ACTA SCIENTIFIC NUTRITIONAL HEALTH

Volume 3 Issue 7 July 2019

Review Article

Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques

Monther Tarawneh* and Ossama Embarak

HCT, Alfujairah, UAE

***Corresponding Author:** Monther Tarawneh, HCT, Alfujairah, UAE.

Received: May 27, 2019; **Published:** June 20, 2019

Abstract

Heart disease is one of the significant reason of death and disability. The shortage of Doctors, experts and ignoring patient symptoms lead to big challenge that may cause death, disability to the patient. Therefore, we need expert system that serve as an analysis tool to discover hidden information and patterns in hear disease medical data. Data mining is a cognitive procedure of discovering the hidden approach patterns from large data set. The available massive data can used to extract useful information and relate all attributes to make a decision. Various techniques listed and tested here to understand the accuracy level of each. In previous studies, researchers expressed their effort on finding best prediction model. This paper proposes new heart disease prediction system that combine all techniques into one single algorithm, it called hybridization. The result confirm that accurate diagnose can be taken by using a combined model from all techniques.

Keywords: Hybrid; Heart Disease; Mining

Introduction

Heart disease is considered the main reason for death in the world. The heart disease diagnosis is the process of detecting or predicting heart disease from patient's records. Doctors may not able to diagnose patient properly in a short time, especially when the patients suffer from more than one disease. Therefore, heart disease diagnose is a complex task that require experience and knowledge. Improper diagnose may cause death, or disability to the patient. Disease prediction Model can support medical professionals and practitioners in predicting heart disease. The huge amount of data that can be collected using digital devices (by the patient itself of in hospital) can used with data mining techniques to diagnose patient and predict diseases. This paper analyses various types of classification and prediction techniques used in heart disease prediction. Also, propose a hybrid approach that combine all techniques into a single one to combine all functions and produce accurate diagnose.

Related work

Huge amount of medical data generated by healthcare devices is large and complex to be analyzed by traditional methods. Data mining is used to improve the process by discovering patterns and features in large complex data. Several techniques have been presented to give accurate medical diagnose for various diseases. Data mining techniques are uses to eliminate interrelated and redundant data and to find hidden pattern from data since the data are high dimensional. Most known datamining tasks are association rules, feature selection, classification, clustering, prediction and sequential patterns.

Classification techniques are widely used in healthcare, since they are capable of processing large set of data. The common used techniques in healthcare are Naïve Bayesian, support vector machine, Nearest neighbor, decision tree, Fuzzy logic, Fuzzy based neural network, Artificial neural network, and genetic algorithms[1]. achine learning with classification can be efficiently applied in medical applications for complex measurements. Modern classification techniques provide more intelligent and effective prediction techniques for heart disease [2]. Many studies have been stationing on prediction of heart diseases; by applying different data mining techniques based on some feature selection techniques in order to get accurate classification using the optimal features set.

ANFIS is a heart disease prediction model based on coactive neuro-fuzzy inference system[3]. The model diagnosed disease by using various techniques include neural network adaptive capabilities, fuzzy logic qualitative approach and genetic algorithm. ANFIS was evaluated in term of classification accuracy and training data, the result showed a great potential in heart disease prediction with a very little mean square error. ANFIS is an adaptive neuro fuzzy inference system to train the neural network in order to predict heart diseases and cancer in diabetic patients based on some factors like age, obesity and some other factors related to life style [4]. The input nodes in neural network are constructed based on the input attributes and the hidden nodes are used to classify given input based on training dataset. GAFL use genetic algorithm and fuzzy logic [5]. The role of genetic algorithm was to provide the optimal solution for the features selection problem to help the diagnosing system. Fuzzy logic role was to develop a classification model using

fuzzy inference system. The accuracy of this model was 86% using stratified k-fold technique with specificity and sensitivity 0.9 and 0.8 respectively. With adaptive group-based-k-nearest-neighbor algorithm (AGKNN) help along with feature subset selection using PSO. This model reduced the cost of various medical tests and helps the patients to take a preventative measures well in advanced and showed very good prediction accuracy compared to traditional methods. According to this experiment 0.4% of people are under the risk of both cancer and heart disease if they are having diabetes.

A simple and reliable features selection method proposed to determine heartbeat case using weighted principal component analysis (WPCA) method [6]. In pre-processing stage they enlarged the ECG signal's amplitude and eliminated the noises. The total accuracy for this model was 93.19%. A hybrid classification technique by [7] which is a correlation based filter selection algorithm and support vector machine classifier, were features ordered according to their absolute correlation value with respect to the class attribute. The top k-features have been selected from an ordered features list. Classification accuracy was measured using SVM classifier with and without extended features of the reductive dataset. This model observed high accuracy in the case of three of five high dimensional used datasets with very less number of features up to 100% using the proposed hybrid technique with square extended feature.

