Aman Kumar, Tarandeep Singh / International Journal of Engineering Research and Applications (IJERA)
ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.088-093

Data Mining, knowledge discovery With Neural Network Support Architecture: A study

Aman Kumar[*], Tarandeep Singh[**]
[*](M.tech(CSE), Graphic Era University,Dhradun,Uttarakhand,India
[**](Research Scholar ,SU, Rajsthan, India)

ABSTRACT
Organizations gathering massive data warehouses in which data is stored as historical facts. But very few companies have been capable to recognize the concrete value stored in it. The main requirements of these companies are asking is how to extract the meaningful information. The only response in positive direction is Data mining. Several technologies available to data mining practitioners, including Artificial Neural Networks(ANN), Regression Analysis, and Decision Trees. Many practitioners are guarded of Neural Networks due to their black box nature, even though they have proven themselves in many situations. The application of neural networks in the data mining has turn out to be wider. Even though neural networks may have multifaceted structure, long training time, and uneasily reasonable demonstration of results, neural networks have high acceptance ability for noisy data and high accuracy and are preferable in data mining.

In this paper the key process of data mining based on neural networks is studied up to a level, and ways to achieve the data mining based on neural networks are also studied. This paper also shows the major drawback of neural network model and addresses issues the effective ways of using rule extraction with the introduction of DNN (Descriptive Neural Network) as the solution of that major blockage and for further scope to expend the things.

I. INTRODUCTION
Data mining is the expression used to explain the process of extracting the desired and meaningful value information from a database/data-warehouse. A data-warehouse is a location where information is stored as the historical facts. The type of data stored depends mainly on the type of particular organization. Data mining techniques can predict the future trends and actions to support the decision of analysis done formally [1] i.e. by analyzing the whole database system of the organization the data mining techniques can answer the different problems such as "Which customer is most likely to payment to the another insurance broker, why", and other parallel situations.

Some data mining techniques can also determinate some traditional situations which takes much time, this is because that they can quickly look around the whole database and find some meaningful information experts unnoticed. There are some essential things which are required such as high-quality data, the "right" data, an adequate sample size and the right technique. There are several tool available in the hand of data mining researchers like decision tree, regression analysis and the resemblance of neural network i.e. ANN (Artificial Neural Network).[2]

II. DATA MINING AND ARTIFICIAL NEURAL NETWORK (ANN)
An **artificial neural network**, is a mathematical model or computational model based on biological neural networks, a parallel processing network which generated with simulating the image sensitive thinking of human, according to the features of biological neurons and neural network and by simplifying, summarizing and refining. It is an interconnected group of artificial neurons and uses the thought of non-linear mapping, the method of parallel processing and the structure of the neural network itself to express the associated knowledge of input and output. A neural network has to be configured such that the application of a set of inputs produces (either "direct" or via a relaxation process) the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a *priori* knowledge. Another way is to "train" the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.[3] However, due to the complex structure and long training time the application of neural networks for data mining are not positive. But due to high affordability to noise data and low error rate the continuously advancing and optimization of various network training algorithms with small training time and continuously improving various network pruning algorithms and rules extracting algorithms for data mining.

II(a). FORMATION: KDD PROCESS BY ANN

Neural Network method is one of the seven methods of data mining which is mainly used for classification, clustering, feature mining, prediction and pattern recognition. An essentially neural network is a distributed matrix structure and imitates the neurons based on Hebb learning rules. Its used repeated iteration method for calculating the weights of distributed neural networks connected by training data mining.

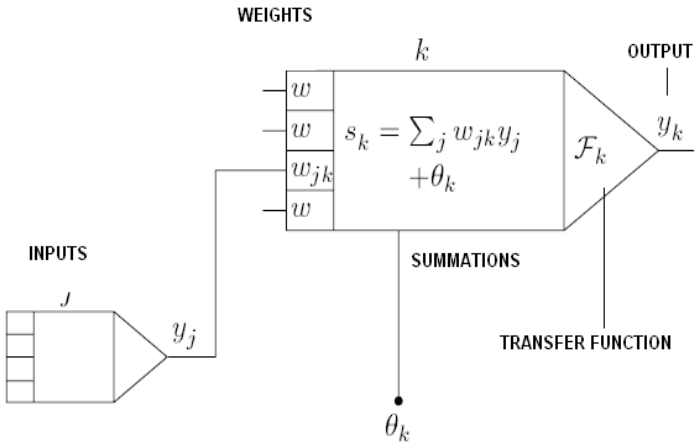


Figure (b) The basic components of an artificial neural network. The propagation rule used here is the "standard" weighted summation for information processing.[3]

