Asian Journal of Information Technology 16 (7): 639-644, 2017
ISSN: 1682-3915

A Hybrid Heart Disease Prediction System using Evolutionary Learning Algorithms

[1]S. Mohandoss, [2]V. Sai Shanmuga Raja and [1]S.P. Rajagopalan
[1]Department of Computer Applications, Dr. MGR Educational and Research Institute University, Chennai, India
[2]Department of Computer Science and Engineering, Shanmuganathan Engineering College, Arasampatti, Pudukottai, Tamil Nadu, India

Abstract: Cardiovascular illness remains the greatest reason for deaths worldwide and the heart disease prediction at the early stage is significance. In this study, we propose a hybrid heart disease prediction system using evolutionary learning algorithms like cascaded neural network and Genetic algorithm. It is used for heart disease prediction at the early stage utilizing the patient's therapeutic record. The results are compared with the known supervised classifier Support Vector Machine (SVM). During classification, 13 attributes are given as input to the CNN classifier to predict the risk of heart illness. The proposed framework can be used as a guide by the doctors to predict the disease in a more productive way. The effectiveness of the classifier is tried utilizing the records gathered from 270 patients. The outcomes demonstrate that the Genetic based CNN classifier can anticipate the probability of patients with coronary illness in a more effective manner.

Key wards: Prediction, anticipate, SVM, algorithms, significance, attributes

INTRODUCTION

The analysis of disease is an essential employment in medication. The social insurance industry gathers gigantic measure of human services information and afterward they are mined to find concealed data for successful basic leadership. Cardiovascular ailments allude to any ailment that influences the cardiovascular framework. Therapeutic analysis is viewed as a huge errand that should be completed exactly and proficiently. The computerization of the same would be exceptionally gainful.

An intelligent heart disease prediction framework worked with the guide of information mining strategy like choice trees, guileless bayed and neural system was proposed by Palaniappan and Awang (2008). The outcome represented the unconventional quality of each of the procedures in fathoming the goals of the predetermined mining destinations.

Fuzzy model was used to optimize the parameters. Patil and Kumaraswamy (2009) introduced k-means clustering algorithm to extract the data appropriate to heart attack from the warehouse. In addition the pattern vital to heart attack were selected on basis of the computer significant weight age. Ordonez (2006) used association rules to improve heart disease prediction. Association rules were applied on a real data set contacting medical records of patient with heart disease and the risk factors were identified. Srinivas *et al*. (2010) applied data mining techniques to predict heart attack. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart attack. Based on the calculated significant weight age the frequent pattern having value greater than a predefined threshold were chosen for the valuable prediction of heart attack.

The main objective of this researcher is to build up an intelligent heart disease prediction system utilizing genetic algorithm based CNN with back propagation training algorithm using historical heart disease databases to settle on smart clinical choices which traditional decision supportive networks cannot. A few PC supported determination techniques have been proposed in the writing for the conclusion of heart attacks. A smart coronary illness forecast framework worked with the guide of information mining strategy like decision trees, Naive Byes and neural system was proposed by numerous specialists. The outcome outlined the comprehensive quality of each of the procedures in attaining the targets of the predefined mining destinations. It encouraged the foundation of indispensable information, e.g., patterns associated with coronary illness.

One of these promising strategies is Artificial Neural Networks (ANNs) which rise as a well performing method

Corresponding Author: S. Mohandoss, Department of Computer Applications, Dr. MGR Educational and Research Institute University, Chennai, India

for coronary illness expectation (Raut and Dudul, 2010). It is an exceedingly successful measure utilized as a part of classification problems and to tackle numerous critical issues like signal enhancement, identification and prediction of signals. ANNs has a vital component, since, it adapts to complex data preparing in data mining process. This makes it possible that the ANNs are connected in situations where there is difficult to make a strict mathematical model, however has an adequately illustrative set of samples. The other important characteristics of neural networks is their ability to sum up info data and to give correct answers for new information which makes them effective in taking care of complicated classification problems (Priya *et al*., 2013).

