International Journal of Information Sciences and Application.
ISSN 0974-2255 Volume 4, Number 1 (2012), pp. 1-9
© International Research Publication House
http://www.irphouse.com

Rule based Classification of BSE Stock Data with Data Mining

Anil Rajput [1] , S.P. Saxena [2], Ramesh Prasad Aharwal [3] and Ritu Soni [4]

[1]Principal, Bhabha Engineering Research Institute-MCA, Bhopal-26, India
E-mail: drar1234@yahoo.com
[2]HOD, T.I.T. Engineering college-MCA, Bhopal, India
[3]Asstt. Prof., Department of Mathematics and Computer Science,
Govt. P.G. College Bareli (M.P.), India:
E-mail: ramesh_ahirwal_neetu@yahoo.com
[4]Asstt Prof., Department of Computer Science, N.R.I. Group of Institute
Bhopal (M.P.) India

Abstract

Data mining aims to discover hidden knowledge, unknown patterns, and new rules from large databases that are potentially useful and ultimately understandable for making crucial decisions. This paper focuses for the rule based classification model of historical BSE stock data with data mining techniques. In this Paper we have used decision tree and rule induction method with the help of data mining software. We have to create classification rules and induction rules with the use of J48 and PRISM classifiers under WEKA software.

Keywords: Decision tree, Stock market, Data mining, PRISM, Moving average, WEKA

Introduction

Data mining is able to uncover hidden patterns and predict future trends and behaviors in financial markets [2].Data mining is the process of extraction of interesting (nontrivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data [3]. There is still a nontrivial gap between general principals of Data Mining and domain specific, effective Data Mining tools for particular applications. Data mining techniques have been profitably have to shown to generate high forecasting accuracy of stock price movement. Financial stock market is a complex, non inactive, noisy, disordered, nonlinear and dynamic system but it does not follow random walk process. There are many factors that may cause the rise and fall of financial market movement. Predictions of stock market price and its direction are quite difficult. In answer to such difficulty, data mining techniques and data mining software also have been applied for this task.

Stock Market Prediction Techniques

Commonly, techniques on stock market prediction can be divided in two main categories. They are traditional techniques and intelligent techniques. The traditional techniques can be further categorized into fundamental and technical analysis. today, intelligent techniques have become the main focus of the current researches and have widely applied in stock market prediction. We have used some important indicator functions such as moving average (MA), Exponential moving average (EMA). Another indicator functions are used in this paper for transforming new attributes which are used for data mining algorithm, Such as DM+, DM- .

Learning Environment

The Weka Data Mining Environments were used for carrying out the experiments. Waikato Environment for Knowledge Analysis, called shortly WEKA, is a set of state-of-the art data mining algorithms and tools to in-depth analyses. The author of this environment is University of Waikato in New Zealand. The programming language of WEKA is Java and its distribution is based on GNU General Public License. The general setup used is as follows:

Data mining

Fayyad (1997) viewed data mining as a single step in the knowledge discovery in databases (KDD) process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data [4]. Data mining involves applying different operations such as classification, prediction, clustering and Association rule mining. During the implementation of these operations, specific algorithms need to be constructed.

Data mining refers to the process of finding interesting patterns in data that are not explicitly part of the data [9]. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., kmeans clustering is unsupervised) [5]. In this research we used supervised learning. There are different data mining methods are available in the literature of data mining. Some of them are applicable in the Finance domain.

Data Mining Techniques

Author Norashikin [6], described data mining operations as belongs to several categories like classification, clustering, association and prediction. Besides that, researchers have widely used data mining operations such as classification, clustering and time series prediction in [7]. Among a variety of data mining techniques that have been used in finance, we mainly focus on one commonly used technique, namely Rule induction method.

