Medical Diagnosis System based on Artificial Neural Network

May Kywe Zin Soe
University of Technology (Yatanarpon Cyber City)
Dr. Chaw Su Win
University of Technology (Yatanarpon Cyber City)

ABSTRACT - Nowadays, computerized systems are becoming very popular in medically related areas. One of the major problems in medical field is setting the diagnosis. Medical diagnosis system is a system which can diagnose diseases through checking out the input symptoms. In this paper, an intelligent medical diagnosis system for three diseases: diabetes mellitus, heart disease and liver disorder is presented by applying Back Propagation Neural Network Learning Algorithm. This system aims to effectively identify a disease as infected or non-infected according to user input data. The required data for the diagnosis system are obtained from UCI machine learning repository. The accuracy of the proposed system for diabetes mellitus, heart disease and liver disorder are 83%, 64% and 57% respectively.

KEYWORDS: Artificial Neural Network, Medical Diagnosis, Back Propagation Neural Network Learning Algorithm.

I. INTRODUCTION

Artificial Neural Network (ANN) is often used as a powerful discriminating classifier in medical diagnosis tasks for early detection of diseases. ANN is excellent in recognizing diseases since there is no need to provide a specific algorithm on how to identify the disease. Neural network learns by example so the details of how to recognize the disease is not needed [6]. Artificial Neural Networks are applied to medical diagnosis mainly for the task which is based on the input features to assign the patient to one of the classes [16] [17].

In this system, the assessment of diabetes mellitus, heart disease and liver disorder is implemented by using ANN as classifier. The proposed system intends to analyze the capability of ANN for diagnosis of diseases according to their set of associated symptoms. The system also aims to support the physicians by identifying disease depending on the set of input symptoms.

Today, a number of researches are going on in the area of medical diagnosis applications by using ANN [18] [19]. Many researchers have been studied on the prediction of various intense diseases and mild diseases like lung cancer, pancreatic cancer, hypertension, diabetes and so on.

II. LITERATURE REVIEW

There are several reviews related with the applications of ANN in medical diagnosis.

Rahul Samant, et.al. (2013) [11] has used LM Back Propagation training algorithm in Feed Forward Network and decided the best configuration for the diagnosis of hypertension in the form of number of hidden layers and hidden neurons. 13 input parameters have been taken including age, pulse, systolic blood pressure, diastolic blood pressure, Serum Proteins, Serum Albumin, Hematocrit, Erythrocyte Sedimentation Rate, Serum cholesterol, Serum Triglycerides, Whole Blood Viscosity, Plasma Viscosity and Red Cell Aggregation.


Rehana Parvin (2014) [20] developed by integration of fuzzy inference system(FIS) and fuzzy schema design and implementing it by SQL in three different health care context; heart disease, diabetes mellitus and liver disorder. The comparison of prediction capability between implemented FIS in MATLAB and Neural Network for the same dataset is shown.

Soni et al. (2011) [13] designed what they called an “Intelligent and Effective Heart Disease Prediction System using weighted associative classifiers”. They implemented the system using Java as front end and MS Access as the back end tool. In case of prediction the authors considered two cases: people with heart disease and people with no heart disease.
III. BACKGROUND THEORY

A. Artificial Neural Network

Artificial neural network (ANN) is the simulation of human brain and is being applied to an increasingly number of real world problems [1]. Neural networks are typically arranged in input layer, one or more hidden layers and output layer [3]. There is no specific rule that dictates the number of hidden layer [5]. A network with one hidden layer is sufficient to solve most tasks [6]. Each layer in a layered network is an array of processing elements or neurons [3]. The number of neurons in input layer and output layer is defined by problem [6]. The number of hidden neurons should be in the range between the size of the input layer and the size of the output layer. If the number of neurons in the hidden layer is more, the network becomes complicated [5]. Artificial Neural Network uses supervised learning to classify input data into desired output [4]. Neural networks are used to get high accuracy and objectivity of medical diagnosis [2]. ANN provide a powerful tool to help doctors to analyze, model and make sense of complex clinical data across a broad range of medical applications.

