Neural Network in Data Mining

Karan S Nayak

Dept. of Electronics & Telecommunicaion, Dwarkadas J. Sanghvi College of Engg., Mumbai, India

Abstract

Data Mining means mine data from huge amount of data. It is beneficial in every field like business, engineering, web data etc. In data mining classification of data is very difficult task that can be solving by using different algorithms. In this paper different neural networks are explain that help in classification of data and help in made supervised and unsupervised learning. This paper appreciates application of neural network in area of data mining.

Keywords

Data Mining, Single Layer Perceptron, Supervised, Unsupervised, Multilayer Perceptron, Back Propagation, Feed Forward, Recurrent, Data Clustering, Rules Assessment.

I. Introduction

Data mining is the term used to describe the process of extracting value from a database. A data-warehouse is a location where information is stored. The type of data stored depends largely on the type of industry and the company. Many companies store every piece of data they have collected, while others are more ruthless in what they deem to be “important”. Consider the following example of a financial institution failing to utilize their data-warehouse. Income is a very important socio-economic indicator. If a bank knows a person’s income, they can offer a higher credit card limit or determine if they are likely to want information on a home loan or managed investments.

Data mining is also known as Knowledge Discovery Data (KDD) [6]. It analyze large amount of data. It has relationship with other areas like neural network, database and business intelligence. There are different type of learning mechanisms in the data mining supervised and unsupervised learning. Supervised learning means classification of data and unsupervised learning means clustering. Different methods are used to classify the data in data mining like decision trees, nearest neighbour, neural network. Neural network play significant role in data mining. Neural network consist of different node with weighted inputs, it is constructive in classification of complex data. Advantage of data mining is that it can construct and learn boundaries for large number of attributes. In this paper different algorithms are explain with examples that help in construct classification in data mining.

II. Learning System By Neural Networks

In data mining two learning methods are used one is classification and other is clustering. Classification mean Supervised learning in which data can determine by predetermine method. This learning system is predictive. There are different techniques for this naïve Bayes, decision tree and neural networks. For example on the base of old data we can predict eligibility of person for particular job. With neural network we can predict the value and find error between actual and desire outputs then adjust the new weight. In supervised learning different methods of neural network are there they are backpropagation, hopified, and recurrent method. Unsupervised learning is known as clustering, it is not depend on the historic data but this method is depend on the good examples from the similar data. In this grouping of data is done. Different types of clustering methods are there partitional clustering on base of distance by k-mean, by density and hierarchal.

III. Classification by Different Types of Neural Networks

A. Single Layer Neural Network

It is also known as single layer perceptron. It is design by Frank Rosenblat. It is sum of all weighted inputs and doesn’t require any prior knowledge. It is used to classify two classes by linear separable method and on thebase of zero or one. In data mining it is use to classify two kind of data by decision boundary in two classes. Its example is below:

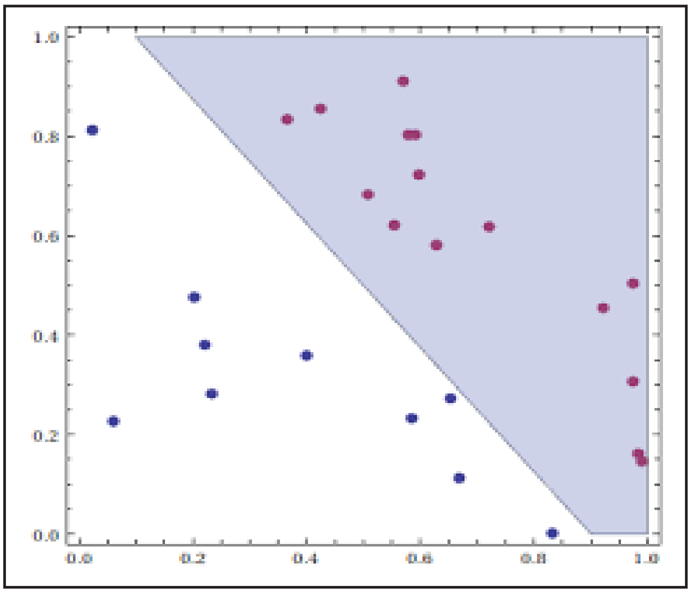

Fig. 1:

Red dots denote one class and blue dots refer to other class. Single perceptron classify two types of patterns. Single layer perceptron in data mining work for OR, AND and COMPLEMENT functions. It represent as follows. Example of AND function:

Table 1:

X	Y	Output
0	0	0
0	1	0
1	0	0
1	1	1

It represent as follows: In fig (a) representation of AND function is there when both inputs are true then output is true otherwise it is false. In fig (b) two classes are classify one with output 1 and other with output zero the class the line that divide two classes is decision boundary. This can be work in data mining for making decision between two classes as example below. Neural network is represent in the below: [2]

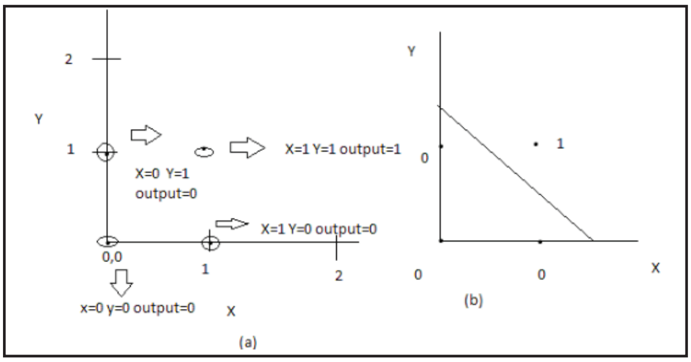

Fig. 2:

Some basic terms used in this are:

1. **Inputs:** In diagram shown below, the inputs are x0,x1.
2. **Bias:** Bias is function that give input as one.
3. **Summation:** It is represent as ∑ symbol i.e summation of weighted inputs.
4. **Activation Function:** This is function that decide whether output will be one or zero.
5. **Decision Boundary:** It will classify two classes, as in below diagram decision boundary will be (x0*w0+x1*w1+w2=0). And two classes are (one is x0*w0+x1*w1+x2>0 and second x0*w0+x1*w1+w2<=0).

Some basic steps are:

1. Value of Input may be multidimension I=(x0,x2……… xn).
2. Input value of weights W=(w0,w1,w2……….wn)
3. Threshold=t .
4. Summation weighted inputs may be taken p=∑ wi*xi.
5. If p>0, Then output =1.
6. Else output=0 .
7. But if classification is incorrect then wi= wi+ class[i]*xi and then calculate again with new wi.

Table 2:

Sr. no.	Name	Age	Salary	Bank Loan
1	Karan	50	10,000	No
2	Jimit	30	30,000	Yes
3	Sumedh	23	15,000	No
4	Dhruv	69	0	No
5	Anish	74	0	No
6	Yash	25	40,000	Yes

In this example two inputs one is age and other is salary are two inputs, here for yes condition is (age<66 AND salary>15000). Classification of these two classes is given below :

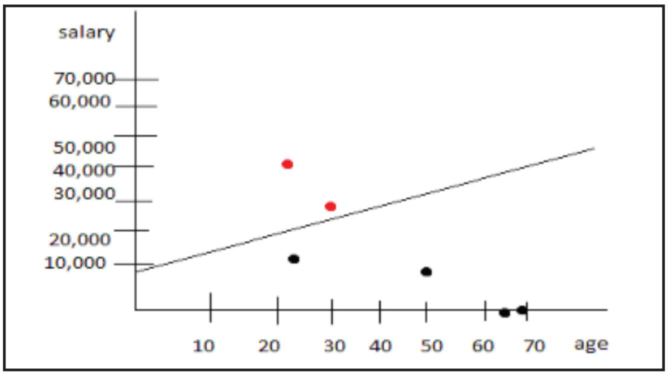

Fig. 3:

Here black dots represent the one class in which like AND function one condition is false and red dots represent other class in which both conditions are true.

B. Multi-Layer Neural Network

It is also known as Multilayer perceptron neural network. It is consist of two layers, three layers. In single layer perceptron we can solve simple problems like linear problem with one decision boundary. But for non-linear problems, we can join two or more single layer perceptrons known as multilayer perceptron. It will classify complex problems, in two layer perceptron neural network classify problem with two decision boundaries and three layer perceptron calculate more complex problems. There are different neural networks. Common are: Feedforward (backpropagation) Associative Recurrent Feedforward-backpropagation method: Unlike single layer neural network it consists of three layer-input layer, intermediate layer or hidden layer and output layer. Intermediate layer process the output of previous layer [4].

IV. Benefits of Hidden Layer

It calculate the complex problems, by choosing several inputs from the previous layer.

