International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Using Association Rule Mining: Stock Market Events Prediction from Financial News

Shubhangi S. Umbarkar[1], Prof. S. S. Nandgaonkar[2]

[1]Savitribai Phule Pune University, Vidya Pratishtan's College of Engineering, Vidya Nagari, Baramati, Pune, 413133, India

[2]Savitribai Phule Pune University, Vidya Pratishtan's College of Engineering, Vidya Nagari, Baramati, Pune, 413133, India

Abstract: ***Ability to predict direction of stock price accurately is very crucial for market dealers or investors to maximize their profits. Decision-making such as whether to buy, sell or hold of shares for investor in stock market is also another difficult task. Data mining techniques have been successfully shown to generate high forecasting accuracy of stock price movement and corresponding signals. Prediction of stock price is the activity of determining future state of the stock price by using various techniques. In presented work Data Mining Technique such as Association Rule Mining is used for prediction of stock market. Prediction is depends on technical trading indicators and closing prices of the stock. Rules are defined according to signal generated by each technical trading indicator and mapped across the current date query to generate the signals like buy, sell or holds the shares.***

Keywords: Stock market prediction, Decision making, Association rule Mining, Data mining technique, Naïve Bayes.

1. Introduction

Stock market prediction is burning topic in the field of finance. Due to its business increment, it has attracted often aid from educator to economics sector. It is impossible to give the prediction of prices of stock market because of stock prices are changed by every second. Market stock prediction has ever been a subject of curiosity for most investors and business analyst. In today's information-driven domain, more individuals try to keep record up-to-date with the current developments by reading informative news items on the web [1]. Many factors are responsible for the fluctuations of the market movement. The main factors are financial condition, political circumstances, trader's opportunity and other unforeseen events. Therefore, predictions of stock market price and its tracks are difficult. Investor always tries to find an effective way to find the good stock and right timing to buy or sell of the stock. For the all companies the stock market is extremely important area. By the behavior of the investor the stock price is determine and by using appropriate information investor determines stock prices to predict how the market will act or react.

In order to predict the nature of the stock market computing techniques need to be combined. Distinct methods of financial analysis have been improved, and traditional financial methods are varied, as the time elapsed. To predict future state of the market there are two methods can be used. First method is fundamental analysis that includes statistical data of a company which is most important part of the theoretical investigation and includes reports, financial status of the company, the balance sheets etc. It also includes reasoning of marketplace aggregation, capableness and assets of visitor, the contention, import/export loudness, production indexes, price statistics, and the daily information or rumors about company. The second method is technical analysis [2]; it includes different techniques for prediction of the stock market. Data mining techniques are used for extracting information from large databases, which is a wide technology that help to focus on the most important information in data repositories with great possibilities. Data mining techniques such as association rule mining is used for making automated decisions such as buy, hold, or sell an shares. Neural network [3] is another significant method for stock predictions because of its ability to deal with fuzzy, unsure and inadequate data, which may fluctuate rapidly in very short period.

Numerous studies showed that there is a direct relation between the stock markets and financial news [4] because financial market is motivated by information and an important source of information is news, which comes from different communicating media by verities of channels. News include general news articles, company releases, news of global economy that can affect the performance of stock market and used them to predict the future stock prices and direction of the stocks. As the number of information sources is increasing day by day, which may results in high volumes of news therefore; there is a need to extract the important information from news article for prediction. News articles are mainly unstructured and found in World Wide Web that needs to convert them into structured form to analyze patterns in these articles.

In decision-making, process selecting, preprocessing all the relevant information is a challenging task [5]. Natural Language Processing and Data Mining methods such as text classification can be used for extracting the news information for crating feature vectors. Extraction was unknown, implicit and potentially useful information from data in databases previously, which is an effective way of data mining. It is generally known as knowledge discovery in database (KDD). In our approach, association rule mining is used for prediction of stock market. By using the historical data of the closing prices of stocks, the future state of the shares is determined and appropriate signal is generated i.e. whether to sell, buy or holds the shares.

The paper is structured as follows. First, the related work on stock market prediction and news extraction is presented in section II. Then the implementation details are described in

section III. Data Set and Results are described in Section IV. Conclusion is describes Section V. Future Scope described in Section VI.