A prototype intelligent heart disease prediction system (IHDPS) developed [8] using decision tree, Naïve Bayes and neural network data mining techniques and utilizing 909 records with 15 features from Cleveland heart diseases dataset. IHDPS can predict the likelihood of patients getting a heart disease and enables a significant knowledge such as relationship between related heart disease medical factors in addition to its ability to answer a complex "what if" queries. The result showed that each of used data mining techniques has its own strength side in grasping the defined mining goals objectives. Naive bayes was the best classifier to predict heart disease with accuracy 95% and to identify the significant influence and relationship in the medical input associated with predictable heart disease. Decision tree showed the best accuracy 99.61% to identifying the impact and relationship between the medical attributes in relation to the heart disease predictable state. for the last two goals Naïve Bayes was the best to identify heart disease patients characteristics and to determine the attributes values that differentiate nodes favoring and disfavoring the predictable state for patients with and without heart disease .IHDPS was exploited DMX (data mining extension) query language for model creation, training, prediction and content access, and evaluated using lift chart and classification matrix method. The most effective model to predict heart disease patients was Naïve Bayes with 86.53% accuracy followed by neural network with less than 1% difference and decision tree get the best accuracy for predicting patients with no heart disease since the model was evaluated on data set contain patients with and without heart disease.

Two phases experiment has been done to understand how machine learning techniques can help in comprehending the level of risk associated with heart disease using information gain and gain ratio feature selection techniques [9].The 1st phase is to applying feature selection techniques on commonly used 13 attributes. The 2nd phase is to use 75 attributes related to heart anatomic structure such as heart blood vessels from four heart disease datasets and to test the used feature selection techniques accuracy using Naïve Bayes, decision tree, support vector machine, logistic regression, and multi-layer perception and adaboost classifiers. The use of 13 attribute was not enough to understand the risk level of heart disease unlike the use of 75 attribute with feature selection techniques which improve the classification accuracy. Decision tree with adaboost get the best accuracy 98% on three datasets.

A dataset with 303 instances and 54 features exploited to study the role of feature selection and data mining techniques in diagnosis of coronary artery disease [10]. Information gain and confidence were used to determine the features effectiveness and select the optimal set of them. SVM, Naïve Bayes, bagging algorithm and neural network data mining classification techniques was used to evaluate the proposed feature selection techniques in addition to confusion matrix to detect the sensitivity, specificity and accuracy. Researchers were used LAD (left anterior descending), LCD (left circumflex) and RCA (right coronary artery) recognizers to create three new features in order to recognize if the three major coronary arteries are blocked. The achieved accuracy of this experiment was 98.08%.

A novel approach for normal and coronary artery disease conditions automated detection using heart rate signal Introduced by [11]. Discrete wavelet transformation (DWT) was used to decompose the heart rate signals into frequency sub-hand, in total the 3rd level of detail coefficient will be 76 in number. Principle component analysis, linear discriminate analysis and independent component analysis dimensionality reduction techniques were applied on DWT extracted coefficient and reduced the dimensionality from 67 to 10 features (coefficient). The remaining features applied to SVM, Gaussian Mixture Model, Probabilistic Neural Network and KNN classifiers. The accuracy of this approach was 96.8%.

Back Propagation algorithm methods [12] compares classification techniques. The author efficiency and deliver high accuracy from the heart disease prediction. A comparative analysis of accuracy on heart disease prediction [13] used Naïve Bayes and SVM, logistic regression. The highest accuracy (80%) of heart disease prediction is SVM. Various types of clustering techniques compared on heart disease prediction [14], it shows that the best performance accuracy of cluster algorithm make density based cluster. Mai and her partners [15] compare 3 classification algorithm c4.5, j4.8 and

nagging algorithm and conclude that the best performance algorithm in heart disease prediction with 81.41% accuracy is bagging algorithm. The study was based on UCL repository heart disease data set.

Combining techniques in a hybrid model may result high accuracy. A hybrid approach that combine genetic and Naïve Bayes [16] produce high accuracy against others techniques, the authors investigate most used techniques and chose the best two. Another hybrid machine learning model comprising of genetic algorithm, SVM and regression analysis [17]. As we can see that each algorithm used by each technique has some functions that help in heart disease prediction. To get the best result, we can combine the output of each algorithm and compare, it is called hybridization [18].

All researchers made a major effort in present an optimal heart disease diagnosis system based various techniques. The achievement of high classification accuracy remains the common objective of all those different models which are expanded from simple to complex features types from supervised to unsupervised and semi-supervised features selection and from simple to more advanced techniques.

