ISTANBUL UNIVERSITY –
JOURNAL OF ELECTRICAL & ELECTRONICS ENGINEERING

YEAR : 2003 (707-712)
VOLUME : 3
NUMBER : 1

CLASSIFICATION AND PREDICTION IN DATA MINING WITH NEURAL NETWORKS

Serhat ÖZEKES[1] **Onur OSMAN[2]**

[1, 2] Istanbul Commerce University,
Ragıp Gümüşpala Cad. No: 84 Eminönü 34378 ISTANBUL

[1]e-mail: serhat@iticu.edu.tr [2]e-mail: oosman@iticu.edu.tr

ABSTRACT

In this paper Neural Networks (NN) are drawn in data mining for classification and prediction. Back propagation is used as a learning algorithm. Data mining is one of the hottest current technologies of the information age. As computer systems getting cheaper and its power increases, the amount of collected and processed data available increases. Data mining is a process to discover the patterns and trends in large datasets. In our simulation, financial data set is evaluated. The expectations of bank results and our proposed Neural Network results are compared and some differences are obtained.

Keywords: *Data Mining, Neural Network, Classification, Prediction, Back-Propagation*

1. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data. If we see the data that have been stored in our datawarehouses and databases like a mountain, the gems are buried within the mountain. The problem is that there are also lots of non-valuable rocks and rubble in the mountain that need to be mined through and discarded in order to get to that which is valuable. The trick is that both for mountains of rock and mountains of data you need some powerful tools to unearth the value of the data [1]. Data mining helps end users to extract useful business information from large databases. Here the interesting word is large. If the databases were small, we wouldn't need any new technology to discover useful information.

Data mining is a multidisciplinary field bridging many technical areas such as databases technology, statistics, artificial intelligence, machine learning, pattern recognition and data visualization methods.

Data mining is an essential step in the process of Knowledge Discovery in Databases (KDD). Knowledge Discovery consists of an iterative sequence of steps [2]:

1- Data Cleaning : Removing noise and inconsistent data.
2- Data Integration: In some cases multiple data sources may be combined.
3- Data Selection: The data relevant to the analysis task are retrieved from the database
4- Data Transformation: Data is transformed into forms appropriate for mining.

Received Date : 12.08.2001
Accepted Date: 22.12.2002

5- Data Mining: An essential process where intelligent methods are applied in order to extract data patterns.
6- Pattern Evaluation: Identifying the truly interesting patterns representing knowledge based on some interestingness measures.
7- Knowledge Presentation: Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data mining tasks can be classified into two categories: Descriptive and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea. Predictive data mining allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered form the database.
Data mining models can be categorized according to the tasks they perform:

1- Classification and Prediction
2- Clustering
3- Association Rules

Classification and prediction is a predictive model, but clustering and association rules are descriptive models.

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification is the task of examining the features of a newly presented object and assigning it to one of a predefined set of classes. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have [3,4].
The most popular classification and prediction methods are these [1,5]:

1- Neural Networks
2- Decision Trees
3- Genetic Algorithms
4- K-Nearest Neighbor
5- Memory Based Reasoning
6- Naive-Bayes

There are different type of neural network studies [7-16]. In this study we realized the classification and prediction model with back propagation, which is the most popular neural network learning algorithm. The field of neural networks was originally kindled by psychologists and neurobiologists who sought to develop and test computational analogues of neurons. A neural network is a set of connected input/output units where each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class of the input samples. The back propagation algorithm performs learning on a multilayer feed-forward neural network.

A detailed description of the back propagation process can be find in section 2. In section 3, Training, Target and Test sets are explained. In section 4 prediction results of our work are given. At the final section the paper is concluded.

2. THE BACK PROPAGATION ALGORITHM

The error signal at the output of neuron j at iteration n is defined by

$$e_j(n) = d_j(n) - y_j(n) \quad (1)$$

neuron j is an output node [6]. The instantaneous value of the error energy for neuron j can be defined as $\frac{1}{2}e_j^2(n)$. Correspondingly, the instantaneous value $\mathcal{E}(n)$ of the total error energy is obtained by summing $\frac{1}{2}e_j^2(n)$ over all neurons in the output layer; these are the only "visible" neurons for which error signals can be calculated directly. We may thus write

$$\mathcal{E}(n) = \frac{1}{2}\sum_{j \in C} e_j^2(n) \quad (2)$$

where the set C includes all the neurons in the output layer of the network [6]. Let N denote the total number of patterns (examples) contained in the training set. The average squared error energy is obtained by summing $\mathcal{E}(n)$ over all n and then normalizing with respect to set size N, as shown by

$$\mathcal{E}_{av} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}(n) \quad (3)$$