B. Back Propagation Algorithm

Back Propagation is a supervised neural network learning algorithm. Back propagation networks are fully connected, layered, feed forward networks, in which activations flow from the input layer through the hidden layer and then to the output layer [14]. The back propagation algorithm performs learning on a multilayer feed forward neural network repeatedly. Back propagation learns by iteratively processing a set of training data comparing the network’s prediction for each sample with actual known class label. Classification by back propagation is the most applicable neural network learning algorithm in various fields [15]. The training steps of the proposed algorithm are as follows:

Step1-Initialize weights and biases: Firstly, the weights and biases are initialized to small random numbers.
Step2-Feed the training data: The input data of the system is fed into the input layer of the network.
Step3-Propagate the inputs forward: The net input and output of each unit in the hidden layer and output layer are computed. To compute the net input to the unit, the input values and its corresponding random weights are multiplied and then summed including biases. Given a unit in a hidden layer or output layer, the net input, \( I_j \) is

\[
I_j = \sum w_{ij} O_i + \theta_j \tag{1}
\]

Where, \( w_{ij} \) is the weight of the connection from unit i in the previous layer to unit j; \( O_i \) is the output of unit i from the previous layer; \( \theta_j \) is the bias of the unit.

The activation function used in this system is logistic or sigmoid function and the output of unit j is computed as:

\[
O_j = \frac{1}{1+e^{-I_j}} \tag{2}
\]

This function maps a large input domain into smaller range of 0 to 1.

Step4-Back propagate the error: The error is propagated backward by updating the weights and biases to get the network prediction result. For a unit j in the output layer, the error is computed by:

\[
Err_j = O_j (1 - O_j) (T_j - O_j) \tag{3}
\]

Where \( O_j \) is the actual output of unit j and \( T_j \) is the known target value of unit j.

The error of unit j in the hidden layer is:

\[
Err_j = O_j (1 - O_j) \sum_k Err_k w_{jk} \tag{4}
\]

Where \( w_{jk} \) is the weight of the connection from unit j to unit k in the next layer and \( Err_k \) is the error of unit k.

Step5-Update weights and biases: The weights and biases are updated to achieve the output values that are closer to the target output. Weights are updated by:

\[
w_{ij}^{new} = w_{ij}^{old} + (l)Err_j O_i \tag{5}
\]

Where, l is the learning rate.

Biases are updated by:
\[ \theta_i^{\text{new}} = \theta_i^{\text{old}} + (1) \cdot \text{Err}_i \quad (6) \]

**Step 6 - Repeat and apply terminating condition:** Training stops when all weights in the previous epoch were as small as to below some specific threshold, the percentage of tuples misclassified in the previous epoch is below some threshold or a prespecified number of epochs have expired.

**C. Explanation of Three Diseases in the Proposed System**

**Diabetes Mellitus**

Diabetes mellitus is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. This occurs because either the pancreas can’t produce enough insulin or the cells in body have become resistant to insulin. Diabetes is a chronic, lifelong condition that affects the body’s ability to use the energy found in food. Diabetes Mellitus has become a common health problem nowadays, which would affect people and lead to various complications like visual impairment, cardiovascular disease, leg amputation and renal failure if diagnosis is not done in the right time [10]. As of 2015, an estimated 415 million people had diabetes worldwide.

**Heart Disease**

Any disorderliness that affects the heart from infection to genetic defects and blood vessel disease is referred to as heart disease [8]. There is an increase in death rate yearly as a result of heart diseases. One of the major factors that cause this increase is misdiagnoses on the part of medical doctors or ignorance on the part of the patient [7]. Heart disease is a serious disease and its diagnosis at early stage remains challenging task [9]. According to WHO and CDC, heart disease is the leading cause of death in UK, USA, Canada and Australia.

**Liver Disorder**

The liver is a resilient organ, able to function even when damaged and also to regenerate itself. The liver is involved in a number of roles of converting food into energy and eliminating alcohol and poisons from the blood. There are many different conditions that can affect the proper functioning of the liver. If any of the normal functionality fails, it is considered to be a liver disorder. If liver disease is left untreated it can be fatal. Incidence of liver disorder is growing and it is now estimated that it affects some two million people in the UK alone, where it is the fifth largest cause of death.

**IV. OVERVIEW OF THE PROPOSED SYSTEM**

**Fig 1: Program flow of the proposed system**
The main idea of this system is to construct the neural network model, which will perform the presumptive diagnosis of diabetes mellitus, heart disease and liver disorder. The dataset is acquired from UCI machine learning repository. Firstly, the input data from the dataset is transformed into normalized values. And then, the network is trained with back propagation algorithm. When training is finished, the network is evaluated with the test dataset. In diagnosis phase, the user can check whether he/she is infected with the disease or not.