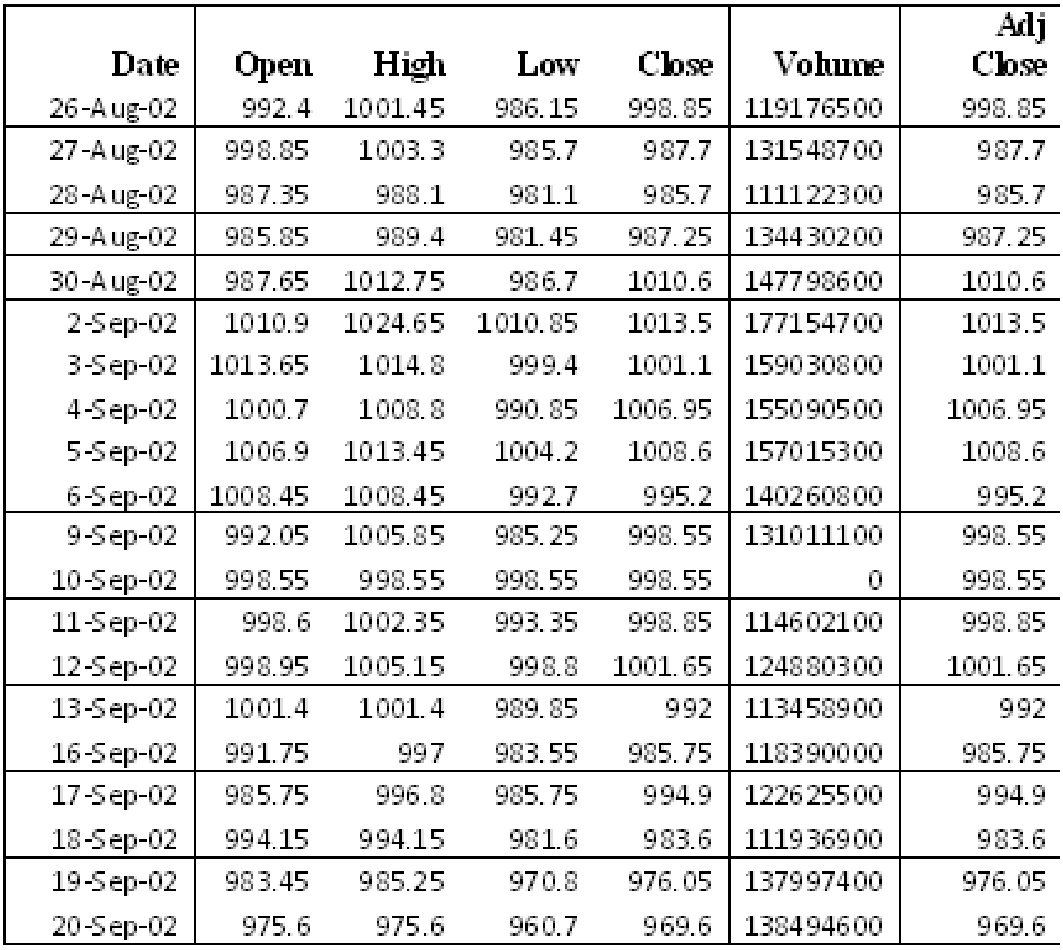
Rule Induction

Rule induction models belong to the logical, pattern distillation based approaches of data mining. Based on data sets, these techniques produce a set of if-then rules to represent significant patterns and create prediction models. Such models are fully transparent and provide complete explanations of their predictions. One commonly used and well-known type of rule induction is the family of algorithms that produce decision trees. A decision tree, which is usually constructed using a training data set, consists of hierarchically organized sets of rules. It is a simple recursive structure for representing a decision procedure in which a new instance is classified into one of the predefined classes. In decision trees, instances are represented as feature vectors containing a list of attribute-value pairs. Each internal node represents a decision attribute-value test. Each branch represents an outcome of the test, and each leaf node denotes a decision class. The decision tree technique is based on a divide-and-conquer approach to the classification problem. It works in a top-down manner: at each stage, it seeks an attribute that separates classes the best to split on, and then recursively processes the partitions resulted from the split. The basic principle is to maximize the entropy of the split subsets, while recursive partitioning is designed to minimize the expected cost of misclassification.

