Classification of Heart Disease Dataset using Multilayer Feed forward backpropogation Algorithm

Miss. Manjusha B. Wadhonkar¹, Prof. P. A. Tijare² and Prof. S. N. Sawalkar³

¹ M.E Computer Engineering, ² Associate Professor, ³ Assistant Professor
Computer Science and Engineering Department
Sipna College of Engineering

ABSTRACT
In this study, for the classification of heart diseases dataset, multilayer feed forward network with backpropogation algorithm is introduced. Artificial Neural Network (ANN) is widely used data mining method to extract patterns. Data mining is the process of automating information discovery. Its main aim is to find relationships in data and to predict outcomes. Classification is one of the important data mining techniques for classifying given set of input data. Many real world problems in various fields can be solved by using classification approach such as business, science, industry and medicine. For analysis the medical data related to Heart diseases is considered and analyzed using artificial neural network (ANN). To perform classification task of medical data, the neural network is trained using back propagation algorithm with momentum. To increase the efficiency of the classification process parallel processing approach is also applied on each neuron in different layers.

Keywords: Heart disease Dataset, Backpropogation Algorithm, Artificial neural network, MLP, Knowledge data discovery.

1. INTRODUCTION
Data mining also called knowledge data discovery is the process of analyzing data from different perspective and summarizing it into useful information. Knowledge Discovery in database is concerned with the development of methods and techniques for making sense of data. KDD is the process of mapping low-level data into other forms that might be more compact, more abstract or more useful. Main aim of data mining is to uncover relationship in data and predict the outcome. [1]

Data mining extract the patterns in the process of knowledge discovery in the database. As the dataset has grown in size and complexity, new emerging field of data mining provides new techniques and methods which help to analyse and understand large bodies of data. Data mining involves some common methods that are association rule learning, clustering, classification, regression, summarization and sequential pattern matching. Classification is one of the important techniques of data mining. Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts. [2]

In classification, inputs are given a set of data, called a training set, where each record consists of several fields or attributes. These attributes are continuous, coming from an ordered domain, or categorical, coming from an unordered domain. One of the attributes, called the classifying attribute, indicates the class to which each dataset belongs. The objective of classification is the method to build a model of the classifying attribute based upon the other attributes which are not from the training data set. [3]

Data mining is an interdisciplinary field. For the implementation of classification, techniques from other disciplines may also be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high performance computing which will lead to intelligent, low cost solution. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing[3]. Artificial neural network is one of the widely used techniques for extraction of patterns in data mining. Artificial neural networks has some advantages such as it automatically allow arbitrary nonlinear relations between the independent and dependent variables, and allows all possible interactions between the dependent variables. Due to above said advantages of ANN the use of neural network technique is adopted for the classification of dataset.

Neural network approach is proposed for the efficient classification of medical data. Neural network is trained using backpropagation algorithm. To increase the efficiency of classification process, parallel processing is implemented at each node in the network. After training the neural network performance of the network is analysed with various set of test data.

1.1. RELATED WORK
The researchers in the medical field identify and predict the diseases besides proffering effective care for patients with
the aid of data mining techniques. The data mining techniques have been utilized by a wide variety of works in the
literature to diagnose various diseases including: Diabetes, Hepatitis, Cancer, Heart diseases and the like. Information
associated with the disease, prevailing in the form of electronic clinical records, treatment information, gene
expressions, images and more; were employed in all these works. In the recent past, the data mining techniques were
utilized by several researchers to present diagnosis approaches for diverse types of heart diseases [4].

Analysis of different data mining techniques that can be employed in automated heart disease prediction systems.
Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient
and effective heart disease diagnosis. The analysis shows that Neural Network with 8 and 13 attributes has shown the
approximate accuracy of 81% so far. Moreover, in combination with Genetic Algorithm and 6 attributes, Decision Tree
has shown 99.2% efficiency. Thus in this study for efficient classification of dataset ANN is proposed as a method of
classification in data mining.

The overall organization of the paper includes introduction and related work in section 1. Section 2 presents
fundamental issues of artificial neural network for classification. In section 3 classification of heart disease dataset is
analyzed using neural network. Section 4 presents application of neural network in medical and other fields. Section 5
concludes the paper.

2. NEURAL NETWORK

Neural network is a computing system made up of a number of simple, highly interconnected processing elements,
which process information by their dynamic state response to external inputs. (Dr. Robert Hecht-Nielsen). An Artificial
Neural Network (ANN) commonly referred as neural network is an information processing paradigm that is inspired by
the way biological nervous systems such as the brain, process information. Structure of Neural network is composed of
a large number of highly interconnected processing elements (PE) called as neurons. This structure basically consist of
inputs, which are multiplied by weights (strength of the respective signals), and then computed by a mathematical
function which determines the activation of the neuron and one more function computes the output of the artificial
neuron (sometimes in dependence of a certain threshold). Training or learning is the process of adjusting weights on
input connections to get the specific desired output. Resemblance of ANN with biological neural network (BNN) gives
opportunity to user to apply parallel processing concept on each neuron at each layer.

