Data Mining using Artificial Neural Network Rules

Pushkar Shinde
MCOERC , Nasik

Abstract - Diabetes patients are increasing in number so it is necessary to predict , treat and diagnose the disease. Data Mining can help to provide knowledge about this disease. The knowledge extracted using Data Mining can help in treating and preventing the disease. Artificial Neural Network (ANN) can be used to create an classifier from the data. The neural network is trained using backpropagation algorithm The knowledge stored in the neural network is used to predict the disease. The knowledge stored in neural network is extracted using NeuroRule method. The knowledge extracted is in form of human readable rules which can used to analyze the disease amd in turn help in treating the disease.

Keywords : Diabetes, data mining, neural network rule extraction, NeuroRule

I. INTRODUCTION

Diabetes is an chronic disease i.e. a long term disease which can be controlled but not cured. The number of patients suffering from diabetes is increasing. This has caused various researchers to work on controlling the disease. Diabetes is a condition due to malfunction in the pancreas. The pancreas doesn't produce enough insulin to control the sugar level in the blood. Data Mining can be used to extract knowledge from medical data. The knowledge extracted can help in predicting, diagnosing and preventing diabetes.

Data mining is the process of discovering useful knowledge in data and also finding the inter-relation pattern among the data [1].

Data mining is used to extract new knowledge from existing data. The knowledge is hidden in the data, which is extracted using data mining. Data Mining requires large amount of data.

Medical data can be used for data mining. The Pima Indian diabetes data set is used for performing data mining. The data set has been created for female patients with attributes like number of times pregnant, age, BMI, blood pressure, plasma glucose level. Attributes like this can give more knowledge about their effect on diabetes.

The medical industry is among the most information intensive industries. Medical data keep growing on a daily basis. From this data useful knowledge should be extracted to provide quality health care .With the help of data mining methods, useful patterns and relationships of information can be found within the data, which can be utilized for diagnosis, prediction and detections of the trend of the disease [1].

ANN has been applied in many applications with remarkable success. For example, ANN have been successfully applied in the area of speech generation and recognition, handwritten character recognition , vision and robotics .

ANN can be used to extract knowledge from the data. The knowledge extracted using ANN is stored in the form of neural network. This knowledge is not comprehensible. So ANN is also called as an “black box”. ANN can be used as an classifier, but doesn't easily provide information about the classification decision. To overcome this limitation an technique called NeuroRule is used as an decompositional technique. This technique extracts knowledge from the neural network as simple human readable rules. These rules provide further knowledge about the disease. These rules also help in justifying the neural networks classification decision. These rules provide information about diabetes which can be used in treating, diagnosing and preventing diabetes.

In this paper an approach named Artificial Neural Network Rules (ANNR), i.e. ANN training preceded by rules extraction method. This approach is helpful for utilizing the power of ANN in data mining applications where comprehensibility is as important as the generalization ability. In other words this method overcome the "black box nature" of ANN. This method is tested on diabetic data set.

This paper is organized as follows. In section 2, we briefly explain the basic concepts of ANN. Section 3 presents the rule extraction, section 4 gives the experimental result and followed by conclusion in section 5.

II. NEURAL NETWORK

A three layer neural network having thirty two neurons in the input layer, nine neurons in the hidden layer, one neuron in the output layer is considered (shown in Figure 1). The input data is binarized and given as input to input layer. Each attribute is clustered into an range and four clusters are formed for each attribute.

Backpropagation algorithm is used to train the neural network. Backpropagation algorithm is an gradient descent algorithm. The algorithm initializes the weights of the neural network to a random value initially. The network activation values are calculated by multiplying the weights with the input neurons. The activation function used is an sigmoid function which is calculated using hyperbolic tangent of the summation. To determine the optimal weights for a classification and prediction problem is, to minimize the error. The most common method to define a cost function is, as the root mean square error function between the networks predicted output (Yk) and the expected output (Ok). This is commonly known as the root mean square error (RMS error) cost function.
is used., and it is computed using equation (1)

$$E(p) = \sqrt{\sum_{k=0}^{P}(Ok - Yk)^2}\,/P$$

where P is the total number of patterns in the data set,
is the output units, Ok is the target value at the kth output neuron for pth sample and Yk is the actual output at the kth output neuron for the pth sample.

