Heart Disease Prediction (Medical Diagnosis) using Data Mining

Anupam Choudhary\textsuperscript{a}, Asad Siddiqui\textsuperscript{b}, Huzaifa Vakil\textsuperscript{c}, Sohel Tharani\textsuperscript{d}, Zain Momin\textsuperscript{e}

\textsuperscript{a}Student, Rizvi College of Engineering, University of Mumbai, Mumbai
\textsuperscript{b,c,d,e}Asst. Prof and HOD, Dept. of Computer Engineering, Rizvi College of Engineering, Mumbai

\textsuperscript{b}asad-siddiqui@hotmail.com, \textsuperscript{c}vakilhuzaifa@gmail.com, \textsuperscript{d}sohel.tharani786@gmail.com, \textsuperscript{e}zmomin29@gmail.com

Abstract: Medical domain application development is a rapidly expanding area of research. These applications have been very useful in designing clinical support systems because of the ability to find patterns in medical data. Important application of such system is the diagnosis of diseases in the heart, which has a high mortality rate all around the world. One of the data mining techniques used for medical data are neural networks. Model selection for a neural network, which automatically obtains knowledge from the patient’s clinical data, uses selection of optimal number of hidden nodes, relevant input variables and optimal connection weights. In this paper, the use of Multi-Layer Feed Forward Neural Network that incorporates Genetic Algorithm and Back Propagation network (BPN) for heart attack projection is shown. GA initializes and optimizes the connection weights of neural network. Risk factors such as age, family history, diabetes, hypertension, high cholesterol, tobacco smoking, alcohol intake, and physical activity are utilized by system. Patients with heart disease have a lot of these negative risk factors thus allowing system to come to a conclusion. This system based on such factors will allow a doctor to find an early diagnosis of heart disease before complex and expensive and time-consuming traditional methods. This system was created in MATLAB and has up to 95.02% accurate results.

Key Words: Multi-Layer Feed Forward Neural Network, Genetic Algorithm, Back Propagation Network.

1. INTRODUCTION

The extraction of patterns that are hidden and relationships from databases that are huge in size, Data mining joins statistical analysis, machine learning and databases. In several areas of medical services, that includes effectiveness of surgical procedures, medical tests, all of the medication and the relations that are present in clinical and diagnosis data, in all of these, data mining is applies.

Medical diagnosis is a process that is very complex and requires many years of experience in that expertise. Heart attack is one of the main causes of death throughout the world. Hence there are huge databases pertaining to the causes of heart attack, symptoms of it and hence it gives a huge opportunity for data science and machine learning to hop on to this opportunity and make predictions and inferences hence to arrive on a conclusion that can help further people suffering from heart diseases or those people that are showing symptoms pertaining to heart disease. Most of the times symptoms also include other side effects such as discomfort in the shoulders, back pain, shortness of breath, nausea, light headedness.

We have proposed the classification of heart patients in which our system will make use of machine learning and find significant patterns in heart patients to help them in further treatment. We have applied intervals with equal binning all over the data we have made use of the neuro genetic network using back propagation algorithms. Pre-Processing of the data is done to make the data efficient and useful to make predictions and inferences.

Our idea was inspired by Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg that made use of Neural Network and Genetic Algorithm. The study is improved by us by using hybrid system in data mining. The objective of our Proposed system is to build an intelligent system that is focused on heart diseases using previous databases. 13 attributes were proposed but in the end it was deemed necessary to include only those factors that actually contributed to the heart disease conclusions incorporating optimal model construction time.

2. DATA MINING TECHNIQUES

Data Mining techniques are used to analyze, explore and extract medical data using high level instructions in order to find random and not known patterns. Data mining techniques are used by the researchers so that they can diagnose many fatal diseases such as diabetes, stroke, heart disease and cancer.
and these techniques are proven to give good accuracy for heart disease. Data mining techniques such as naïve bayes, neural network, bagging, kernel density, decision tree and support vector machine for prediction and diagnosis of heart disease are used by researchers. Some of the systems have shown that neural based learning classifier to classify data mining tasks performs equivalently to supervised learning classifier. IEHPS intelligent and effective heart attack prediction system was built with the help of data mining and neural networks and it expressed that in order to mine the frequent patterns, extracting significant patterns for heart disease prediction using K-means clustering and used MAFIA algorithm can be used. Polatat et al., developed system using hybrid fuzzy and k-nearest neighbour approach for the prediction of heart disease, which had 87% accuracy in diagnosis [1]. System’s using hybrid fuzzy and k-nearest neighbour method for the prediction of heart disease which was able to provide 87% accuracy in diagnosis was developed by Polatat et al. In another system where neural network was implemented showed 94.02% accuracy in diagnosis of heart disease. Using genetic algorithm and CANFIS, Latha and Subramanian were able to propose an intelligent heart disease prediction system which has very low mean square error. Keeping in mind the various techniques discussed, this paper proposes a system using neural network and genetic algorithm in order to predict the risk of heart disease. To optimize neural network weights genetic algorithm is used. We have used such hybrid techniques on risk factors for prediction of heart disease. The main objective of this system is to use in clinical decision support as well as to use as a risk indicator for people to keep track of their health in the future.