Artificial Neural Network (ANN) has the ability to learn complex issue in nonlinear situations. Even though numerous ANN procedures are effectively connected to heart attack prediction framework, the fixed number of hidden neurons has not been well practiced. Thus, we propose a novel technique to anticipate the cardiovascular disease by deciding the size and topology of the network naturally using genetic algorithm. A cascade neural network comprises of a cascade design in which hidden neurons are added to the network each one in turn and don't change after they have been included. The architecture adapts rapidly since it requires no back propagation of error signals through the associations of the network. The performance of the proposed model is tested and compared with the well known SVM classifier.

Literature review: Huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods hence calls for technological interventions, so as to simplify management of those data. The decision making can be improved by using data mining in discovering patterns and trends in large amounts of complex data. Several ways are carried out in finding efficient technique of medical diagnosis for various diseases. Popular data mining tasks are association rules, classification, clustering, prediction and sequential patterns.

Classification techniques are capable of processing a large amount of data. Classification is one of the most widely used methods of data mining in healthcare organization. The common classification techniques used in healthcare are bayesian networks, support vector machines, nearest neighbor method, decision trees, fuzzy logic, fuzzy based neural networks, artificial neural network, Genetic algorithms (Kumari and Godara, 2011).

In computer-aided heart disease diagnosis methods where the data is obtained from some other sources and is evaluated by computer based applications. Computers have usually been used to build knowledge based clinical decision support systems which used the knowledge from medical experts and transferring this knowledge into computer algorithms was done manually. This process is time consuming and really depends on the medical expert's opinion which may be subjective. To handle this problem, machine learning techniques have been developed to gain knowledge automatically from examples or raw data. Medical diagnosis is an important but complicated task that should be performed accurately and efficiently and its automation would be very useful.

By Bhuvaneswari and Kalaiselvi (2012) use Naive Bayes classifier in medical applications. Two of the well-known algorithms are used in data mining classification are Backpropagation Neural Network (BNN) and Nave Bayesian (NB) calculate the priors, the probability of the object among all objects based on the previous experience. Bayesian technique is constructed on the probability concept. The posterior from the prior is calculated by bayes rules. Depending on the precise nature of the probability model, Naive Bayes classifiers is used to trained very efficiently in a supervised learning setting.

By Shouman *et al*. (2012), combine different classifiers through voting to outperform other single classifiers. Decisions of multiple classifiers are associated by using aggregation technique called as voting. The idea of applying multiple classifier voting is to divide the training data into smaller equal subsets of data and building a classifier for each subset of data. The results show that applying voting could not enhance the k-nearest neighbor accuracy in the diagnosis of heart disease.

By Kumar (2013), proposed a method that uses components of fuzzy logic like fuzzification, advanced fuzzy resolution mechanism and defuzzification. Fuzzification is a process to transfer crisp values into fuzzy values. In the analysis of heart disease a fuzzy resolution mechanism uses predicted value with five layers, each layer has its own nodes. The results are tested with cleveland heart disease dataset. Fuzzy resolution mechanism was developed using MATLAB. Defuzzification process converts the fuzzy set into discrete values.

MATERIALS AND METHODS

Genetic Based Cascaded Neural Network (GBCNN): A CNN consists of a cascade architecture in which hidden neurons are added to the network one at a time and do not change after they have been added. It is called a cascade

because the output from all neurons already in the network feed into new neurons. As new neurons are added to the hidden layer, the learning algorithm attempts to maximize the magnitude of the correlation between the new neurons output and the residual error of the network which we are trying to minimize.

CNN are "self-organizing" networks. Cascade correlation network training is quite robust and good results usually can be obtained with little or no adjustment of parameters (Dheeba and Padma, 2007). The network begins with only input and output neurons. During the training process, neurons are selected from a pool of candidates and added to the hidden layer.

CNN consists of a layer of input units, one or more layers of hidden units and one output layer of units. The number of input and output units depends upon the application and requires experimentation to determine the best number of hidden units.

