Implementation of Back-Propagation Algorithm For Renal Datamining

M S S Sai msssai@gmail.com
Asst. Professor
Dept of MCA
Hindu College PG Courses
Guntur, AP, India

P.Thrimurthy profpt@rediffmail.com
Professor
Dept. of Computer Science & Engg.
ANU, Guntur, AP, India

Dr.S.Purushothaman dr.s.purushothaman@gmail.com
Professor
Sun College of Engineering and Technology
Nagerkoil, India

Abstract

The present medical era data mining place a important role for quick access of appropriate information. To achieve this full automation is required which means less human interference. Therefore automatic renal data mining with decision making algorithm is necessary. Renal failure contributes to major health problem. In this research work a distributed neural network has been applied to a data mining problem for classification of renal data to have for proper diagnosis of patient. A multi layer perceptron with back propagation algorithm has been used. The network was trained offline using 500 patterns each of 17 inputs. Using the weight obtained during training, fresh patterns were tested for accuracy of diagnosis.

Keywords: Datamining, Renal data, Back-propagation algorithm, Diagnosis.

1. INTRODUCTION

Two types of databases are available in medical domain. The one is a dataset acquired by medical experts, which are collected for a special research topic. These data have the following characteristics: (1) The number of records are small. (2) The number of attributes for each record are large, compared with the number of records. (3) The number of attributes with missing values are very few. This type of databases is called p-databases(prospective databases). The analysis of those data is called prospective analysis in epidemiology, because data collection is triggered by the generated hypothesis. Statistical analysis has been usually applied to these datasets [l-7].

The second type is a huge dataset retrieved from hospital information systems. These data are stored in a database automatically without any specific research purpose. Usually, these databases only include laboratory tests, although researchers in medical informatics are discussing how to store medical image, and physical examinations as electronic patient records [8-11]. These data in hospital information system (HIS) have the following characteristics: (1) The number of records are very huge. (2) The large number of attributes for each record (more than several hundred).(3) Many missing values will be observed. (4) Many temporal sub-records are stored for each record (patient). This type of databases is called r-databases(retrospective databases). The analysis of these data is called retrospective analysis in epidemiology, because data will be analyzed after data collection. Those data will lose any good features which prospective data holds and even statistical techniques do not perform well. This type of data is very similar to business databases. Concerning p-databases, data will be prepared with a hypothesis generated by medical experts very carefully. Thus, the quality of data is very high, and any data analysis technique will be applicable and useful. Only the problem with p-databases is that the number of measurements is very large, compared with the number of records. Thus, data reduction or rule induction will be useful to detect the important attributes for analysis. On the other hand, as for r-databases, there are many difficult issues for data analysis.

1.1 Renal systems

The renal system consists of all the organs involved in formation and release of urine. It includes the kidneys, ureters, bladder and urethra. Initially, it is without specific symptoms and can only be detected as an increase in serum creatine. As the kidney function decreases, renal failure is a serious medical condition affecting the kidneys. When persons suffer from renal failure, their kidneys are not functioning properly or no longer work at all. Renal failure can be a progressive disease or a temporary one depending on the cause and available treatment options.

The kidneys are glands that are located in the abdominal region just above the pelvis on either side of the body. When functioning normally, the kidneys separate and filter excess water and waste from the blood stream. The kidneys are responsible for producing urine, which is used to flush away the toxins. The kidneys maintain a healthy balance of fluids and electrolytes, or salt compounds, in the body. In renal failure the kidneys undergo cellular death and are unable to filter wastes, produce urine and maintain fluid balances. This dysfunction causes a build up of toxins in the body which can affect the blood, brain and heart, as well as other complications. Renal failure is very serious and even deadly if left untreated.

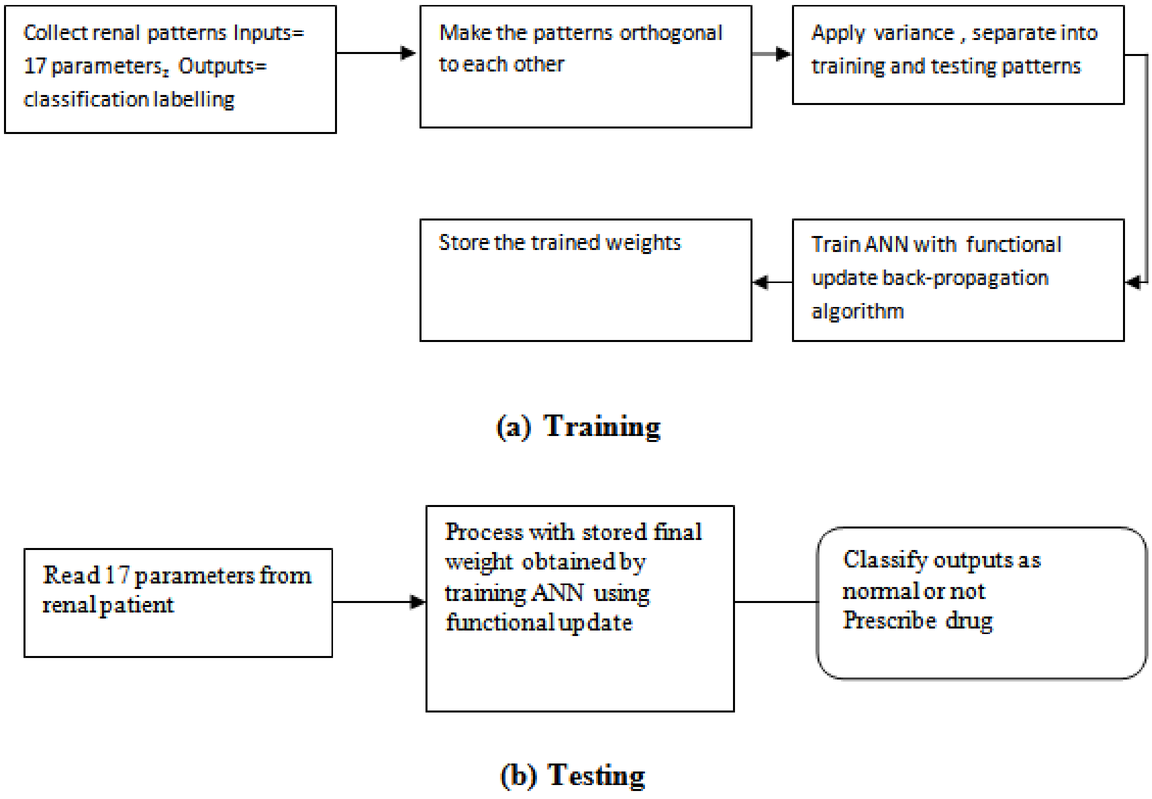
The quantity and complexity of data acquired, time-stamped and stored in clinical databases by automated medical devices is rapidly and continuously increasing. As a result, it becomes more and more important to provide clinicians with easy-to-use interactive tools to analyze huge amounts of this data. These tools would serve different purposes, such as supporting clinical decision making, evaluating the quality of the provided care, and carrying out medical research. The specific clinical context is in the domain of hemodialysis, where clinicians have to deal with huge amounts of data automatically acquired during the hemodialytic treatment of patients suffering from renal failure.