Data Collection

Considering all kinds of available data source, we chose Yahoo! It is because Yahoo! provided Indian, American’s and other countries stock market data, so that we didn’t need to collect the stock data from different places, this could ensure a uniform style format of data, which could facilitate our data pre-processing. Besides, it is free to download the data from Yahoo! which is also another reason that we chose to use it. Data were collected from [10]. Common fields of stock market datasets are Date, open, high, low, close, volume and adj. close. Dataset format is shown as following.

Table 1: Raw data



Date	Open	High	Low	Close	Volume	Adj Close
26-Aug-02	992.4	1001.45	986.15	998.85	119176500	998.85
27-Aug-02	998.85	1003.3	985.7	987.7	131548700	987.7
28-Aug-02	987.35	988.1	981.1	985.7	111122300	985.7
29-Aug-02	985.85	989.4	981.45	987.25	134430200	987.25
30-Aug-02	987.65	1012.75	986.7	1010.6	147798600	1010.6
2-Sep-02	1010.9	1024.65	1010.85	1013.5	177154700	1013.5
3-Sep-02	1013.65	1014.8	999.4	1001.1	159030800	1001.1
4-Sep-02	1000.7	1008.8	990.85	1006.95	155090500	1006.95
5-Sep-02	1006.9	1013.45	1004.2	1008.6	157015300	1008.6
6-Sep-02	1008.45	1008.45	992.7	995.2	140260800	995.2
9-Sep-02	992.05	1005.85	985.25	998.55	131011100	998.55
10-Sep-02	998.55	998.55	998.55	998.55	0	998.55
11-Sep-02	998.6	1002.35	993.35	998.85	114602100	998.85
12-Sep-02	998.95	1005.15	998.8	1001.65	124880300	1001.65
13-Sep-02	1001.4	1001.4	989.85	992	113458900	992
16-Sep-02	991.75	997	983.55	985.75	118390000	985.75
17-Sep-02	985.75	996.8	985.75	994.9	122625500	994.9
18-Sep-02	994.15	994.15	981.6	983.6	111936900	983.6
19-Sep-02	983.45	985.25	970.8	976.05	137997400	976.05
20-Sep-02	975.6	975.6	960.7	969.6	138494600	969.6

Data Pre-processing

Sometimes, data in the stock market is dirty, this means the data may be: incomplete: lacking attribute values, deficient certain attributes of interest, or containing only aggregate data noisy. Therefore, if no quality data, then the result will be no quality too. Good result must be based on quality data. For example, duplicate or missing data may cause incorrect or even misleading statistics. Thus, in this section, it is important that the missing values have been filled, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data Preparation

Before implementing the data mining algorithm, we did some preparation work, it was collecting the financial data. Because we were using the technical analysis way to do the forecasting, but not the fundamental analysis, when we prepared our data, we were not looking for the financial statement like the annual reports of the company nor the interest rate of the country. What we needed are the historical stock quotes.

Intuitively, based on the EMH, the price of the stock yesterday is going to have the most impact on the price of the stock today. Thus as we go along the time-line, data-points which are nearer to today's price point are going to have a greater impact on today's price. We add one more attribute to the above attributes; this attribute will act as our label for predicting the movements of the stock price. This attribute will be called "Indicator" and will be dependent on the other available attributes. For our experiments we use the EMA (Exponential Moving Average) as the indicator function.

Data Representation

After selecting the features or attributes to be included in the training and testing sets and proper data segments for encoding, the focus is on the representation of the data in the datasets. The data sets in this study are further transformed into ten days closing prices and the training entries are based on these scaled closing prices values.

The un-processed data were full of noises, which mean there were instances that would impede our later data processing, and would affect the result of the data mining if we didn't clear away the noises. Because the un-processed data fie is in comma separate value format (.csv) which can be open by the Microsoft Excel.

