International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

Stock Market Prediction From Financial News: A survey

Shubhangi S. Umbarkar
PG Student
Department Of Computer Engineering
VPCOE,Baramati,Savitribi Phule University
Baramati,India

shubhangishriramumbarkar@gmail.com

Abstract: A number of investors on financial markets are growing day by day. Investors need to continuously observe financial news for market events to deciding buy and sell equities due to the upcoming news of market sensitivity. For taking information about financial market, investors always prefer news and financial markets are motivated by news information, which is comes from different media agencies through a various channels. Information is time-sensitive, especially in the circumstance of financial markets, selecting and processing all the relevant information in a decision-making process, such as whether to buy, hold, or sell shares is difficult task. There are various techniques used for prediction of stock market like Data mining, ontology learning, machine learning, artificial neural network (ANN), decision tree etc. Due to uncertainty nature of financial market investors are confuse to take the significant decision about investment. Stock market prediction is the ongoing research field in the data mining. This paper is about to discuss different techniques, challenges related to prediction of stock market.

Keyword- Stock market prediction, Data mining, artificial neural network (ANN), Descision making, machine learning, ontology learning

I.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. The content of news items reflect past, current, and upcoming world conditions, and thus news contains valuable information for various purposes. Being alert of ongoing marketplace situations is of paramount importance for investors and traders, who require to creating knowing decisions that could have an evidentiary impact on definite aspects specified as profits and marketplace perspective. However, due to the ever expanding of information, it is virtually impractical to keep evidence of all future applicable news in a regulated trend. It is need to do the automatically extracting news items by means of computers that would alleviate effort that are required for manually processing of news information.

Financial markets are motived by information. There are many sources of information. The most important source of information is news which are comes from different communication media through a various channels. The increasing number of information sources resulting in high volumes of news. Manually processing of such huge information is very tedious task. Information about financial markets is time sensitive. Selecting, processing of the relevant news information in decision-making process is challenging job. Data mining tools such as ViewerPro tool can be use for automatically extract the market event[1].

II. RELATED WORK

Jethro Borsje,Frederik Hogenboom and Flavius Frasincar[2] introduced lexico-semantic patterns and lexico-syntactic patterns methods for extraction of financial event from RSS news feeds. 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.

Advantages

- It achieved the accuracy.
- Most of the news item headlines are not associated with a buy event and thus the rule engine has successfully ignored a large portion of such news items in each run.
- It reduced the error rate of the event detector.

Limitations

- It is impossible to remove individuals or instances of properties from an ontology using SWRL.
- On the use of SWRL, there is incompleteness of good documentation.

Wouter IJntema, Jordy Sangers[3] makes the use of text rule based method that uses lexico semantic pattern for learning ontology instances from text that helps domain experts for maintaining ontology population process. Ontology is used for retrieving relevant news items in a semantically in efficient way. Authors used Hermes Information Extraction Language (HIEL) which apply the semantic concepts from ontology and used to evaluate for extracting events and relations from news. Hermes Information Extraction Engine (HIEE) also has implemented. The Hermes news portal (HNP) is a stand-alone application and java based tool, which gives an opportunity to use the various Semantic web technology. The overall framework of Hermes Information Extraction Engine (HIEE) is divided into two parts i.e. preprocessing stage and rule engine. In preprocessing few steps are implemented like tokenization, sentence splitting, and Part-Of-Speech (POS) tagging. Then the Hermes Information Extraction Rule Engine compiles the rule in rule compiler and matches the rule using rule matcher to the text after preprocessing the news information. They have showed that the lexico-semantic patterns are superior than lexico-syntactic patterns with respect to efficiency and effectively. Pattern-based information extraction techniques are mainly focused by authors.

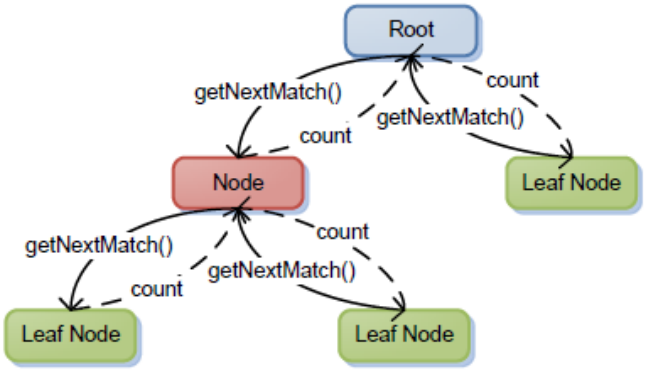

Figure1: Rule Tree

Advantages

- As they have mainly focused on pattern-based information extraction techniques that often require less training data and help users to gain more insight into why a certain relation found.
- Hermes Information Extraction Language (HIEL) enhanced the expressivity of the rules.
- Lexico-semantic rules requires significantly less time than creating equally performing lexico syntactic rules.