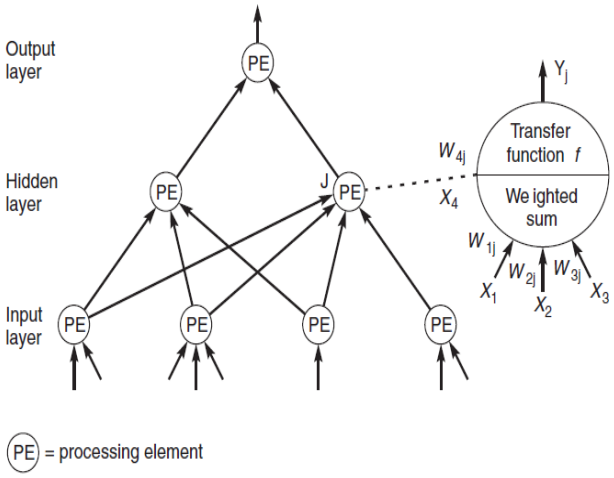
when we connected the units which provide an adaptive contribution. The total input to unit k is simply the weighted sum of the separate outputs from each of the connected units plus a bias or offset term $\boldsymbol{\theta_k}$.

$$s_k(t) = \sum_j w_{jk}(t)\, y_j(t) + \theta_k(t). \qquad (1)$$

The contribution for positive w_{jk} is considered as an excitation and for negative w_{jk} as inhibition.
In some cases more complex rules for combining inputs are used. in which a distinction is made between excitatory and inhibitory inputs .We call units with a propagation rule (1) sigma units.

The Neural Network method used following major structure

Feed-forward neural networks (FF networks) are the most popular and most widely used models in many practical applications. They are known by many different names, such as "multi-layer perceptions."



Fig(a) .A feed forward network with one hidden layer and one output.

Mathematically the functionality of a hidden neuron is described based on rule (1) sigma units.

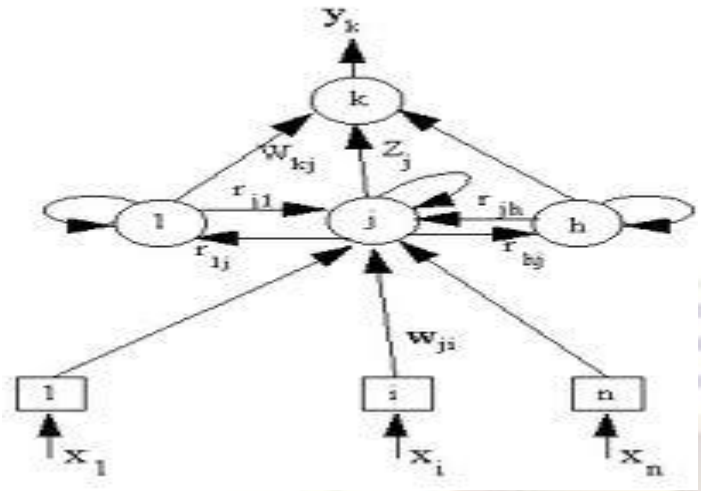
$$\sigma\left(\sum_{j=1}^{n} w_j x_j + b_j\right)$$, $\mathbf{b_j}$ bias or offset term.

where the weights $\{w_j, b_j\}$ are symbolized with the arrows feeding into the neuron.[11]

Recurrent neural networks This is the basic architecture developed in the 1980s: a network of neuron-like units, each with a directed connection to every other unit. Each unit has a time-varying real-valued activation. Each connection has a modifiable real-valued weight. Some of the nodes are called input nodes, some output nodes, the rest hidden nodes. [3]

The learning algorithms includes the following procedure [6]

1. Initialize weights ($w_1, w_2, \ldots w_k$) with random values and set other parameters.
2. Read array (input) and the desired output.
3. Compute the actual output via the calculations, working forward through
the layers.
4. Compute the error.
5. Change the weights by working backward from the output layer through the hidden layers.



Fig(b) .A Recurrent network with one output.