Data Transformation

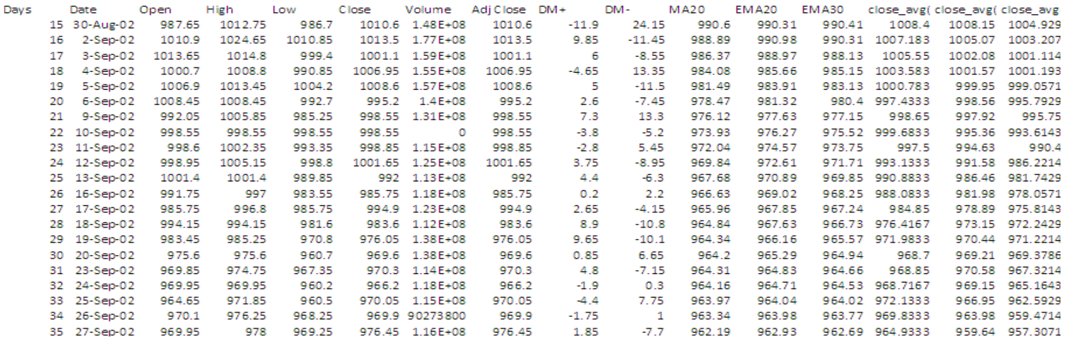
Before the data mining, we did one more data preparing step; that is data transformation. In this step, we create more attributes on our stocks' data, and we also included the class labels in the data file. This stage is important because our data mining methods required this kind of data to work. There are more 20 attributes and 2 class labels we added to our pre-processed data. Each attribute is a TRUE/FALSE question.

For example:
1. Close > EMA20
2. Vol_Avg(1) > Vol_Avg(3)

The 21 TRUE/FALSE-question attributes are generated based on the pre-processed data, which means we would use the combination among OPEN, HIGH, LOW, CLOSE, DM+, DM-, EMA20 and EMA30.

Besides generating the attributes, we also to generate the class labels to our data with the help of Microsoft Excel, because class label is one important element for classification type data mining methods. Table 2 shows the attributes and their values. Table 3 present the transformed attributes. These attributes were transformed from the technical indicators and auxiliary attributes which are shown the right hand side of the table 2.

Table 2: Preprocessed Data.



