IOSR Journal of Engineering (IOSRJEN)
ISSN: 2250-3021 ISBN: 2878-8719 PP 09-12
National Symposium on engineering and Research

Data Mining Using Artificial Neural Network Tree

Pushkar Shinde
MCOERC , Nasik

Abstract: - *Proper diagnosis, classification and prediction of diabetes are essential due to the increasing prevalence of the disease and the increasing cost to control it. Data Mining can help researchers to find knowledge regarding this disease which will in turn help in finding cure for the disease. Appropriate discovery of knowledge from historical data for this disease would be a valuable tool for clinical researchers. The main purpose of data mining is to gain insight of the data, and extract knowledge (inter-relational patterns) from the data. Applying data mining techniques in diabetic data can facilitate systematic analysis. Artificial Neural network (ANN) has already been applied in a variety of domains with remarkable success. However, it has not has been well utilized in data mining because of the "black box" nature. In this paper we present a method of using ANN in data mining and overcoming the "black box" nature using Decision Tree (DT).*

Keywords: - *Diabetes, data mining, rule extraction, neural network, decision tree*

1. INTRODUCION

Diabetes have remained the focus of many clinical studies due to the increasing prevalence of the disease and the increasing cost to control it. Diabetes is a chronic disease that has life threatening complications . The word chronic indicates a long term disease which may be controlled but not cured. Diabetes is a condition due to malfunction in pancreas. That is the pancreas is not producing enough insulin to control the sugar level in the body. This causes the sugar to remain in the blood. The proper diagnosis, classification and prediction of diabetes are therefore essential. Appropriate discovery of knowledge from historical data for this disease would be a valuable tool for clinical researchers.

Data mining is the process of discovering useful knowledge in data and also finding the inter-relation pattern among the data [2]. It is an automated discovery of strategic hidden patterns (useful information) in large amounts of raw data using intelligent data analysis methods [4].

Data mining is targeted for discovery of new knowledge instead of testing the assumptions that are suggested by the user (hence automated). In short the major objective of data mining is to acquire knowledge (meaningful patterns and trends) from large amounts of data that are already in the data but were previously unseen.

The medical industry is among the most information intensive industries. Medical data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year [5]. The ability to use these data to extract useful information for quality healthcare is crucial. 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 diseases [6-7].

ANN has been applied in many applications with remarkable success [8]. For example, ANN have been successfully applied in the area of speech generation and recognition [9], handwritten character recognition [10], vision and robotics [11]. Although most of the results that can achieved through the application of neural networks are remarkable [12] and have frequently found to outperform traditional approaches, there has not been a rush to accept neural networks as a valid problem solving technique in the data mining, despite a great amount of drive to do so. This is because the main purpose of data mining is to gain ~nsight of data and extract knowledge from it, that IS, the knowledge is explicit to human being. But in ANN, this is the main limitation. It is regarded

as a "black box", the knowledge accumulated is not comprehensible. In other words, neural networks will output a result, but do not explain how it has been achieved. Without an explanation of how a result has been reached, it is unlikely that the result will be accepted as valid. A critical factor especially data mining for medical diagnosis is the transparency in the method used to reach a decision for a given medical conditions. In other words it is a necessity to be able to explain how a diagnosis is reached.

Decision Tree (DT) is one of the most popular approaches for machine learning [6]. DT are tree shaped structures that represent sets of decisions. It explains how a decision or diagnosis is reached. The DT approach can generate rules. To have the advantages of ANN and DT methods, it is important to combine them together [2].

In this paper an approach named Artificial Neural Network Tree (ANNT) , i.e ANN training preceded by DT 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. ANNT approach does not try to directly improve the comprehensibility of ANN but it is used as a pre-process for DT rule extraction This method is tested on diabetic data set [5].

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

2. NEURAL NETWORK

A three layer neural network having eight neurons in the input layer, nine neurons in the hidden layer, one neuron in the output layer is considered (shown in Figure 1). The hidden neurons are activated using *tanh* activation function with linear output units.

To determine the optimal weights for a classification and prediction problem is, to minimize the cost function. The most common method to define a cost function is, as the sum of the squared differences between the networks predicted output (Yk) and the expected output (tk). This is commonly known as the sum squared error (SSE) cost function. Here SSE cost function augmented by a weight decay

is used., and it is computed using equation (1)

$$E(w) = \frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{K}(t_{k,p} - y_{k,p})^2 + \frac{1}{2}\alpha w^2 \qquad (1)$$

where *P* is the total number of patterns in the data set,

is the output units, tk"p is the target value at the kth

output neuron for pth sample and is the actual output at the kth output neuron for the pth sample and w is the weight.

The weights' decay *a* is a parameter that controls

the value of the weights and can contribute to hinder

overfit to the noise in the data set [13].

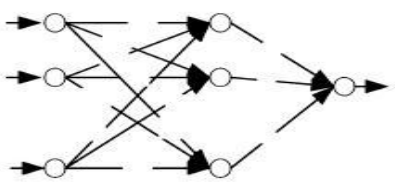

Figure 1:- Three layer Neural Network

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 17 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 0.001 in

this work.