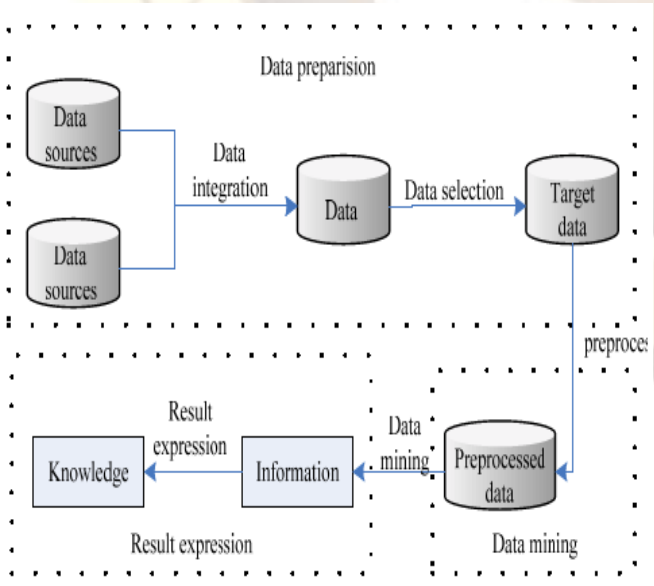
III. NEURAL NETWORK BASED MINING PROCESS

General data mining process contain three main sections

Data Preparation

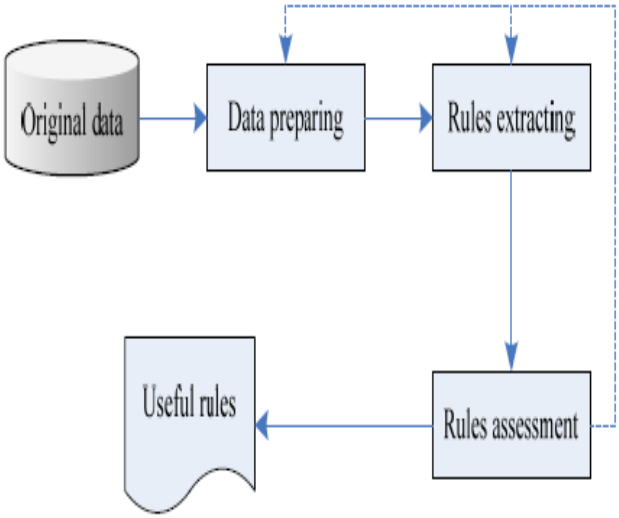
- Mining to particular data
- Expression and Interpretation of results based on particular base pattern

As shown in fig (C)



Fig(c): General Mining Process

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown in Fig. (d).



Fig(d): mining process based on neural network

A. DATA COLLECTION & PREPARATION

Data preparation is the first important step in the data mining where mining data to make it fit specific data mining method and crucial role in the entire data mining process. It mainly includes the following four processes.

1) DATA CLEANSING

Data cleaning, also called *data cleansing* or *scrubbing*, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. [13]

2) DATA OPTION & DATA PREPROCESSING

Selecting the proportion of data range to feeding it and preprocessing is to superior process the clean data which has been selected.

3) DATA EXPRESSION

Transform the data after preprocessing 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. Establishing 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. Fig.(e) gives the conversion of the three data types. The symbol "Apple" in the figure can be transformed into the corresponding discrete numerical data by using symbol table or Hash function. Then, the discrete numerical data can be quantified into continuous numerical data and can also be encoded into coding data.[4]

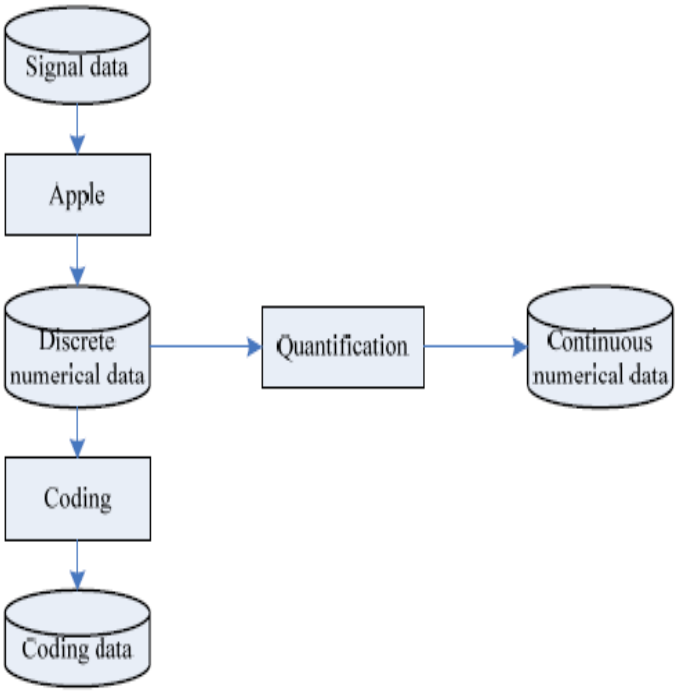


Fig. (e) Data expression and conversion in data mining based on neural network

***B.* RULES EXTRACTING**

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 (Partial-RE) and full rules extracting algorithm (Full-RE), are simple and most used..