2. PROBLEM DEFINITION

The problem is to implement an intelligent data mining concept for the huge amount of renal data. As the number of patients is growing rapidly due to food habits and other deficiencies in the body, renal failure plays predominantly in the life of patient. Quick diagnosis and telemedicine requires immediate solution for a patient. This can be achieved properly only from the knowledge gained from the experts with regard to diagnosing methods.

Renal data such as person age in terms of years, male / female, Edema, Oliguri, Normochronic, Urgent, Hypertension, Diabetics, Family History, Polymer Chain Reaction, Obesity, Hemoglobin, Cholostral, Creatine have been collected for 1000 patients. In this research work, back-propagation algorithm is used to implement data mining. BPA is a supervised algorithm to train an artificial neural network. It is an intelligent method for mining information meaningfully and quickly.

3 SCHEMATIC ARCHITECTURE


Collect renal patterns Inputs= 17 parameters, Outputs= classification labelling
Make the patterns orthogonal to each other
Apply variance , separate into training and testing patterns
Train ANN with functional update back-propagation algorithm
Store the trained weights
Read 17 parameters from renal patient
Process with stored final weight obtained by training ANN using functional update
Classify outputs as normal or not Prescribe drug

(a) Training

(b) Testing

FIGURE.1: Renal data mining

4 ARTIFICIAL NEURAL NETWORKS

A neural network is constructed by highly interconnected processing units (nodes or neurons) which perform simple mathematical operations, Fortuna et. al [12]. Neural networks are characterized by their topologies, weight vectors and activation function which are used in the hidden layers and output layer, Lippmann [13]. The topology refers to the number of hidden layers and connection between nodes in the hidden layers. The activation functions that can be used are sigmoid, hyperbolic tangent and sine, Yao and Fang [14]. The network models can be static or dynamic Hush and Horne [15]. Static networks include single layer perceptrons and multilayer perceptrons. A perceptron or adaptive linear element (ADALINE), Widrow [16] refers to a computing unit. This forms the basic building block for neural networks. The input to a perceptron is the summation of input pattern vectors by weight vectors. In Figure 2, the basic function of a single layer perceptron is shown.

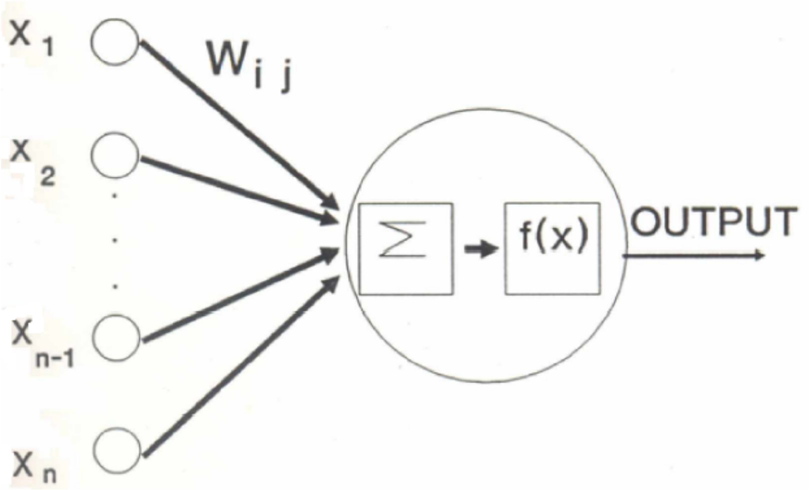


FIGURE 2:. Operation of a neuron

In Figure 3, a multilayer perceptron is shown schematically. Information flows in a feed-forward manner from input layer to the output layer through hidden layers. The number of nodes in the input layer and output layer is fixed. It depends upon the number of input variables and the number of output variables in a pattern. In this work, there are six input variables and one output variable. The number of nodes in a hidden layer and the number of hidden layers are variable. Depending upon the type of application, the network parameters such as the number of nodes in the hidden layers and the number of hidden layers are found by trial and error method, Hirose et. al [17]

