An Economic Forecasting Based on Association Rules and Neural Network

Sarjon Defit, Mohd Noor Md Sap
Faculty of Computer Science and Information System
University Technology of Malaysia , KB. 791
80990 Johor Bahru, Malaysia
Telp: (607)-5576160, Fax: (607) 5566155
{sarjon@fsksm.utm.my},{mohdnoor@fsksm.utm.my}

Abstract
The world economic environment is currently undergoing drastic changes, such as economic forecasting. It is a complex and challenging task because of the following reasons: i) there is no economic model which carries conviction, ii) economic time series are intrinsically very unreliable and generally have poor signal to accurate results, iii) non stationary and iv) non linearity. In this paper we propose an economic forecasting based on association rules and neural network. Our study concludes that our model can improve the prediction ability by using frequent itemsets and predicted attribute as input fields. In this paper, we explain the important of data mining and prediction, association rule mining, the proposed economic forecasting model, testing and experimental results, model performance and conclusion.

Keywords: Economic Forecasting, Neural Networks, Data Mining, Association Rules

1. Introduction

With the wide applications of computers, database technologies and automated data collection techniques, large amount of data have been continuously collected into databases. It creates great demands for analyzing such data and turning them into useful knowledge. Obviously, such large amount of data is far beyond human capabilities to analyze using traditional manual methods of data analysis. This gives rise to the significant need for new techniques and tools to assist human intelligently and automatically in analyzing such gigabytes or even terabytes of data to get useful information. One of the available techniques is data mining.

Data mining is defined as the automatic extraction of patterns, rules, previously unknown and potentially useful from large amount of data in databases and using it to make crucial business decision (Sarjon, D., Mohd, N., 2001). The result of the extracted knowledge can be applied to information management, query processing, decision process, process control and many other application (Yongjian, F., 1996; Sarjon, D., Mohd, N., 2001).

The world economic environment is currently undergoing drastic changes, such as economic forecasting. It is a complex and challenging task due to the following reasons (ChengYi, S., et.al., 1996):

i) There is no economic model which carries conviction
ii) Economic time series are intrinsically very unreliable and generally have poor signal to accurate results
iii) Non stationary, and
iv) Non linearity

In the past research, a multitude of promising forecasting methods for predicting stock price from numeric data have been developed. These methods include statistics, ARIMA (Auto Regression Integrated Moving Average), Box-Jenkins, and neural networks (Ikuo, M., 1991; Clarence N.W, T., 1993; Kazuhiro, K., 1995; Wuthrich, B., et.al., 1998; Alexandra, I.C., and Toshio, O., 1998; Sheng-Chai, C., et.al., 1999), stochastic models (ChengYi, S., et.al., 1996).

Sheng-Chai, C., et.al (Sheng-Chai, C., et.al., 1999) developed a forecasting approach for stock index future using Grey theory and neural networks. This approach applied Grey forecast model to predict the next day's stock index future and grey relationship analysis to filter the most important quantitative technical indices. Finally, a recurrent neural network (the modified from back propagation neural network) is developed to train and predict the prices trend of stock index future. In this model, the most important quantitative technical indices is generated by using statistical techniques. It is time consuming and much harder to know which of quantitative technical indices really interest the user.

Duffie (ChengYi, S., et.al., 1996) built stochastic models of stock market based on stochastic differential equation, but they have some disadvantages as follows:

i) There is a function used in the models which represent the influence on stock prices of various factors, including corporation factors, macro economy factors, political factors and psychological factors of investors. The function is very difficult to be decided or cannot be decided at all.
ii) The models cannot be used for prediction

Based on the above mentioned problems (ChengYi, S., et.al., 1996; Sheng-Chai, C., et.al., 1999), at the present stage we investigate ways to make use of the association rules and neural networks in economic forecasting. In this research, association rule technique is used to generate the frequent itemsets, and finally a neural networks is employed to forecast the next day's prices.

The rest of the paper is organized as follows. The proposed economic forecasting is given in section 2 and testing and experimental results insection 3. The model performance and conclusion are given in section 4 and 5 respectively.

2. The Economic Forecasting Model

At the present stage, we propose an economic forecasting based on frequent itemsets and neural networks (EFFIN) model. The general architecture of EFFIN model is shown in figure 2.1. This model consists of two sub modules; association rule mining and prediction. In the first module, we generate frequent itemsets from large databases. And, the following module is made to construct the prediction. The briefly description of each module is given in the following sections.