Days	Date	Open	High	Low	Close	Volume	Adj Close	DM+	DM-	MA20	EMA20	EMA30	close_avg(	close_avg(	close_avg
15	30-Aug-02	987.65	1012.75	986.7	1010.6	1.48E+08	1010.6	-11.9	24.15	990.6	990.31	990.41	1008.4	1008.15	1004.929
16	2-Sep-02	1010.9	1024.65	1010.85	1013.5	1.77E+08	1013.5	9.85	-11.45	988.89	990.98	990.31	1007.183	1005.07	1003.207
17	3-Sep-02	1013.65	1014.8	999.4	1001.1	1.59E+08	1001.1	6	-8.55	986.37	988.97	988.13	1005.55	1002.08	1001.114
18	4-Sep-02	1000.7	1008.8	990.85	1006.95	1.55E+08	1006.95	-4.65	13.35	984.08	985.66	985.15	1003.583	1001.57	1001.193
19	5-Sep-02	1006.9	1013.45	1004.2	1008.6	1.57E+08	1008.6	5	-11.5	981.49	983.91	983.13	1000.783	999.95	999.0571
20	6-Sep-02	1008.45	1008.45	992.7	995.2	1.4E+08	995.2	2.6	-7.45	978.47	981.32	980.4	997.4333	998.56	995.7929
21	9-Sep-02	992.05	1005.85	985.25	998.55	1.31E+08	998.55	7.3	13.3	976.12	977.63	977.15	998.65	997.92	995.75
22	10-Sep-02	998.55	998.55	998.55	998.55	0	998.55	-3.8	-5.2	973.93	976.27	975.52	999.6833	995.36	993.6143
23	11-Sep-02	998.6	1002.35	993.35	998.85	1.15E+08	998.85	-2.8	5.45	972.04	974.57	973.75	997.5	994.63	990.4
24	12-Sep-02	998.95	1005.15	998.8	1001.65	1.25E+08	1001.65	3.75	-8.95	969.84	972.61	971.71	993.1333	991.58	986.2214
25	13-Sep-02	1001.4	1001.4	989.85	992	1.13E+08	992	4.4	-6.3	967.68	970.89	969.85	990.8833	986.46	981.7429
26	16-Sep-02	991.75	997	983.55	985.75	1.18E+08	985.75	0.2	2.2	966.63	969.02	968.25	988.0833	981.98	978.0571
27	17-Sep-02	985.75	996.8	985.75	994.9	1.23E+08	994.9	2.65	-4.15	965.96	967.85	967.24	984.85	978.89	975.8143
28	18-Sep-02	994.15	994.15	981.6	983.6	1.12E+08	983.6	8.9	-10.8	964.84	967.63	966.73	976.4167	973.15	972.2429
29	19-Sep-02	983.45	985.25	970.8	976.05	1.38E+08	976.05	9.65	-10.1	964.34	966.16	965.57	971.9833	970.44	971.2214
30	20-Sep-02	975.6	975.6	960.7	969.6	1.38E+08	969.6	0.85	6.65	964.2	965.29	964.94	968.7	969.21	969.3786
31	23-Sep-02	969.85	974.75	967.35	970.3	1.14E+08	970.3	4.8	-7.15	964.31	964.83	964.66	968.85	970.58	967.3214
32	24-Sep-02	969.95	969.95	960.2	966.2	1.18E+08	966.2	-1.9	0.3	964.16	964.71	964.53	968.7167	969.15	965.1643
33	25-Sep-02	964.65	971.85	960.5	970.05	1.15E+08	970.05	-4.4	7.75	963.97	964.04	964.02	972.1333	966.95	962.5929
34	26-Sep-02	970.1	976.25	968.25	969.9	90273800	969.9	-1.75	1	963.34	963.98	963.77	969.8333	963.98	959.4714
35	27-Sep-02	969.95	978	969.25	976.45	1.16E+08	976.45	1.85	-7.7	962.19	962.93	962.69	964.9333	959.64	957.3071

Table 3: Experimental Attributes (Transformed data).

Attributes	Description	Values
a1	Close_avg(3)>close_avg(5)	TRUE/FALSE
a2	Close_avg(5)>close_avg(7)	TRUE/FALSE
a3	Open_avg(3)> Open _avg(5)	TRUE/FALSE
a4	Open_avg(5)>Open_avg(7)	TRUE/FALSE
a5	Open_avg(3)>Open_avg(7)	TRUE/FALSE
a6	low_avg(3)>low_avg(5)	TRUE/FALSE
a7	Low_avg(5)>low_avg(7)	TRUE/FALSE
a8	High_avg(3)>High_avg(5)	TRUE/FALSE
a9	High_avg(5)>High_avg(7)	TRUE/FALSE
a10	DM+_avg(3)>DM+_avg(5)	TRUE/FALSE
a11	DM+_avg(5)>DM+_avg(7)	TRUE/FALSE
a12	DM-_avg(3)>DM-_avg(5)	TRUE/FALSE
a13	DM-_avg(5)>DM-_avg(7)	TRUE/FALSE
a14	DM+>DM+_avg(3)	TRUE/FALSE
a15	DM->DM-_avg(3)	TRUE/FALSE
a16	Close>close_avg(3)	TRUE/FALSE
a17	Open>open_avg(3)	TRUE/FALSE
a18	High>High_avg(3)	TRUE/FALSE
a19	Low>Low_avg(3)	TRUE/FALSE
a20	EMA20>EMA20_avg(3)	TRUE/FALSE
a21	EMA20_avg(3)>EMA20_avg(5)	TRUE/FALSE
A22	EMA20_avg(5)>EMA20_avg(7)	TRUE/FALSE
CLASS10_2PER	Stock price will be increase 2 percent within 10 days	BUY/SALE
CLASS15_3PER	Stock price will be increase 3 percent within 15 days	BUY/SALE