2. Related Work

This section provides a brief background survey of this area. Over the past two decades, many measurable changes have assumed in the surroundings of business markets. The utilization of strong communication and trading facilities has enlarged the scope of option for investors. Prediction of stock market event is an important and sensitive motive that has attracted researchers. Financial market is mostly motivated by news information, which is announced in newspaper and in different communication media. Day by day information regarding finance is increasing widely. Financial market information is time sensitive. Uncertainty is the main characteristic of all stock markets, which is related to their future state. This feature is undesirable and unavoidable for the investor whenever the stock market is selected for the investment. To reducing this uncertainty is specially challenging task. Stock market prediction is one of the best option to reduce uncertainty. Stock market prediction includes uncovering market trends, planning investment ,investment strategies, determining the perfect time to purchase the stocks and what stocks to purchase.

The lexico-semantic patterns and lexico-syntactic patterns methods for extraction of financial event from RSS news feeds[6]. Lexico-semantic patterns used for financial ontology that leveraging the commonly used lexico-syntactic patterns to a higher abstraction level by enabling lexico-semantic patterns to identify more and more relevant events than lexico-syntactic patterns from text. The semantic web used to classify the news item. Semantic actions allow to up-dating the domain knowledge. Semantic Web Rule Language (SWRL) is responsible for implementation of the action rule. Triples paradigms are used for defining lexico-semantic information extraction patterns that resemble simple sentences in natural language. Event rule engine used to allow rules creation, financial event extraction from RSS news feed headlines, and ontology updates. The rule engine does the following actions.

- Mining text items for patterns,
- Creating an event if a pattern is found,
- Determining the validity of an event by the user,
- Executing appropriate update actions if an event is valid.

The engine consists of multiple components. The first component is rule editor, using the editor user can construct the rule. Second component is event detector, which is used for mining text items for the lexico-semantic patterns occurrence for the event rules. The third component validation environment using this component user can determine the validation of the event and can modify the event if event detector made an error. In addition, the last component is action execution engine which is used to perform the updating the rule, finding the event which are associated with that rule, if and only if the event is valid. The effectiveness of the above work is tested with the help of precision, recall, and F1 measures.

The comparison of lexico-semantic patterns and lexico syntactic patterns is studied in [7]. In this work, ontology is used for retrieving relevant news items in a semantically enhanced way. They have used Hermes Information Extraction Language (HIEL), which apply the semantic concepts from ontology and used to evaluate in the context of extracting events and relations from news. Hermes Information Extraction Engine also has implemented that compiles the rules in the rule compiler and matches the rules to the text using the rule matcher after preprocessing the news corpus. They also showed that the lexico-semantic patterns are superior to lexico-syntactic patterns with respect to efficiency and effectiveness. Pattern-based information extraction techniques are mainly focused.

A number of prototypes for predicting the short-term market reaction to news based on text mining techniques are described in[8] and some of them are explained below:

2.1 Prototype developed by Wthrich et al.

This prototype attempts to predict the 1-day trend of five major equity indices such as the Dow Jones, the Nikkei, and the Straits Times. According to a 3-category model, the information of this prototype were labeled. The first and second category contains news articles followed by 1-day periods and is associated with increasing or decreasing of equity index at least 0.5%. The remaining news articles contains the third category. The threshold is of +/- 0.5% was chosen for trading sessions so that one-third part of it roughly falls in each of the three categories. During its operational phase the prototype categorized all newly published articles. The numbers of news articles in each category were counted and depending on where the most news articles were appoint to, the prototype activated for the corresponding index a buy recommendation, a short recommendation, or advised to do nothing.

2.2 Prototype developed by Lavrenko et al.

The prototype Enalyst was developed around 2000 at the University of Massachusetts Amherst [9][10][11]. Enalyst aims to predict stocks in very short-term i.e. intraday price trends of a subset by investigating the homepage of YAHOO.

2.3 Prototype developed by Thomas et al.

This prototype was developed at the Robotics Institute of Carnegie Mellon University between 2000 and 2004 [12][13]. This prototype mainly focused on forecasting the instability. In this strategy once news is published that may increase instability the market is temporarily exit for particular stock. Then the decision of re-entering into market is depends on technical indicators. Genetic Programming [14] is used for interpreting the future state of the stock market. Based on parallel search, natural selection and historical data it generates the decision tree for taking decision. Evolution based Functional Link Artificial Neural Network (FLANN) [15] is used to predict the Indian stock market. The model is based on the Back-Propagation (BP) algorithm and Differential Evolution (DE) algorithm. The FLANN is a single layer, single neuron architecture which

has the capability to form complex decision regions by creating non-linear decision boundaries.