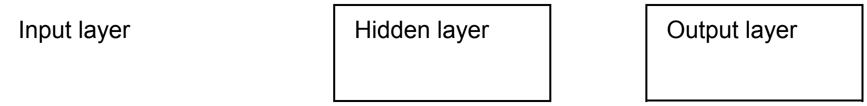


FIGURE 3: Multilayer perceptron

In most of the applications one hidden layer is sufficient. The activation function which is used to train the ANN, is the sigmoid function and it is given by:

(1)

where

f (x) is a non - linear differentiable function,

where

N_n is the total number of nodes in the n^{th} layer

W_{ij} is the weight vector connecting i^{th} neuron of a layer with the j^{th} neuron in the next layer.

θ is the threshold applied to the nodes in the hidden layers and output layer and

P is the pattern number.

In the first hidden layer, x_i is treated as an input pattern vector and for the successive layers, x_i is the output of the i^{th} neuron of the proceeding layer. The output x_i of a neuron in the hidden layers and in the output layer is calculated by :

(2)

For each pattern, error E (p) in the output layers is calculated by :

(3)

where

M is the total number of layer which include the input layer and the output layer,

N_M is the number of nodes in the output layer.

$d_i(p)$ is the desired output of a pattern and

$X_i^M(p)$ is the calculated output of the network for the same pattern at the output layer.

The total error E for all patterns is calculated by :

(4)

where

L is the total number of patterns.

4.1 Disadvantages of steepest-descent method

The number of cycles required for E to reach the desired minimum is very large. The E does not reach the desired minimum due to some local minima whose domains of attraction are as large as that for the global minimum. The algorithm converges to one of those local minima and hence learning stops prematurely or the value diverges. The updating of weights will not stop unless every input is outside the significant update region. The significant update region is from 0.1 to 0.9. Due to this, the output of the network will be approaching either 0.0 or 1.0. This requires a large number of iterations for the convergence of the algorithm.

5 Functional update method (FUM)

In classification problems, input patterns can be grouped into classified subset and misclassified subset for any given weights, Huang [18] The input patterns are said to be misclassified if the error `D' in the output layer is greater than 0.5 The input patterns are said to be classified if D is less than 0.5. Weights are modified only when D is greater than 0.5. The functional update algorithm used is as follows :

Step 1 : Initialize the weights randomly.

Step 2 : Present a pattern with new inputs and desired outputs.

Step 3 : Compute network output by Equation (2).

Step 4 : Determine V^n, the set of valid update data in the output layer for the i^{th} output node by :

$$0.5 < D < 1 - \in \qquad (5)$$

where

$\in$ is the error fixed by the programmer

If V^n is empty, i.e. not even one node in the output layer does satisfy Equation (5), go to step 8. Otherwise go to step 5.

Step 5: Compute the objective function E (p) by :

(6)

Step 6 : In BPA algorithm with FU, adapt weights by using equations given in Table 1.
Step 7 : Repeat by going to step 3.
Step 8 : Change the sigmoid function of the output neuron to the signum function

The main advantage of FUBPA is that it will stop as soon as the misclassified set is empty. The flow chart for FUBPA is given in Figure 4

6. DESCRIPTION OF EXPERIMENTS

6.1 Experimental set-up

Renal data such as person age in terms of years, male / female, Edema, Oliguri, Normochronic, Urgent, Hypertension, Diabetics, Family History, Polymer Chain Reaction, Obesity, Hemoglobin, Cholostral, Creatine have been collected for 1000 patients. The collected data are given in Table 1. A total of 17 parameters about renal organ have been collected from 1000 patients.