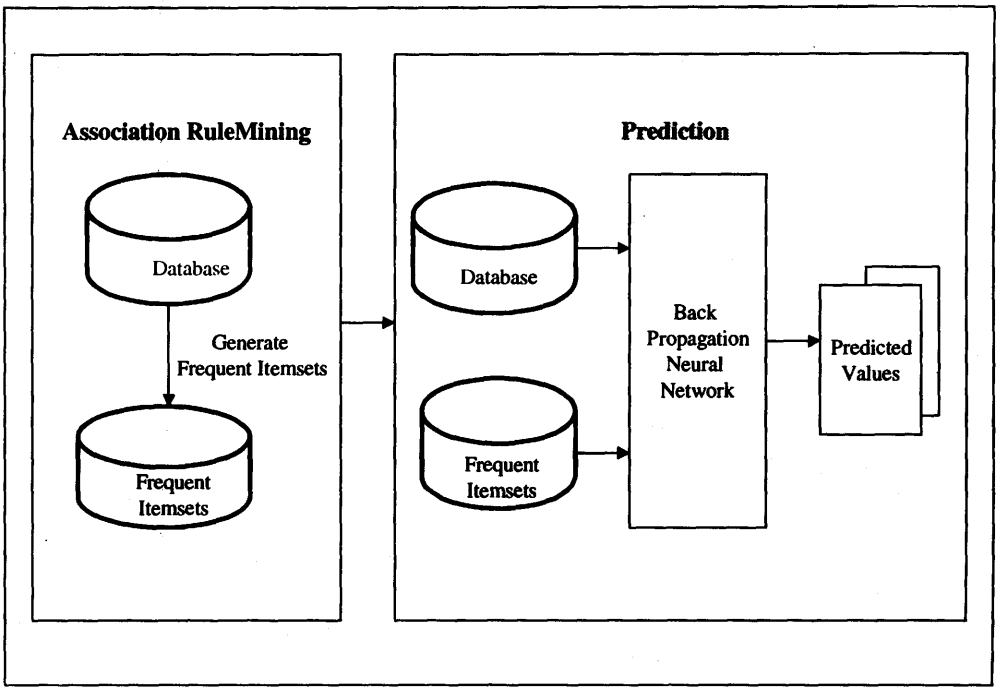


Figure 2.1: The General Architecture of Economic Forecasting

2.1 Assocation Rule Mining

Association rule mining has attracted great attention in database research communities in recent years. It is a form of data mining to discover interesting relationships among attributes. The discovered rules may help marketing, decision making, and business management to make business decisions (David, W.C., Vincent, T.Ng., et.al (1996); Hua, Z., (1998); Juchen, H., Ulrich, G. et.al (2000); Ke, W., Yu, H., et.al (2000); Bing, L., Wynne, H., et.al (1999). A formal definition is given below.

Definiton 2.1: An association rule is a rule in the form of

$$A_1, A_2,\ldots,A_m \rightarrow B_1, B_2,\ldots,B_n \qquad (1)$$

where A_i and B_j are predicates or items.

The basic model of association rules is as follows. First of all, a set of items is referred as an itemset. The itemset that contains k items is a k-itemset. The left hand side of the rule (the part of $A_1, A_2,\ldots,A_m$) is known as the body of the rule, and the righ hand side is the head of the rule. The frequency f of an itemset is the total number of transactions that contain the itemset and $|T|$ is the number of transactions. Support σ is a major measure of rule interestingness. Given a rule

$A \rightarrow B$, the support σ is the probability for a transaction to contain $A \cup B$. By the concept of itemset frequency, the computation of support can be defined by the following equation:

$$\sigma(A \rightarrow B) = \frac{f(A \cup B)}{|T|} \tag{2}$$

An itemset is frequent if the support is no less than a minimum support threshold.

2.1.1. Generating Frequent Itemsets

Frequent itemsets mining plays an essential role in many important data mining tasks. An interesting and influential algorithm for generating frequent itemsets is Apriori algorithm which is proposed by Agrawal and Srikant. (Xinfeng, Y., John, A.K. (1998); Christian, H. (1998); Eui, H., Sam, H., et.al (1999). The detailed Apriori algorithm is given in the algorithm 2.1.