The instantaneous error energy $\mathcal{E}(n)$, and therefore the average error energy $\mathcal{E}_{av}$, is a function of all the free parameters (i.e., synaptic

weights and bias levels) of the network. For a given training set, $\mathcal{E}_{av}$ represents the cost function as a measure of learning performance. The objective of the learning process is to adjust the free parameters of the network to minimize $\mathcal{E}_{av}$. To do this minimization, we use an approximation similar in rationale to that used for the derivation of the LMS algorithm. We consider a simple method of training in which the weights are updated on a pattern-by-pattern basis until one epoch, that is, one complete presentation of the entire training set has been deal with.

$$\Delta w_{ij}(n) = \eta \delta_j(n) y_i(n) \tag{4}$$

in this equation $\delta_j(n)$ is the local gradient [6]. Local gradient points to required changes in synaptic weights.

We get the back-propagation formula for the local gradient $\delta_j(n)$ as

$$\delta_j(n) = \varphi_j'(v_j(n)) \sum_k \delta_k(n) w_{kj}(n) \tag{5}$$

neuron j is hidden. Figure 1 shows the signal-flow graph representation of Eq.(3), assuming that the output layer consists of m_L neurons.

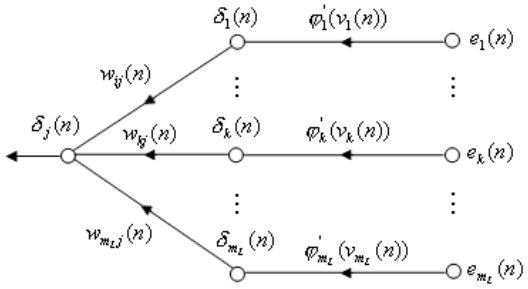

Figure 1 Signal flow graph of a part of the adjoint system pertaining to back-propagation of error signals.

The factor $\varphi_j'(v_j(n))$ involved in the computation of the local gradient $\delta_j(n)$ in Eq.(5) depends solely on the activation function associated with hidden neuron j. The remaining factor involved in this computation, namely the summation over k, depends on two sets of terms. The first set of terms, the $\delta_k(n)$, requires knowledge of the error signals $e_k(n)$, for all neurons that lie in the layer to the immediate right of hidden neuron j, and that are directly connected to neuron j. The second set of terms, the $w_{kj}(n)$, consists of the synaptic weights associated with these connections. We may redefine the local gradient $\delta_j(n)$ for hidden neuron j as

$$\delta_j(n) = -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \tag{6}$$

$$= -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} \varphi_j'(v_j(n)), \tag{7}$$

neuron j is hidden. The induced local field $v_j(n)$ produced at the input of the activation function associated with neuron j is therefore

$$v_j(n) = \sum_{i=0}^{m} w_{ij}(n) y_i(n) \tag{8}$$

where m is the total number of inputs (excluding the bias) applied to neuron j [6]. The synaptic weight w_{j0} (corresponding to the fixed input $y_0 = +1$) equals the bias b_j applied to neuron j. Hence the function signal $y_j(n)$ appearing at the output of neuron j at iteration n is

$$y_j(n) = \varphi_j(v_j(n)) \tag{9}$$

Next differentiating Eq.(9) with respect to $v_j(n)$, we get

$$\frac{\partial y_j(n)}{\partial v_j(n)} = \varphi_j'(v_j(n)) \tag{10}$$

where the use of prime (the right-hand side) signifies differentiation with respect to the argument [6].

3. A FINANCIAL APPLICATION IN DATA MINING

The aim of our study is to make predictions on the loans already granted by a bank. The bank wants to improve their services. For instance the bank managers have only vague idea, who is a good client (whom to offer some additional services) and who is a bad client (whom to watch

carefully to minimize the bank loses). Fortunately, the bank stores data about their clients, the accounts (transactions within several months), the loans already granted, the credit cards issued.

This is a real world data set which contains data about accounts whose contracts are finished and whose contracts are still running. This paper describes designing a neural network using the data of the accounts whose contracts are finished. After the learning phase, the network is used to make predictions on the accounts whose contracts are still running. Our aim is to predict whether the clients will pay the loan back or not. The data about the clients and their accounts consist of following tables in the data set (see figure 2):

- **account** table (4500 objects) - each record describes static characteristics of an account,
- **client** table (5369 objects) - each record describes characteristics of a client,
- **disposition** table (5369 objects) - each record relates together a client with an account i.e. this relation describes the rights of clients to operate accounts,
- **permanent order** table (6471 objects) - each record describes characteristics of a payment order,
- **transaction** table (1056320 objects) - each record describes one transaction on an account,
- **loan** table (682 objects) - each record describes a loan granted for a given account,
- **credit card** table (892 objects) - each record describes a credit card issued to an account,
- **demographic data** table (77 objects) - each record describes demographic characteristics of a district.

