International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 4, August 2019, pp. 2675~2681
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i4.pp2675-268

Predicting heart ailment in patients with varying number of features using data mining techniques

T. R. Stella Mary, Shoney Sebastian
Department of Computer Science, Christ University, India

Article Info

Article history:

Received Nov 12, 2017
Revised Mar 4, 2019
Accepted Mar 12, 2019

Keywords:

Data mining
Heart ailments
Naïve bayes classifier
Random forest
Random tree

ABSTRACT

Data mining can be defined as a process of extracting unknown, verifiable and possibly helpful data from information. Among the various ailments, heart ailment is one of the primary reason behind death of individuals around the globe, hence in order to curb this, a detailed analysis is done using Data Mining. Many a times we limit ourselves with minimal attributes that are required to predict a patient with heart disease. By doing so we are missing on a lot of important attributes that are main causes for heart diseases. Hence, this research aims at considering almost all the important features affecting heart disease and performs the analysis step by step with minimal to maximum set of attributes using Data Mining techniques to predict heart ailments. The various classification methods used are Naïve Bayes classifier, Random Forest and Random Tree which are applied on three datasets with different number of attributes but with a common class label. From the analysis performed, it shows that there is a gradual increase in prediction accuracies with the increase in the attributes irrespective of the classifiers used and Naïve Bayes and Random Forest algorithms comparatively outperforms with these sets of data.

Corresponding Author:

T. R. Stella Mary,
Department of Computer Science,
Christ University, India
Email: stellarichardz94@gmail.com

1. INTRODUCTION

Data mining involves the exploration of huge datasets in order to mine unknown, verifiable, relationships, knowledge and possibly helpful data from information [1, 2]. It involves various steps in converting a raw data into a data with useful information. This complete process involves various steps such as data cleaning, integration, transformation, selection, mining, evaluation of pattern and representation of knowledge. This has made data mining find its applications in the fields of healthcare, business, education, media, weather analysis, bio-informatics and many more.

In healthcare, data mining has gained wide interest and importance and has become the most efficient means in diagnosing patients with various ailments. Medical data is rich in information but lacks in terms of tools and technologies to extract the useful patterns and knowledge from them. Almost all the healthcare industries today all streaming out a lot of huge data as in, data pertaining to patients, diagnosis of various ailments, patient records in electronic form, data excavated from medical devices etc. This huge amount of data is processed and analyzed to extract useful patterns which helps in fast diagnosis of ailments and which even aids in cost reduction.

Among the various ailments, heart ailment is one among the main reason for death of human beings all around the world in last decade by the WHO. The European public health alliance reported that over 41% of all deaths were caused by strokes, heart attacks, and other chronic diseases [3]. As indicated by the American Heart Association, more than 7 million Americans have suffered a heart ailment in their lifetime.

At the point when plaque is worked inside the coronary supply routes, they limit these veins making them unfit to convey oxygenated blood to the heart muscle causing the outstanding manifestations of CAD, for example, chest torment (angina) and shortness of breath [4, 5]. There are many kinds of cardiovascular diseases which includes the coronary heart diseases (CHD), stroke, intrinsic heart, fringe vein, rheumatic heart, provocative heart sickness and hypertensive coronary illness [6-8]. The extreme reasons for heart illness are tobacco utilization, physical latency, an undesirable eating regimen and customary use of liquor [9].

Medical diagnosis is an important and complicated task as it must ensure accurate results and need to be executed precisely. Not all doctors are completely well versed in their various sectors and there is lack of proper diagnosis. This paves way for and the stream of data mining to ease the tasks and which is capable of performing the diagnosis and generate the results accurately with less human interference and less investment both in terms of money and time [10].