Algorithm 2.1: (Apriori) Generating frequent itemsets using an iterative approach

Input : Transaction database D and minimum support threshold Min_supp
Output : The set of frequent itemsets L in D
Method

1. *$k = 1$; $L = 0$;*
2. *Candidate 1-itemsets C_1 = {all the distinct values in item attribute}*
3. *Compute frequent 1-itemsets L_1. L_1 = gen_frequent ($1, C_1$);*
4. *Repeat*
   $K = k + 1$;
   Generate Candidate k-itemsets C_k = gen_candidate (k, L_{k-1});
   Compute frequent k-itemsets L_k = gen_frequent (k, C_k);
   $L = L \cup L_k$;
   Until $L_k = 0$;

Function gen_frequent (k, C_k)
For each itemset $c \in C_k$ do
$c.frequency = 0$;
For each transaction $t \in D$ { % scan D for counts
C_t = subset (C_k, t); % get the subsets of $t \in C_k$
For each candidate $c \in C_t$
$c.frequency$ ++;
}
return L_k = { $c \in C_k$ | $c.support \geq Min\_supp$}

```
Function gen_candidate (k, L_{k-1})
        C_k = 0;
        For each itemset l_1 ∈ L_{k-1}
                For each itemset l_2 ∈ L_{k-1}
                If (first k-2 items in l_1 and l_2 are same
                        but the last item are different) then {
                        if ∃(k-1)-subset s of c, s ∈ L_{k-1} then
                                delete c
                        else add c to C_k;
        }
return C_k;
```

Figure 2.2 : The Apriori Algorithm

Step 2 and 3 of apriori find the frequent 1-itemsets, L_1. Then step 4, divided into two sub-steps, generates frequent k-itemset. First, candidate k-itemsets C_k are generated by the gen_candidate procedure, which joins the frequent (k-1)-itemset L_k and eliminates those having a (k-1)-subset that is not frequent. Second, frequent k-itemsets L_k are generated from C_k in procedure gen_frequent a described below: once all the candidates have been generated, the database is scanned; for each transaction, a subset function is used to find all subsets of the transaction that are candidate, and the count for each of these candidates is accumulated; finally, all those candidates satisfying minimum support comprise L_k.

2.2. Prediction

Figure 2.3 shows the neural network prediction model. In this model, we employed Back Propagation neural Network. The inputs are predicted attributes, i.e., NYSE_Comp and frequent itemsets that achieved from the first step.

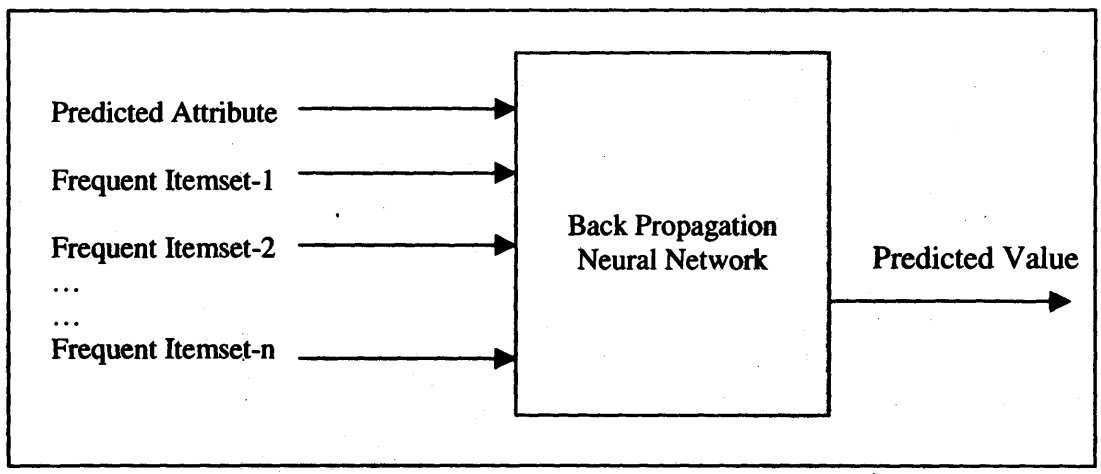


Figure 2.3 : The Neural Network Prediction Model

3. Testing The EFFIN MODEL

We have studied and tested our model using financial market data, as is shown in table 3.1, which contains 250 records. This data contains the weekly close price data from August 6, 1993 to May 15, 1998. The attributes of Financial Market data are as follows:

Financial_Market_Data (*Friday_Date,Dow_Industrial,Dow_Industrial_Volume, NYSE_Comp, NYSE_Comp_Volume, S&P 100, S&P 100_Volume, S&P 500, S&P 500 Volume, TSE, TSE_Volume, P_Income, Unemployment, Energy, Savings, Loans, M1, M2, M3, Exchange, 30Day, 90Day, 180Day, 1Year, 5Year, 10Year*)

These attributes are categorized into three classifications;

i) The Friday Date which indicates the Friday date August 6, 1993 to May 15, 1998

ii) The weekly close price and volumes of the six major stock indices. It consists of Dow_Industrial, Dow_Industrial_Volume, Dow_Transport, Dow_Transport_Volume, NYSE_Comp, NYSE_Comp_Volume, S&P 100, S&P 100_Volume, S&P 500, S&P 500_Volume, TSE and TSE_Volume. These attributes are used as predicted attribute.

iii) The indicators that affect the movement of the next weekly close price. It consists of Unemployment, Energy, Savings, Loans, M1, M2, M3, Exchange, 30Day, 90Day, 180Day, 1Year, 5Year and 10Year. These attributes are used as predictors.

Table 3.1: The Financial Market Samples Data

NYSE_COMP	Pincome	Unemployment	Energy	Saving	.....	10Year
249.7	5556.9	6.8	103	1206.6		5.85
253.25	5556.9	6.8	103	1208.1		5.78
255.4	5556.9	6.8	103	1209.8		5.66
256.22	5556.9	6.8	103	1212.7		5.51
255.93	5556.9	6.8	103	1207.3		5.41
254.66	5558.1	6.7	102.6	1209		5.28
254.18	5558.1	6.7	102.6	1210		5.35
256.29	5558.1	6.7	102.6	1218.2		5.44
255.81	5558.1	6.7	102.6	1210.8		5.33
260.48	5580.3	6.8	105	1209.3		5.33
257.06	5580.3	6.8	105	1211.1		5.24
259.38	5580.3	6.8	105	1214.6		5.31
254.2	5580.3	6.8	105	1218.7		5.44
257.57	5603.7	6.6	104	1214.5		5.66
255.53	5603.7	6.6	104	1214.2		5.68
255.61	5603.7	6.6	104	1217.1		5.71
256.74	5603.7	6.6	104	1219.5		5.83
256.93	5603.7	6.6	104	1215.7		5.8
257.79	5793.4	6.5	103.3	1218.6		5.71
258.49	5793.4	6.5	103.3	1218.7		5.82
259.08	5793.4	6.5	103.3	1225.7		5.79
260.34	5793.4	6.5	103.3	1216.8		5.77

Table 3.1: The Financial Market Samples Data (Cont)

262.9	5562	6.6	102.5	1216.3		5.85
263.12	5562	6.6	102.5	1224.9		5.69
265.42	5562	6.6	102.5	1228.4		5.74
261.21	5562	6.6	102.5	1231.6		5.74
261.31	5562	6.6	102.5	1226		5.8
...	...	...	...	...		...
574.78	7157.5	4.7	103	1460.5		5.61

The economic forecasting is done as follows:

a. Generate frequent itemsets of predictors from table 3.1. Figure 3.1 shows frequent itemsets with minimum support threshold = 9%.

C1

Itemset	Support
Pincome	4%
Unemployment	2%
Saving	9%
Energy	3%
Loans	9%
M1	8%
M2	10%
M3	10%
Exchange	7%
30Day	7%
90Day	6%
180Day	6%
1Year	4%
5Year	8%
10Year	8%

Compare Candidate Support With Minimum Support

L1

Itemset	Support
Saving	9%
Loans	9%
M2	10%
M3	10%

Generate C2 From L1

C2

Itemset	Support
{Saving, Loans}	9%
{Saving, M2}	8%
{Saving, M3}	9%
{Loans, M2}	8%
{Loans, M3}	9%
{M2, M3}	9%

