easy to detect the heart disease because it requires skilled knowledge or experiences about heart failure symptoms for an early prediction. Most of the medical dataset are dispersed, widespread and assorted. However, data mining is a robust technique for extracting invisible, predictive and actionable information from the extensive databases. In this paper, by using info gain feature selection technique and removing unnecessary features, different classification techniques such that KNN, Decision Tree (ID3), Gaussian Naïve Bayes, Logistic Regression and Random Forest are used on heart disease dataset for better prediction. Different performance measurement factors such as accuracy, ROC curve, precision, recall, sensitivity, specificity, and F1-score are considered to determine the performance of the classification techniques. Among them, Logistic Regression performed better, and the classification accuracy is 92.76%.**

Keywords: Data mining; KNN; Decision Tree (ID3); Gaussian Naïve Bayes; Logistic Regression; Random Forest, Heart Disease

I. INTRODUCTION

The amount of data in the medical industry is increasing day by day. It is a challenging task to handle a large amount of data and extracting productive information for effective decision making. For this reason, medical industry demands to apply a special technique which will provide fruitful decision from a vast database. Data mining is an exciting field of machine learning and thus capable of solving this type of problem very well. For solving various kinds of real-world problems, data mining is a novel field for discovering hidden patterns and the valuable knowledge from a large dataset. Because it is very strenuous to extract any useful information without mining large database. In brief, it is an essential procedure for analyzing data from various perspectives and gathering knowledge. However, heath care industry is another field where a substantial amount of data collected using different clinical reports and patients manifestations.

Nowadays, people can face any heart failure symptoms at any stage of a lifetime. But old people face this type of problem rather than the young people. Data mining classification techniques can discover the hidden relationship along correlated features which plays a consequential role in predicting the class label from a large dataset. By using those hidden patterns along with the correlated features, it is straightforward to detect heart disease patients without any support of medical practitioners. Then, it will act as an expert system for separating patients with heart disease and patients with no heart disease more accurately with lower cost and less diagnosis time.

II. LITERATURE REVIEW

Researchers have conducted numerous studies related to the diagnosis of heart disease using different data mining techniques in recent years. Sellappan Palaniappan et al. [2] developed an IHDPS (Intelligent Heart Disease Prediction System) based on classification techniques such as Naïve Bayes, Neural Network and Decision Tree. It was web-based and implemented in .NET platform. The accuracy of Naïve Bayes = 86.12%, Neural Network = 85.68% and Decision Tree = 80.4%. Anchana Khemphila et al. [3] compared the performances of Decision Tree, Logistic Regression and Artificial Neural Network for detecting the heart disease patients. Their result showed that the accuracy of Decision Tree = 79.3%, Artificial Neural Network = 80.2% and Logistic Regression = 77.7%. Marjia Sultana et al. [4] analyzed the performances of KStar, J48, SMO, Bayes Net and Multilayer Perceptron for classifying heart disease patients using the standard Cleveland Heart Disease data and some collected data. The accuracy of standard data for KStar = 75.1852%, J48 = 76.6667%, SMO = 84.0741%, Bayes Net = 81.1111% and MLP = 77.4074%. Shan Xu et al. [5] compared the performances of Random Forest, C4.5, SVM, Bayes, AdaBoost for classifying the cardiovascular heart disease patients. The authors considered wrapper based CFS Subset Evaluation method and found the highly correlated features with the classification goal. Their results showed accuracy for Random Forest = 91.6%, C4.5 = 89.6%, SVM = 89.2%, Bayes = 85.2% and AdaBoost = 82.8%. Seyedamin Pouriyeh et al. [6] compared performances of Decision Tree, Naïve Bayes, KNN, Multilayer Perceptron, Radial Basis Function, Single Conjunctive Rule Learner and Support Vector Machines. They also tuned different values of k for determining the performance of KNN. The accuracy for Decision Tree = 77.55%, Naïve Bayes = 83.49%, KNN (k=1) = 76.23%, KNN (k=3) = 81.18%, KNN (k=9) = 83.16%, KNN (k=15) = 83.16%, MLP = 82.83%, RBF = 83.82%, SCRL = 69.96% and SVM = 84.15%. Peter C. Austin et al. [7] said that conventional Logistic Regression method provides better performance than Classification and Regression Trees. Hend

Mansoor et al. [8] compared performances of Logistic Regression and Random Forest method for predicting the risk level of heart disease patients. They used National Inpatient Sample Data over the years 2011-2013 and found that Logistic Regression model provides better accuracy than Random Forest. The accuracy for Logistic Regression model was = 89% and random Forest = 88%. M. A. Jabbar et al. [9] considered different feature selection measures and measured the performance of Naïve Bayes for diagnosis of heart disease patients.