3. KNOWLEDGE REPRESENTATION

Learning by ANN can be thought of as the process of collecting knowledge which has been gathered or accumulated by training process. In general this knowledge is very difficult to interpret because it is distributed into the weight set and is represented in analog form. One method of converting the knowledge into comprehensible form is rule extraction. In this paper data mining is addressed using rule extraction. In other words rule extraction is the process of interpreting the knowledge of trained ANN into comprehensible form to the user and at the same time it is useful for gaining insights (relations and pattern) into the training data set.

3.1. DECISION TREE

DT offers a structured way of decision making [1415]. A DT is characterized by an ordered set of nodes. Each of the internal nodes is associated with a

decision function of one or more features.. DT approach can generate *if -then* rules. Specific DT methods include Classification and Regression Trees (CART), Chi Square Automatic Interaction Detection (CHAID), ID3 and C4.5. In this paper C4.5 which is the extension of ID3 is utilized . C4.5 Decision Tree is based on Information Theory, that is it uses information theory to select features which give the greatest information gain or decrease of entropy [16]. Information gain is the informational value of creating a branch in a

decision tree based on the given attribute using entropy theory.

3.2.ANNT RULE EXTRACTION

To extract the rules using ANNT, first we train the ANN. The trained ANN will result in a set of weights being produced which represents the knowledge. To extract the knowledge we then decompose the ANN using DT method. From the weights and activation pattern obtained from hidden to output layer of ANN, output- hidden tree was built and intermediate rules was generated. Next we map the tree to input layer. Consequently with each link of the tree, we build an input-hidden tree and generate input rules. The final rule is obtained by substituting the input rules in the intermediate rules.

The final rules will describe the relationship between

input pattern and the target (output). See [7] for more detail on this algorithm.

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).[2]

4.2. ANNT RULE EXTRACTION

First the network is trained and rule is extracted from this trained network using DT method. 5 networks were trained and the result are shown in table 1. Using the weight and activation pattern DT was build for output-hidden layer (figure 2) and we translate the DT to intermediate rules *(if and then* rules). The accuracy of the rule extracted from each of this network is shown in table2.

```
col1 <= 0.990771
|  col7 <= -0.927961: 0
|  col7 > -0.927961
|  |  col1 <= 0.957992
|  |  |  col7 <= -0.095586
|  |  |  |  col9 <= 0.999986: 0
|  |  |  |  col9 > 0.999986
|  |  |  |  |  col1 <= 0.179149
|  |  |  |  |  |  col7 <= -0.261264
|  |  |  |  |  |  |  col1 <= -0.372984: 1
|  |  |  |  |  |  |  col1 > -0.372984
|  |  |  |  |  |  |  |  col3 <= 0.995197: 0
|  |  |  |  |  |  |  |  col3 > 0.995197: 1
|  |  |  |  |  |  col7 > -0.261264: 0
|  |  |  |  |  col1 > 0.179149: 0
|  |  |  col7 > -0.095586: 1
|  |  col1 > 0.957992: 0
col1 > 0.990771
|  col9 <= 0.018302: 1
|  col9 > 0.018302
|  |  col7 <= 0.999932: 0
|  |  col7 > 0.999932: 1
```

Figure 2 : example of output-hidden tree of network 1.

Col l represents as number of times pregnant, col2 plasma glucose, col3 is blood pressure followed by triceps skin and insulin in col4 and 5, col 6 is body mass index, col7 diabetes pedigree and col8 is age. From this output-hidden DT, intermediate rules was formed. Example of the intermediate rules is shown in figure 3.

```
IF col1<=0.990771 AND col7>-0.927961 ... then 1
```

Figure 3 : Part of the intermediate rules

For each of hidden-output leaves, a hidden – input tree was built. The part of hidden-input tree (figure 4) is obtain from hidden -output leaves where if col l <=0.99077 [2].

```
col5 <= 148
|  col3 <= 30: 0
|  col3 > 30
|  |  col5 <= 108
|  |  |  col8 <= 58: 1
|  |  |  col8 > 58
|  |  |  |  col3 <= 74: 0
|  |  |  |  col3 > 74: 1
```

Figure 4 : Part of the DT from hidden to input layer.

Input rules was obtain by translating the hiddeninput tree. The input rules was then substitute in the intermediate rules (obtained from output-hidden), to
get the final rules.

A part of the final rules obtained from the above experiments, which have been changed to user friendly rules is shown in figure 5.

IF hours serum insulin <=148 AND diastolic blood pressure <=30 THEN not having diabetes
IF hours serum insulin <=148 AND diastolic blood pressure >30 AND hours serum insulin <=108 AND age<=58 THEN having diabetes.

IF hours serum insulin <=148 AND diastolic blood pressure >30 AND hours serum insulin <=108 AND age>58 AND diastolic blood pressure <=74 AND age<=61 THEN having diabetes.

Figure 5 : Part of the final rule which have changed to user friendly rules.

5. CONCLUSION

In this paper, with the proposed ANNT approach, ANN can be utilized in data mining application. First

ANN was trained, then the rules are extracted from the ANN using DT method. Rule extraction from ANNT method does not depend on the structure or the training methods. It also can work on continuous and discrete data. ANNT 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 ANNT will be on reducing the redundant rules.

REFERENCES

[1] Y. Bengio, 1M. Zurada.. Introduction to the Special Issue on Neural Networks for Data Mining and Knowledge Discovery, *IEEE Tran. on. Neural Networks*, vol.ll, no. 3, pp. 545-549. 2000.

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

[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.