3. DISEASE DATASET

Databases in hospitals and clinics have collected large amounts of data about patients which are being virtually unused, this information is being used in our system. Details of patients have been taken. The data set has been created based on 13 attributes/risk factors. Thus, we use this patient database to predict heart disease. Dataset was obtained from UCI[1]. The data of 50 people was collected from surveys done by American Heart Association [2].

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Risk Factors</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Age (Years)</td>
<td>20-34 (0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35-50 (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>51-60 (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;60 (3)</td>
</tr>
<tr>
<td>2.</td>
<td>Sex</td>
<td>Male (0), Female (1)</td>
</tr>
<tr>
<td>3.</td>
<td>Blood Cholesterol</td>
<td>Below 200 mg/dL – Low (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200-239 mg/dL - Normal (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>240 mg/dL – High (3)</td>
</tr>
<tr>
<td>4.</td>
<td>Blood Pressure</td>
<td>Below 120 mmHg – Low (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>120 to 139 mmHg – Normal (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Above 139 mmHg – High (3)</td>
</tr>
<tr>
<td>5.</td>
<td>BP Treatment</td>
<td>Yes (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (2)</td>
</tr>
<tr>
<td>6.</td>
<td>Hereditary</td>
<td>Yes (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (2)</td>
</tr>
<tr>
<td>7.</td>
<td>Smoking</td>
<td>Yes (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (2)</td>
</tr>
<tr>
<td>8.</td>
<td>Alcohol Intake</td>
<td>Yes (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (2)</td>
</tr>
<tr>
<td>9.</td>
<td>Physical Activity</td>
<td>Low (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Normal (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High (3)</td>
</tr>
</tbody>
</table>
10. Diabetes | Yes (1)  
| No (2)

11. Diet | Poor (1)  
| Normal (2)  
| Good (3)

12. Obesity | Yes (1)  
| No (2)

13. Stress | Yes (1)  
| No (2)

Output | Heart Disease | Present (0)  
| Not Present (1)

4. MULTI-LAYER FEED FORWARD NEURAL NETWORK (MLFFNN)

The Supervised learning technique where the network is learned using known output is called as Neural Network Learning.

A. General Structure of MLFFNN

A neural network also consists of multiple layers. These layers are interconnected artificial neurons, as shown in Figure 1. In neural network, this type of neurons are sometimes called as a “node” or unit, all these terms have same meaning, and they are interchangeable. A multilayer feed forward neural network consists of various layers. It includes a layer of input units, one output layer and one or more layer of hidden units. A neural network without a layer of hidden units are called as a Perceptron. However, a Perceptron can only represent linear functions, so it isn’t powerful enough for the kinds of applications to be solved. [3]

![Figure 1: Typical structure of a Multilayer Feedforward Neural Network](image)

Multi-Layer Feed Forward Neural Network has a layer of processing units (i.e., the hidden units) in addition to the output units, therefore, this structure is called multilayer. In this network, the information moves in only one direction i.e. forward. It can move from the input nodes, through the hidden nodes (if any) and to the output nodes. Therefore, it is called as Feed Forward. In the network, there are no cycles or loops[4]. The connections can never skip a layer because there are never any backward connections. The layers are fully connected to each other, therefore, all units of one layer is connected to all units of next layer.

B. MLFFNN Training and Propagation
In this section, using a Heart Disease dataset, the neural network is trained by feed forward neural network model and back propagation learning algorithm with the parameter as a weight.

**The Back-Propagation Training Algorithm**

**Step 1: Initialization**

In this step, we set threshold levels and all the weights of the network to random numbers which are uniformly distributed inside a small range.

\[
\left( -\frac{2.4}{F_i} + \frac{2.4}{F_i} \right)
\]

Where \( F_i \) = total number of inputs of neuron \( i \) in the network

Neuron-by-neuron basis is used for weight initialization.