III. METHODS FOR PREDICTION AND CLASSIFICATION

A. K-Nearest Neighbors (KNN)

KNN classifies the test data using the training set directly. To classify any test data, it first calculates *K* value, which denotes the number of K-Nearest Neighbors. For each test data, it calculates the distance between all the training data and then sorts the distance. Then by using majority voting, class label will be assigned to the test data. The equation that measures the Euclidean distance is given below:

$$D_e = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

B. Decision Tree (ID3)

Decision Tree (ID3) is a greedy algorithm that follows recursively top-down greedy approach [13]. It selects the attribute with the highest information gain [1]. Assuming P_i, the probability such that $x_i \in D$, exists to a class C_i, and is predicted by $|C_i , D|/|D|$ [13]. To classify a tuple in D, the expected information is needed, and the following equation estimates it:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{2}$$

In Eq. 2 *Info (D)* is the average amount of information needed to identify C_i of an instance, $x_i \in D$ and the objective of Decision Tree is to divide repeatedly, *D,* into sub data sets $\{D_1, D_2 ,..., D_n\}$ [13]. The following equation computes the *InfoA(D)*:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j) \tag{3}$$

Now, the information gain is given below:

$$Gain(A) = Info(D) - Info_A(D) \tag{4}$$

C. Gaussian Naïve Bayes

When all the data values of any particular dataset are numeric, then Gaussian Naïve Byes is used. It follows a normal distribution. Mean, and standard deviation are used to define the probability density function. It calculates the mean and standard deviation for each attribute of the dataset. After calculating this, when any test data pattern comes, then by using the mean and standard deviation calculate the probabilities for each test data. It assigns a class label to the test data which probability is close to 1. The necessary equations are given below:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

$$\partial = \frac{1}{(n-1)}\sum_{i=1}^{n} (x_i - \mu)^2 \tag{6}$$

$$f(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\partial} e^{-(x-\pi)^2} \tag{7}$$

D. Logistic Regression

The logistic regression algorithm is based on the following logistic function:

$$p = \frac{1}{(1 + e^{-x})} \tag{8}$$

At the training stage, for each instance x_1, x_2, x_3 ..... x_n the logistic coefficients will be b_0, b_1, b_2 ..... b_n. The coefficients values are estimated and updated by stochastic gradient descent.

$$value = b_0 x_0 + b_1 x_2 + \ldots\ldots + b_n x_n \tag{9}$$

$$p = \frac{1}{(1 + e^{-value})} \tag{10}$$

Now, the coefficients values are updated by the using the following equation:

$$b = b + l * (y - p) * (1 - p) * p * x \tag{11}$$

Initially, all the coefficients values are 0 and *y* is the output value for each training instance. Where *l* is the learning rate, *x* is biased input for b_0 and is always 1. It updates the coefficients values until it predicts correct output at training stage.

E. Random Forest

Random forest is an ensemble classification method based on Decision Tree. At the training stage, it produces a vast number of trees and creates a forest of Decision Trees. At the testing stage, each tree of the forest predicts a class label for each data. When each tree predicts a class label, then the final decision for each test data depends on majority voting. Which class label gets the majority of votes this label assumes to be the correct label assigned to the test data. This process is repeated for each of data in the dataset.

IV. EXPERIMENTATION

In this research, Cleveland Heart Disease dataset from UCI machine learning repository was used [10]. This dataset consists of total 14 attributes and 303 records. This experiment was divided into four phases.

A. Data Preprocessing

Data preprocessing involves data cleaning, handling missing values, handling inconsistent data, dimensionality reduction, feature selection, etc. As it is mentioned earlier, this experimental dataset contains 14 attributes, such that age, sex, cp, trestbps, chol, fbs, restcg, thalach, exang, oldpeak, slope, ca, thal, class. For effective decision making info gain feature

selection method was used and selects the highly correlated features with the classification goal. Using the feature selection method 10 attributes out 14 attributes were selected such that sex, cp, oldpeak, restcg, thalach, exang, slope, ca, thal, class that were highly correlated with classification goal.

B. State of Art

We have taken the paper of Anchana Khemphila et al. [3], and Seyedamin Pouriyeh et al. [6] as our two base papers because their work has the similarity with our objective. The authors of both papers have used all the 14 attributes for prediction.