Proposed method

The proposed Method contains three phases as depicted in figure 1. Starts with preprocessing phase where data filtered and classified before any processing. The output of this phase goes into number of classification techniques where these techniques evaluated tom eliminate low performance one. Then we combine the result and look at the patient history to give a decision (negative/positive) of heart attack. The steps of each phase described in details below.

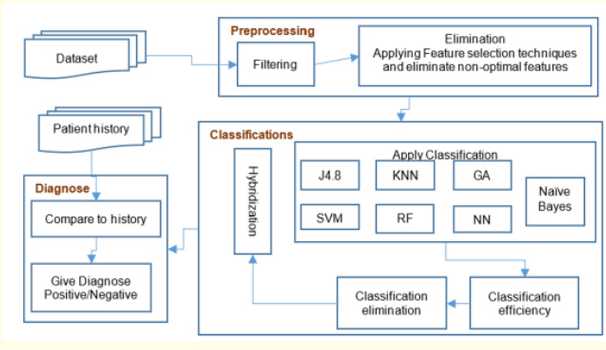

Figure 1: Overview of the proposed system.

Phase 1: Preprocessing

- Pass the data set through a filtering which replace all missing values with a value inferred from the column values using the concept of mutation in evolutionary algorithm. This a very important step especially for real dataset.
- Apply feature selection techniques (Information gain, gain ratio, reliefF, symmetrical uncertainty, and oneR feature selection) on the data set to eliminate non-optimal features. Features with zero rank or clearly low in comparison to others should be eliminated, only 2 features eliminated.

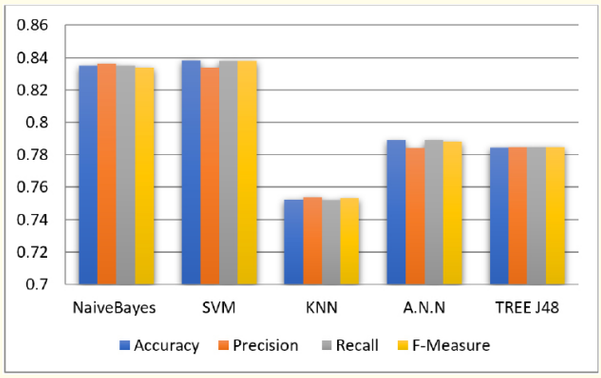

Figure 2: Classification of the original data.

Phase 1: Classification

- Apply number of classifications techniques on the output of the first phase.
- Classification accuracy, precision, recall and f-measure will be used to evaluate the efficiency of the used techniques, Figure 2 shows the classification results of the original data.
- Eliminate low efficiency algorithms based on the evaluations from previous step. This process done by comparing the values of accuracy, precision, recall and f-measure for each feature to determine the consistency of the classification on the data set. We notice that Naïve Bayes and SVM always perform better than others and never been eliminate, tree decision eliminated a couple of times. Where KNN is most of the time get eliminated.
- Apply Hybridization, where we combine the results from the chosen Classification.

Phase 3: Diagnose

- If the patient history is available, compare the result with the patient history. Patient profile can available in the hospitals or clouded by the health services in the country.
- Give a diagnose result with either positive or negative heart attack diagnose

Results

Our experiment applied on publicly available heart disease data set from UCI machine learning repository. Cleveland heart disease dataset contain fourteen numerical features and 303 instances [19]. This data set usually uses for researches about heart disease, we used this data set due its reasonable number of features and instances as well as its free from noisy data and contain a very little

missing data. We achieved accuracy of 89.2%. we were able to reduce the number of features from 14 to 12 without loss in the accuracy as computed by Naïve Bayes, SVM, KNN, NN, J4.8, RF, and GA on the whole data set.

We notice that Naïve Bayes and SVM are always give better accuracy than other classification techniques, where the last one is KNN. However, the hybridization approach is recommended since each technique has its own functions that help in heart disease prediction, and combining these techniques will combine all functions on all features. The performance study of different data mining algorithms in prediction of Heart disease as given below in the figure below.

The increase in water use efficiency due to treating the sandy soil with CKD could be attributed to the effect of CKD on increasing the water holding capacity of the treated soil and decreasing the water evaporation from the soil surface, hence. It increased the water available to the plants. The available water will be used in producing more plant materials and then more water use efficiency.

The proposed system can used in hospitals to help doctors make a quick diagnose or test new ones on some cases. Students in medical colleges may wish to use this system to learn and test their learning.