Compare Candidate Support With Minimum Support

L2

Itemset	Support
{Saving, Loans}	9%
{Saving, M3}	9%
{Loans, M3}	9%
{M2, M3}	9%

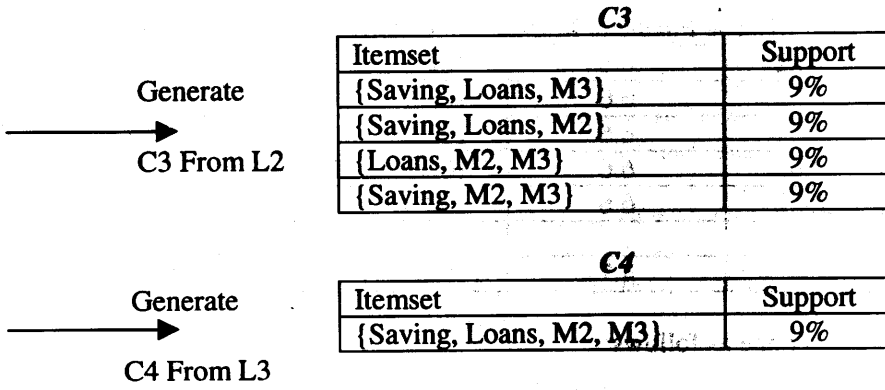


Figure 3.1 : Frequent Itemset With Minimum Support = 9%

b. Construct the prediction. The EFFIN model use a Back Propagation Neural Network. The raw data consists of the weekly close price of the stock indices, i.e., NYSE_Comp. The input are frequent itemsets and predicted item, i.e., NYSE_Comp. The output of the network is the closing stock price of the following week (next week's close). We have tested data samples with 250 train data, 250 test data , momentum equivalent to 50, learn rate equivalent to 50 and verify rate equivalent to 5. The results of prediction is given in table 3.2.

Table 3.2: The Differences Between Actual and Predicted Values of NYSE_Comp

Date	Actual	Predicted	Difference
08/06/93	249.7	254	4
08/13/93	253.25	254	1
08/20/93	255.4	254	1
08/27/93	256.22	254	2
09/03/93	255.93	254	2
09/10/93	254.66	254	1
09/17/93	254.18	254	0
09/24/93	256.29	254	2
10/01/93	255.81	254	2
10/08/93	260.48	254	6
10/15/93	257.06	254	3
10/22/93	259.38	254	5
10/29/93	254.2	254	0
11/05/93	257.57	254	3
11/12/93	255.53	254	1
11/19/93	255.61	254	2
11/26/93	256.74	254	3
12/03/93	256.93	254	3
12/10/93	257.79	254	4
12/17/93	258.49	254	4
12/24/93	259.08	254	5
12/31/93	260.34	254	6
01/07/94	262.9	254	9

Table 3.2: The Difference Between Actual and Predicted Values (Cont)

01/14/94	263.12	254	9
01/21/94	265.42	254	11
01/28/94	261.21	254	7
02/04/94	261.31	254	7
...	...	...	...
05/15/98	574.78	582	7

4. Model Performance

The performance ability of our model is measured based on the following criteria (Sarjon, D., Mohd, N., (2001):

a. Normalized RMS Error. Normalized RMS Error indicates the Root Mean Square (RMS) Error for the entire testing data. This error is also called Standard Error of Estimate which is defined as:

$$Normalized RMS Error = \sum \frac{(Actual \text{Predicted})^2}{Number of \text{Prediction}} \quad (3)$$

The smaller of the Normalized RMS error value indicates that the prediction is better.

b. Unexplained Variance which indicates what portion of the target value is not explained by the prediction value. The Unexplained Variance is defined as follow:

$$Unexplained Variance = \frac{(Actual RMS Error)^2}{Variance of \text{Target Column}} \quad (4)$$

The smaller of the Unexplained Variance value give a better prediction.

c. Correlation Coefficient. It is a number between zero and one which indicates how well the prediction is correlated to the actual. A value of one indicates perfect predictions, and a value of zero indicates no relationship between prediction and target. The Correlation Coefficient is defined as follow:

$$Corr. Coefficient = \sqrt{(1 - Unexplained Variance)} \quad (5)$$

In the following, we demonstrate the experimental results of three major stock indices using all frequent itemsets as input fields. Table 4.1 shows the results of experiment.