1. Different neuron process at same time, thus it is parallel process.

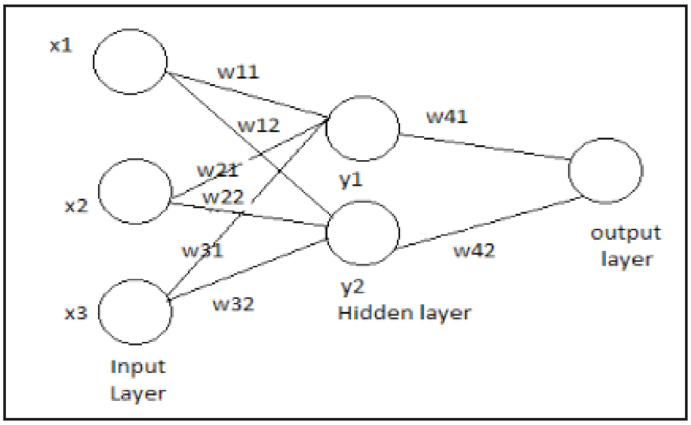

Fig. 4:

Here input layer consist of different nodes, each node represent particular attribute. Like age and qualification of person decide eligibility of candidate for particular job. Output layer give answer in terms of 0 or 1, like eligibility of person. Potency of input is decided on base of weight, it is crucial part in neural output of neural network, summation function and transformation function. Summation function is sum of weighted inputs ∑ xi*wi and transformational function that compute the activation level neurons. There are different transformational functions one is sigmoidal also known as logarithmic activation that is S-shaped produce output as 1 or Sometimes threshold value can be used, let threshold value is θ and if value more than θ then output will be 1 otherwise it will zero. Sigmoidal function can denoted as 1/1+e-y where y is output. It will solve the problem of XOR function that is calculated by Feedforward Neural Network, in this network combination of two neural network done one is AND function Neural Network and other is OR. According to McCulloch pitts method it is represent as below:

In this bias b1=-1.55 and bias b2= -.45 and bias b3=-.45 and activation here means activation function θ (), that is one if output more than zero, and zero otherwise.

Neuron 1 acts as AND gate and neuron 2 act as OR gate here and w11=w12=w21=w22=1, truth table for neuron 1 is given below:

Table 3:

X1	W11	X2	W21	X1*W11+X2*W21+bias	Neuron 1
0	1	0	1	0+0-1.55=-1.55<0	0
0	1	1	1	0+1-1.55=-.55<0	0
1	1	0	1	1+0-1.55=-.55<0	0
1	1	1	1	1+1-1.55=.45>0	1

Neuron 2 acts as OR gate represent as below by the truth table below:

Table 4:

X1	W11	X2	W21	X1*W11+X2*W21+bias	Neuron 2
0	1	0	1	0+0-.45=-.45<0	0
0	1	1	1	0+1-.45=.55>0	1
1	1	0	1	1+0-.45=.55>0	1
1	1	1	1	1+1-.45=1.55>0	1

Combination of these two tables constructs XOR that is represents below:

Table 5:

X1	X2	N1	W31	N2	W32	N1*W31+ N2*W32+Bias	Neuron 3
0	0	0	-3	0	2	0+0-.45=-.45<0	0
0	1	0	-3	1	2	0+2-.45=1.55>0	1
1	0	0	-3	1	2	0+2-.45=1.55>0	1
1	1	1	-3	1	2	-3+2-.45=-1.45<0	0

So in this way multi-layer feedforward method construct outputs with two decision boundaries, it is very helpful in data mining where we have to classify on the base of two decision boundaries. Now graphic representation of this graph will below.

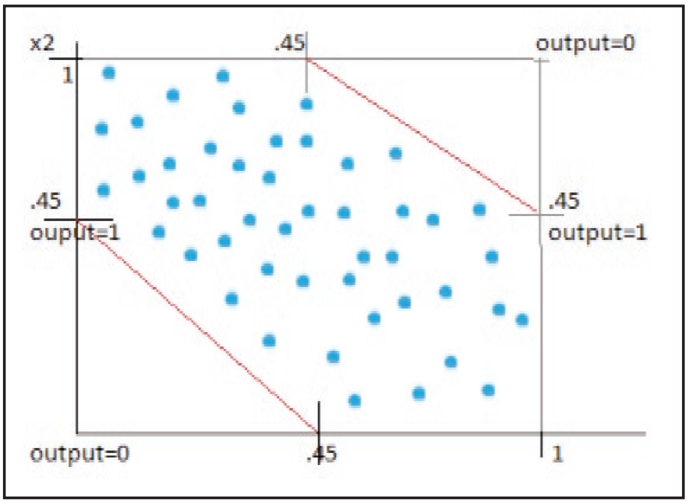

Fig. 5:

Here area with blue dots represent output =1 and area above and below represent output=0

V. Back Propagation Algorithm:

It is also known as back-error propagation use as supervised learning. It consist more than or equal to two hidden layers. It is simple algorithm to find error, basic steps of Backpropagation algorithm with example is given below.