Limitations

- Lexico semantic rules exploit the inference capabilities of ontology.
- Ontology is not update automatically.

M.A. Mittermayer and G.F. Knolmayer[4] implemented several prototypes for predicting the short-term market reaction to news based on text mining techniques.

Prototype developed by Wthrich et al. Cho99[5], ChWZ99[6], WCLP98][7]. This prototype attempts to predict the 1-day trend of five major equity indices such as the Dow Jones, the Nikkei, the FTSE, the Hang Seng, and the Straits Times. According to a 3-category model the documents of this protocol were labeled. News articles followed by 1-day periods are fell into first (second) category which are associated with at least 0.5% increasing equity index. For trading sessions, the threshold is of +/- 0.5% was chosen so that one-third part of it roughly fell in each of the three categories. Prototype categorized all newly published articles during its operational phase.

From each category the numbers of news article were counted and depending on where the most news article were assign to the prototype triggered to buy or to sell or to do nothing recommendations for the corresponding index. Term Frequency times Inverse Document Frequency (TFxIDF) use as feature selection technique.

Advantages

- When prototype was tested then the result is 40% for Straits Times and 46.7% for FTSE was correct.

Limitations

- The simulated performance results of this prototype cannot be achieved in reality.

Prototype developed by Lavrenko et al. LSLO00a[8], LSLO00b[9]. The prototype Enalyst was developed around 2000 at the University of Massachusetts. The goal is to forecast stocks in very short-term i.e. intraday price trends of a subset by analyzing the homepage of YAHOO finance where news articles are published. In this prototypes 5-category model are as follows:

- Surge
- Slight
- No Recommendation
- Slight-
- Plunge

The author first segmented the stock price according to time series with a linear regression into small trend windows. If the news articles published in h hours before start of a trend window with price trend slop ≥ 0.75 were put in the surge category. Slight+ category is assign, if news article belongs to a slope between 0.5 to 0.75.and other category were assigned accordingly. Naive Bayes used to train the classifier. If an incoming news article was assigned to the categories Surge or Slight+ then the prototype triggered a buy recommendation. The short recommendation triggered if an articles are assigned to slight- and plunge category.

Advantages

- When the prototype was tested based on 10 minute stock price data between mid of March and April 2000 with investment of U.S. Dollar (USD) 10,000 in each roundtrip. Then after testing period of 40 days, by performing about 12,000 transactions, 280,000 USD is achieved.

Limitations

- The author included only those 127 U.S. stocks that showed largest positive or negative price. Such selection leads to bias towards highly volatile stock.
- This protocol is very unrealistic.

Prototype developed by Thomas et al. This prototype was developed at the Robotics Institute of Carnegie Mellon University between 2000 and 2004 SeGS02[10], SeGS04[11]. This prototype mainly focused on forecasting the volatility. The author developed strategy that, once news is published, market is temporarily exit for particular stock that may increase volatility. Then reentering decision is depends on technical indicators.

Advantages

- The result of strategy improved return.

Limitation

- Vital information regarding the simulation missing.

Prototype developed by Elkan/Gidfalvi .The aim of this prototype is to forecast stock price trends regarding the publication of a news article into windows of influence. This prototype divides the stock price time series. Foe example if the window is ranging from 0 to 20, it means that in 20 minutes most of the price adjustment occurs. The documents were labeled according to a 3-category model in the learning phase. The first and second category consists of news leading to a price increase or decrease at least with 0.2% during the window of influence. The remaining news fell into the third category. By using MI (Mutual Information) as selection criterion, feature definition was done automatically. The learning phase was finished by training a Naïve Bayes classifier. A virtual roundtrip is performed if the prototype sorts an incoming article into one of the first two categories. The asymmetric exit strategy was applied for triggering the market exit.

Advantages

- This prototype achieved a performance of 10 bps per roundtrip.

Prototype developed by Peramunetilleke/Wong.

The prototype developed at the University of New South Wales was developed in collaboration with currency traders from UBS around 2001 PeWo02[17]. It contains more than 400 features, each one consisting of two to five words that are combined with the logical operator AND. Also this dictionary has not been made accessible to the public. The authors create a rule-based classifier based on the three categories in the learning phase i.e. “dollar up>0.23%”, “dollar down >0.23%”, and “dollar steady” based on threshold 0.23%. Profit per roundtrip do not provide by the authors.

Advantages

- The system gives 50% right predictions.
- Decision speed is significantly improved.
- The achieves decision quality.

Limitations

- A random traders achieved only 33.3% right decision.