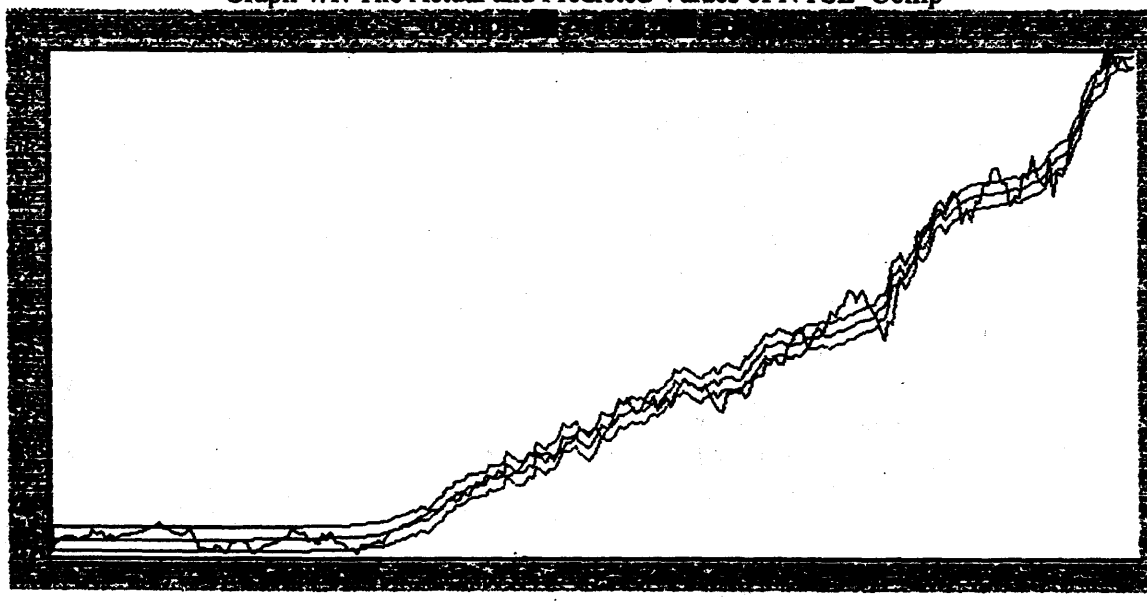
Table 4.1 :The experimental Results of Three Major Stock Indices

Predicted Attribute	Normalized RMS Error	Unexplained Variance	Correlation Coefficient
NYSE_Comp	2.42	0.01	1.00
Dow_Transport	2.80	0.01	0.99
Dow_Industrial	2.41	0.01	1.00

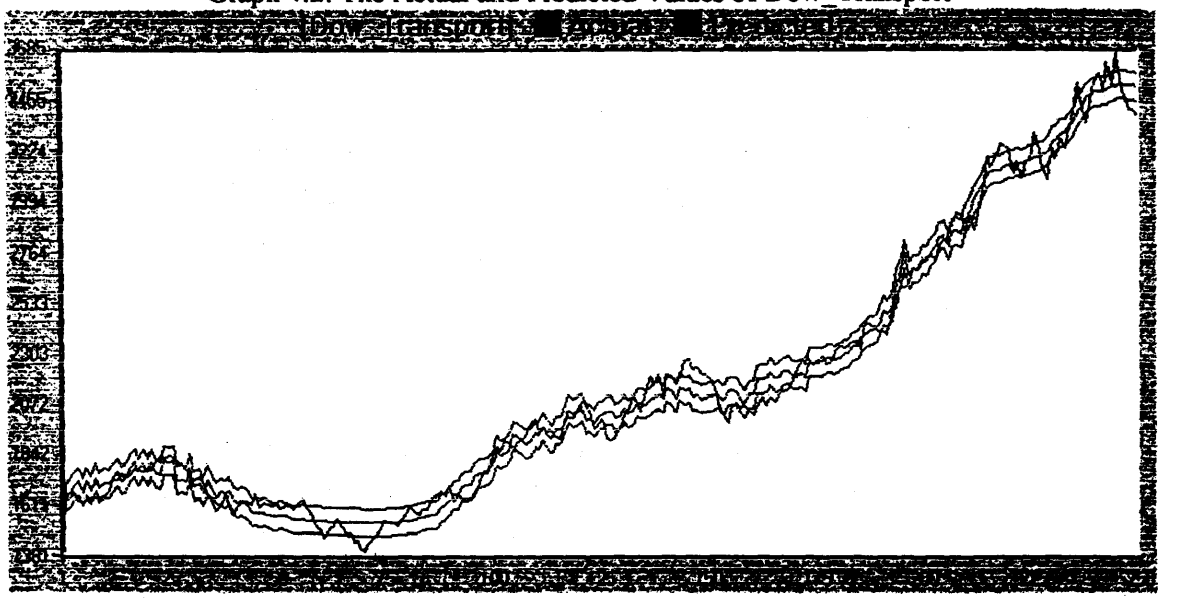
Table 4.1 shows the experimental results of three major stock indices, called NYSE_Comp, Dow_Transport and Dow_Industrial. Our prediction model gives a small normalized RMS error and unexplained variance. This prediction also gives correlation coefficient greater than 0.9. For example, NYSE_Comp gives normalized RMS error equivalent to 2.42, unexplained variance equivalent to 0.01 and correlation coefficient equivalent to 1.00. The achieved results indicate that our prediction could give a better prediction and can improve the ability of prediction.