- Let training vector= t(1,1,1), associate target=(.2,.5,.2), output from layer 4 and 5=(.5,.5) and activation function used here is logistic function denoted as f(x)=1/1-exp-x
- First of all calculate the actual output of three nodes Node 1=.4*.5+.5*.5=.2+.25=.45 and f (0.45)=1/1-exp-.45=.305 Node2=.2*.5+.2*.5=.10+.10=.20 and f (.20)=1/1-exp-.2=.284 Node3=.5*.5+.4*.5=.2+.25=.45 and f (.45)= 1/1-exp-.45=.305
- Calculate the error for each output unit Err=target-output, here target = (.2, .5, .2) First output unit= .2-.305= -.105 Second output unit=.5-.284= .216 Third output unit=.2-.305= -.105
- Error can occur in the hidden layer that can be calculated as Err[j]=∑i=1to n (w[i]*Err[i]) Err [4]=-.4*-.105+.2*.216+.5*-.105=-.042+.043-.053=-.052 Err [5] = .5*-.105-.2*.216-.4*-.105= .053-.0413-.042= .052
- Now new connection with new weight from 1 to 4 and consider learning rate α= .30 Calculation of this is done by formula as below w[j][i]= w[j][i]+(α* activation[j]*Err[i]*F(e[i])*(1-F(e[i]) w[4][1]= w[4][1]+(.30*.5*-.105*.305*(1-.305) = .4-.003098=.397

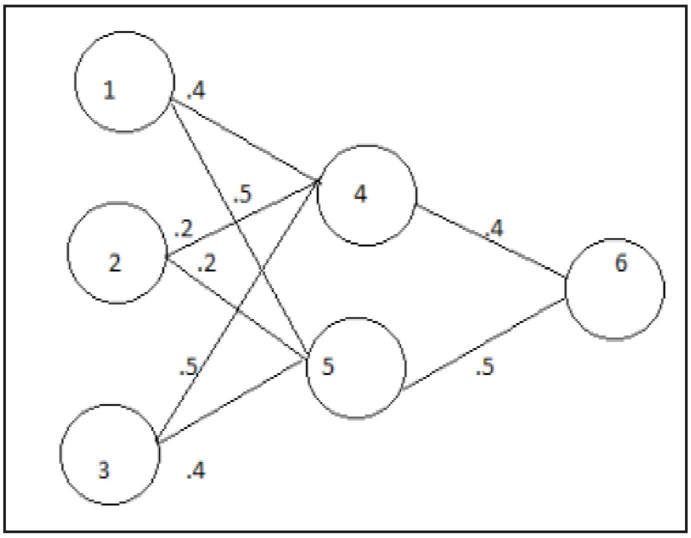

Fig. 6:

Role of This Model in Data Mining

This model help to solve the static problems like classification and generalization. It also solves the problem of decision boundaries.

VI. Recurrent Model

This model as compare to single layer and multi-layer neural network help in solve dynamic problems, these problems are time dependent like to make account of process number, sequence number and forecasting marketing data, in the stock market.In their simplest form this are just neural networks with feedback loop.The previous times hidden layer and final output are fed back as part of the input to the next time steps hidden layer. This is represents as below:

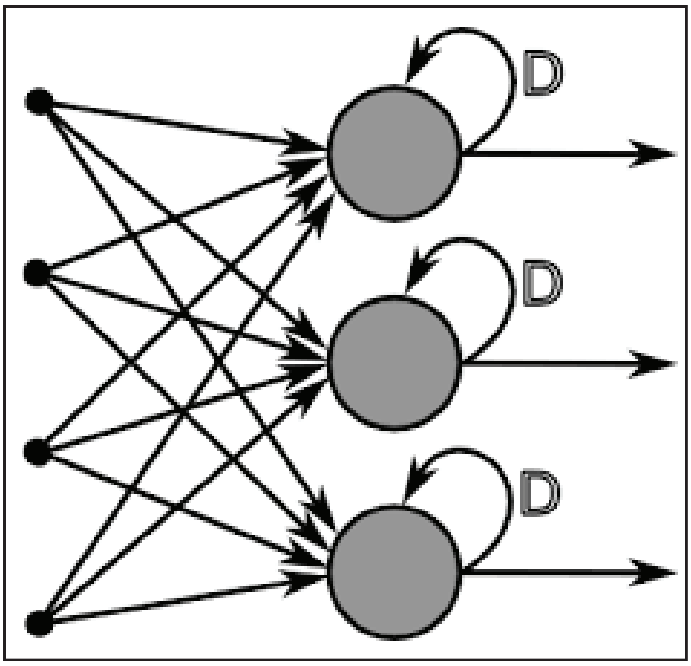

Fig. 7:

VII. Application of Neural Network in Data Mining

Neural network in data mining identify the fraud detection in tax and credit card. It also detects the bankrupt person. Forecasting is another application by which we can predict the data of future on the base of historic data. Forecasting is done in foreign exchange, stock market, loan approval, and change in economics. It can also predict the nature of employees in the firm. Customer behaviour can also analyze that either customer can spend money [6].

VIII. Data Mining Process Based On Neural Networks

Data mining process can be composed by three main phases: data preparation, data mining, expression and interpretation of the results, data mining process is the reiteration of the three phases. The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown below :

A. Data Preparation

Data preparation is to define and process the mining data to make it fit specific data mining method. Data preparation is the first important step in the data mining and plays a decisive role in the entire data mining process. It mainly includes the following four processes:

1. Data Clustering

Data cleansing is to fill the vacancy value of the data, eliminate the noise data and correct the inconsistencies data in the data.

2. Data Option

Data option is to select the data arrange and row used in this mining.

3. Data Pre-processing

Data pre-processing is to enhanced process the clean data which has been selected.

4. Data Expression

Data expression is to transform the data after pre-processing into the form which can be accepted by the data mining algorithm based on neural network. The data mining based on neural network can only handle numerical data, so it is need to transform the sign data into numerical data. The simplest method is to establish a table with one-to-one correspondence between the sign data and the numerical data. The other more complex approach is to adopt appropriate Hash function to generate a unique numerical data according to given string. Although there are many data types in relational database, but they all basically can be simply come down to sign data, discrete numerical data and serial numerical data three logical data types.

B. Data Mining

There are many methods to extract rules, in which the most commonly used methods are LRE method, black-box method, the method of extracting fuzzy rules, the method of extracting rules from recursive network, the algorithm of binary input and output rules extracting (BIO-RE), partial rules extracting algorithm (PartialRE) and full rules extracting algorithm (Full-RE).

C. Rules Assesment

1. Although the objective of rules assessment depends on each specific application, but, in general terms, the rules can be assessed in accordance with the following objectives:
2. Find the optimal sequence of extracting rules, making it obtains the best results in the given data set;
3. Test the accuracy of the rules extracted;
4. Detect how much knowledge in the neural network has not been extracted; Detect the inconsistency between the extracted rules and the trained neural network [5].

IX. Conclusion

In this paper, we present research on data mining based on neural network. At present, data mining is a new and important area of research, and neural network itself is very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The combination of data mining method and neural network model can greatly improve the efficiency of data mining methods, and it has been widely used. It also will receive more and more attention.

References

[1] Sushmita Mitra, Sankar K. Pal, Pabitra Mitra,“Data Mining in a Soft Computing Framework: A Survey”, IEEE Transactions on Neural Networks; (January 2002, Vol. 13, No. 1)

[2] Mark W. Craven, Jude W. Shavlik,"Using Neural Networks for Data Mining".

[3] Prof. S. Sudarshan CSE Dept, IIT Bombay Most slides courtesy: Prof. Sunita Sarawagi, School of IT, IIT Bombay.

[4] Xianjun Ni,"Research of Data Mining Based on Neural Networks.

[5] Uwe Lämmel,"Artificial Neural Networks and Data Wismar Business School Mining.

[6] Feng Jiansheng,"KDD and its applications", BaoGang techniques. 1999(3): pp. 27-31.

[7] Researchers William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus.

[8] David Hand,"Principles of Data Mining [M]", Massachusetts Institute of Technology, 2001.

[9] Arjun K. Pujari,"Data Mining Techniques".

Karan Nayak, student in Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai, India.