A vector of predictor variable values x_1, x_2, ... , x_p is presented to the input layer. In addition to the predictor variables there is a constant input of 1.0 called the bias that is fed to each of the hidden and output neurons, the biasis multiplied by a weight and added to the sum going into the neuron.

For regression type of problems there is only a single neuron in the output layer. Each output neuron receives values from all of the input neurons and all of the hidden layer neurons. For classification problems, a sigmoid transfer function is used as the activation function.

In the neural network, the hidden neuron can influence the error on the nodes to which their output is connected. It can greatly degrade the generalization capability of the neural network which leads to the significant deviation in prediction result to the problem. To overcome this an approach is proposed which is able to find minimum number of hidden nodes. The neural network training problem consists in determining the synaptic weights of a neural network to get the desired output for a set of input vectors. As the Genetic algorithm is able to find global optimize solution to the problem it can be used for the initialization of neural network weights. Thus, the proposed method with genetic-neural approach can be used to design system for the heart disease prediction.

This hybrid system uses back propagation algorithm for learning and training the neural network. The multi-layer neural network is optimize by calculating the number of nodes in hidden layer to minimize the over fitting which causes the overestimation of complexity of the target problem that leads to significant deviation in prediction. As the initialization of the neural network weights is a blind process which makes it difficult to find

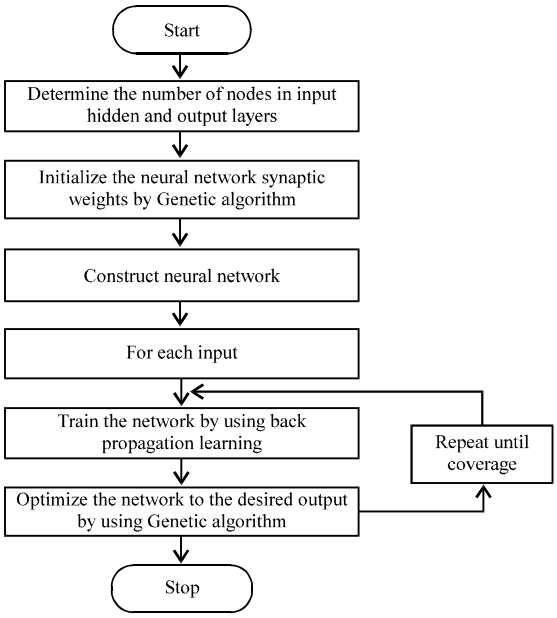

Fig. 1: Flowchart of genetic based CNN for heart disease prediction

out globally optimized initial weights and the network output would run towards local optima, hence, the overall tendency of the network to find out a global solution is greatly affected. So, the problem of local optimum solution is solved by optimizing the initial weights of neural network. For this a genetic algorithm is used which is specialized for global searching. Thus, the system uses the back propagation algorithm to train the network by using the weights optimized by genetic algorithm. Error is calculated using Eq. 1 to measure the differences between desired output and actual output that has been produced in feed forward phase. Error then propagated backward through the network from output layer to input layer as represented below. The weights are modified to reduce the error as the error is propagated:

$$\text{Error} = 1/2(\text{Output}_{\text{desired}} - \text{Output}_{\text{actual}})^2 \quad (1)$$

This process will be repeated iteratively until convergence is achieved (targeted learning error or maximum number of iteration). The genetic-neural approach for heart disease prediction is used to test data to the optimum value and predict whether the patient have a heart disease or not. The neural network with genetic algorithm approach for heart disease prediction is shown in Fig. 1.

The neural network uses the genetic algorithm fitness function to initializes the weights that makes it possible to have global optimal convergence. Neural network

architecture is constructed by identifying the input and output layer neuron along with number of hidden layers and hidden nodes identification. As Genetic algorithm is adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics which can be used to initialize the neural network weights. Thus, genetic-neural network takes advantage of global optimization of genetic algorithm for initialization of neural network. Then the genetic algorithm fitness function is used to predict the heart disease.