The effectiveness of our model can be also measured based on the relevance feedback of the prediction. This can be done by calculating the differences between the actual and predicted values (refer to section 3(b)). The actual and predicted values of three major stock indices are depicted in graph 4.1, 4.2 and 4.3 respectively.

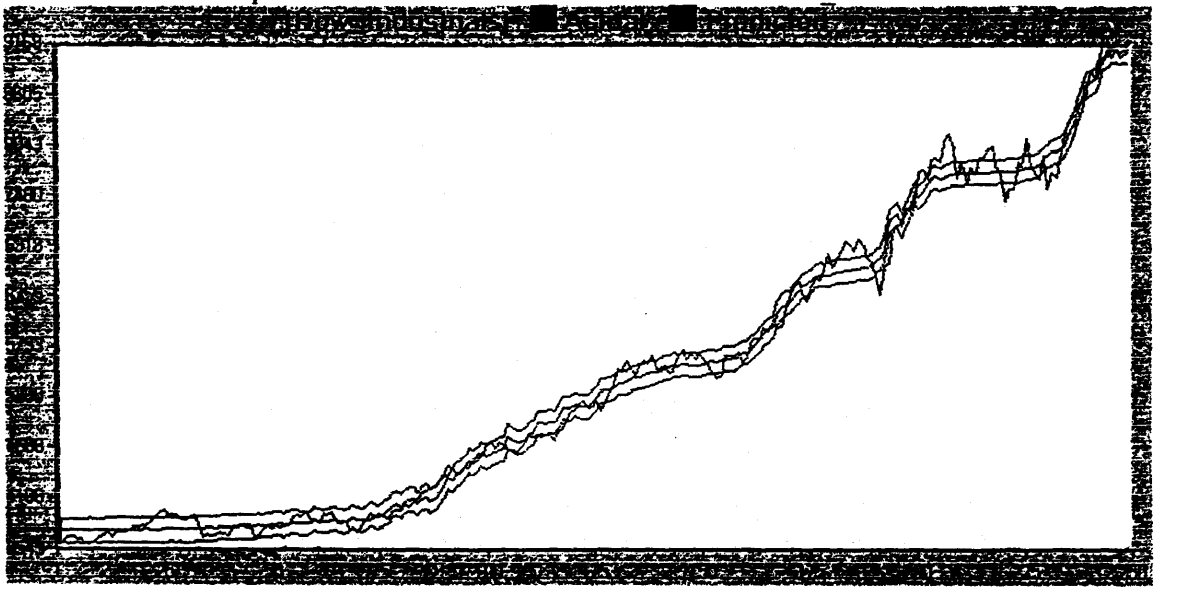
Graph 4.1: The Actual and Predicted Values of NYSE_Comp



Graph 4.2: The Actual and Predicted Values of Dow_Transport



Graph 4.3: The Actual and Predicted Values of Dow_Industrial



The previous graphs, graph 4.1, 4.2 and 4.3 show the actual and predicted values of three major stock indices, called NYSE_Comp, Dow_Transport and Dow_Industrial respectively. These graphs have smooth graphs. Based on the graphs above, we conclude that our model give us a better prediction.

5. Conclusion

Association rule is one of the data mining techniques in data mining that has received a great deal of attention. Today, the mining of such rules is still one of the most popular pattern discovery methods in knowledge discovery in databases. At the present stage, we propose an economic forecasting which consists of two main step called association rule mining and prediction. This model has been implemented and tested using financial market data. The experimental results show that association rules and neural network can improve the ability of making a prediction and could give a better prediction.