***C.* RULES ASSESSMENT**

The rules can be assessed in accordance with the following objectives.

(1) Obtains the best results in the agreed data set, by find the optimal succession of extracting rules,

(2) Extracted rules should tested for accuracy.

(3) The knowledge in the neural network has not been extracted is amounted and measured;

(4) The inconsistency between the extracted rules and the trained neural network should be detected.

IV. IMPLEMENTATION AND ANALYSIS: A COMPARATIVE STUDY WITH LOGISTIC REGRESSION AND ANN TECHNIQUES OF DATA MINING

A typical application of neural networks to predict bankruptcy of companies using the same data and a similar experimental design as used by Wilson and Sharda. For comparative analysis, the performance of neural networks is contrasted with logistic regression. Some financial ratios used here to realizing that more sophisticated inputs to the neural network model should only enhance its performance.

These ratios are as follows:

***X*1**: WC/TA
***X*2**: RE/TA
***X*3**: EBIT/TA
***X*4**: MVE/TB
***X*5**: S/TA

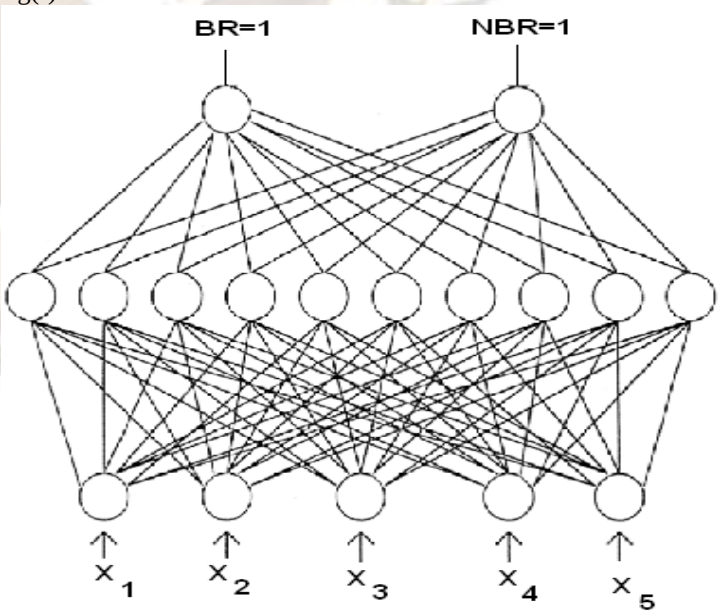
TA-Total Assets, **WC**-Working Capital
S-Sales, **RE**-Retained Earnings, **EBIT**-Earnings Before Interest and Taxes, **MVE**-Market Value Of Equity, **TB**- Total Debt

Step 1 consists of collecting relevant data. Data used for the bankrupt firms are from the last financial statements issued before the firms declared bankruptcy. Thus, the prediction of bankruptcy is to be made about 1 year in advance.

Step 2 requires us to break the data set into a training set and a testing set. .i.e a training set of 20 patterns can be created by randomly setting 20 records from the collected set. A set of 20 other patterns/records can be created as a test set. We do not really know the actual proportion of firms going bankrupt; the 80/20 and 90/10 cases should be close and 50/50 as base rate. Thus, a total of 60 distinct training and testing data set pairs were generated from the original data. To summarize, neural networks and logistic regression models are developed using training sets of equal proportions of firms to determine the classification function but are evaluated with test sets containing 50/50, 80/20, and 90/ 10 base rates.

Steps 3 to getting ready for a neural network experiment.

A neural network software package that implements the aforementioned back propagation training algorithm to construct and test trained neural network models. As show in fig(f)



Fig(f): A Typical Prediction Model of Neural Network for Bankruptcy

Step Next refers to the actual neural network training. In training the networks in this example, a heuristic back propagation algorithm was used to ensure convergence. For correct classifications, a testing threshold of 0.49 used. Hence, the output node with a value over 0.5 used to assess network provided a correct classification. If output neurons provided output levels either less than 0.5 or greater than 0.5 were automatically treated as misclassifications.