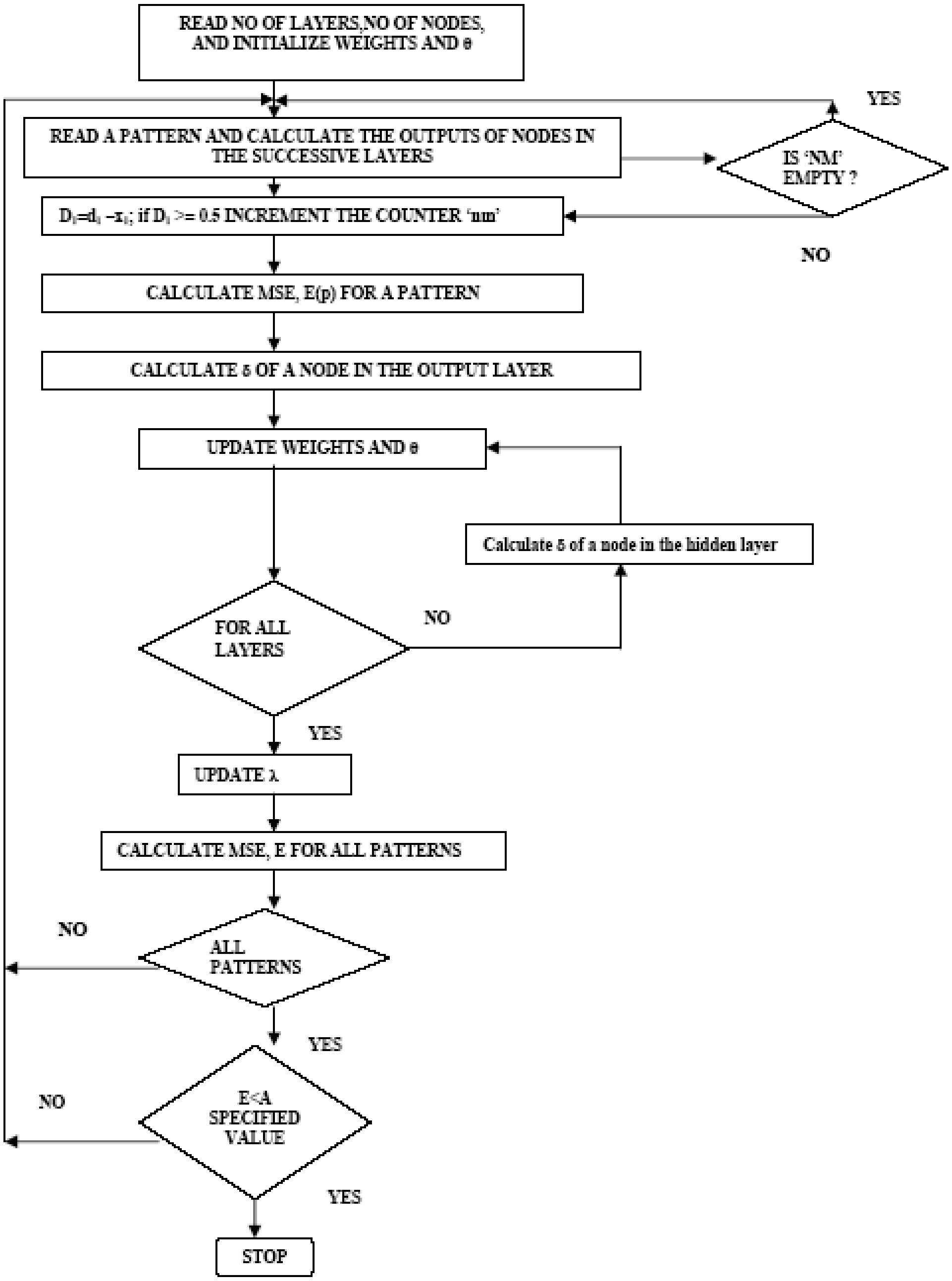

READ NO OF LAYERS,NO OF NODES, AND INITIALIZE WEIGHTS AND θ
YES
READ A PATTERN AND CALCULATE THE OUTPUTS OF NODES IN THE SUCCESSIVE LAYERS
IS 'NM' EMPTY ?
Dᵢ=dᵢ –xᵢ; if Dᵢ >= 0.5 INCREMENT THE COUNTER 'nm'
NO
CALCULATE MSE, E(p) FOR A PATTERN
CALCULATE δ OF A NODE IN THE OUTPUT LAYER
UPDATE WEIGHTS AND θ
Calculate δ of a node in the hidden layer
FOR ALL LAYERS
NO
YES
UPDATE λ
CALCULATE MSE, E FOR ALL PATTERNS
NO
ALL PATTERNS
YES
NO
E<A SPECIFIED VALUE
YES
STOP

FIGURE.4: Functional update Back Propagation Algorithm(FUBPA)