C. Results and Discussion

To conduct the experiment, five classification algorithms were used. Classification algorithms were implemented in Anaconda Python (Spyder 3.6). This experiment was performed based on 10-fold cross-validation, and 50% of data were selected for training, and 50% of data were selected for testing. Each classification algorithm was implemented into two cases. First, they used all of the 14 attributes of the dataset and then they used the selected 10 attributes from the dataset. For K-Nearest Neighbors K=10 and for Decision Tree, the maximum depth of the tree was=30. The confusion matrices for the classifiers with 14 and 10 features are shown in Table I and Table II respectively.

TABLE I. CONFUSION MATRICES OF CLASSIFIERS USING 14 ATTRIBUTES

Classification	Confusion Matrix		Accuracy
KNN	TP = 61	FN = 21	69.73%
	FP = 25	TN = 45	
Decision Tree(ID3)	TP = 74	FN = 08	87.5%
	FP = 11	TN = 59	
Gaussian Naïve Bayes	TP = 75	FN = 11	88.15%
	FP = 07	TN = 59	
Logistic Regression	TP = 77	FN = 08	89.5%
	FP = 08	TN = 59	
Random Forest	TP = 81	FN = 04	88.12%
	FP = 13	TN = 54	

TABLE II. CONFUSION MATRICES OF CLASSIFIERS USING 10 ATTRIBUTES

Classification	Confusion Matrix		Accuracy
KNN	TP = 62	FN = 23	71.05%
	FP = 21	TN = 46	
Decision Tree(ID3)	TP = 74	FN = 12	90.4%
	FP = 03	TN = 63	
Gaussian Naïve Bayes	TP = 80	FN = 05	90.78%
	FP = 09	TN = 58	
Logistic Regression	TP = 76	FN = 06	92.76%
	FP = 05	TN = 63	
Random Forest	TP = 79	FN = 06	92.1%
	FP = 06	TN = 61	

From the above performance comparison table, it can be said that the accuracy of each classification algorithm increased while 10 attributes are used instead of 14 attributes. Also in both of the cases, Logistic Regression performed better than the other classification algorithms that were implemented in the prediction of heart disease patients in this paper. Graphical representation of accuracy for each algorithm in both cases is given below:

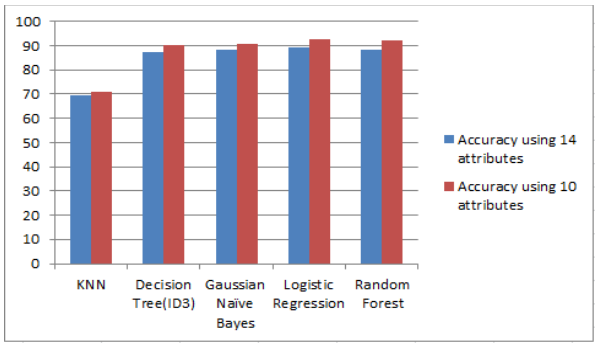

Fig. 1. Graphical representation of the accuracy for each classifier

TABLE III. CLASSIFICATION REPORT OF CLASSIFIERS

Classification	Precision	Recall	Specificity	F1-Score	Roc Area
KNN(using 14 attributes)	0.7	0.7	0.64	0.7	0.695
KNN(using 10 attributes)	0.71	0.71	0.68	0.71	0.72
Decision Tree(ID3 using 14 attributes)	0.88	0.88	0.84	0.88	0.88
Decision Tree(ID3 using 10 attributes)	0.91	0.90	0.95	0.90	0.89
Gaussian NB(using 14 attributes)	0.88	0.88	0.89	0.88	0.94
Gaussian NB(using 10 attributes)	0.91	0.91	0.87	0.91	0.94
Logistic Regression(using 14 attributes)	0.89	0.89	0.88	0.89	0.96
Logistic Regression(using 10 attributes)	0.93	0.93	0.88	0.93	0.96
Random Forest(using 14 attributes)	0.89	0.89	0.80	0.89	0.96
Random Forest(using 10 attributes)	0.92	0.92	0.91	0.92	0.96

Based on the performance evaluation measures as shown in Table III it can be said that the performance of each classification algorithm was increased while 10 attributes are used. From the above experimental results, it can also be said that using all the 14 attributes for heart disease prediction are not so useful. It decreases the performance of the classifier as it contains irrelevant attributes. Using the attribute selection method for removing the irrelevant attributes, it increases the classifiers' performance, and logistic regression showed better performance in prediction of heart disease patients.

D. Behavioural Analysis

Every classification algorithm has its own characteristics. For diverse characteristics, the output of each classification algorithm was different in the prediction of heart disease patients. Reasons, why these classification algorithms behaved like this for this particular dataset like this, is given below:

K-Nearest Neighbors: In this method, K- Nearest Neighbors showed poor performance because KNN classifies test data directly from the dataset, no training was performed before testing.