To create a new hidden unit, we begin with a candidate unit that receives trainable input connections from all of the network's external inputs and from all pre-existing hidden units. The goal of this adjustment is to maximize S:

$$S = \sum_o |\sum_p (V_p - \overline{V})(E_{p,o} - \overline{E_o})| \quad (2)$$

Where:

- o = The output unit
- p = Number of patterns in the training set
- V_p = The candidate units value at p
- $E_{p,o}$ = The residual error of all the training pattern at the output unit
- $\overline{V}$ and $\overline{E_o}$ = The values of V and E_o averaged over all patterns

GBCNN training: The training is done using back prop algorithm. Back prop uses a gradient descent method to update the weights. The algorithm 1 used to train the dataset is discussed:

Algorithm 1:

Step 1: Initialize the input and output units based on Genetic algorithm for the problem defined. The input and the output neurons are fully connected
Step 2: Train the network with input and output neurons until the residual error no longer decreases
Step 3: Select a temporary unit (Candidate unit) connected with the input unit and find the residual error
Step 4: Train this network unit S (Eq. 2) no longer improves
Step 5: Connect the temporary unit with the output unit and freeze its weights
Step 6: Train the input, output and the hidden unit until the residual error is minimized
Step 7: Repeat the step 2 to step 6 until the net error falls below a given value

Support vector machine: SVM is a set of related supervised learning method used in medical diagnosis for classification and regression. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. SVM is called maximum margin classifiers and it can be efficiently perform non-linear classification using kernel trick. An SVM Model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a large margin gap that is as wide as possible (Burges, 1998). Given labeled training data as data points of the form:

$$M = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \quad (3)$$

where, y_n = 1/-1, a constant that denotes the class to which that point x_n belongs. n = number of data sample a p-dimensional real vector. The SVM classifier first maps the input vectors into a decision value and then performs the classification using an appropriate threshold value. To view the training data, we divide (or separate) the hyper plane which can be described as:

$$\text{Mapping: } w^T.x+b = 0 \quad (4)$$

Where:

w = A p-dimensional weight vector
b = A scalar

The vector w points perpendicular to the separating hyper plane. The offset parameter b allows increasing the margin. In the absence of b, the hyper plane is forced to pass through the origin and restricts the solution.

RESULTS AND DISCUSSION

Comparison of results: The training and testing accuracy obtained using GBCNN is 78.55 and 85%, respectively. The training and testing set accuracy of SVM classifier is 75 and 82%, respectively. The training and testing set accuracy it is proved that the GBCNN classifier outperforms the other existing strategies in the literature. Moreover, it is observed that the proposed classifier achieved comparable performance over the testing and training set of all the patient records. The training set accuracy of the GBCNN classifier is shown in Fig. 2.

In cascaded correlation neural network the epochs got stopped when the desired accuracy got the system will stop immediately. The accuracy of the CNN is 85% and which is increased almost 3% compared to SVM. Form Fig. 3 and 4 it is noticed that the training set accuracy of the CNN classifier is high compared to SVM classifier. The training set accuracy of proposed SVM classifier is shown in Fig. 3.

Sensitivity, specificity and accuracy are the commonly used statistical measures to illustrate the medical diagnostic test and especially, used to enumerate how the test was good and consistent. Sensitivity evaluates the diagnostic test correctly at detecting a positive disease. Specificity measures how the proportion of patients without disease can be correctly ruled out.