Age = [49 40 35 28 36 30 43 40 60 48 30 25 42 25 55 50 65 31 75 50 60 29 62 55 21 42 40 60 70 56 22 27 30 46 22 65
65 25 62 40 62 27 57 42 40 32 32 25 60 39 57 61 30 37 60 56 44 45 30 30 30 52 57 37 13 25 26 45 42 24 36 63 67 64 48
55 67 60 51 74 34 53 70 56 66 40 60 55 20 53 58 55 64 54 49 65 52 28 40 59 53 48 40 35 48 35 65 65 54 28 51 22 57 19
60 60 48 45 35 65 60 55 75 50 53 60 72 33 60 59 60 74 51 45 42 13 58 58 63 18 28 20 59 40 50 40 60 45 46 52 48 27 72
62 59 45 25 60 32 47 25 45 55 45 26 20 45 38 35 66 52 60 47 43 70 41 50 40 34 70 56 49 67 66 54 46 68 55 26 55 40 54
30 62 70 65 41 42 65 49 55 30 50 48 56 45 61 41 65 48 43 70 53 51 50 25 33 49 55 52 60 25 42 40 54 17 70 40 42 70
52];
Sex = [1 0 1 1 0 1 1 1 0 1 1 0 1 0 0 1 1 1 1 1 1 0 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 0 0 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1
1 1 0 0 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 0 0 1 0 1 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1 1
1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1 0 1 0 1 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 1 1 1 1 0 0 1 1 1 1 1 0 1 0 1 1 0 1 0
1 1 1 1 1 0 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 0 1 1 0 1 0 0 0 0 0 0 0 1 1];
Ede = [1 1 1 1 1 1 1 0 1 1 0 1 1 1 0 0 0 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 1 0 1 0
1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 0 1 1 1 0 1 0 1 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 0
1 0 1 1 1 1 1 1 1 0 1 0 1 0 1 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1];
Puf = [1 1 1 1 1 1 1 0 1 1 0 1 1 1 0 0 0 1 1 0 1 0 1 0 1 1 1 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 0 1 0 0 1 0 1 1 0 1 1 1 0 1 0
1 1 1 1 1 1 1 1 1 0 1 0 1 1 0 0 0 0 0 1 1 1 0 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 0
0 0 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 0 1 0 1 1 1 1 1 1 1 1 0 0 1 1 0 0 0 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 1 0 1 1 1 0 1 1
1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0 0 1 0 1 1 1 0 1 1];
Oli = [1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 0 0 1 1 1 0 1 1 1 1 0 0 1 1 1 0 1 1 1 1 0 1 0 1 0 1 0 1 1 0 0 1 0 1 1 0 1 1 1 1 1 0 0 1 1
0 1 1 1 1 1 0 1 1 1 0 1 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 1 1 0 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1 1 1 1 0 0 1 0 1 0 0 0 0 1 0 0 1 1 1 0 0
0 0 1 1 1 0 0 1 1 0 1 1 0 1 1 1 0 0 0 0 1 1 0 0 1 1 0 0 0 0 1 0 0 0 1 0 1 0 0 0 1 1 1 1 1 0 0 1 0 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1
1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 1 0 1 0 1 0];
Pol = [0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0];
Noc = [0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
1 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 0];
Urg = [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];
Hes = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];
Hyp = [0 0 0 1 0 1 0 0 1 1 1 1 1 1 1 0 0 1 1 0 0 1 0 0 1 1 1 1 1 0 0 1 0 1 0 1 0 1 1 0 1 1 1 1 0 1 1 1 0 0 0 1 0 1 1 2 1 1 1 1 0
1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 1 1 2 0 2 1 1 1 1 1 1 3 0 0 1 0 0 1 0 1 0 2 0 0 1 1 1 1 1 0 0 1 1 1 2 0 1 0 1 0 2 1 1
1 1 1 1 1 1 1 1 0 1 3 1 1 0 1 0 0 1 0 1 1 1 3 1 1 1 1 0 3 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 0 1 1 1 0 0 3 3 1 1 0 1 2 1 1 1 1
1 1 1 2 1 1 1 1 2 1 0 0 1 2 0 2 1 0 1 0 1 1 1 1 0 1 1 1 2 1 1 1 1 1 1 0 0 0 1 0 1];
Dia = [3 0 0 0 0 0 0 0 0 3 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 1 3 0 0 0 0 0 0 0 0 0 1 3 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 1 0 0
1 0 3 0 0 0 1 1 0 0 0 0 0 0 0 3 0 1 3 0 1 0 0 2 0 0 2 0 0 0 0 0 3 3 0 0 0 0 0 1 0 3 0 0 0 0 2 2 0 0 0 3 0 2 1 1 0 0 1 2 0 2 3 2 1
0 0 0 2 1 0 1 0 0 0 2 0 2 0 0 0 2 2 0 0 1 3 0 0 0 0 3 0 2 0 0 3 0 0 0 3 0 0 0 0 0 0 0 0 3 0 1 3 0 0 1 3 0 0 0 0 0 0 0 0 3 3 0 0 0 0 1
2 0 0 2 2 0 0 2 0 0 0 0 1 3 0 0 0 1 1 1 0 3 2 1 0 0 1 0 2 0 1 0 0 0 0 2 1 0 1 3];
Tob = [0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 1 0 1 1 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 1 1 0 0 0
0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 0 0 0 1 0 0 0 0 0 1 1 0 1 1 0 1 1 1 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1