3. Implementation Details

The technique of Market Event Prediction uses information (historical) system for the prediction of future market state and for making the useful decision i.e. whether to buy, sell or holds the shares at a perfect time. In previous works event extraction from financial news was done manually, which is very tedious job therefore, there is need to tackle this problem by changing the information selection criteria. Here we consider the closing prices of the shares of respective day which reduces effort that are required to processing the news information. For interpreting the future state of stock market, the technical trading indicator is very important. Technical trading indicator such as Simple Moving Average (SMA), the Bollinger Band (BB), the Exponential Moving Average (EMA), the Rate of Change (RoC), Momentum (MOM), and Moving Average Convergence Divergence (MACD) and relation- ship is calculate which are train by Nave Bayes classifier. The overall framework is shown in figure.1

Dataset i.e. closing prices of shares is collected from mcxindia.com. Collected dataset is used for calculating technical trading indicator and for defining the rule. Then we used Naive Bayes algorithm for training of technical trading indicator for generating the signals. **The following steps are executed for signal generation.**

3.1 Technical Trading Indicator

3.1.1 Simple Moving Average (SMA)

Simple Moving Average is the most important technical trading indicator that will gives the averages of last 20 days of the price of the stock and is calculated as:

$$M_i = \frac{\sum_{j=1}^{N} P_i}{N} \quad (1)$$

Where P_i represents the price on day i. When the price is exceed than the moving average in a downward movement then sell signal is generated and when the price is less in a upward movement then buy signal is generated.

3.1.2 Bollinger Bands (BB)

The Bollinger bands is a another technical indicator which makes two bands i.e. upper band and lower band around a moving average and that are based on the standard deviation of the price. The bands will wide and narrow if and only if the volatility is high and low respectively.

$$L = M - 2 \times \sigma_M \quad (2)$$

$$U = M + 2 \times \sigma_M \quad (3)$$

Where, σ_M is for the volatility of moving average M. When the price is under the lower band a buy signal will generated, as an oversold situation and when the price is greater than the upper band a sell signal will generated at an overbought situation.

3.1.3 Exponential Moving Average (EMA)

By using a short and a durable average, the exponential moving average (EMA) identify trends. The new trend is starts if the averages cross each other. As an example if the short-term average is set at 5 days and the long-term average at 20 days then EMA is calculated as follows.

$$E_i = \frac{2}{N+1} \times (P_i - E_{i-1}) + E_{i-1} \quad (4)$$

Where P_i represents the on a day i, and N is the total number of day. A buy signal is generated if the short term average crosses the long term average upwards and if the short term average crosses the long term average in downwards a sell signal is generated.

3.1.4 Rate of Change (RoC)

Rate of change technical indicator is very important which calculates the difference between the closing price of the current day i and the closing price of 10 days earlier. It is calculated as:

$$C_i = \frac{P_i - P_{i-10}}{P_{i-10}} \quad (5)$$

A sell signal is generated if the R_oC starts decreasing above 0 and a buy signal is generated if RoC starts increasing below 0.

3.1.5 Momentum

This indicator is same like as RoC and uses same formula. It will generate the buy signal when the momentum crosses the zero level instead of after a peak. A sell signal is generated when the R_oC crosses the zero level downwards.

3.1.6 Moving Average Convergence Divergence (MACD)

The moving average convergence divergence is another technical indicator that subtracts two exponential averages from each other. If there are two exponential averages namely as 12 and the 26-day exponential average then MACD will calculated as:

$$D_i = E[12]_i - E[26]_i \quad (6)$$

When the MACD reaches zero level in an upward motion a buy signal is generated and when a MACD breaks through the zero in downward motion sell signal is generated.