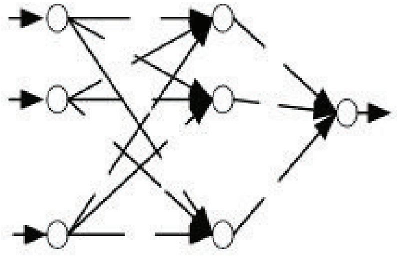


Figure 1:- Three layer Neural Network

An approach to determine the weights minimize the cost function is to use the stochastic algorithm: gradient descent with back-propagation. This is one of the simplest network training algorithms and is also known as steepest descent. With gradient descent, the initial weight vector Wold is often chosen at random, then with each iteration the weights are updated such that we move a distance in the opposite direction of the error function gradient in each iteration (1).

The weights are updated at each iteration as follows:

$$W_{new} = W_{old} + \left(-\eta \frac{\delta E(w)}{\delta w} \right) \qquad (2)$$

Where n is the is called the learning rate. A large learning rate leads to rapid learning but the weights may oscillate while lower learning rates leads to slower learning. The learning rate is fixed at 0.01 in
this work.

III. RULE EXTRACTION

The neural network once trained contains knowledge about diabetes. The neural network is used to classify and predict the class of the input person. The knowledge that is hidden in the neural network can be extracted using decompositional technique like NeuroRule[2]. The NeuroRule technique computes the hidden neurons that activate the output neuron. The hidden neurons that have weights greater than an active threshold value are considered. Once the hidden neurons are selected the input neurons that activate the hidden neurons are selected and using this information the rules are extracted.

The rules were extracted from the neural network by identifying the hidden neurons that have activation level above the threshold. The input-hidden rules for the corresponding hidden neurons are identified. NeuroRule technique is used for rule extraction. To facilitate the rule extraction the values of the numeric attributes were discretized. Each of the six attributes with numeric values was discretized by dividing its range into subintervals. For example the plasma glucose value is having range 81 to 199 is divided into four subintervals. The rules were extracted such that there are positive and negative rules. The positive rules are given such that they give the attribute values causing diabetes. Negative rules are given which give attribute values for not having diabetes.

IV. EXPERIMENTAL RESULTS

4.1. DATA SET

The Pima Indians Diabetes database [8] is used to test the proposed procedure. The total data is 768 from which 461 samples (60%) are randomly chosen and used as training patterns and tested with 307 instances (40%) of the same data set. Each instances
sample represents eight attributes of female patients
of Pima Indian heritage. The eight attributes are namely, number of times pregnant, plasma glucose concentration (2 hours in an oral glucose tolerance test), Diastolic blood pressure (mm Hg), Triceps skin
fold thickness (mm), 2 hours serum insulin (mU/ml),
body mass index (weight in kg/(height in m)2 Diabetes pedigree function, Age (years) .

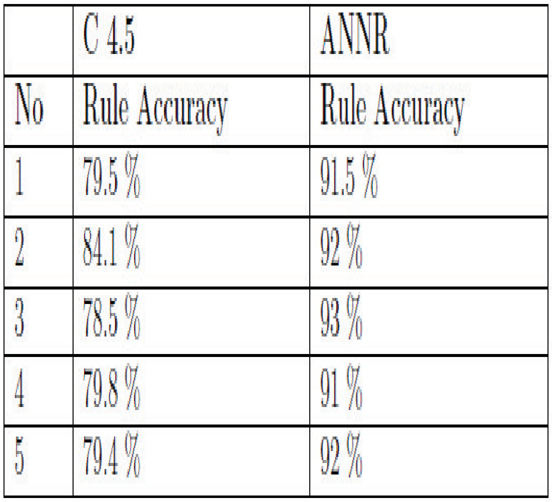
4.2. TRAINED NEURAL NETWORK

First the network is trained using backpropagation algorithm. The trained network uses sigmoid activation function. The activation threshold has been taken as 0.7 and accordingly the RMS error has been calculated for the neural network.