1 0 0 1 0 1 0 0 1 0 1 1 0 0 1 0 0 0 1 1 1 0 0 1 1 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 1 1 1 0 0 0 1 0 1 1 0 0 0 0 0 0 1 0
0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 0 0 0 0 1 0 1 0];
Fam = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];
PCR = [2.2 2.65 2.0 1.31 4.04 0.4 .65 .22 1.42 1.7 2.7 .65 2.19 1.89 .87 .66 .68 1.1 2.4 3.43 2.09 1.76 2.64 .06 .45 .27 2.7
3.18 1.15 2.78 1.71 1.1 2.4 1.76 .16 .36 1.93 4.2 1.5 1.8 1.6 1.69 1.51 .85 .98 2.2 3.92 .7 .3 .55 2.4 .18 3.56 .3 2.15 3.83
.74 .18 1.95 2.8 2.4 1.17 1.78 3.78 1.34 1.14 3.38 3.18 .27 1.42 1.5 .3 1.2 2.4 2.0 .98 1.0 .30 1.1 2.0 1.2 2.1 .23 .38 .5 .3
1.94 1.0 1.0 .12 .2 2.2 1.51 .97 1.3 1.8 .72 3.0 .31 2.7 2.1 .76 2.44 1.47 1.17 1.4 .75 2.02 1.67 .37 1.1 2.0 1.08 1.82 2.0
1.9 2.3 1.2 2.13 1.37 .04 .56 .57 1.85 2.8 1.22 .25 1.28 .3 1.46 2.1 .44 1.87 .91 1.2 1.3 .25 .58 2.5 1.67 1.53 2.28 1.9 .95
2.74 .62 .28 2.1 1.35 2.5 2.7 2.6 1.18 .03 .98 .58 2.8 .7 2.61 .5 1.95 3.25 1.1 .76 .9 1.9 .8 .4 1.44 .65 1.4 1.62 2.6 2.01
1.65 2.39 .28 .5 .4 .9 .16 2.1 1.73 3.7 3.4 3.0 1.77 1.5 .7 .8 3.5 2.74 .18 1.91 .02 3.62 1.4 .22 1.2 .3 2.9 .4 3.9 2.44 1.9 1.1
.4 2.08 1.8 2.0 2.06 4.29 2.6 1.55 .7 1.3 2.4 1.6 .07 1.71 .4 2.51 .4 .2 .14 1.6 2.62 1.3 .32 .4 1.52];
Obs = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];
Hem = [13.2 10.8 7.2 12.0 12.1 13.8 13.0 9.8 12.8 13.7 11.4 8.0 8.8 11.6 12.1 12.0 8.8 13.0 10.2 10.8 7.1 7.8 10.8 13.0 13.0 12.6 4.8 11.0 11.5 12.6 8.8 11.0 3.8 9.8 10.2 11.0 9.2 5.2 10.8 12.0 12.6 12.6 9.1 11.5 13.2 11.0 8.8 8.2 11.0 13.0 12.8 12.1 11.0 10.8 9.7 7.2 12.6 11.5 13.8 7.8 3.8 11.0 10.2 8.8 9.4 12.6 8.8 11.0 12.6 12.8 13.9 10.8 12.0 9.8 9.8 8.8 13.2 11.0 13.0 9.2 11.5 8.2 9.8 15.0 14.2 11.0 9.2 7.8 7.8 11.5 12.0 9.8 11.0 13.2 5.2 10.2 10.2 7.2 9.8 7.8 8.2 10.8 11.8 9.2 9.8 12.0 8.2 8.2 8.8 10.2 12.06 12.8 13.0 8.2 12.8 11.5 9.8 9.8 14.2 15.0 11.4 11.5 10.3 9.2 7.8 13.9 12.8 14.0 11.0 9.4 8.2 12.6 10.2 8.4 9.2 8.2 9.0 6.0 16.8 7.1 9.4 4.8 8.8 10.2 8.4 13.2 13.6 5.2 10.8 9.8 10.8 8.8 10.2 10.8 9.2 13.0 6.6 13.9 7.2 7.8 9.8 11.5 7.8 11.4 11.8 8.9 8.2 6.6 7.2 9.4 9.2 12.8 10.3 12.8 12.1 9.2 9.8 6.8 11.4 8.8 12.0 6.8 8.8 8.8 7.8 11.0 12.2 13.9 8.7 10.8 10.8 8.7 10.8 10.2 7.8 10.2 11.0 12.0 8.2 10.2 7.8 13.8 10.2 12.0 9.2 14.2 7.8 5.8 9.2 9.8 9.8 9.8 7.8 9.2 11.0 9.8 6.8 8.8 12.6 6.8 8.8 9.2 5.2 9.2 9.8 6.0 8.2 10.3 7.2 7.2 7.2];
Cre = [330 88 180 299 115 875 88 340 180 280 270 290 185 88 88 370 240 105 490 710 180 360 550 130 88 85 466 550 140 105 85 105 369 260 88 250 88 330 280 160 200 330 190 470 430 330 330 380 200 732 88 250 220 350 410 410 190 260 200 360 369 95 430 95 600 270 350 90 85 260 320 120 380 622 166 290 157 170 530 825 380 320 270 300 88 536 350 710 190 565 170 360 340 88 888 580 470 350 510 1500 320 470 130 710 809 95 350 550 120 732 200 334 88 180 80 84 200 390 1300 110 140 210 450 220 704 450 200 380 110 450 290 310 80 510 270 228 210 501 651 501 1478 430 105 316 320 360 88 563 387 350 290 431 378 88 280 220 1300 88 825 600 430 88 410 430 500 732 500 600 600 550 510 100 410 430 320 180 310 616 360 825 90 607 756 290 123 120 149 475 176 210 237 457 510 280 86 175 192 254 250 853 324 260 184 404 157 572 422 1540 202 298 280 352 114 86 289 501 1170 271 80 184 642 80 1082 457 175 448 369 219 589 289 271];
Chol = [175 178 176 206 206 172 190 190 175 180 196 210 380 196 180 182 186 168 176 179 169 185 172 176 185 172 179 176 185 185 236 196 166 190 168 200 172 188 170 182 196 195 189 168 190 195 166 166 180 169 165 192 182 179 206 170 180 208 188 191 166 183 196 260 196 196 180 185 190 180 208 196 191 198 166 188 182 177 182 166 165 186 188 201 188 170 190 186 192 196 186 182 200 180 186 181 185 182 165 162 186 180 165 195 177 210 179 170 192 188 166 185 200 195 236 141 275 195 161 172 165 173 185 188 166 162 195 199 178 185 163 164 206 190 188 180 188 165 172 180 168 176 190 150 171 182 178 175 173 170 181 170 172 168 190 178 170 168 173 185 180 172 179 190 188 196 184 152 189 188 192 168 190 172 168 215 177 166 188 151 196 180 196 181 178 211 185 190 168 146 190 175 185 195 168 172 168 190 156 188 186 166 190 200 170 161 155 172 165 172 180 190 160 186 167 167 152 162 193 180 148 210 160 190 162 182 182 172 168 165 176];