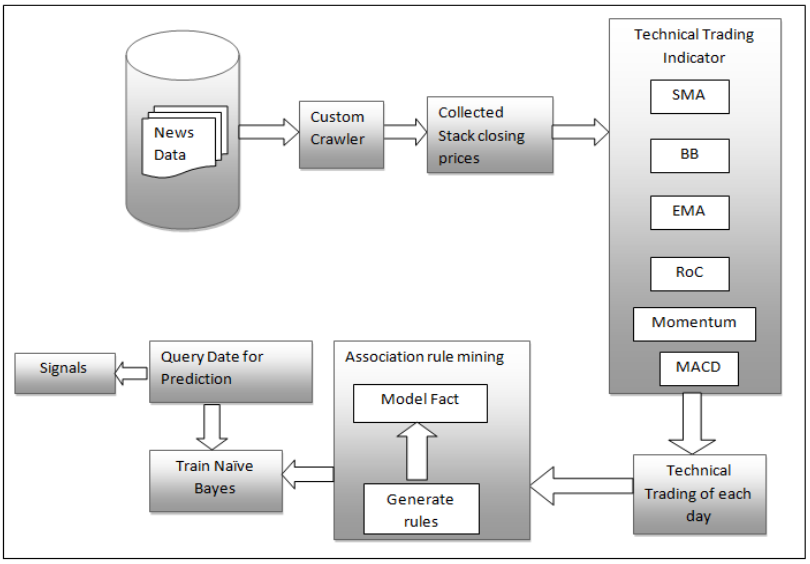

Figure1: System Architecture

3.2 Association Rule Mining

3.2.1 Generate Rules

When technical trading values are collected then, the technical trading values for each day are modeled. Depending upon the prediction date given by the user the mapping is done and rules are generated. Each rule has one fact. For example "If Share Price=164.00 and SMA=166.25 then Signal=Buy. In this Price=164.00 and SMA=166.25 is a rule and Signal=Buy is a fact. Similarly, for each day all technical indicators generate rules.

3.2.2 Facts

Facts are nothing but the Buy/Sell/Hold signals generated by the rules. These facts give the prediction to user whether to buy/sell/hold the shares. When we receive the multiple facts from the multiple rules the probability of all facts are calculated and the corresponding signal is predicted to the user.

3.2.3 Generate signal

After the rule mapping process the corresponding prediction for the shares i.e. Sell/Buy/Hold is provided to the user. According to this prediction, the user decides his/her strategy.

3.3 The Naive Bayes Model

Naïve Bayes classifier is used to train the technical indicator. Rules which are generated by using all technical indicator values are trained by the Naïve Bayes. Bayesian classifier is based on Bayes theorem. Naive Bayesian classifiers assume that the effect of technical indicator values on a given class is independent of the values of the other technical indicator. This assumption is called class conditional independence. The naive Bayesian classifier works as follows:

Let S be a training set of technical indicator like SMA, BB, EMA, RoC, Momentum, MACD with their class labels and there are k classes, C_1, C_2, C_3....., C_n. Each technical indicator is represented by an n-dimensional vector, $X=fx_1$, $x_2,...,x_n$ depicting n measured values of the n attributes, A_1, A_2, A_3....., A_n, respectively. Given a technical indicator X, the classifier will predict that X belongs to the class having the highest probability of the similarity, conditioned on X. That is X is predicted to belong to the class C_i if and only if $P(C_iX) > P(C_iX)$ for $1 \le j \le m$; $j \neq i$.

Bayes theorem:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} (7)$$

3.3.1 Naïve Bayes Algorithm

Learning Phase: Given a training set **S**,

1. For each target value of c_i $(c_i = c_1,\cdots,c_L)$

$\hat{P}(C = c_i) \leftarrow$ estimate $P(C = c_i)$ with examples in **S**;

2. For every attribute value a_{jk} of each attribute

x_j $(j = 1,\cdots,n; k = 1,\cdots,N_j)$

3. $\hat{P}(X_j = a_{jk} | C = c_i) \leftarrow$ estimate $P(X_j = a_{jk} | C = c_i)$

with examples in **S**;

Output: conditional probability for elements x_i , $N_j \times L$

Test Phase: Given an unknown instance $X' = (a'_{1,\ldots,}a'_n)$

Look up to assign the label c^* to **X'** if

$$[\hat{P}(A'_1|c^*)...\hat{P}(A'_n|c^*)] > [\hat{P}(A'_1|c)...\hat{P}(A'_n|c)]\hat{P}(c),$$
$$c \neq c^*, c = c_1, ..., c_n$$

4. Result and Dataset

The proposed method can be evaluated in the context of two different data sets collected from xignite[1], mcxindia[2].

