ISSN 2277-3061

DATA MINING USING ARTIFICIAL NEURAL NETWORK POS-NEG RULES

Pushkar Shinde , Dr. Varsha Patil
MCOE&R , Nasik
Pushkar.shinde79@gmail.com
MCOE&R , Nasik
Varsha.patil@gmail.com

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 Pos-Neg sensitivity method. The knowledge extracted is in form of sensitivity analysis to analyze the disease and in turn help in treating the disease.

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

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 11, No. 6
editor@cirworld.com
www.cirworld.com, member.cirworld.com

1. INTRODUCION

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 Pos-Neg sensitivity analysis 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 Pos-Neg 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 pos-neg rule extraction, section 4 gives the experimental result and followed by conclusion in section 5.

2. 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 normalized and given as input to input layer. Each attribute is normalized into an range 0 to 1.

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)

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

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.

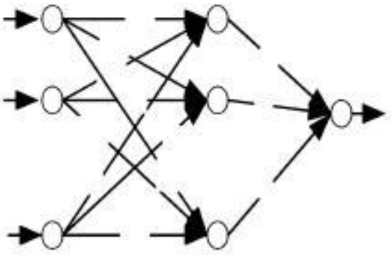


Figure 1:- Three layer Neural Network

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.

3. POS-NEG 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 Pos-Neg sensitivity analysis. The Pos-Neg technique computes the hidden neurons that activate the output neuron. The hidden neurons that have weights greater than an active threshold value are considered. The sensitivity analysis of input neurons on the hidden neurons is done. The input neurons are sorted as per their weights and the input neurons that activate the output neuron are considered for Pos rules. The input neurons that classify the output neuron as non-diabetic are considered for Neg rules.

3.1 Pos Neg Rules

The neural network that is constructed contains connections that are influencing the output neuron to classify the result as diabetes class. These input neurons are considered to be pos neurons. These pos neurons are extracted from neural network as positive sensitivity rules like given in Figure 3.

The input neurons that classify the output as non diabetes class is considered to be neg neurons. These neurons are extracted as negative sensitivity rules like given in figure 4.

4. 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%

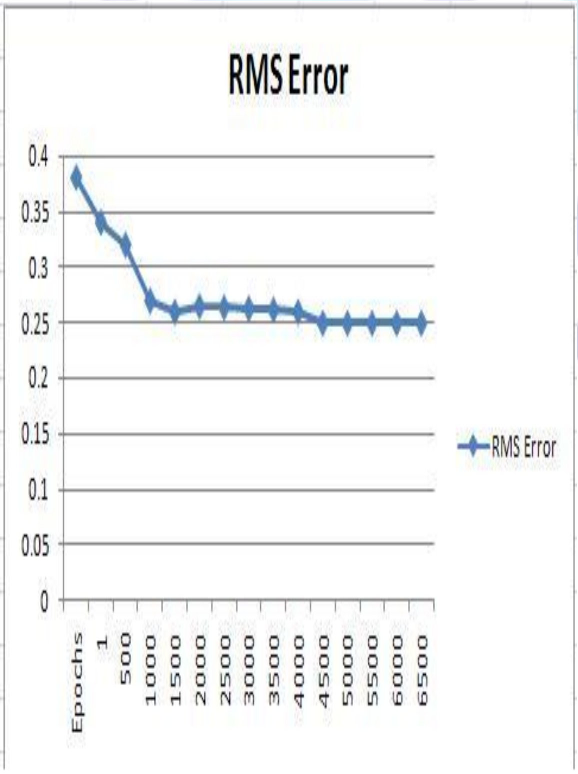
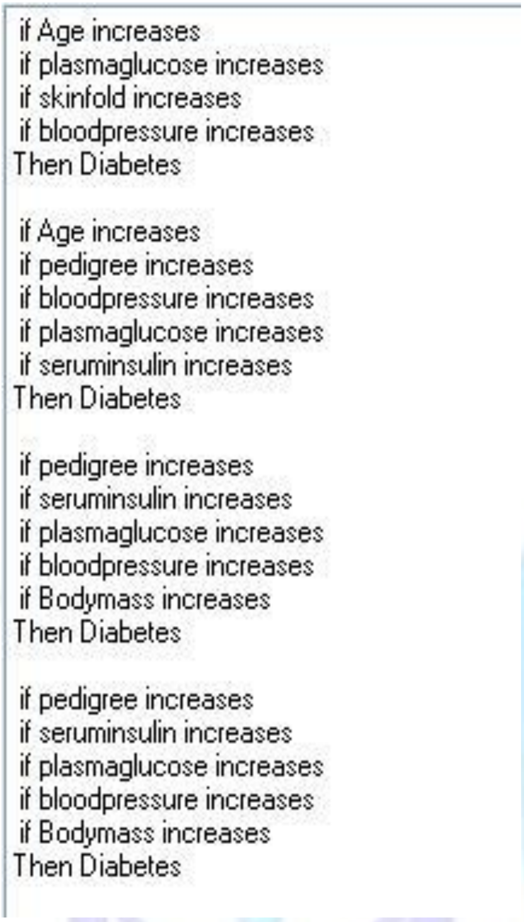