To compare the performance of the neural network against using classical statistical techniques, a logistic regression approach was implemented by SYSTAT, a statistical software package. Table (i) represents the average percentage of correct classifications provided by the two different techniques. Neural networks correctly classified 97.5 percent of the holdout cases(the testing sets contained an equal number of the two cases), whereas logistic regression was correct 93.25 percent of the time. Neural networks classified at a 95.6 percent correct rate, whereas logistic regression correctly classified at a 92.2 percent rate(the testing sets contained 20,070 bankrupt firms).

	Test fractions					
	50/50		**80/20**		**90/10**	
Base Criteria	**ANN**	**CLR**	**ANN**	**CLR**	**ANN**	**CLR**
Total % of acceptable classification	97.5a	93.25	95.6a	92.2	95.68b	90.23
success rate of classification(bankrupt)	97.0a	91.90	92.0	92.0	92.5	95.0 (p =.282)
success rate of classification (Nonbankrupt)	98.0a	95.5	96.5a	92.25	96.0b	89.75

Table(i):Performance of ANN over logistic Regression

ANN-Artificial Neural Networks
CLR-Classical logistic Regression

a_p<.01
b_p<.05

V. MAJOR DRAWBACK AND A SOLUTION APPROACH: AS THE INTRODUCTION OF DNN (DESCRIPTIVE NEURAL NETWORK)

The major drawback of ANN it is unable to explain the knowledge implanted in trained neural networks. Extracting rules from trained neural networks is one of the solutions to this problem. The DNN is a neural network embedded with business rules that have been discovered from formerly trained networks. The architecture of DNN is not only decided by training examples but also by hidden rules extracted from trained networks. One of the aim of our descriptive technique is to create and use innovative a second layer of rule extraction techniques to clarify the hidden rules from formerly trained neural networks. This will enable us to explain the mechanism of a neural network forecasting model.DNN that is expected to make more accurate and explainable forecasts. The DNN system to traditional neural network is similar to econometrics to regression analysis. One of advantages of neural networks applied in the forecasting domain is that high forecasting accuracy may be achieved without knowing domain knowledge.[14]

VI. CONCLUSION

Hence, data mining is an innovative and important area of research, and neural network itself is very appropriate for solving the tribulations of data mining because its distinctiveness of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance as this paper analysis shows. The combination can greatly improve the efficiency of data mining methods, and it has been widely used. The thinking can further be expended as this paper show the major blockage of ANN, by making the DNN as the appropriate successor of formerly trained ANN we can make it as the exact solution for KDD.

ACKNOWLEDGEMENTS

This work is the short review or the study of the data mining approach with neural network based on the information over the research papers and books.

REFERENCES

[1] Berson, “Data Warehousing, Data-Mining & OLAP”, TMH .

[2] Bradley, I., Introduction to Neural Networks, Multinet Systems Pty Ltd 1997

[3] Ben Krose ,Patrick van der Smagt" An introduction to neural network" ,Eighth edition 1996

[4] Jiawei Han and Micheline Kamber "Data Mining:Concepts and Techniques" second edition

[5] Gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166.

[6] Bod´en, M. and Wiles, J. (2000). Context-free and context-sensitive dynamics in recurrent neural networks. Connection Science, 12(3).

[7] Berry, J. A., Lindoff, G., Data Mining Techniques, (Wiley Computer Publishing, 1997 ISBN 0-471-17980-9).

[8] Cynthia Krieger: 'Neural network in data mining ', 1996

[9] Bhavani , Thura-is-ingham,"Data-mining Technologies ,Techniques tools & Trends", CRC Press

[10] Fayyad, Usama, Ramakrishna " Evolving Data mining into solutions for Insights", communications of the ACM 45, no. 8

[11] Fausett, Laurene (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, New Jersey, USA.

[12] Haykin, S., Neural Networks, Prentice Hall

[13] Erhard Rahm, Hong Hai Do University of Leipzig, Germany" Data Cleaning: Problems and Current Approaches" http://dbs.uni-leipzig.de

[14] J. T. Yao, Department of Computer Science," Knowledge Based Descriptive Neural Networks" University or Regina Regina, Saskatchewan, CANADA S4S 0A2 Inc., 1999

[15] Xianjun Ni "Research Of data mining based on neural networks" Worlds Academy of Science, Engineering and technology 2008