In real time there are several tests taken to predict and diagnose patients with heart diseases which in turn gives the accurate results but turns out to be time consuming and cost effective. Meanwhile in automated approach people aim at reducing the attributes which in turn reduces the time taken but lacks in accuracy since an extra care has to be taken to ensure that a patient is suffering from heart ailment providing a detailed analysis is done in all the perspectives. Keeping this in mind that most of the features affecting heart disease must be considered, a predictive analysis is done to see how the prediction accuracies vary with very few and maximum features taken. This paper aims at providing a more efficient approach towards the prediction of heart ailments in patients with the most important attributes affecting the patients that leads to heart disease and analyze the data as to how the prediction accuracies fluctuate with minimal and maximal attributes taken into consideration using the data mining approach. Numerous studies have been performed with the aim of efficient diagnosis and prediction of heart ailment. Most of the authors have applied different data mining methods for prediction and have come up with different probabilities of accuracies for different classifiers.

El-Bialy et al., [11] depicts that heterogeneity datasets gives in heterogeneity data types, based on this it combines four datasets with same class label namely the Cleveland heart disease, Hungarian heart disease, V.A. heart disease and Statlog project heart disease with 13 features from UCI repository and applies C4.5 and Fast Decision tree in order to construct decision tree for each dataset and compare the accuracies in WEKA. Out of which five common features are selected and are integrated into a new dataset, on which the two models are applied again, which results in the four most commonly occurring features (ca, age, cp, thal) with highest information gain. This indicates that these four features are the most important features for the prediction of heart disease.

Chaurasia [12] uses three data mining algorithms CART, ID3 and Decision table on the Cleveland heart disease dataset with 11 attributes out of 13 attributes. Theses 3 classifiers are implemented on WEKA with 10-fold cross validation. A detailed result in terms of time taken to build the model, correctly and incorrectly classified instances, error rate, confusion matrices and accuracies are analyzed for all the three classifiers and finally depicts that CART out performs from the other two classifiers with an accuracy of 83.49%. Further the importance of each attribute is analyzed with Chi-squared test, Information gain and Gain ratio test which depicts that cp attributes impacts the output the most.

Dangare and Apte [13] analyses the Cleveland heart disease dataset with 3 classifiers namely the Neural Networks, Decision trees and Naïve Bayes algorithms. To this existing dataset two more attributes obesity and tobacco are included which makes up for the next dataset. Initially the classifiers are applied to the first dataset with 13 attributes and the same classifiers are applied to the next dataset with 15 attributes. Finally, a detailed analysis of their confusion matrices and accuracies are compared between the two datasets which depicts that for the dataset with 13 attributes Neural Network outperforms with 99.25% accuracy and for the dataset with 15 attributes again Neural Network outperforms with 100% accuracy. They conclude that Neural Network algorithm outperforms in both the cases and has relatively higher accuracy with 15 attributes.

Anbarasi et al., [14] objective was to decrease the number of attributes for which they play out the analysis on the Cleveland heart disease dataset with 13 attributes. Feature subset selection is performed on the dataset using the Genetic algorithm which results with 6 attributes on which the classifiers are applied. The classifiers used here are Decision tree, Naïve Bayes and Classification via Clustering. Further while comparing among all the classifiers with their corresponding confusion matrices and accuracies, Decision tree outperforms with an accuracy of 99.2%.

Rohilla and Gulia [15] performs the analysis on the Cleveland heart disease dataset with 11 attributes by applying the classifiers on a 10-fold cross validation process. The different classifiers used are Naïve Bayes, Bagging, ID3, J48, Simple Cart, Logistic Regression and REPTREE algorithms. Comparatively the Decision tree classifier ID3 outperforms with an accuracy of 88% and with a minimum error. Hence it concludes that the Decision tree method that is ID3 outperforms when compared to all other classifiers.

Tamilarasi and Porkodi [16] aims at providing a cost effective automated system in prediction of heart ailment in patients for which the a dataset with 13 attributes are considered. They even specify the risk factors of heart disease which are smoking, cholesterol, obesity and lack of physical exercise. To analyze this dataset 6 classifiers namely the Naïve Bayes, KNN algorithm, J48, CART, ANN and SMO are applied to see which performs comparatively better in terms of correctly and incorrectly classified instances and the error rate. Hence, it concludes that the K-nearest neighbor algorithm outperforms with an accuracy of 100%.