Decision Tree (ID3): At training stage, it converted the continuous value's data into categorical values and given a range. When test data pattern contained values out of this given range, the classifier performance was affected and thus predicts wrong class label.

Gaussian Naïve Bayes: At the training stage, it calculated the mean and standard deviation of each attribute. This mean and standard deviation were used to calculate the probabilities for the test data. For this reason, some attributes values are too big or too small from the mean. When testing data pattern contains those attributes values, it affects the classifier performance and sometimes gives wrong output label.

Logistic Regression: At the training stage, Logistic Regression algorithm estimated coefficient values by using stochastic gradient descent. The model can be trained for a fixed or as much as no of epochs by using stochastic gradient descent. Coefficients values are updated until the model predicts the correct class label for each training data.

Random Forest: Random Forest is an ensemble classification method which is based on Decision Tree algorithm. This algorithm takes a portion of the dataset and then builds a tree, repeat this step for creating a forest by combining the generated trees. At the test stage, each tree predicts a class label for each test data and majority values of the class label is assigned to the test data. Therefore, it showed reasonable performance than conventional decision tree algorithm for this data.

V. CONCLUSION AND FUTURE WORK

This paper compares the performances of the classification algorithms in the prediction of heart disease. It tries to find out the best classifier for this task. In the experimental dataset, 14 attributes were used. But all the attributes are not equally emphasized for detecting heart disease. For this reason, a feature selection method was presented that removes the irrelevant attributes which are not highly correlated with the other features used for classification. Each classification algorithm gives a noticeable performance while using the selected 10 attributes instead of 14 attributes in the prediction of heart disease. Among the studied classifiers, Logistic Regression performs better than other classification algorithms. Binary class problem is solved to identify whether the patient has heart disease or not. It is recommended to solve the multiclass problem for detecting heart disease by dividing heart disease patients into various classes. This system will be customized to predict not only the presence or absence of heart disease but also to predict the risk factor of heart failure to take extra care of those patients at an early stage and avoid heart failure. Real-time data from different hospitals may be collected for detecting heart disease patients and compute the effectiveness of classifiers for more consistent diagnosis of heart disease patients.

REFERENCES

[1] M. Kamber and P. J. Han, *Data Mining Concepts*, *and Techniques*, 3rd ed., 2012.

[2] S. Palaniappan, R. Awang, " Intelligent Heart Disease Prediction System Using Data Mining Techniques," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, no. 8, August 2008.

[3] A. Khemphila, V. Boonjing "Comparing Performances of Logistic Regression, Decision trees, and Neural Networks for Classifying Heart Disease Patients," *2010 IEEE International Conference on Computer Information Systems and Industrial Management Systems(CISIM)*, pp. 193-199, 2010.

[4] M. Sultana, A. Haider and M. S. Uddin, "Analysis of Data Mining Techniques for Heart Disease Prediction," *3rd International Conference on Electrical Engineering and Information Communication Technology* (*ICEEICT*), 2016.

[5] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu, "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework," *2017 IEEE 2nd International Conference on Big Data Analysis*, 2017.

[6] S. Pouriyeh, S. Vahid, G. Sannino, G. D. Pietro and H. Arabnia, J. Gutierrez, "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease," *22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth,* 2017.

[7] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee, "Using Methods from Data Mining and Machine Learning Literature for Disease Classification and Prediction: a Case Study Examining Classification of Heart Failure Subtypes," *Journal of Clinical Epidemiology 66 (2013) pp. 398-407*, 2013.

[8] H. M. Islam, Y. Elgendy, R. Segal, A. A. Bavry and J. Bian, "Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach," *Journal of Heart & Lung,* pp. 1-7, 2017.

[9] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Computational Intelligence Technique for Early Diagnosis of Heart Disease," *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, 20th March 2015.

[10] UCI Machine Learning Repository. [Online]. Available: http://archive.ics.edu/ml/datasets/heart+disease.

[11] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network-Based Data Mining in Prediction of Heart Disease Using Risk Factors," *2013 IEEE Conference on Information and Communication Technologies,* 2013.

[12] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, May 2010.

[13] A. A. Pathan, M. Hasan, M. F. Ahmed, and D. M. Farid, "Educational Data Mining: A Mining Model for Developing Students Programming Skills," *8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA),* 2014

[14] S. Fathima and N. Hundewale, "Comparison of Classification Techniques- Support Vector Machines and Naive Bayes to predict the Arboviral Disease-Dengue," *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 2011.