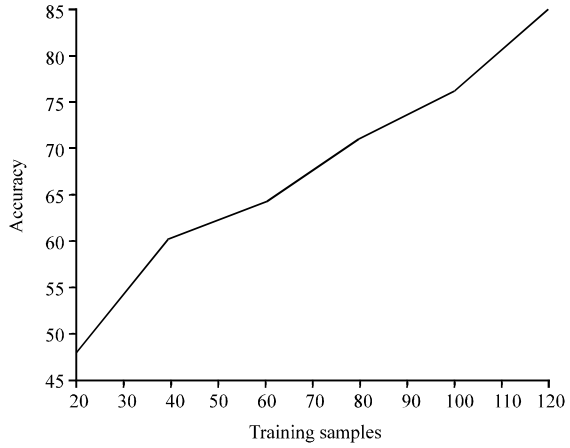

Fig. 2: GBCNN training set accuracy

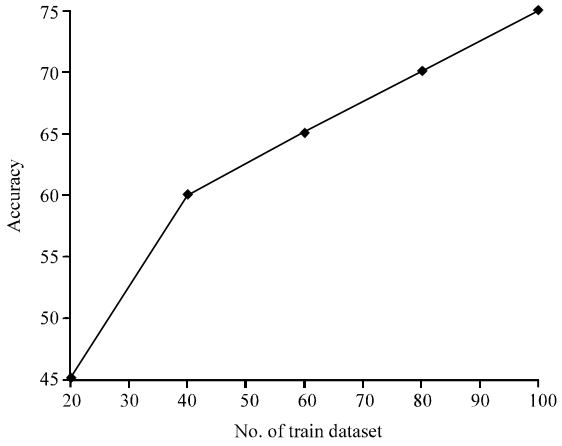

Fig. 3: Training set accuracy of heart disease dataset using SVM

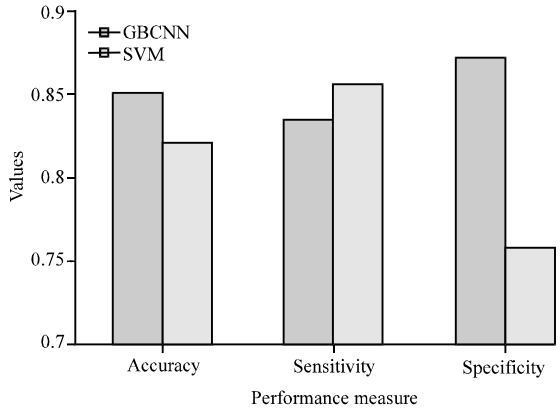

Fig. 4: Performance measure of SVM and CNN

The performance of the proposed heart attack prediction system is also analyzed by calculating the sensitivity and specificity. When a single test is performed, the person may in fact have the disease or the person may be disease free. The test result may be positive, indicating the presence of disease or the test result may be negative, indicating the absence of the disease.

Table 1: Performance measures for GBCNN and SVM

Measures	GBCNN	SVM
Accuracy	0.850	0.820
Sensitivity	0.835	0.855
Specificity	0.871	0.757

The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease. A high sensitivity is clearly important where the test is used to identify a serious but treatable disease. The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease. Sensitivity, specificity and accuracy were calculated by the formula given below:

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

Table 1 shows the specificity, sensitivity and accuracy of the discussed classifiers. Accuracy measures correctly figured out the diagnostic test by eliminating a given condition.

The specificity of the GBCNN classifier is comparatively high compared to SVM and hence, the GBCNN classifier can predict the patient without disease. But the sensitivity of the SVM is little bit high compared to GBCNN and indicated the ability of disease prediction is high for SVM. But the overall accuracy is high for GBCNN. Hence, the GBCNN classifier is suggested for heart disease prediction with higher accuracy.

The performance measures such as accuracy, specificity and sensitivity in percentage of the GBCNN classifier is compared with the SVM classifier. The comparative performance measure is shown in Fig. 4. It is clearly proved that the proposed system with GBCNN gives higher accuracy, specificity and sensitivity. False positive rate and true positive rate gives relative trade-off between true positive and false positive:

$$\text{True Positive rate} = TP/(TP+FN)$$

$$\text{False Positive rate} = FP/(FP+TN)$$

Table 2 shows true positive rate and false positive rate for SVM and GBCNN. The true positive rate there is

Table 2: True positive rate and false positive rate

Variables	True positive rate	False positive rate
GBCNN	0.850	0.164
SVM	0.812	0.221

a small difference between two classifiers but the false positive rate of CNN is significantly less compared to SVM classifier. Hence, GBCNN can be suggested for disease prediction in accurate manner compared with other classifier.