**Step 2: Activation**

In this step, we activate the back propagation neural network. Applying \( x_1(p), x_2(p), x_n(p) \) and desired outputs \( x_d,1(p), x_d,2(p), \ldots, y_d,n(p) \).

(a) **Calculate the actual outputs in the hidden layer of the neurons:**

\[
y_j(p) = \text{sigmoid} \left[ \sum_{i=1}^{n} x_i(p) \cdot w_{ij} - \theta_j \right]
\]

Where sigmoid is the sigmoid activation function, and \( n \) is the number of inputs of neuron \( j \) in the hidden layer.

(b) **Calculate the actual outputs in the output layer of the neurons:**

\[
y_k(p) = \text{sigmoid} \left[ \sum_{j=1}^{m} x_{jk}(p) \cdot w_{jk}(p) - \theta_k \right]
\]

Where \( m \) is the number of inputs of neuron \( k \) in the output layer.

**Step 3: Weight Training**

In this step, we update the weights propagating backward errors associated with output neurons in back-propagation network.

(a) **Calculate the error gradient in the output layer for the neurons:**

\[
\delta_k(p) = y_k(p) \cdot (1 - y_k(p)) \cdot e_k(p)
\]

\[
e_k(p) = y_{d,k}(p) - y_k(p)
\]

Calculate the weight corrections:

\[
\Delta w_{jk}(p) = \alpha \cdot y_j(p) \cdot \delta_k(p)
\]

Update the weights at the output neurons:

\[
w_{jk}(p + 1) = w_{jk}(p) + \Delta w_{jk}(p)
\]
Step 4: Iteration

In this step, we increase iteration p by one, then go back to step 2 and repeat the process until the selected error criterion is satisfied.

Learning

![Diagram of Learning of Weight Adjust](image)

Figure 2. Learning of Weight Adjust

Back Propagation algorithm is used to train the network [5]. In back-propagation algorithm, the general idea is to use gradient descent to update the weights to minimize the squared error between the target output values and network output values. Then by using gradient descent, each weight is adjusted according to its contribution to the error. It can find that in your text and other sources, but it won’t go into the actual derivations here. Starting with last set of weights, this process occurs iteratively for each layer of the network. It works back towards the input layer with output calculations, so the weight update depends on the type of problem.

5. GENETIC ALGORITHM

This algorithm is highly appreciated for solving optimization, search and machine learning problems. A genetic algorithm is a heuristic search that is inspired by Charles Darwin’s theory of natural evolution. This algorithm is an optimization technique that reflects the process of natural selection and natural genetics [6]. In natural selection, only the fittest individuals are selected for reproduction in order to produce next generation offspring.
Pseudo Code:

BEGIN
INITIALIZE population with random solution
EVALUATE each candidate
REPEAT Until (termination condition is satisfied DO)
SELECT parents
RECOMBINE pairs of parents
MUTATE the resulting offspring
SELECT individuals or the new
generation
END

Genetic Algorithm are Selection, Cross Over, Mutation, Accepting.

a) According to the fitness, select two parent chromosomes from a population.
b) Crossover of parents with crossover probability to form new offspring.
c) At each locus, Mutation occurs with a mutation probability mutate new offspring.
d) In Accepting, Place new offspring in the new population.
6. WEIGHT OPTIMIZED GENETIC ALGORITHM

We will be addressing the problem ANN based heart disease prediction using a genetic algorithm. Genetic algorithm will be used to optimize weights in different layers of the network. Both the weight and biased are satisfactorily trained as appose to conventional ANN. The difference between genetic algorithm ANN and conventional ANN is also satisfactorily.

The interaction between the neurons are defined by their weights. Thresholds are the minimum value that will get a neuron to fire. Matrix can be used to describe the weighted connection between neurons. For example, take the connection between the two layers of the neural network shown below.

![Neural Network Diagram]

The numbers displayed on the lines joining the neurons are their weights. Every neuron in the first layer is connected to every neuron in the second layer. A total of six connections are there in the above example. Hence these weights can be represented as a 3X2 matrix.

Format of weight matrix

\[
W = \begin{bmatrix}
  w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\
  w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\
  w_{S,1} & w_{S,2} & \cdots & w_{S,R}
\end{bmatrix}
\]

The destination neuron of the weight are indicated by the rows of the matrix W whereas the columns represent which source is the input for that weight. Thus, W12 represents the strength of signal from second input element to the first neuron.