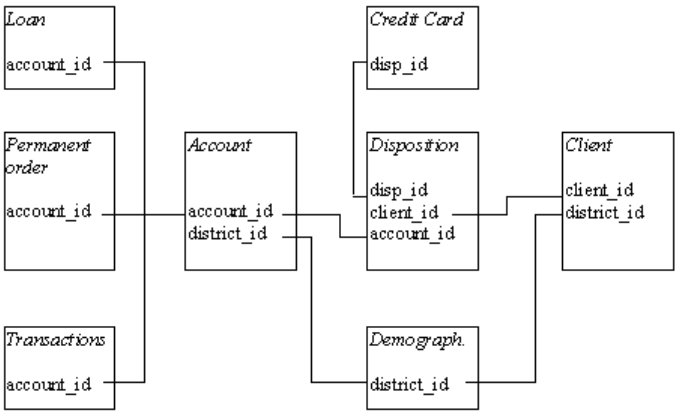

Figure 2 Relations of the tables in our data set.

As a training sub-task we look for the descriptions of the loans in the Loan table. There are 4 types of loan status: A if contract has been finished without problems, B if contract finished but loan not paid, C for running contract, OK so far, and D for running contract, where client is in debt. The data analysis is performed using account parameters evaluated 1 year back from the back payment if the loan is successful (loan status A and C). If the loan is unsuccessful (loan status B and D) the data analysis is carried out using account parameters evaluated 1 year back from the month where loan has been created. We use A and B states (234 objects) as the training set. After the network's learning phase has completed, we use network to make predictions on C and D states (448 objects), whether they will be A or B as soon as the contracts are finished.

In the input layer of the network we have eleven inputs (see figure 3). The inputs are :

1- Minimum Amount: is the minimum amount paid/received.
2- Maximum Amount: is the maximum amount paid/received.

3- Average Amount: is the average of the amounts paid/received.
4- Minimum Balance: is the minimum balance in the account..
5- Maximum Balance: is the maximum balance in the account.
6- Average Balance: is the average of balances.
7- Loan Amount: is the amount of money granted by the bank.
8- Loan Duration: is the duration of the loan.
9- Card Type: is the type of the card (junior, classic or gold)
10- Sex: is the sex of the client
11- Age: is the age of the client.

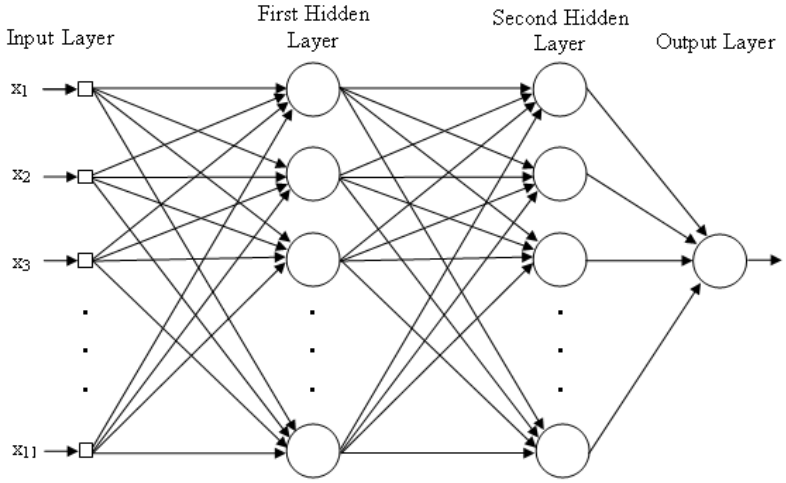

Figure 3 Architectural graph of the network.

Selection, calculation and transformation of these inputs are the preprocessing steps of our work.

4. NEURAL NETWORK PREDICTION RESULTS

In the result of our predictions 365 accounts whose loan status is C seems to be A at the end of the contract. This means 365 clients who are not in trouble now, seems to pay their loans without problem. This is an expected result. Similarly, 41 clients whose loan status is D seems to be B at the end of the contract. This means 41 clients who are in trouble now, seems to not pay their loans back. This is also an expected result too.

After applying our proper NN, in the test data 38 accounts whose loan status is C seems to be B at the end of the contract. This means 38 clients who are not in debt while the contract is running, seems to not repay their loan when the contract is finished. And similarly 4 accounts whose loan status is D seems to be A at the end of the contract. This means 4 clients who are in debt while the contract is running, seems to repay their loan when the contract is finished.