1) A closing prices of the shares at the end of the day takes from xignite[1].
2) The another source is barchartondemand[2] which is used for collection of company name, closing price of the shares closing prices of the shares.

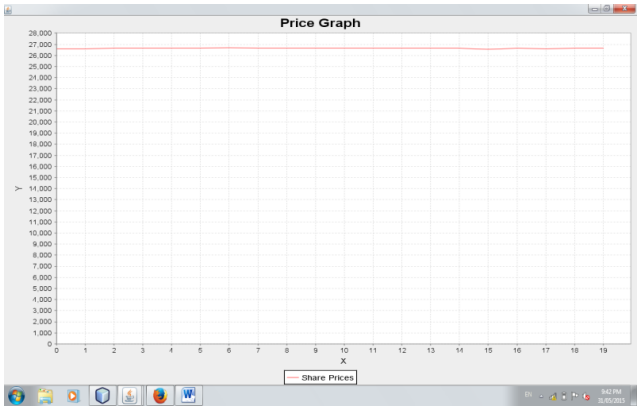

Figure 2: Graph of Share prices

[1].http://www.xignite.com/product/XigniteNews/api/legacy/1/GetStckHeadlines.
[2].http://www.mcxindia.com

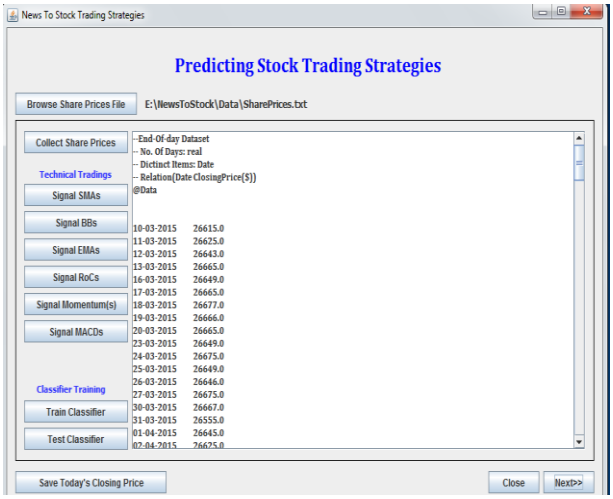

Figure 3: Collected Share Prices

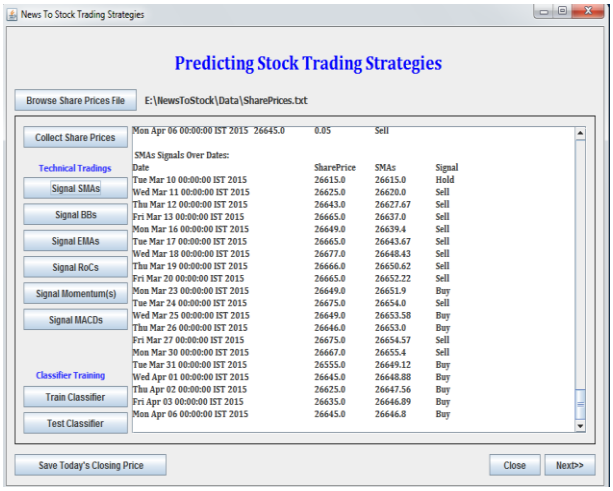

Figure 4: Signal ganrataed by Simple Moving Avarege (SMA)**Figure 5:**Final Predicted Signal