Table 1 Sample Renal data collected from patients

Target outputs
Classification =[3 1 2 2 1 1 1 1 2 3 1 2 2 2 3 1 2 1 2 2 2 2 2 1 1 3 3 2 2 1 1 2 2 2 1 2 3 2 2 1 2 2 3 2 1 1 1 2 1 1 1 1 1 2 2 3 2 1 1 2 2 2 2 3 2 2 2 3 3 2 1 1 1 2 2 1 3 1 3 3 2 3 1 1 3 1 2 3 2 2 1 2 2 3 3 1 2 2 1 2 3 2 3 2 1 2 2 3 3 2 1 2 3 2 3 3 3 2 1 2 3 2 3 3 3 3 1 1 2 3 3 2 3 2 2 2 3 2 2 2 2 1 3 3 2 1 3 3 2 2 2 2 3 1 1 1 2 3 2 1 2 3 2 2 2 2 2 2 2 2 3 2 3 3 2 2 3 3 2 1 2 1 2 2 2 3 3 2 2 1 1 2 3 1 2 3 3 2 2 3 2 2 1 2 2 3 1 2 2 2 2 2 2 3 3 2 2 2 3 2 3 2 2 2 2 2 2 3 2 2 3 3 2 2 2 2 2 3 2 3 2 2 2 2 2 2 2 1 2 2 1 1 2 2 2 3 1 2 2 3 2 2 3 3 2 1 2 2 2 2 2 2 2 3 2 1 2 2 2 2 1 2 2 2 2 1 2 2 1 3 2 2 1 2 3 2 2 2 1 2 2 1 2 3 3 3 2 2 3 2 3 2 2 2 1 2 2 2 3 2 2 2 2 2 2 2 1 2 3 2 3 2 2 2 2 2 2 2 3 2 2 2 3 2 2 3 2 3 2 2 3 2 2 3 3 2 2 2 2 2 3 3 3 2 2 2 2 2 3 3 2 2 2 2 2 2 2 3 3 2 2 2 3 3 3 2 1 1 2 2 3 3 3 3 3 2 2 2 2 2 3 2 1 2 3 2 2 2 2 1 2 2 2 2 2 1 3 2 3 3 2 3 2 2 2 3 2 2 1 2 1 2 2 1 2 2 2 3 3 2 3 2 3 3 3 2 3 2 2 3 2 3 3 3 2 2 2 2 2 2 2 2 3 3 3 3 3 3 2 2 3 2 2 2 1 3 3 2 2 2 3 2 3 3 1 3 2 3 2 2 2 1 3 2 2 2 1 3 3 2 2 2 1 2 2 2 2 2 2];

Table 2 Corresponding target outputs used for training the ANN

6.2 Data Preparation

The data include information on the dialysis prescription, data electronically collected during each dialysis treatment, laboratory tests, pharmacy records, patient diagnosis and demographic data. Before each session, the patient was weighed and her/his blood pressure (systolic and diastolic) registered while sitting (supine pressure), and when possible, while standing. The weight and blood pressure measurements are repeated at the end of the session. The levels of sodium, bicarbonate, potassium, calcium, and glucose in the dialysis solution are recorded.

Total time for the dialysis session, blood flow rate, total volume of blood processed, dialysis flow rate, and the overall average pressures at the arterial and venial side of the blood pump were another set of collected values. A set of measurements was collected by the dialysis machine every twenty minutes or on request. This set includes systemic blood pressure, pulse, blood flow rate, arterial and venial blood pump pressures, trans-membrane pressure, and the rate of ultrafiltration. To reduce data noise, averages were computed over the fifteen readings taken by the machine during the dialysis session.

The demographic and outcomes data set contains the patient's date of birth, gender, and race; the date(s) of death, kidney transplant, and transfer into or out of the dialysis center. The final portion contains the diagnosis codes for the primary and secondary diagnoses. Differences between each patient's average post and pre systemic blood pressures were calculated for all four combinations of systolic and diastolic pressures and supine and standing positions. The pulse pressures (determined by the difference between the systolic and diastolic blood pressures) were calculated for both pre and post conditions for both supine and standing positions. Differences between the supine and standing pressures were also calculated for both systolic and diastolic blood pressure and for the pre and post dialysis conditions. Some new features were added to the data set by using the concept of data transformation. Averages were computed for each patient for all variables to form a single representative record (aggregate data set). Initial data mining focused on a selected group of long-term dialysis patients with at least fifteen or more visits.

6.3 Selection of data

Selection of patterns for training the neural network is important as they should be representative of all the patterns collected during machining. Therefore, statistical techniques have been used to select the patterns out of 500 patterns collected during the experiment. The number of classes selected are two. Patterns with maximum variance VE$_i^2$ are selected. The maximum VEi^2 of a pattern is calculated by:

(7)

where

nf is the number of features.

7 RESULTS AND DISCUSSIONS

Data mining has been carried out using an approach of partial individual visit data set mining. The grouping of features for partial data sets was prepared, keeping in mind medical relevance between these features (e.g. dialysis chemical solution, weight, blood pressure, difference in blood pressure (i.e. pulse pressure), etc. Eleven different combinations were determined to form trial data sets. These eleven data sets were mined separately using rough set based and decision-tree based data mining algorithms. Each data subset produced two sets of rules (classifiers), one each from the two data mining algorithms. Thus in all there were twenty-two classifiers capable of predicting the outcomes for new patients. These classifiers were developed to perform multi-angle, highly reliable (parallel redundancy concept in reliability engineering), robust, accurate decisions/predictions. The classifiers can be combined to form a single classifier, which could be used for prediction of new patients or individual classifiers could come with their own prediction and these predictions, could be combined by using voting/weighted-voting schemes. There was considerable increase in the prediction accuracy of individual visit over the aggregate data set

7.1 Medical Significance

The significant features identified by data mining algorithms are as follows diagnosis, time on dialysis, deviation from target weight, blood pressures ranges for different patients, calcium and potassium levels in dialysis solution, total blood volume, blood flow rate, venial pressures. Table 2 gives the classification performance and Table 3 gives the amount of misclassification for different number of nodes in the hidden layer of the network

