American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-7, Issue-2, pp-278-283
www.ajer.org
Research Paper Open Access

Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction

Md. Fazle Rabbi[1], Md. Palash Uddin[1], Md. Arshad Ali[1], Md. Faruk Kibria[2], Masud Ibn Afjal[1], Md. Safiqul Islam[2]and Adiba Mahjabin Nitu[1]

[1](*Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200, Bangladesh)*

[2](*Department of Electrical and Electronic Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200, Bangladesh)*

Corresponding author: Md. Fazle Rabbi

ABSTRACT: *Heart disease might be one of the foremost causes to death. Because of the lack of skilled knowledge or experiences of real-life practitioners about heart failure symptoms for an early prediction, it is not an easy task to detect the disease. Consequently, computer-based prediction of heart disease may play a significant role as a pre-stage detection to take proper actions with a view to recovering from it. However, the choice of the proper data mining classification method can effectively predict the early stage of the disease for being recurred from it. In this paper, the three mostly used classification techniques such as support vector machine (SVM), k-nearest neighbor (KNN) and artificial neural network (ANN) have been studied with a view to evaluating them for heart disease prediction using Cleveland standard heart disease dataset. The experimental result shows that the classification accuracy using SVM (85.1852%) outperforms that of using KNN (82.963%) and ANN (73.3333%).*

KEYWORDS*-Data Mining, SVM, KNN, ANN, Heart Disease Prediction, Classification Techniques*

Date of Submission: 13-02-2018 Date of acceptance: 28-02-2018

I. INTRODUCTION

In recent years, the volume of computerized medical data is increasing rapidly [1]. However still, it is a complex task to manipulate the bulky amount of the data for extracting knowledge from it. However, data mining technique, an important filed of machine learning, can be used to explore the meaningful information from such kind of medical data repository. Furthermore, the data mining technique can broadly be applied for versatile applications such as classification, clustering, regression, prediction etc. [2].

Nevertheless, the heart disease is a vital issue to be fixed for sound human life. Though, real-life consultants can be able to predict the disease with an enormous number of tests and requiring a huge processing time, sometimes, their prediction may be incorrect because of lack of skilled knowledge and proper experiencesregarding this [1]. Consequently, it is obvious that computer-based prediction of heart disease can be more effective and time saving way for the better humanity. Consequently, data mining classification techniques are broadly applied to discover the early stage of the heart disease prediction. Since the development of the efficient classification technique is growing rapidly for various types of classification tasks, it is important to choose the appropriate classification approach for effective heart disease prediction [3]. From this motivation, in this paper, the mostly used classification techniques e.g., support vector machine (SVM), *k*-nearest neighbor (KNN), and artificial neural network (ANN) have been studied and compared for heart disease prediction using Cleveland heart disease dataset.

II. BACKGROUND REVIEW

There are several works on the heart disease predictions performed by different researchers. The authors in [3] presented heart disease prediction using several classification techniques in which Bayes classifier shows an accuracy of 86.12% while ANN shows an accuracy of 85.68% and decision tree learning shows an accuracy of 80.4%. On the other hand, in [4] the performances of decision tree learning (79.3%), logistic regression (77.7%) and ANN (80.2%) have been compared for predicting a patient of the heart disease. The authors in [5] performed an analysis of the performances of K-Star (75.1852%), J48 (76.6667%), SMO (84.0741%), Bayes Net (81.1111%) and MLP (77.4074%) for predicting the heart disease patients. In [6] the performances of Random Forest (91.6%), C4.5 (89.6%), SVM (89.2%), Bayes classifier (85.2%), AdaBoost (82.8%) have been compared for predicting the cardiovascular heart disease patients. The authors in [1] have compared the performances of decision tree learning (77.55%), Naïve Bayes classifier (83.49%), KNN (k=1: 76.23%, k=3: 81.18%, k=9: 83.16%, k=15: 83.16%), MLP (82.83%), RBF (83.82%), Single Conjunctive Rule Learner (69.96%) and SVM (84.15%). In [7] it has been analyzed that the conventional logistic regression approach can provide better result than regression trees. The author in [8] has equated the performances of logistic regression and random forest approach for predicting the risk level of the heart disease patients in which the logistic regression (89%) technique can provide better performance than the random forest (88%). Also, the authors in [9] have considered different feature selection approaches and measured the performance of the Naïve Bayes for the diagnosis of heart disease patients. However, in this paper, the prediction performance of the widely used classifiers such as SVM, KNN and ANN has been analyzed and compared using standard Cleveland heart disease dataset.