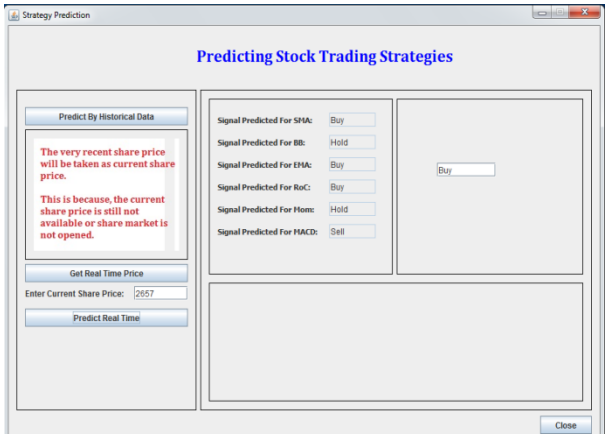

Figure 5: Final Signal

5. Conclusion

The presented method for prediction of stock market considers only closing prices of the shares instead of textual information about the stocks. This reduces the efforts that are required for the extraction of news information. The trading strategies consider technical trading indicators, which are used to generate the superior returns. The technical trading indicator give decision that is more appropriate. Data mining technique such as association rule miming and Naïve Bayes algorithm generates significant signals within the polynomial time. It also increased the accuracy of the prediction system by accepting the accurate closing prices of the stock.

6. Future Scope

The future work will focus on including more technical indicators which will generate the trading strategies. The interaction between events occurring within the same day, or within finer time intervals will be considered.

7. Acknowledgment

I express great many thanks to Prof. S.S.Nandgaonkar for her great effort of supervising and leading me, to accomplish this fine work. Also to college and department staff, they were a great source of support and encouragement. To my friends and family, for their warm, kind encourages and loves. To every person gave us something too light my pathway, I thanks for believing in me.

References

[1] Sesia J. Zhao, Wagner, Chen Huaping, "Review of Prediction Market Research: Guidelines for Information Systems Research"
[2] Hellström, T., Holmström, K., "Predicting the Stock Market, Technical Report Series" IMATOM- 1997-07, (1998).
[3] Schoeneburg, E.(1990), "Stock Price Prediction Using Neural Networks: A Project Report", Neurocomputing, vol. 2, pp. 17-27.
[4] Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J., Lam, W., Daily prediction of major stock

indices from textual www data, in KDD, (1998), pp. 364–368.
[5] Wijnand Nuij, Viorel Milea, Frederik Hogenboom, "An Automated Framework for Incorporating News into Stock Trading Strategies" IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 4, april 2014
[6] Shubhangi S. Umbarkar, "Stock Market Prediction From Financial News: A survey," IJERGS vol. 06, ISSN 2091-2730
[7] W. IJntema, J. Sangers, F. Hogenboom, and F. Frasincar, "A Lexico-Semantic Pattern Language for Learning Ontology Instances From Text," J. Web Semantics: Science, Services and Agents on the World Wide Web, vol. 15, no. 1, pp. 37-50, 2012.
[8] M.-A. Mittermayer and G.F. Knolmayer, "Text Mining Systems for Market Response to News: A Survey," technical report, Institute of Information Systems University of Bern.
[9] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J.: Mining of Concurrent Text and Time Series. In: Proceedings 6th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining. Boston 2000, pp.37-44.
[10] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J.: Language Models for Financial News Recommendation. In: Proceedings 9th Int. Conference on Information and Knowledge Management. Washington 2000, pp. 389-396.
[11] Ogilvie, P.; Schmill, M.: Ænalyst - Electronic Analyst of Stock Behavior. Project Proposal 791m, Department of Computer Science, University of Massachusetts, Amherst.
[12] Seo, Y.; Giampapa, J.A.; Sycara, K.: Text Classification for Intelligent Portfolio Management. Technical Report CMU-RI-TR-02-14, Robotics Institute, Carnegie Mellon University, Pittsburgh.
[13] Seo, Y.; Giampapa, J.A.; Sycara, K.: Financial News Analysis for Intelligent Portfolio Management. Technical Report CMU-RI-TR-04-04, Robotics Institute, Carnegie Mellon University, Pittsburgh.
[14] F. Allen and R. Karjalainen. \Using Genetic Algorithms to Find Technical Trading Rules,"J. Economics, vol. 51, no. 2, pp. 245-271, 1999.
[15] Puspanjali Mohapatra, Alok Raj,Indian Stock Market Prediction Using Differential Evolutionary Neural Network Model International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 2 Issue 4 (July 2012)

Author Profile

Shubhangi S. Umbarkar. Received her B.E. degree in Computer science from university of Amravati in 2013. she is currently working toward the M.E. Degree in Computer Engineering from University of Pune. She has attended number of workshops on Research Methodology, Cyber Security, Latex, Scilab, Computer Vision etc. Also workshop on Image Processing, Computer Network, conducted by IIT, Bombay remote center at VPCOE, Baramati.