Pattekari and Parveen [17] built a prototype system with a model built with the classifiers namely Decision trees, Naïve Bayes and Artificial Neural Networks. A real time website was developed called the Heart Disease Prediction System (HDPS) where in it retrieves the data stored in the database and predicts if the patient had heart disease using the trained dataset.

Akhiljabbar et al., [18] developed an efficient algorithm using the genetic algorithm approach known as the Associative Classification algorithm in order to predict patients with heart ailment. In this approach all the attributes are analyzed and measured in terms of the Gini Index which in turn gives weightage for each of the attributes which helps in determining the most important factors affecting the heart disease. Then the Z statistics is used to analyze the wellness of the rule obtained. Finally, based on the rules generated the classifiers are built and the accuracy of the dataset is measured.

2. RESEARCH METHOD

All most all the hospitals today maintain vast amount of healthcare information obtained from patients, from electronic devices, medical diagnosis etc. This huge amount of data consists a lot of useful hidden data in them which is very essential in diagnosing various ailments with a minimum number of medical tests and make the diagnosis more intelligent. Many a times we limit ourselves with minimal attributes that are required to predict a patient with heart disease. By doing so we are missing on a lot of important features that are main causes for heart diseases.

Taking this into consideration, an approach is proposed wherein all the attributes responsible for the heart ailment in patients are considered and analyzed in a step by step procedure to see how the attributes play a vital role in predicting heart ailment in patients. Hence, the pre-dominant objective is to provide a more efficient approach towards the prediction of heart ailments with the most important attributes causing heart ailment and analyze the data, as to how the prediction accuracies fluctuate with minimal and maximal attributes taken into consideration using the data mining approach.

For which a benchmark heart ailment dataset i.e. Cleveland heart disease dataset along with it two more minimal datasets with same class label are used. Medical terminologies such as sex, chest pain type, and age etc., making up for 7 attributes in first dataset, 10 attributes in second dataset and 13 attributes in the benchmark dataset are used. To improvise the results, step by step approach is followed in increasing the number of attributes and analyze to see if the prediction accuracies reduced or gradually increased. The data mining classification methods used to classify the datasets are Decision Trees (Random Forest and Random Tree) and Naive Bayes are used.

This proposed approach finds its application in predicting the heart ailments in patients as it considers all the important attributes or features responsible in causing a heart ailment. It reduces the several tests and time taken to predict and diagnose patients with heart diseases and is also cost effective.

2.1. Data classifiers used

The different data mining classifiers used are Naïve Bayes, Decision trees Random Forest and Random Tree.

2.1.1. Naïve Bayes

This algorithm has its basis from the Bayes theorem and follows the concept of conditional independence i.e., it considers an attribute value of a given class to be independent of other attributes values. The main objective is to maximize the posterior probability in predicting the class [19]. The Bayes theorem is as follows:

Let X={x1, x2, ..., xm} be a set of m attributes. In Bayesian, X is considered as evidence and H be some hypothesis that the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given the evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as

$$P(H|X) = P(X|H)\,P(H) / P(X)$$

Naïve Bayes classifier performs relatively better with nominal data but not with numeric data [20].

2.1.2. Random tree

It uses the concept of divide and conquer in order to build decision trees [21]. At each node of the tree, the algorithm takes up a feature that further divides the data into subsets with the help of a splitting algorithm [22]. Every predicting node leads to a class or decision rule. In order to classify a new item, it should build a decision tree taking into consideration the existing items in the training set. With the help of information gain it finds the dependent variable that clearly discriminates other instances. Now among the possible values of this variable if any of the value turns out to be the target value or leaf node and we can terminate that branch. Hence we follow this method until we get the decision rules of different combination of features leads to the target value.

2.1.3. Random forest

It is a classifier which consists of n number of decision trees where in each individual result is depicted as a separate tree. It was derived from the random decision of forest that was proposed by Tin Kam Ho of Bell Labs in 1995 [23]. This approach does the selection of attributes randomly in order to build the decision trees but with restricted fluctuations. Multi-classifiers are the after effect of combining a few independent classifiers and sets of classifiers responsible for broadening the execution have been depicted [24].