Table 1: Neural Network training performance

	In Neurons :32	**Hidden Neurons: 8**	**Output Neurons: 1**	**Learning rate : 0.01**
No	**Epoch**	**Time (s)**	**Train data accuracy**	**Test data accuracy**
1	5000	150	90%	90%
2	5000	160	89%	90%
3	5000	150	90%	90%
4	5000	153	89.5%	89%
5	10000	250	90.5%	90%

Table 2 : Extracted Rule Accuracy



	C 4.5	ANNR
No	Rule Accuracy	Rule Accuracy
1	79.5 %	91.5 %
2	84.1 %	92 %
3	78.5 %	93 %
4	79.8 %	91 %
5	79.4 %	92 %

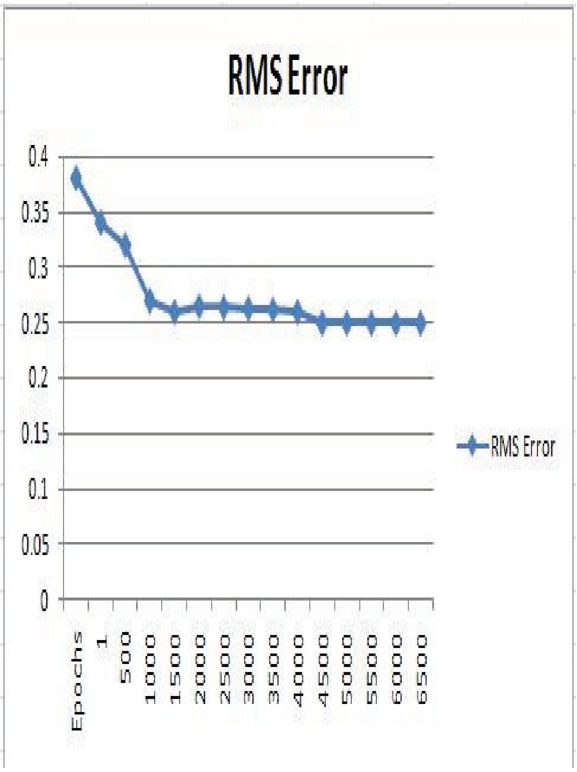
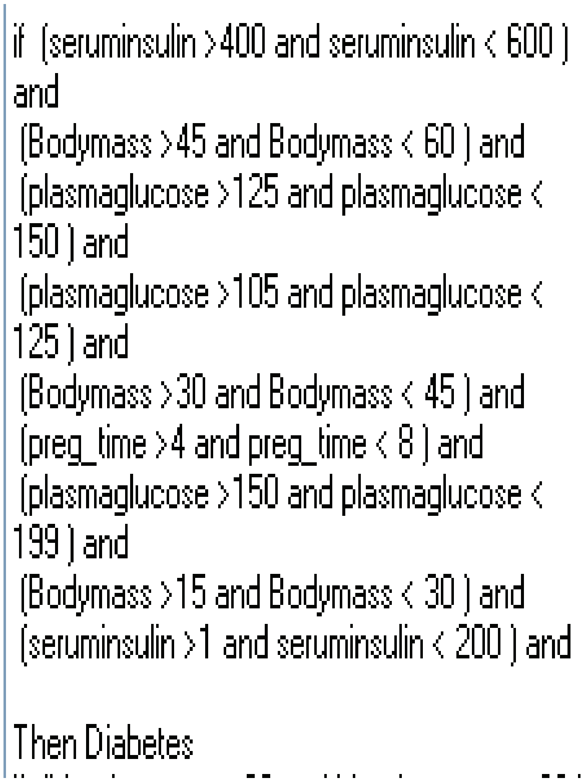


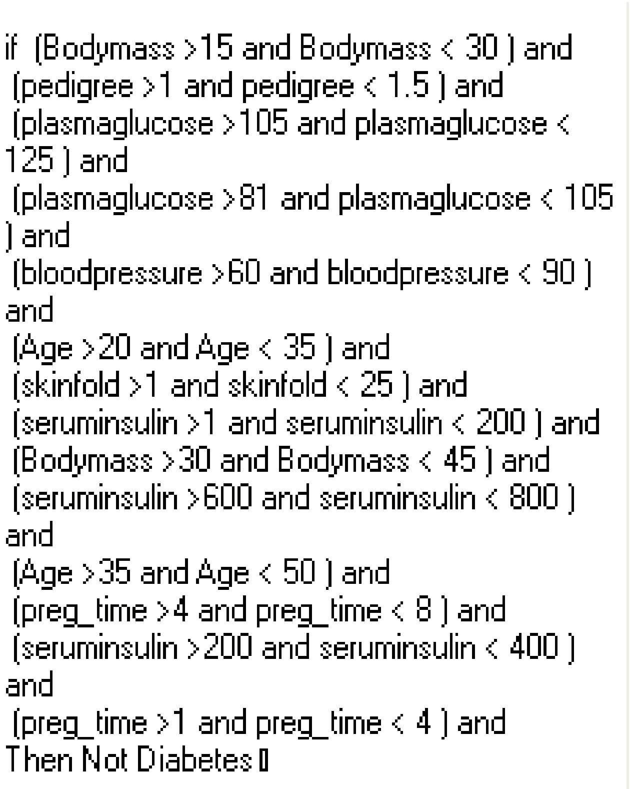
Figure 2 : RMS error



```
if  (seruminsulin >400 and seruminsulin < 600 )
and
 (Bodymass >45 and Bodymass < 60 ) and
 (plasmaglucose >125 and plasmaglucose <
150 ) and
 (plasmaglucose >105 and plasmaglucose <
125 ) and
 (Bodymass >30 and Bodymass < 45 ) and
 (preg_time >4 and preg_time < 8 ) and
 (plasmaglucose >150 and plasmaglucose <
199 ) and
 (Bodymass >15 and Bodymass < 30 ) and
 (seruminsulin >1 and seruminsulin < 200 ) and

Then Diabetes
```