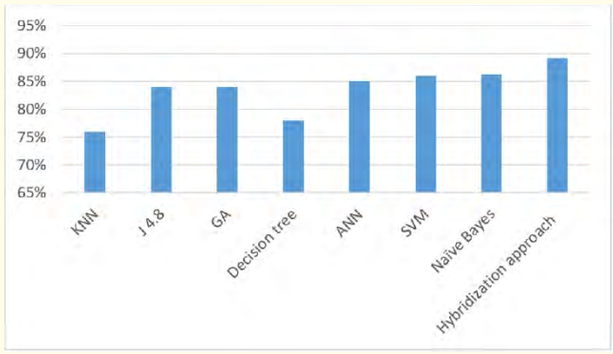

Figure 3: The performance study of different data mining algorithms.

Conclusion

Patient's electronic medical records usually contain relevant, irrelevant and redundant features. Doctors face a challenge to predict and diagnose heart diseases quickly and correctly due to this inefficient number of features. Different studies investigates feature selection techniques role in extracting the optimal set of features to being used in predicting and diagnoses heart disease with various methodologies and different classification accuracy.

Here we have investigated most of the used classification techniques in data mining and apply them on Cleveland data set. Some of the techniques perform better always such as Naïve Bayes and SVM, while others depend on the selected features. The main goal in this paper is to investigate available data mining techniques to predict heart disease and compare them, then combine the result from all of them to get most accurate result. The focus was on the classification and prediction methods. The accuracy of the algorithms can improved by hybridization or combining algorithm into single powerful algorithm. The new algorithm can be used as expert system in hospitals to help doctors in diagnose heart disease quickly and save life. Also, can used for education purpose in medical schools.

Future Work

In future, we are planning to introduce an efficient Remote heart disease prediction system to monitor and predict the heart disease based on the patient data collected from remote devices. Better accuracy is the main goal.

Bibliography

1. Kumari M and S Godara. "Comparative study of data mining classification methods in cardiovascular disease prediction". 2.2 (2011): 2229-4333.

2. Purusothaman G and P Krishnakumari. "A survey of data mining techniques on risk prediction: Heart disease". *Indian Journal of Science and Technology* 8.12 (2015).

3. Parthiban L and R Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm". *International Journal of Biological, Biomedical and Medical Sciences* 2008. 3.3 (2008).

4. Kalaiselvi C and G Nasira. "Prediction of heart diseases and cancer in diabetic patients using data mining techniques". *Indian Journal of Science and Technology* 8.14 (2015).

5. Santhanam T and E Ephzibah. "Heart disease prediction using hybrid genetic fuzzy model". *Indian Journal of Science and Technology* 8.9 (2015): 797-803.

6. Yeh YC., *et al*. "A reliable feature selection algorithm for determining heartbeat case using weighted principal component analysis". in System Science and Engineering (ICSSE), 2016 International Conference on. IEEE (2016).

7. Dubey VK and AK Saxena. "Hybrid classification model of correlation-based feature selection and support vector machine". in Current Trends in Advanced Computing (ICCTAC), IEEE International Conference on. IEEE (2016).

8. Krishnaiah V., *et al*. "Heart disease prediction system using data mining technique by fuzzy K-NN approach". *in Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI)* 1(2015): 371-384.

9. Dominic V., *et al*. "An effective performance analysis of machine learning techniques for cardiovascular disease". *Applied Medical Informatics* 36.1 (2015): 23-32.

10. Alizadehsani R., *et al*., "A data mining approach for diagnosis of coronary artery disease". *Computer Methods and Programs in Biomedicine* 111.1 (2013): 52-61.

11. Giri D., *et al*., "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform". *Knowledge-Based Systems* 37 (2013): 274-282.

12. Al-Milli N. "Backpropagation neural network for prediction of heart disease". *Journal of theoretical and applied information Technology* 56.1 (2013): 131-135.

13. Dbritto R., *et al*. "Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods". *International Journal of Applied Information Systems* (IJAIS)–ISSN (2016): 2249-0868.

14. Pandey AK., *et al*., "Datamining clustering techniques in the prediction of heart disease using attribute selection method". *heart disease* 14 (2013):16-17.

15. Shouman M., *et al*. "Using decision tree for diagnosing heart disease patients". in Proceedings of the Ninth Australasian Data Mining Conference-Volume 121. Australian Computer Society, Inc. (2011).

16. Singh N., *et al*. "Heart Disease Prediction System using Hybrid Technique of Data Mining Algorithms *International Journal of Advance Research, Ideas and Innovations in Technology* 4.2(2018).

17. Agrawal A., *et al*., "Disease Prediction Using Machine Learning". (2018).

18. Shirwalkar N., *et al*., "Human Heart Disease Prediction System Using Data Mining Techniques". *International Journal of Pure and Applied Mathematics* 120.6 (2018): 499-506.

19. Cleveland Database: https://archive.ics.uci.edu/ml/datasets/heart+Disease.

Volume 3 Issue 7 July 2019