7. HEART DISEASE PREDICTION USING GENETIC OPTIMIZED MLFFNN

The first step is to take the dataset received from UCI, dataset consists of patient information, patient history etc. The second step is to perform preprocessing technique, which removes noisy and anomalous data. The third step is to perform min-max normalization, so that negative and high values are removed/reduced. The fourth step is to train the NN using BPN. The fifth step is to predict the heart disease using the neuro-genetic algorithm. The final step is the diagnosis if the patient has heart disease or not.

Steps in Neural network with genetic algorithm are:

1. System initialization.
2. Risk factors extraction from patient details.
3. Selection process starts by assigning weights to risk factor attributes.
4. Each attribute fed to hidden layer of NN for training.
5. Adjust values to absolute value.
6. If training process satisfied, go to feature subset selection.
7. Use testing function of prediction of heart disease patients
8. Else if not satisfied, use GA processes then adjust weights and go to step 3.
9. Result, if patient has heart disease or not
Sample gene type for weight initialization is:
\{w_{11},w_{12},...,w_{1n},w_{21},w_{22},...,w_{2n},w_{n1},w_{n2},...,w_{nh}\}

Weights from input to hidden layer is
\{w_{11},w_{12},...,w_{nh}\}

Weight from hidden to output layer is
\{w_{11},w_{12},...,w_{nm}\}

The disease data ‘x’ of thr connection GA loop is
\{w_{11},w_{12},...w_{1n},w_{21},w_{22},...,w_{2n},w_{n1},w_{n2},...,w_{nm},b_1,b_2\}

Here, b_1 and b_2 are Bias 1 and Bias 2 respectively. After the initialization stage, the genetic reproduction, crossover, mutation are applied to the output.

8. RESULT AND ANALYSIS

In this section, the experimental results of the heart attack disease system for prediction are explained. Here, the evaluation of the sensitivity, specificity and accuracy occurs by comparing the performance of the proposed system with neural network-based system.

In the proposed system, we find the risk factors of heart attack patients and obtained results are evaluated with namely sensitivity, specificity, and accuracy. After detecting a positive disease, Sensitivity evaluates the diagnostic test correctly. By eliminating a given condition, Accuracy measures correctly figured out diagnostic test. In order to find these metrics, we first compute some of the terms like, True positive (TP), True negative (TN), false positive (FP) and false negative (FN).

\begin{align}
\text{Sensitivity} &= \frac{TP}{TP+FN} \quad (a) \\
\text{Specificity} &= \frac{TN}{TN+FP} \quad (b) \\
\text{Accuracy} &= \frac{(TN+TP)}{(TN+TP+FN+FP)} \quad (c)
\end{align}

The confusion matrix shows the number of samples which have been classified into the two correctly/falsely classes of C1 and C2. The entries of this matrix are used to explain the performance measures [7]. In the confusion matrix, the correctly classified number of samples of class C1 falls under true positive (TP); false negative (FN) is the number of the samples of class C1 which have been falsely classified as C2; and false positive (FP) is the number of the samples of class C2 which have been falsely classified as C1.

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Actual Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td>C2</td>
<td>False Negative (FN)</td>
</tr>
</tbody>
</table>

The genetic optimized NN is trained and tested using sample of 50 patient data. Utilizing accuracy, sensitivity and specificity, the performance of the system is compared with the neural network-based system. In the True Positive value 34 and True Negative is 15. Then False positive 1 and False Negative 0. The accuracy is given by the 95.02%.

Based on the implementation result Multi-Layer Feed Forward Neural Network and Genetic Network evaluate the best accurate performance.

BPN is widely used in the learning algorithm in Neural Network for the many applications. However, BP learning depends on weights in the MLFFNN. Due to this, GA has been used to obtain the optimal parameter value and weight for the BP learning. So that the performance of GA is increased better than the MLFFNN.
9. CONCLUSION

The proposed heart system has taken 50 patients suffering from heart diseases for machine learning and intelligent systems, weighted based on frequency in the datasets. The usage of multi-layer feed forward neural network optimized with genetic algorithm should be adjusted by adjusting the variable and given the better improved results were compared with other neural-based systems to get a measure of accuracy, sensitivity and specificity. This work also demonstrates about GA-NN prediction by improving rate at which hidden neurons are optimized. By using this, predictions and conclusions become more clear and accurate.

10. REFERENCES