2.2. Architectural diagram

Figure 1 shown architectural diagram of proposed work.

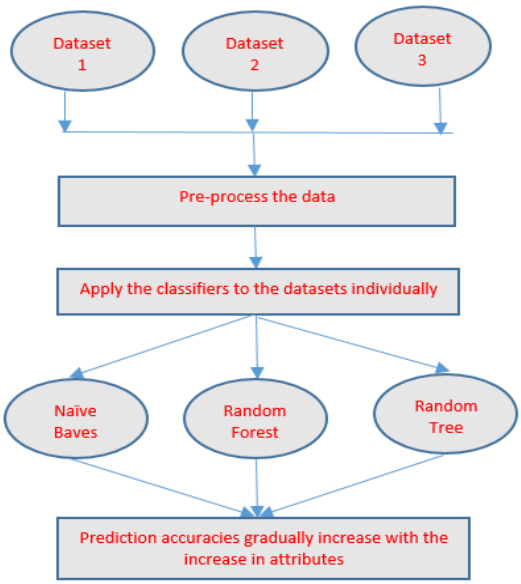

Figure 1. Architectural diagram of proposed work

3. RESULTS AND ANALYSIS

In Waikato Environment for Knowledge Analysis (WEKA) is used for analyzing and applying the different classifiers on the datasets because of its excellence in discovering, analysis and predicting efficient patterns [25]. A detailed analysis is done in Weka wherein Table 1 represents the results obtained by applying the classifiers on a dataset of 7 attributes. Table 2 depicts the results obtained by applying the classifiers on a dataset of 10 attributes and Table 3 depicts the results obtained by applying the classifiers on a dataset of 13 attributes. Figure 2 depicts the gradual increase in the prediction accuracies along with the increase in the attributes.

The initial dataset (Dataset-1) consists of 7 attributes making up for 300 records namely age which is continuous attribute, sex which is categorical attribute, exercise induced agina which is nominal attribute, depression induced by exercise forming the categorical attribute, slope of peak exercise again a categorical attribute, number of major vessels a nominal attribute and finally the defect type which is categorical again.

The second dataset (Dataset-2) consists of all the first dataset attributes along with additional 3 more attributes namely the chest pain type which is a categorical attribute, cholesterol which is continuous and

resting ECG measure again categorical attribute totally making up for 290 records. Dataset-1 and Dataset-2 are collected by taking into consideration the various heart ailment attributes through a survey done with the patients. Finally, the third dataset Cleveland Heart Disease dataset consists of 303 records with 13 attributes is used. Predictable attribute (num) is a common class label for all the three datasets wherein,

Value = 0: depicts that the patient has no heart disease
Value = 1: depicts that the patient has heart disease.

All the dataset are analyzed and the classifiers are applied on a basis of 70% split i.e., 70% of dataset is used as training data and 30% as testing data.

Step – 1: Pre-processing of datasets.

Dataset-1 with 7 attributes didn't have any missing values and so does the second dataset. But the third Dataset-3 had missing values which was rectified by using the Weka filter ReplaceMissingValues. Since the class variable is the same for all the datasets it is converted from numeric to nominal using the filter numeric to nominal under the unsupervised attribute filters in order to make few of the classifiers like Naïve Bayes and Decision trees to be applicable. Further the multi-class classification in the third benchmark dataset is converted two-class classification. this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [2, 5].

Results are analyzed by applying the three classifiers Naïve Bayes, Random Tree and Random Forest on all the three datasets to see if there are fluctuations in the prediction accuracies after applying them to build the models.

Step – 2: Apply the 3 classifiers on the dataset with 7 attributes.

Initially, the Dataset-1 with 7 attributes is pre-processed and the class label is converted to nominal using the filter Numeric to Nominal on which the 3 classifiers are applied. It is evident from Table 1 that the Random forest algorithm outperforms with 7 attributes.