Experimental Setup

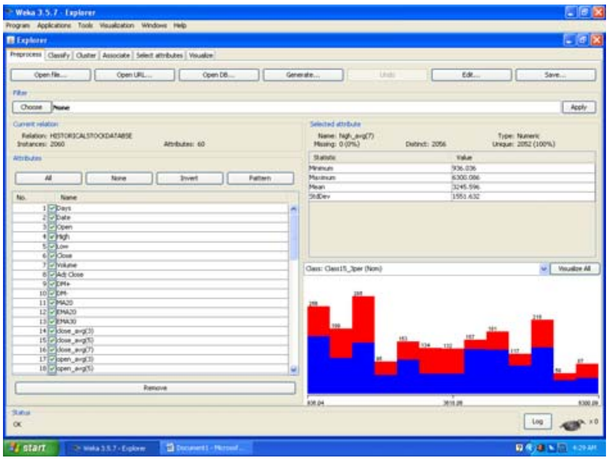
Rule-Based Classification

Rule-based classification is another type of supervised learning data mining method. The learned model of the rule-based classification is represented as a set of IF-THEN rules. This algorithm provides mechanisms that generate rules by concentrating on a specific class at a time and maximizing the probability of the desired classification.

PRISM Algorithm

Prism algorithm generates rules for each class by looking at the training data and adding rules that completely describe all tuples in that class. PRISM identifies a rule that covers many instances in the class, separates out the covered instances, and continues the process with the rest. It measures the success of a rule by the formula: p/t, where t is a total number of instances, and p is number of those that are positive.

Screen Shots



Filter: None

Relation: HISTORICALSTOCKDATABSE
Instances: 2060 Attributes: 60

Name: high_avg(7) Type: Numeric
Missing: 0 (0%) Distinct: 2056 Unique: 2052 (100%)

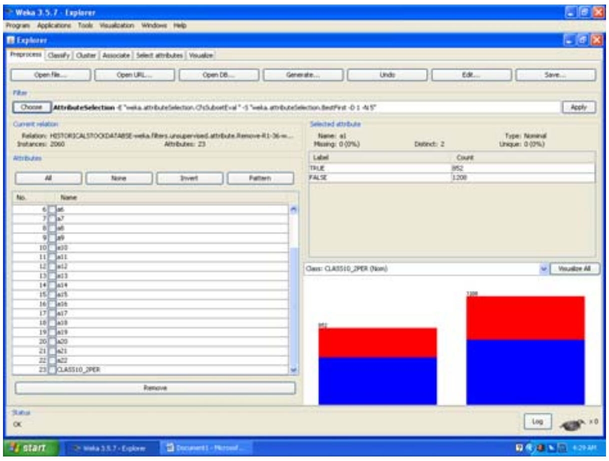
Statistic	Value
Minimum	936.036
Maximum	6300.086
Mean	3245.596
StdDev	1551.632

No.		Name
1	☑	Days
2	☑	Date
3	☑	Open
4	☑	High
5	☑	Low
6	☑	Close
7	☑	Volume
8	☑	Adj Close
9	☑	DM+
10	☑	DM-
11	☑	MA20
12	☑	EMA20
13	☑	EMA30
14	☑	close_avg(3)
15	☑	close_avg(5)
16	☑	close_avg(7)
17	☑	open_avg(3)
18	☑	open_avg(5)

Class: Class15_3per (Nom)

Status: OK

Figure 1: Weka Explorer for Experiment



Filter: AttributeSelection -E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Relation: HISTORICALSTOCKDATABSE-weka.filters.unsupervised.attribute.Remove-R1-36-w...
Instances: 2060 Attributes: 23

Name: a1 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Label	Count
TRUE	852
FALSE	1208