SL. No	No of Hidden layers	Classifications		
		I	II	III
		53	116	62
1	5	49	92	59
2	6	49	92	59
3	7	51	90	58
4	8	51	90	58
5	9	52	88	60
6	10	51	92	59
7	11	51	92	59
8	12	50	92	57
9	13	48	95	60
10	14	52	82	56
11	15	52	82	56
12	16	50	98	59
13	17	46	99	60
14	18	50	91	53
15	19	50	92	51
16	20	49	78	33
17	21	41	96	60

TABLE 3: Effect of nodes in hidden layer and percentage of classification

8 CONCLUSION AND FUTURE SCOPE OF WORK

This work addresses the problem of recognition of visual types of renal artery lesions from radiological signs. Important issues are related to this work, in particular the determination of a visual type independent of the observer. To evaluate the extent to which the result of the classification is objective, we need to establish a ‘significant cases database as well as to justify and validate the quantification scheme used in the domain. Another aspect of this work is to provide a conceptual description of normal and abnormal aspects of a renal artery that can be integrated into a more general medical decision making systems.

The most significant result obtained from this research was to demonstrate that data mining, data transformation, data partitioning, and decision-making algorithms are useful for survival prediction of dialysis patients. The potential for making accurate decisions for individual patients is enormous and the classification accuracy is high enough (above 75–85%) to warrant use of additional resources and conduct further research. Data transformation increased the classification accuracy by approximately 11%. Analyzing and comparing the data mining rule sets produced a list of significant parameters, such as the diagnosis, total dialysis time, potassium, calcium and sodium levels, deviation from target weight, arterial pressure, post-dialysis pulse rate supine, difference between post- and pre-supine.

SL. No	No of Hidden layers	Mis classifications		
		I	II	III
		53	116	62

1	5	4	24	3
2	6	4	24	3
3	7	2	26	4
4	8	2	26	4
5	9	1	28	2
6	10	2	24	3
7	11	2	24	3
8	12	3	24	5
9	13	5	21	2
10	14	1	34	6
11	15	1	34	6
12	16	3	18	3
13	17	7	17	2
14	18	3	25	9
15	19	3	24	11
16	20	4	38	29
17	21	12	20	2

TABLE 4 Effect of no. of nodes in hidden layer and misclassification

REFERENCES

1 Altman, D. 1991. Practical Statistics for Medical Research, Chapman and hall.

2 Kleinbaum, D.G., Kupper ,L.L.(eds.) 1982. Epidemiologic Research: Principles and Quantitative Methods, John Wiley &: Sons, New York.

3 Tsumoto, S. G5: Medzcine, In: Kloesgen, W. and Zytkow, J. (eds.) Handbook of Knowledge Dicovery and Data Mining.

4 Van Bemme1,J. and Musen, M. A.1997. Handbook of Medical Informatics, Springer-Verlag, New York.

5 Y Shahar & MAMusen, 'Knowledge-based Temporal Abstractioin in Clinical Domains' Artif. Intell. In Med. 8, 1996, pp.267-298.

6 Ming-Syan Chen, Jiawei Han and Philip S. Yu. Data Mining: An Overview From a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, Vol. 8(6), December 1996, pp. 866-883.

7 Pena-Mora, F. & Hussein, K. 1998, Interaction Dynamics in Collaborative Civil Engineering Design Discourse: Applications in Computer Mediated Communication.Journal of Computer Aided Civil and Infrastructure Engineering, Vol. 14, pp. 171-185

8 Usama Fayyad. 1997, Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases. Proceedings of the 9th International Conference on Scientific and Statistical Database Management (SSDBM '97). Olympia, WA, pp. 2-11.

9 Usama M. Fayyad. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, October 1996, pp. 20-25.

10 M. S. Sousa, M. L. Q. Mattoso and N. F. F. Ebecken.(1998). Data Mining: A Database Perspective. COPPE, Federal University of Rio de Janeiro, pp.1-19.

11 Themistoklis Palpanas. Knowledge Discovery in Data Warehouses. ACM Sigmod Record. vol. 29(3), September 2000, pp. 88- 100.

12 Fortuna L, Graziani S, LoPresti M and Muscato G (1992), Improving back propagation learning using auxiliary neural networks, Int. J of Cont. , 55(4), pp 793-807.

13 Lippmann R P (1987), An introduction to computing with neural nets, IEEE Trans. On Acoustics, Speech and Signal Processing Magazine, V35, N4, pp.4.-22

14 Yao Y L and Fang X D (1993), Assessment of chip forming patterns with tool wear progression in machining via neural networks, Int.J. Mach. Tools & Mfg, 33 (1), pp 89 -102.

15 Hush D R and Horne B G (1993), Progress in supervised neural networks, IEEE Signal Proc. Mag., pp 8-38.

16 Bernard Widrow (1990), 30 Years of adaptive neural networks: Perceptron, madaline and back-propagation, Proc. of the IEEE, 18(9), pp 1415 - 1442.

17 Hirose Y, Yamashita K Y and Hijiya S (1991), Back-propagation algorithm which varies the number of hidden units, Neural Networks, 4, pp 61-66.

18 Shih-Chi Huang and Yih-Fang Haung (1990), Learning algorithms for perceptrons using back-propagation with selective updates, IEEE Cont. Sys. Mag., pp 56-

19 Manal Abdel Wahed, Khaled Wahba (2004), Data Mining Based-Assistant Tools for Physicians to Diagnose Diseases ,IEEE Trans, pp 388-391.

20 Adam E. Gaweda, Alfred A. Jacobs and Michael E. Brie (2003), Artificial Neural Network-based Pharmacodynamic Population Analysis in Chronic Renal Failure, IEEE Tans, pp 71-74