Table 1. Accuracies achieved with 7 attributes

Evaluation criteria	Classifiers		
	Naïve Bayes	Random Forest	Random Tree
Time taken to build model (in sec)	0	0.25	0
Correctly classified instances	71	72	68
Incorrectly classified instances	20	19	23
Accuracy (%)	78.022%	79.1209%	74.7253%

Step – 3: Apply the 3 classifiers on the dataset with 10 attributes.

Secondly, the Dataset-2 with 10 attributes is pre-processed and the class label is converted to nominal using the filter Numeric to Nominal on which the 3 classifiers are applied.

It is evident from Table 2 that the Naïve Bayes algorithm outperforms and the accuracies of all the classifiers are gradually increased with the addition of 3 more attributes chest pain type, cholesterol and ECG measure to the Dataset-1.

Table 2. Accuracies achieved with 10 attributes

Evaluation criteria	Classifiers		
	Naïve Bayes	Random Forest	Random Tree
Time taken to build model (in sec)	0	0.14	0
Correctly classified instances	72	71	67
Incorrectly classified instances	15	16	20
Accuracy (%)	82.7586%	81.6092%	77.0115%

Step – 4: Apply the 3 classifiers on the dataset with 13 attributes.

Finally, the Dataset-3 with 13 attributes is pre-processed and the class label is converted to nominal using the filter Numeric to Nominal on which the 3 classifiers are applied.

It is evident from Table 3 that the Random Forest algorithm outperforms and the accuracies of all the classifiers are gradually increased with the addition of 3 more attributes blood pressure, fasting sugar and thalach to the Dataset-2.

Table 3. Accuracies achieved with 13 attributes

Evaluation criteria	Classifiers		
	Naïve Bayes	Random Forest	Random Tree
Time taken to build model (in sec)	0	0.14	0
Correctly classified instances	76	79	72
Incorrectly classified instances	15	12	19
Accuracy (%)	83.5165%	86.8132%	79.1209%

From the above Table 4 it is evident that with the increase in the number of attributes, the prediction accuracies have gradually increased irrespective of the classifiers used and the dataset values. Moreover, for different datasets different classifiers outperforms i.e., in case of Dataset-1 with 7 attributes Random Forest classifier outperforms with an accuracy of 79.1209%, whereas for Dataset-2 with 10 attributes Naïve Bayes classifier outperforms with an accuracy of 82.7586% and finally for the Dataset-3 with 13 attributes the Random Forest attribute outperforms. Hence, it clearly shows that consideration of all the important attributes or features contribute to a better and accurate prediction of heart diseases in patients.

Table 4. Accuracy fluctuations among the 3 datasets

Datasets	Classifiers		
	Naïve Bayes	Random Forest	Random Tree
Dataset-1	78.022%	79.1209%	74.7253%
Dataset-2	82.7586%	81.6092%	77.0115%
Dataset-3	83.5165%	86.8132%	79.1209%

From the Figure 2 it is evident that there is a gradual increase in the prediction accuracies analyzed on the three datasets with different number of attributes and values using the above three classifiers. It even depicts that in case of Dataset-1 Random forest algorithm outperforms, for Dataset-2 Naïve Bayes algorithm outperforms with an accuracy of 82.75% and for Dataset-3 again Random forest outperforms with an accuracy of 86.81%.

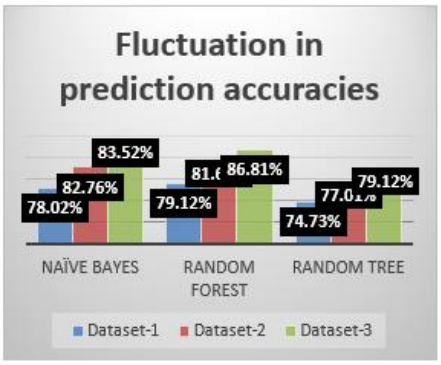

Figure 2. Fluctuations in prediction accuracies

4. CONCLUSION

Finally, to conclude from the results and analysis performed in this paper it is evident that each and every attribute contribute uniquely to the prediction accuracy and hence with the step by step increase in the heart ailment attributes, the prediction accuracy of heart ailment gradually increases irrespective of the classifier used to build the model. Adding to this it is also evident that for different dataset with different values and number of attributes different classifiers out performs each time.