Figure 3: Part of Diabetes rules extracted



```
if  (Bodymass >15 and Bodymass < 30 ) and
 (pedigree >1 and pedigree < 1.5 ) and
 (plasmaglucose >105 and plasmaglucose <
125 ) and
 (plasmaglucose >81 and plasmaglucose < 105
) and
 (bloodpressure >60 and bloodpressure < 90 )
and
 (Age >20 and Age < 35 ) and
 (skinfold >1 and skinfold < 25 ) and
 (seruminsulin >1 and seruminsulin < 200 ) and
 (Bodymass >30 and Bodymass < 45 ) and
 (seruminsulin >600 and seruminsulin < 800 )
and
 (Age >35 and Age < 50 ) and
 (preg_time >4 and preg_time < 8 ) and
 (seruminsulin >200 and seruminsulin < 400 )
and
 (preg_time >1 and preg_time < 4 ) and
Then Not Diabetes
```

Figure 3: Rules for diabetes

VI. CONCLUSION

In this paper, with the proposed ANNR approach, ANN can be utilized in data mining application. First ANN was trained using backpropagation algorithm , then the rules are extracted from the ANN using NeuroRule method. Rule extraction using ANNTR from trained ANN is more accurate. It also can work on continuous and discrete data. ANNR approach was used on diabetes data set, where it shows that this approach could benefit diabetes diagnosis because it could generate rules with strong generalization and comprehensibility ability. The knowledge gained is comprehensible and can enhance the decision making process by the doctors and will be a valuable tool for diabetes researchers. Future work on ANNR will be on reducing the redundant rules.

REFERENCES

[1] S. Kalaiarasi Anbananthen Sainarayanan , Ali Chekima, Jason Teo "Data Mining using Artificial Neural Network Tree", IEEE Conference on Computers , Communications and Signal Processing.

[2] Hongjun Lu , Rudy Setino , Huan Liu , "NeuroRule A connectionist Approach to Data Mining",VLDB Conference

[3] Jiawei Han, Micheline Kamber "Data Mining Concepts and Techniques" .

[4] S.Kalaiarasi Anbananthen, Fabian H.P. Chan,

[5] K.Y. Leong. Data Mining Using Decision Tree Induction of Neural Networks, 3rd Seminar on Science and Technology. Kota Kinabalu : SST. 2004

[6] http://www.1iacc.up.pt/ML/statlog/datasets/diabetes/diabetes.doc.html

[7] Sethi, I.K. Layered Neural Net Design through Decision Trees. Circuits and Systems, IEEE International Symposium. 1990.

[8] A. K. C. Wong and Y. Wang. Pattern Discovery: A Data Driven Approach to Decision Support. IEEE Trans. On Systems, Man and CyberneticsPart C: Applications and Reviews., vol. 33, pp.II4-893.2003

[9] U.M. Fayyad, G.P. Shapiro, and P. Smyth. Advances in Knowledge Discovery and Data Mining. California: Menlo park. 1996.

[10] A. K. C. Wong and Y. Wang. Pattern Discovery: A Data Driven Approach to Decision Support. IEEE Trans. On Systems, Man and CyberneticsPart C: Applications and Reviews., vol. 33, pp.II4-893.2003