References

Alexandra, I.C., and Toshio, O. (1998). " Energy function construction and implementation for stock exchange prediction NNs", 1998 second international conference based intelligent electronic system, April 1998

Bing, L., Wynne, H., et.al (1999). " Mining association rules with multiple minimum supports, ACM SIGKDD, August 1999

C. H. Cai, Ada. W.C. Fu, et.al (1998). " Mining association rules with weighted items", IEEE 1998

Chengyi, S., Xueli, Y, et.al (1996). " Adaptive clustering of stock prices data using cascaded competitive learning neural networks"
Christian, H. (1998). " Online association rule mining", technical report UCB//CSD-98=-1004, department of electrical engineering and computer science, University of california at Barkeley

Clarence, N.W. (1993). " Trading a NYSE stock with a simple artificial neural network based financial trading system", IEEE 1993

Clarence, N.W. and Gerhard, E.W., (1993). " A study of the parameters of a backpropagation stock price prediction model", IEEE 1993

David, W.C., Vincent, T.Ng., et.al. (1996). " Efficient mining of association rules in distributed database", IEEE transaction on knowledge and data engineering, Vol.8, No. 6, December 1996

Eui, H., Sam, H., et.al (1999). " Scalable parallel data mining for association rules", IEEE on knowledge and data engineering, Vol. XX, No. V, 1999

Goebel, M., Le, G. (1999). " A survey of data mining and knowledge", SIGKDD exploration volume 1, issue 1, page 20-23
http://www.cs.ualberta.ca/~zaiane/courses/cmput690/notes/chapter1/index.html

Hua, Z. (1998). " On-Line Analytical mining of association rules", MSc thesis, Simon Fraser University, 1998

Jiawei, H., Chiang, Y. et.al (1999\7). "DBMiner : A system for data mining in relational databases and data warehouse", Proc 1996 Int'l Conf on Data Mining and Knowledge Discovery (KDD'96) Portland, Oregon, August 1996

Juchen, H., Ulrich, G, et.al (2000). " Algorithm for association rule mining : A general survey and comparison", ACM SIGKDD, July 2000

Kazuhiro, K. (1995). "Neural multivariate prediction using event knowledge and selective presentation learning", IEEE 1995

Ke. W., Yu, H., et.al (2000). " Mining frequent itemsets using support constraints", VLDB conference, Cairo Egypt 2000

McLaren, I., (1997). " Data Mining : Finding business value in data". Available online via URL : http://home/clara.net/imclaran/dmpaper.html

Michelin, K., Jiawei, H., et.al (1997). " Using data cubes for meta rule guided mining of multiple dimensional association rules", technical report CMPT-TR-97-10, School of Computing Science, Simon Fraser University, May 1997

Olaru, C., Wehenkel, L. (1999). " Data Mining", IEEE Computer Application in Power, July 1999

Ronkainen, P. (1998). " Attribute similarity event sequence similarity in data mining", PhLic thesis, report C-1998-42, University of Helsinki, Department of Computer Science, October 1998

Sarjon, D., Mohd, N., (2000). " Data Mining : A Preview", Journal of Information Technology", jilid 12, Bil. 2 June 2000

Sarjon, D., Mohd, N., (2000). " Predictive data mining based on similarity and clustering methods", Journal of Information Technology", Jilid 12, Bil. 2 December 2000

Sarjon, D., Mohd, N., (2001). " Attribute Similarity in Predictive Data Mining", SCORED 2001, Kuala Lumpur, 20-21 February 2001.

Sarjon, D., Mohd, N., (2001). "A Stock Price Prediction Based on Association Rules and Neural Network Attribute Similarity in Predictive Data Mining", KMICE 2001, Langkawi, 14-15 May 2001.

Shan, C., (1998). " Statistical approach to predictive modeling in large databases", MSc thesis, Computing Science, Simon Fraser University, March 1998

Sheng-Chai, C., Huang-Pin, C., et.al (1999). " A forecasting approach for stock index future using Grey theory and neural networks", IEEE, 1999

Wei, W., (1999). " Predictive modeling based on classification and pattern matching method", MSc thesis, Computing Science, Simon Fraser University Augustus 1999

Wuthrich, B., Permunetilleke, P. et.al (1998a). " Daily prediction of major stock indices from textual WWW Data".

Wuthrich, B., Permunetilleke, P. et.al (1998b). " Daily stock market forecast from textual web data", IEEE 1998

Xinfeng, Y., and John, A.K. (1998). " Mining association rules in temporal databases, 1998

Yongjian, F. (1996). " Discovery of multiple level rules from large databases". PhD thesis, Simon Fraser University, 1996

Zaiane, R. (1999). " Introduction to data mining", CMPUT690 Principles of Knowledge Discovery in Databases". Available online via URL: http://www.cs.ualberta.ca/~zaiane/courses/cmput690/notes/chapter1/index.html