III. INVESTIGATED CLASSIFICATION TECHNIQUES

3.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning model that is defined as the finite dimensional vector spaces where each dimension characterizes a feature of a particular object. In this way, SVM has been proved as an effective method in high-dimensional space problems. Due to its computational competence on huge datasets SVM is typically used in document classification, sentiment analysis and prediction-based tasks [1], [6], [10].

3.2. K-Nearest Neighbors (KNN)

K-Nearest Neighbor (KNN), a supervised learning model as well, is used to classify the test data using the training samples directly. In KNN, an object is classified by the majority voting of its closest neighbors. Alternatively, the class of a new sample is predicted based on some distance metrics where the distance metric can be a simple Euclidean distance. In the working steps, KNN first calculates k (No. of the nearest neighbors). After that, it finds the distance between the training data and then sorts the distance. Subsequently, a class label will be assigned to the test data based on the majority voting [1].

3.3. Artificial Neural Network (ANN)

The Artificial Neural Network (ANN), also a supervised learning strategy, contains three layers: input, hidden and output. The connection between the input units and the hidden and the output units are based on relevance of the assigned weight of that specific input unit. Usually, if the weight is higher, then it is considered more important. ANN may use linear and sigmoid transfer (activation) functions. Also, the ANNs are suitable for the training of large amounts of data with limited inputs. For multi-layer feed forward ANN, the mostly used learning algorithm is the Backpropagation learning tool [4], [5]. In ANN, the input data records should be separated into three sub-datasets for the purpose of training, validation and testing.

IV. EXPERIMENTAL RESULT ANALYSIS

4.1 Working Procedure

The working procedure of the proposed evaluation scheme of the studied classification techniques for heart disease prediction is illustrated in Fig. 1.

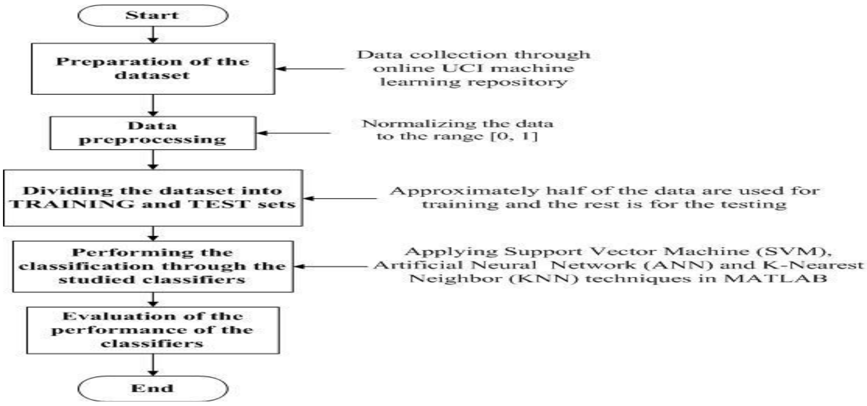

Figure 1: Proposed work flow

4.2 Dataset Description

In this paper, the Cleveland standard heart disease dataset is gathered from the UCI machine learning repository [11]. Although there are total 270 records of 76 different attributes along with the true sample label in the dataset, most of the published experiments has referred to using a subset of 13 attributes. The used 13 attributes with respective explanation is shown in Table 1. In this experiment, approximately half of the data are used for training and the rest is for the testing.

4.3 Experiment Setup and Result Analysis

For the classification of the heart disease dataset using support vector machine (SVM) with RBF kernel, the commonly used MATLAB LibSVM package [13] has been setup. The well-known 10-fold cross validation procedure has been used to select the best C and gamma (g) parameters for the efficient training and testing [12]. Table 2 shows the classification performance using SVM with the best C and g values.