In clinical research a common classification type is a binary classification which predicts whether a disease is present or not. The performance of the classifiers SVM and CNN is compared based on the accuracy, sensitivity and specificity of the results obtained. For prediction of the disease and for the purpose of comparative analysis the standard heart disease data set taken from California University is used. At first, the attributes are collected from the dataset and then classifiers are built based on the input attributes. The true positive rate and false positive rate is computed for both classifier and the obtained results are shown in Table 2.

CONCLUSION

In this study, we have proposed supervised learning algorithm for finding the risk of heart disease of a patient using the profiles collected from the patients. GBCNN has a distinct feature that it does not use a predefined set of hidden units, instead the hidden units gets added up one by one until the error is minimized using genetic algorithm. By exploiting this distinct feature of the GBCNN, a computerized prediction algorithm is developed that are not only accurate but also computationally efficient for heart attack prediction with the proper adaptation of GBCNN classifies, the method can thus evolve an optimum number of hidden units within an architecture space. The results of classification experiment, preformed over data sets obtain from 270 patients shows that GBCNN classifier has achieved better accuracy than SVM classifier. In summary, it's found that the proposed hybrid heart disease prediction system using Genetic Based Cascaded Neural Network (GBCNN) offers substantial improvement in prediction of heart disease. The results of the experiments shows that an improvement in accuracy has been achieved by the proposed system. This implies that the cascaded neural network combined with Genetic algorithm is one of the desirable classifier which can be used as an aid for the physicians to predict the heart diseases in a more efficient way.

REFERENCES

Bhuvaneswari, R. and K. Kalaiselvi, 2012. Naive bayesian classification approach in healthcare applications. Intl. J. Comput. Sci. Telecommun., 3: 106-112.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining Knowl. Discov., 2: 121-167.

Dheeba, J. and A. Padma, 2007. Intelligent adaptive noise cancellation using cascaded correlation neural networks. Proceedings of the International Conference on Signal Processing, Communications and Networking ICSCN'07, February 22-24, 2007, IEEE, India, ISBN:1-4244-0996-9, pp: 178-182.

Kumar, D.A.S., 2013. Diagnosis of heart disease using advanced fuzzy resolution mechanism. Intl. J. Sci. Appl. Inf. Technol., 2: 22-30.

Kumari, M. and S. Godara, 2011. Comparative study of data mining classification methods in cardiovascular disease prediction. Intl. J. Comput. Sci. Technol., 2: 304-308.

Ordonez, C., 2006. Association rule discovery with the train and test approach for heart disease prediction. IEEE. Trans. Inf. Technol. Biomed., 10: 334-343.

Palaniappan, S. and R. Awang, 2008. Intelligent heart disease prediction system using data mining techniques. Proceedings of the International Conference on Computer Systems and Applications, March 31-April 4, 2008, Doha, pp: 108-115.

Patil, S.B. and Y.S. Kumaraswamy, 2009. Intelligent and effective heart attack prediction system using data mining and artificial neural network. Eur. J. Sci. Res., 31: 642-656.

Priya, K., T. Manju and R. Chitra, 2013. Predictive model of stroke disease using hybrid neuro-genetic approach. Intl. J. Eng. Comput. Sci., 2: 781-788.

Raut, R. and S.V. Dudul, 2010. Intelligent diagnosis of heart diseases using neural network approach. Intl. J. Comput. Appl., 1: 97-102.

Shouman, M., T. Turner and R. Stocker, 2012. Applying k-nearest neighbour in diagnosing heart disease patients. Int. J. Inform. Educ. Technol., 2: 220-223.

Srinivas, K., B.K. Rani and A. Govrdhan, 2010. Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng., 2: 250-255.