Figure 2 : RMS error



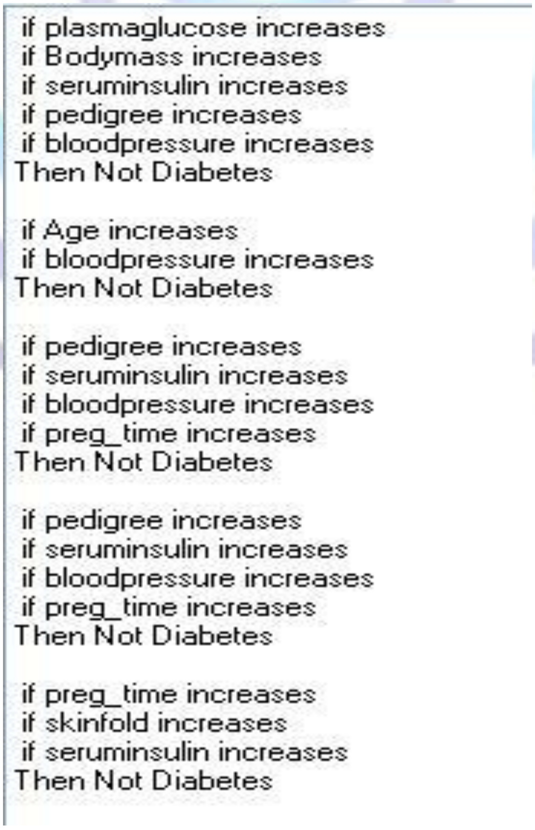
```
 if Age increases
 if plasmaglucose increases
 if skinfold increases
 if bloodpressure increases
Then Diabetes

 if Age increases
 if pedigree increases
 if bloodpressure increases
 if plasmaglucose increases
 if seruminsulin increases
Then Diabetes

 if pedigree increases
 if seruminsulin increases
 if plasmaglucose increases
 if bloodpressure increases
 if Bodymass increases
Then Diabetes

 if pedigree increases
 if seruminsulin increases
 if plasmaglucose increases
 if bloodpressure increases
 if Bodymass increases
Then Diabetes
```

Figure 3: Part of Pos sensitivity Diabetes rules extracted



```
 if plasmaglucose increases
 if Bodymass increases
 if seruminsulin increases
 if pedigree increases
 if bloodpressure increases
Then Not Diabetes

 if Age increases
 if bloodpressure increases
Then Not Diabetes

 if pedigree increases
 if seruminsulin increases
 if bloodpressure increases
 if preg_time increases
Then Not Diabetes

 if pedigree increases
 if seruminsulin increases
 if bloodpressure increases
 if preg_time increases
Then Not Diabetes

 if preg_time increases
 if skinfold increases
 if seruminsulin increases
Then Not Diabetes
```

Figure 3: Part of Neg sensitivity Rules for diabetes

6. 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 sensitivity analysis method. Rule extraction using ANNR 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.

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

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

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

[7] 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

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

[9] 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