For Heart disease, 140 samples are divided into two parts: training data and testing data. Two-third (93 samples) of the dataset is used for training and the rests (47 samples) are used for testing. The input layer of the network consists of 13 attributes such as age, gender, chest pain type, resting blood pressure, etc. These are considered as diagnosis variables to determine whether the patient has heart disease or not.

As in heart disease dataset, the diabetes dataset is partitioned into 306 data for training and 154 data for testing. The eight input data to consider the patient is infected or non-infected with diabetes are number of times pregnant, diastolic blood pressure, body mass index, and so on.

Like the above two datasets, the liver disorder database is segmented into 260 training data and 130 testing data. The six diagnosis variables of the diseases are mean corpuscular volume, alkaline phosphotase, number of half-print equivalents of alcoholic beverages drunk per day, and so on.

**V. IMPLEMENTATION OF THE SYSTEM**

In this system, three layer feed forward network with input layer, one hidden layer and output layer is used. The number of neurons in the input layer is defined by the number of attributes required for diagnosing each disease in the system. So, 8 input neurons for diabetes mellitus, 13 input neurons for heart disease and 6 input neurons for liver disorder are used for diagnosing each disease. At the output layer, two neurons are applied to indicate whether the disease is present or not. The output is implemented as 01 if the disease is present and as 10 if there is absent of the disease. The number of hidden neurons is defined by the mean value of the output nodes and input nodes.

**A. Training Artificial Neural Network**

After the architecture of the network is defined, the network is ready to be trained. As in the proposed algorithm, firstly, the weights and biases are set with small random values between -0.5 and 0.5. The input symptoms that are used to predict the presence or absence of the disease is fed into the input layer of the network. Before feeding the input data, the values are normalized between 0 and 1 because of using sigmoid activation function. And, the system calculates the output of the network using the normalized input values, weights and biases. For each training sample, if the output of the network is closely enough to the target output, the training stops and moves to the next sample. If not, the errors are propagated backwards to know the differences between target output and actual output. And then, the weights and biases are modified as to minimize the differences between the network’s prediction and the actual classification. Then, the output of the network is calculated again by using updated weights and biases. The back propagation process continues until the output of the network is closely enough to the target output. The training process is repeated through all training data for first epoch. The training continues with second epoch, third epoch and so on. The training stops when the neural network gives correct outputs for all training with zero iteration.

**B. Testing Artificial Neural Network**

After finishing the training phase, the trained network is tested with the test data. The network is tested with 154 samples, 47 samples and 130 samples for diabetes mellitus, heart disease and liver disorder respectively.

**C. Evaluation of the System**

The accuracy of the system is calculated by the following equation:

\[
\text{Accuracy} = \left( \frac{\text{Total Number of Matched Records}}{\text{Total Number of Records}} \right) \times 100
\]
The accuracy of diagnosing diabetes mellitus, heart disease and liver disorder is 83%, 64% and 57% respectively. According to the results achieved from this work, the system can be used to effectively identify the infected or non-infected person of diabetes mellitus, heart disease and liver disorder.

Table 1. Accuracy of the Proposed System

<table>
<thead>
<tr>
<th>Disease</th>
<th>Input Neurons</th>
<th>Hidden Neurons</th>
<th>Output Neurons</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes Mellitus</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>83%</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>13</td>
<td>7</td>
<td>2</td>
<td>64%</td>
</tr>
<tr>
<td>Liver Disorder</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>57%</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

Nowadays, people can save their time and energy with the help of technologies. The proposed neural network-based diagnosis system also aims to support people for identifying the patient who is infected or non-infected with heart disease, diabetes mellitus and liver disorder. People can effectively achieve the correct diagnosis of these three diseases because of the advantages of the back propagation neural network learning algorithm. Both the patients and physicians can get enormous benefits by using this diagnosis system.

REFERENCES

[6] Dr. N. Ganesan, Dr.K. Venkatesh, Dr. M. A. Rama, A. Malathal Palani,” Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data”,2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 26
[16] W. David Aha and Dennis Kibler, Instance-based prediction of heart disease presence with the Cleveland database