Future work can be expanded by finding the most appropriate features required for prediction of heart ailment and use other data mining methods such as neural networks, regression etc., for prediction of heart disease.

REFERENCES

[1] H. Yan, et al., "Development of a decision support system for heart disease diagnosis using multilayer perceptron," *Proceedings of the 2003 International Symposium on*, vol. 5, pp. V-709- V-712, 2003.

[2] J. Sandhya, et al., Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques," *International Journal of Engineering and Technology*, vol/issue: 2(4), 2010.

[3] J. Patel, et al., "Heart Disease Prediction using Machine Learning and Data Mining Techniques," Nirma University, Gujarat, India *IJCSC*, vol/issue: 7(1), pp. 129-137, 2016.

[4] Webmd.com, "Risk factors for heart disease," 2014. http://www.webmd.com/heart-disease/risk-factors-heart-disease.

[5] National Heart, Lung, and Blood Institute, "What is coronary heart disease?" 2014. http://www.nhlbi.nih.gov/health/healthtopics/ topics/cad/.

[6] R. Das, et al., "Effective diagnosis of heart disease through neural networks ensembles," *Expert Systems with Applications, Elsevier*, vol. 36, pp. 7675-7680, 2009.

[7] S. Panzarasa, et al., "Data mining techniques for analyzing stroke care processes," *Proceedings of the 13th World Congress on Medical Informatics*, 2010.

[8] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208-217, 2013.

[9] K. Srinivas, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," *IEEE Transaction on Computer Science and Education (ICCSE)*, pp. 1344-1349, 2010.

[10] D. C. J. Ruben, "Data Mining in Healthcare: Current Applications and Issues," 2009.

[11] R. El-Bialy, et al., "Feature Analysis of Coronery Artery Heart Disease Data Sets," *International Conference on Communication, Management and Information Technology (ICCMIT 2015), Procedia Computer Science*, vol. 65, pp. 459-468, 2015.

[12] V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases," *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, vol/issue: 2(4), pp. 56-66, 2013.

[13] C. S. Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," *International Journal of Computer Applications*, vol/issue: 47(10), 2012.

[14] M. Anbarasi, et al., "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *International Journal of Engineering Science and Technology*, vol/issue: 2(10), pp. 5370-537, 2010.

[15] J. Rohilla and P. Gulia, "Analysis of Data Mining Techniques for Diagnosing Heart Disease," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol/issue: 5(7), 2015.

[16] R. Tamilarasi and R. Porkodi, "A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare," *International Journal of Emerging Research in Management &Technology*, vol/issue: 4(3), 2015.

[17] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using naïve bayes," *International Journal of Advanced Computer and Mathematical Sciences*, 2012.

[18] M. A. Jabbers, et al., "Heart Disease Prediction System using Associative Classification and Genetic Algorithm," *International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies*, 2012.

[19] M. Bramer, "Principles of Data Mining," *Springer-Verlag*, 2007.

[20] R. Alizadehsani, et al., "Diagnosis of coronary arteries stenosis using data mining," *J Med Signals Sens*, vol. 2, pp. 153-9, 2012.

[21] J. Nahar, et al., "Computational Intelligence for heart disease diagnosis: A medical Knowledge driven approach," *Expert Systems with Applications*, vol/issue: 40(1), pp. 96-104, 2013.

[22] M. Kumari and S. Godara, "Comparative study of data mining classification methodsin cardiovascular disease prediction," *IJCST*, vol. 2, pp. 304-305, 2011.

[23] Y. E. Shao, et al., "Hybrid intelligent modelling, schemes for heart disease classification," *Applied Soft Computing*, vol. 14, pp. 47-52, 2014.

[24] P. V. A. Makwana, "Identify the patients at high risk of re-admissionin hospital in the next year," *International Journal of Science andResearch*, vol. 4, pp. 2431-2434, 2015.

[25] A. Aziz, et al., "Mining Students' academic Performance," *Journal of Theoretical & Applied Information Technology*, vol/issue: 53(3), 2013.