5. CONCLUSION

This paper concludes that the financial data mining application is constructed with Neural Network. The NN is supervised by using the back propagation algorithm. The constructed network includes; 11 entries in input layer, 11 neurons in the first hidden layer, 11 neurons in the second hidden layer and 1 neuron in the output layer. The inputs consist of data about characteristics, accounts and transactions of the clients. Using these inputs and targets, the training set is produced for 236 clients. With the training set, network performs its learning phase. This NN is applied to new 448 clients whose contracts are still running. Approximately, 10% of financial states of clients are predicted differently when they compared to their current states.

REFERENCES

[1] Fayyad, U., "Mining Databases: Towards Algorithms for Knowledge Discovery", *IEEE Bulletin of the Technical Committee on Data Engineering*, Vol.21 No1, pp.41-48,1998.

[2] Chaudhuri, S., "Data Mining and Database Systems : Where is the Intersection?", *IEEE Bulletin of the Technical Committee on Data Engineering*, Vol.21 No.1, pp. 4-8, March 1998.
[3] Agrawal, R., Imielinski, T., Swami, A., "Database Mining:A Performance Perspective", *IEEE Transactions on Knowledge and Data Engineering*, pp. 914-925, December 1993.
[4] Chen, M., Han, J., Yu P.S., "Data Mining: An Overview from Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8 No.6, December 1996.
[5] Barr, D.S., Mani, G., "Using Neural Nets to Manage Investments", *AI Expert*, February, pp. 16-21, 1994.
[6] Haykin, S., *Neural Networks*, Prentice Hall International Inc., 1999.
[7] Osman N. Ucan, S. Seker and S. Paker, "Jitter Performance of Neural Network Equivalent MPSK Schemes Over Microwave Channels", *International Journal of Communication Systems*, Vol 11, pp. 169-178 May-June 1998.
[8] A.Muhittin Albora, Osman N. Ucan, Atilla Ozmen, Tulay Ozkan, "Separation of Bouguer anomaly map using cellular neural network", *Journal of Applied Geophysics*, 46, pp.129-142, 2001.
[9] Osman N. Ucan, A. Özmen," Performance of Gray Scaled Images Using Segmented Cellular Neural Network Combined Trellis coded Quantization/Modulation (SCNN-CNN CTCQ/TCM) Approach over Rician Fading Channel", *Dogus University Journal*, pp.217- 224, January 2000.
[10] Osman N. Ucan, M. Uysal and A. Ozmen, "Combined TCQ/TCM and Neural Network Modelling (In Turkish)" *Electrical, Electronics and Computer Technologies Conf.* Adana, pp.35, 1998.
[11] A. Karahoca, O. N. Ucan, E. Danacı, "Random Neural Network Approach in Distributed Database Management System", IU-*JEEE*, Vol.1, No. 1, pp,84-110, 2001.
[12] Osman N. Ucan, A. Özmen," Performance of Gray Scaled Images using Quantized Cellular Neural Network Combined Trellis coded Quantization/Modulation (QCNN-CNN CTCQ/TCM) Approach over Rician Fading Channel", *International Conference on Telecommunications*, June 15-18 Kore, 1999
[13] M. Uğur, O. N. Ucan, A. Kuntman, A. Özmen, A. Merev, " Analysing the 2-D surface tracking patterns by using cellular neural networks", *IEEE Int. Power Conf.*, USA. 1999
[14] Osman N. Ucan, A. Özmen" Performance of Cellular Neural Network modeled trellis coded quantization/modulation signals at Rician channels," *IEEE Signal processing and Applications* , Turkey, 1999.
[15] Baran Tander, Osman N. Ucan," 3x3 stable cellular neural network model using PSPICE programming", *Electrical Engineering Chamber* , Gaziantep, 1999.
[16] A.Muhittin Albora, Atilla Ozmen, Osman N. Ucan,"Residual Separation of magnetic fields using a Cellular Neural Network Approach", *Pure Applied Geophysics*, 158, pp. 1797-1818, Sept., 2001.

Serhat ÖZEKES, he was born in Washington D.C., USA on 22 July 1978. He received B.Sc and M.Sc. degree from Marmara University, Technical Education Faculty, Electronics and Computer Education department in 2002. And he is pursuing Ph.D. in the same department. He is working in Istanbul Commerce University Department of Computer Technologies and Programming as a lecturer.

Onur Osman was born in Istanbul, Turkey on 25 September 1973. He graduated from Electrical Engineering Department of Istanbul Technical University in 1994. Then he received M.Sc. honors degree in Electrical Engineering in 1998. He is pursuing Ph.D. at Electrical & Electronics Engineering Department of Istanbul University on Turbo-Trellis Coded Systems. He is in the Editorial Board of the Istanbul University Journal of Electrical and Electronics Engineering. He is working in Istanbul Commerce University as a lecturer and he is married.