No.		Name
6	☐	a6
7	☐	a7
8	☐	a8
9	☐	a9
10	☐	a10
11	☐	a11
12	☐	a12
13	☐	a13
14	☐	a14
15	☐	a15
16	☐	a16
17	☐	a17
18	☐	a18
19	☐	a19
20	☐	a20
21	☐	a21
22	☐	a22
23	☐	CLASS10_2PER

Class: CLASS10_2PER (Nom)

Status: OK

Figure 2: Statistics of Class Attribute

Rules Generated from Prism Algorithm

```
Test mode:    split 66% train, remainder test

=== Classifier model (full training set) ===

Prism rules
----------
If a22 = TRUE  and a20 = FALSE  and a19 = TRUE and a2 = TRUE then BUY
If a22 = TRUE and a4 = FALSE and a9 = TRUE and a6 = TRUE then BUY
If a22 = TRUE and a6 = FALSE and a5 = TRUE and a4 = FALSE then BUY
If a22 = TRUE and a20 = FALSE  and a15 = FALSE and a4 = TRUE and a6 =
FALSE then BUY
If a21 = TRUE  and a20 = FALSE  and a13 = TRUE and a16 = TRUE then BUY
If a22 = TRUE and a2 = FALSE and a13 = TRUE and a7 = TRUE then BUY
If a22 = TRUE and a6 = FALSE and a19 = TRUE and a21 = FALSE and a16
=FALSE then BUY
If a21 = TRUE and a6 = FALSE and a20 = FALSE and a1 = TRUE then BUY
If a21 = TRUE and a4 = FALSE and a7 = TRUE and a11 = FALSE then BUY
If a22 = TRUE and a7 = FALSE and a12 = TRUE and a14 = FALSE and a4 = TRUE
then BUY
….      ….      ….      ….      ….      ….      ….      ….
….
….      ….      ….      ….      ….      ….      ….      ….
….
If a8 = TRUE and a3 = FALSE and a4 = FALSE and a6 = TRUE and a14 = FALSE
then BUY
If a22 = TRUE and a12 = FALSE and a1 = TRUE and a8 = FALSE then BUY
If a22 = TRUE and a8 = TRUE and a15 = TRUE and a13 = FALSE and a6 = FALSE
then BUY
```

Conclusion

This paper present induction rules of BSE (Bombay Stock Exchange) Data which may be useful for stock market investors. These rules are generated from Data mining methods with the help of WEKA machine learning Software. Rules were shown in the above section.

References

[1] Chu, F. and Wang, L. (2003). Gene Expression Data Analysis using Support Vector Machine. IEEE.

[2] Dongsong Zhang and Lina Zhou(2004), Discovering Golden Nuggets: Data Mining in Financial Application, Ieee Transactions On Systems, Man, And Cybernetics—Part C: Applications and reviews, vol. 34, no. 4, pp: 513-521.

[3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine. American Association for Artificial Intelligence.PP- 37-54.

[4] Fayyad, U. (1997). Data Mining and Knowledge Discovery in Database:Implications for Scientific Databases.Proceedings of the IEEE.

[5] Mary K. Obenshain, MAT (2004): Application of Data Mining Techniques to Healthcare Data, Statistics for Hospital Epidemiology,(25),(8), 690-695.

[6] Norashikin binti Ahmad (2003). Pengkelasan Data Kajicuaca Di Dalam Peramalan Hujan Dari Pelbagai Sumber Menggunakan Rangkaian Neural.Universiti Teknologi Malaysia: Master Thesis.

[7] Povinelli, R. J. (2001): Identifying Temporal Patterns for Characterization and Prediction of Financial Time Series Events.J.F. Roddick and K. Hornsby (Eds.): TSDM 2000, LNAI 2007. Springer-Verlag Berlin Heidelberg : 46-61.

[8] Weka: Data mining software in java.

[9] Witten I. H., Frank E. (2005): Data Mining, Practical Machine Learning Tools and Techniques, 2nd Elsevier.

[10] Yahoo finance website http: //finance.yahoo.com