On the other hand, the MATLAB KNN has been applied where Euclidean distance is measured to specify the distance metric between the character vectors. Table 3 shows the classification result for different values of number of neighbors (k).

Finally, the MATLAB multilayered feed-forward Backpropagation ANN has been applied on the dataset where the hidden layer takes the input from the input data and the output layer forms the outputs. The number of the hidden neurons in the proposed network structure is experimentally set to 3 that results the size of the network as 13×3×1. In this network, the 10-fold cross validation process is also used for efficient learning. For this ANN, the total number of the training data is further divided into three subgroup datasets as 45% for training, 5% for validation and the rest for testing purposes. After that, the main testing dataset is used for the testing operation using the trained network. The summarized classification result is shown in Table 4.

Table 1:Description of the Cleveland dataset

No. of the Attribute	Name of the Attribute	Explanation
1.	Age	age in years
2.	Sex	sex (1 = male; 0 = female)
3.	cp	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	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

Table 2: Classification result using SVM

C, g	**Classification Accuracy (%)**
1, 0.1	85.1852

Table 3: Classification result using KNN

k	**Classification Accuracy (%)**
1	79.2593
2	79.2593
3	81.4815
4	82.963
5	80.00
6	79.2593
7	80.00
8	80.00
9	80.7407
10	80.7407

Table 4: Classification result using ANN

Classifier	**Classification Accuracy (%)**
ANN	73.3333

Table 5: Confusion matrices using the classifiers

Classifier	**Confusion matrix**	
SVM	68	13
	7	47
KNN (*k*=4)	67	15
	8	45
ANN	69	29
	6	31

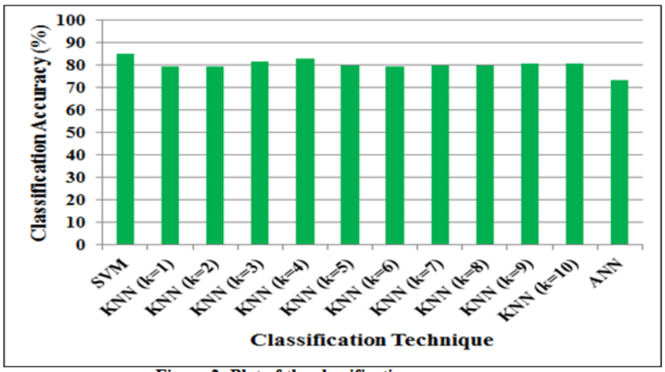

Figure 2: Plot of the classification accuracy

Table 5 shows the confusion matrices produced from each of the classifiers. Now, the graphical representation of the classification result is illustrated in the graph of Fig. 2. From the above performance comparison tables and graph, it can be observed that the classification accuracy using SVM is much effective than that of the KNN and ANN. The reason behind that is that the SVM can tremendously fix the nonlinearity in the dataset for producing better classification performance than KNN and ANN. Also, the KNN has performed better than ANN as it separates the vector splendidly with *k*=4.

V. CONCLUSION

As heart disease is one of the vital causes to death, it should be correctly detected at very early stage to get recovery from it. Sometimes, real-life practitioner may not be able to detect the disease due to some lack of skilled knowledge and proper experiences. Thus, computer-based competently accurate prediction system may be an alternative to detect the heart disease for fixing it immediately. Hence, in this paper, three mostly used data mining classification techniques such as SVM, KNN and ANN have been studied and evaluated using standard Cleveland heart disease dataset. It has been analyzed that RBF kernel based SVM can outperform KNN and ANN on the basis of the classification rate while KNN is also offering better performance than ANN. This comparative study also recommends that the significantly evaluated classifier can be used for real-time prediction of heart disease patients and for predicting the risk factor of heart failure with a view to ensuring additional care so that early-stage heart failure can be avoided. However, more training data whether from hospitals or from domain-experts can be added for increasing the prediction performance of the classifiers. Moreover, diverse feature reduction strategies may also be applied on the dataset for getting improved performance.

REFERENCES

[1]. 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," *IEEE Symposium on Computers and Communication*, 2017.