ANN learns from existing examples. Artificial neural networks are trained based on two training methods 1. Supervised Training- for supervised learning external teacher is available which provide the neural network input data
and actual desired output. The network parameters are adjusted under the combined influence of training vector and
error signal to get that desired output for a given specific input. 2. Unsupervised Training- Input data and computation
function is given to the network and output is calculated. It does not require the desired output. Neural network is an
iterative learning process in which input data cases are given to the network one at a time and weights associated
with input interconnection are adjusted each time [5]. During learning process NN is trained by adjusting weights.
Once the neural network is modeled, then it is ready to be trained for a given specific application. For training process
initial weights are selected randomly.

Structure of neural network consists of three layers, an input layer, one or more hidden layer and the output layer.
Number of hidden layers and number of neurons in each layer strongly depends on the complexity of system studied.
In this study for the classification of dataset, multilayer feed forward system is introduced which contain only forward
paths. In a feed-forward system PE are arranged into distinct layers where each layer receive input from the previous
layer and outputs to the next layer. There is no feedback. The general structure is given in figure 1.

![Figure 1](image)

**Figure 1** fully connected feed forward network [7]

For more complex decision function the inputs are fed into a number of perceptions nodes, each with its own set of
weights and threshold [8]. The outputs of these nodes are then input into another layer of nodes and so on, the output of
the final layer of nodes is the output of the network. Such a network is termed a multi-layer Perceptron (MLP) and the
layers of nodes whose input and output are seen only by other nodes are termed hidden [6]. The connection weights are computed by means of a learning algorithm. The model of each neuron in the network includes nonlinear activation function.

\[
x^i = \frac{1}{1 + \exp(-v^j)}
\]  

(1)

The network exhibits a high degree of connectivity. The change in the connectivity of the network requires change in the population of synaptic connections of their weights. Two types of signals are identified by MLP network. 1. Function Signal and Error signal.

Neural network is trained using backpropogation algorithm. The backpropogation algorithm with a supervised error-correction learning rule, is one of the most popular and robust tools in the training of artificial neural networks. Back propagation passes error signals backwards through the network during training to update the weights of the network. For each training tuple, the weights are modified so as to minimize the mean squared error made in the networks prediction and the actual target value. These modifications are applied in the backwards direction. The actual response \( y_k(n) \) is different from desired output \( d_k(n) \) for \( k \) different neuron. The error signal is generated and is defined as

\[
e(n) = y_k(n) - d_k(n)
\]

(2)

Our objective is to minimize this error using error correction rule.

The weights in the network are initialized to small random numbers (ranging from -1.0 to 1.0 or -0.5 to 0.5). Each unit has a bias associated with it. It is also initialized to small random number. Net input at any such unit \( j \) is computed by multiplying it by its corresponding weight and this is summed.

\[
l_j = \sum w_{ij}o_i + \theta_j
\]

(3)

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 and \( \theta_j \) is the bias of the unit. Bias acts as a threshold.

### 3. ANALYSIS AND CLASSIFICATION OF HEARTDISEASE DATASET

In this paper for the analysis and classification of the medical data related to Heart diseases is considered. This database contains 76 attributes, but we refer to using a subset of 16 of them. The Cleveland database is freely available on the internet. The ‘goal’ field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 3[22]. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 0, 1, 2, 3) from absence (value 0). The names and social security numbers of the patients were removed from the database that is replaced with dummy values. One file has been "processed", that one containing the Cleveland database. This is publicly available dataset on the Internet. Cleveland dataset concerns classification of person into normal and abnormal person regarding heart diseases. Previous analysis used 13 input attributes, in this study two more input attributes are added as smoke and obesity [21].

**Data Representation:**

Number of instances: 450

Number of attributes: 15 and a class attribute

Class:

Class0: Normal Person.

Class1: first stroke

Class2: second stroke

Class3: end of life

**Attribute Description:**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age in years</td>
<td>Continuous</td>
</tr>
<tr>
<td>Sex</td>
<td>(1=male; 0=female)</td>
<td>0, 1</td>
</tr>
</tbody>
</table>
Cp  --Value 1: typical angina  1,2,3,4
    --Value 2: atypical anginal
    --Value 3: non-anginal pain
    --Value 4: asymptotic

Trestbps  Resting blood pressure (in mm Hg)  Continuous

Chol Serum  cholesterol in mg/dl  Continuous

Fbs  (Fasting blood sugar .120mg/dl)  0, 1(1=true; 0=false)

Restecg  electrocardiography results  0, 1, 2
    --Value 0: normal
    --Value 1: having ST-T wave abnormality
    (T wave inversions and/or ST Elevation or depression of>0.05mV)
    --Value 2: showing probable or definite left
    Ventricular Hypertrophy by Estes’ criteria

Thalach  Maximum heart rate achieved  Continuous

Exang  Exercise induced angina (1=yes; 0=no)  0, 1

Oldpeak  ST depression induced by exercise Continuous
    Relative to rest

Slope  The slope of the peak exercise Continuous
    ST segment
    Value 1: up sloping
    Value 2: flat
    Value 3: down sloping

Ca  Number of major vessels (0-3)Continuous
    Colored by fluoroscopy

Thal  Normal, fixed defect, reversible defect  3, 6, 7

Smoke  1=true, 0=false  1, 0

Obesity  1=true, 0=false  1, 0

**Fundamental Steps in Classification of Heart Disease Dataset:**

The workflow of ANN analysis for the classification of dataset is as shown in figure 2 which provide brief overview of fundamental steps that should be followed to apply ANNs for the classification of heart disease dataset.
3.1 Data Collection:
Neural network is trained using suitable dataset of example cases. This dataset is nothing but records of patient’s stores in a database. Database should contain sufficient number of reliable examples to be given as an input to the training network.

3.2 Pre-processing:
Data in the training dataset must be pre-processed before the evaluation by the neural network. Normally data are scaled to be within the interval [0, 1] because the interference function used is logistic one. During pre-processing some data are missing should be removed from the dataset to improve the classification performance.

3.2.1 Data Scaling:
As the input and output to the network is in binary form, each value is converted into the range between 0 to 1 using the following formulae

3.3 Neural Network Training
The neural network is trained with Heart Diseases database by using feed forward neural network model and backpropogation learning algorithm with momentum and variable learning rate. The input layer of the network consists of 15 neurons to represent each attribute as the database consists of 15 attributes. The number of classes is four: 0 – normal person, 1- first stroke, 2- second stroke and 3- end of life. The output layer consists of two neurons to represent these four classes. The backpropogation algorithm with momentum is used to train the neural network during the training process. Several neural networks are constructed with and without hidden layers, i.e., single and multi layer networks and trained with heart disease dataset. Relationship between the number of epochs and the sum of squares of errors during training process for various networks can be observed from the figure. 3and 4[20]

![Figure 2](image)

**Figure 2:** Diagram of fundamental steps in ANN based classification of Heart Disease Dataset

![Figure 3](image)

**Figure 3:** Training the single layer neural net with heart disease dataset [20]

![Figure 4](image)

**Figure 4:** Training the multilayer neural net with heart disease dataset [20]
3.4 Verification:
Heart Disease dataset should be verified by means of a dataset different from the training Dataset.

3.5 Testing of Neural Network
For testing the performance of the net various samples are collected as test data. The test data is given as the input to the trained network and the output of the net is calculated with the adjusted weights. The actual output of the net is compared with the desired output to study the learning ability of the network for classifying the heart disease data. The results for neural network with 13-input attributes and 15 input attributes with two output neuron are tabulated in Table1 [20-21]

<table>
<thead>
<tr>
<th>Input Attributes</th>
<th>Classification Accuracy using multilayer feed forward backpropogation</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>94%</td>
</tr>
<tr>
<td>15</td>
<td>100%</td>
</tr>
</tbody>
</table>

4. APPLICATIONS OF NEURAL NETWORK
Neural networks are applicable to many real world business problems and marketing. Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting that includes sales forecasting, industrial process control, customer research, data validation, risk management, and target marketing.ANN is also used in the following specific paradigms: recognition of speakers in communications, and facial recognition.

4.1 Neural networks in medicine
Artificial Neural Networks (ANN) is currently a ‘hot’ research area in medicine. At present, the research is mostly on modeling parts of the human body and recognizing diseases from various scans (e.g. cardiograms, CAT scans, ultrasonic scans, etc.). Various application of ANN in medicine includes modeling and diagnosing the cardiovascular system, Classification and diagnostic prediction of cancers, diagnosis of urological dysfunctions, diabetes disease diagnosis. They are used in the analysis of medical images from a variety of imaging modalities [10-19]. Applications in this area include tumor detection in ultra-sonograms, detection and classification of micro calcifications in mammograms, classification of chest x-rays and tissue and vessel classification in Magnetic Resonance Images. Neural network as a classification technique of data mining is applied to extract the rules from a disease in medical diagnosis. From this rules it is easy to analyze and predict the disease.

5. CONCLUSION
Classification is an important problem in the rapidly emerging field of data mining. As per wide range of applicability of ANN and their ability to learn complex and nonlinear relationships including noisy or less precise information, neural networks are well suited to solve problems in biomedical engineering. From the analysis, it is observed that multilayer feed forward network with backpropogation algorithm using 15 input attributes gives the highest accuracy [21]. Backpropogation algorithm with momentum and variable learning rate is used to train the networks. To analyze performance of the network various test data are given as input to the network. This binary heart disease dataset classifier can be used to assist physicians to classify the dataset of heart disease. From the neural networks design in this study, it is evident that MLP NNs required a compact architecture as compared to other NNS, in terms of number of hidden nodes required for the classification. Thus the number parameters such as weights and biases required for the designing of MLP NN is sufficiently lower than other. Heart disease dataset is analyzed by considering multilayer neural network.

REFERENCES
[2] Jiawei Han, Micheline Kamber, (2006)”Data Mining: concepts and Techniques“, ELSEVIER