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

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

[4]. A. Khemphila and 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*, pp. 193-199, 2010.

[5]. 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*, 2016.

[6]. 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," *International Conference on Big Data Analysis*, 2017.

[7]. P. C. Austin, J. V. Tu, J. E. Ho, D. Levy and 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*, 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," *International Conference on Engineering and Technology*, 2015.

[10]. 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.

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

[12]. C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, No. 3, pp. 27:1–27:27, 2011.

Md. Fazle Rabbi (rabbi@hstu.ac.bd) received his B.Sc. degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh in 2007. Now, he is pursuing M.Sc. degree in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka, Bangladesh. His main working interest is based on bioinformatics, image processing, data structures and algorithm etc. Currently, he is working as an assistant professor in Dept. of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. He has several scientific research publications in various aspects of Computer Science and Engineering.

Md. Palash Uddin (palash_cse@hstu.ac.bd), a member of IEEE, is presently serving as an Assistant Professor in department of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. Previously, he was a lecturer in the same university and in department of Computer Science and Engineering at Central Women's University, Dhaka, Bangladesh. He is currently pursuing his M. Sc. degree from department of Computer Science & Engineering, Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh. He received his B. Sc. degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh in 2011. His focal research interest is based on the remote sensing image analysis, artificial intelligence based application development, machine learning algorithms, algorithms analysis and design, data mining and combinatorial optimization. He has several national and international journal and conference publications in various fields of Computer Science and Technology. In 2017, he received the "best paper award" on the paper

tilted as "Feature Extraction for Hyperspectral Image Classification" in the prestigious IEEE 5th Region 10 Humanitarian Technology Conference (R10HTC) hosted by Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh on 21-23 December 2017.

Md. Arshad Ali was born in 1986. He received the B.S. degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University (HSTU), Bangladesh in 2007. He worked as an assistant professor in HSTU. Currently, he is a Master's course student in Okayama University, Japan. His research interest includes information security, AES, pseudo-random sequence, and homomorphic encryption. He is a member of IEEE.

Md. Faruk Kibria (mfkibria.eee@hstu.ac.bd) received his B.Sc. degree in Electrical and Electronic Engineering from Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh in 2009. Now, he is pursuing M.Sc. degree in Electrical and Electronic Engineering from Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh. His main working interest is based on Wireless communication, power electronics and renewable energy. Currently, he is working as an assistant professor in Dept. of Electrical and Electronic Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. He has several research publications in various aspects of Electrical and Electronic Engineering.

Masud Ibn Afjal (masud@hstu.ac.bd) received his B.Sc. degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh in 2008. His main working interest is based on the compression for remote sensing image, machine learning, data mining, and database design. Currently, he is working as an assistant professor in dept. of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. Previously, he was a lecturer in the same university and also served in various software development companies of Bangladesh. He has several research publications in different aspects of Computer Science and Engineering.

Md. Safiqul Islam (safiqul_eee@hstu.ac.bd), a member of IEEE, received his B.Sc. degree in Electrical and Electronic Engineering from Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh in 2005. Now, he is pursuing M.Sc. degree in Electrical and Electronic Engineering from Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh. His main working interest is based on Wireless communication, power electronics and renewable energy. Currently, he is working as an assistant professor in Dept. of Electrical and Electronic Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. He has several scientific research publications in various aspects of Electrical and Electronic Engineering.

Adiba Mahjabin Nitu (nitu.hstu@gmail.com) received her B. Sc. (Hons.) and M. Sc. in Computer Science and Engineering from Rajshahi University, Rajshahi, Bangladesh in 2003 and 2004 respectively. She has also obtained her M. Sc. degree in Computer Science from University of Northern British Columbia, Canada in 2015. Her main research interest is based on modelling and simulation for real–time systems, system design, programming, data structures etc. Currently, she is working as an associate professor in Dept. of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. She has several scientific research publications in various fields of Computer Science and Engineering. She is a member of IEEE.

Md. Fazle Rabbi."Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction" American Journal of Engineering Research (AJER), vol. 7, no. 2, 2018, pp. 278-283.