Prototype developed by Fung/Lam/Yu. This prototype was developed around 2002 at the Department of Systems Engineering and Engineering Management of the Chinese University of Hong Kong. The documents are labeled according to approach used in prototype developed by Lavrenko . The price time series are segmented in the first step, around the publication of a news article into time windows. Then the clustering algorithm they have used to divide the sample of time windows into the three most discriminating clusters that are “Rise” ,“Drop” based on steepest positive and negative average slope and the third one was called “Steady”. Instead of programming a prototype the authors used commercial text mining software. For example, the preprocessing of the news articles was performed by IBM's Intelligent Miner for Text and the Support Vector Machine (SVM) was used as classifier.

Advantages of the prototype systems

- None of the above prototypes considers any costs in the performance simulation.
- Systems covers the costs of immediate execution by achieving a gross profit of 10-15 bps.
- Investors are actively involved because of cost effectiveness of prototypes.

Limitations

- The performance studies neglect some important features of the financial markets like transaction costs, limited volume at given prices.
- Practically, the use of prototypes may increase the complexity of the system.

F. Allen, R. Karjalainen and Wijnand Nuij [1][12] used genetic programming to develop optimal trading rule. If the solution is represented in the decision tree or computer program then genetic programming used. Genetic programming uses the principles of parallel search, natural selection and historical data to search for candidate solutions to problems of interest. A computer randomly generates a population of candidate solutions expressible as decision trees to a problem of interest. The rules are required only to be well defined and to produce output appropriate to the problem of interest a buy/sell decision in stock market.

Advantages

- 1) Genetic programming is a multi-dimensional, non-differential, non-continuous, and even non-parametrical so it is useful in problems of decision tree.
- It solves problems with multiple solutions.
- It can solve every optimization problem which can be described with the chromosome encoding.

Limitations

- Genetic programming minimizes but does not eliminate the problem of data snooping by searching for optimal ex ante rules, rather than rules known to be used by traders.
- If generations of the genetic algorithm do not train long enough then the usability would be low.
- It required more time to give the output.

K. Senthamarai Kannan, P. Sailapathi Sekar[13] uses data mining technique for the prediction of stock market. Five methods were combined to predict.

A. Typical Price (TP)

By adding the high, low, and closing prices together, the Typical Price indicator is calculated and then dividing by three. The result is the average, or typical price.

Algorithm:

1. Inputting High, Low, Close values of the daily share

2. Take an output array and add the values of H, L and C

3. Divide the total by 3

$$TP = \left[\frac{H+L+C}{3}\right]$$

Where, H=High; L=Low; C=Close

B. Chaikin Money Flow indicator (CMI)

Chaikin's money flow is based on Chaikin's accumulation/distribution. If the stock closes above its midpoint [(high+low)/2] for the day, then there was accumulation that day, and if it closes below its midpoint, then there was distribution that day. By summing the values of accumulation/distribution for 13 periods and then dividing by the 13-period sum of the volume the CMI was calculated.

The Following formula was used to calculate CMI.

$$CMI=\left[\frac{\text{sum(AD,n)}}{\text{sum(VOL,n)}}\right] \quad AD = VOL\left[\frac{(CL-OP)}{\text{sum(HI-LO)}}\right]$$

AD stands for Accumulation Distribution, Where

n=Period; CL=today's close price; OP=today's open price;

HI=High Value; LO=Low value

C. Stochastic Momentum Index (SMI)

The Stochastic Momentum Index (SMI) is based on the Stochastic Oscillator. The range of SMI is from +100 to -100. The mid point was calculated as [(high+low)/2]. When the close is greater than the midpoint, the SMI is above zero, when the close is less than the midpoint, the SMI is below zero. A buy signal is generated when the SMI rises above -50, or when it crosses above the signal line. A sell signal is generated when the SMI falls below +50, or when it crosses below the signal line.

The Following formula was used to calculate SMI.

$$100\times\left[\frac{[MOV\left[MOV\left[C-\left[5\times[HHV(H,13)+LLV(L,13)]\right]\right],25,E\right],2,E]}{[5\times\left[MOV\left[MOV\left[[HHV(H,13)+LLV(L,13)]\right]\right],25,E\right],2,E]}\right]$$

Where HHV= Highest high value.

LLV = Lowest low value.

E = exponential moving avg.

Using the following formula, exponential moving average

was calculated.

$$EMA=\left[(Price(i)-prevMVG\times\left(\frac{2}{N+1}\right)\right]+prevMVG$$

D. Relative Strength Index

This indicator compares the number of days a stock finishes up with the number of days it finishes down. Usually between 9 and 15 days it is calculated. The RSI has a range between 0 and 100.

RSI=100–(100/1+RS); RS=AG/AL

AG=[(PAG)x13+CG]/14; AL=[(PAL)x13+CL]/14

PAG = Total of Gains during past 14 periods/14

PAL = Total of Losses during past 14 periods/14

Where AG=Average Gain, AL=Average Loss

PAG=Previous Average Gain, CG=Current Gain

PAL=Previous Average Loss, CL=Current Loss

The following algorithm was used to calculate RSI:

Upclose = 0

DownClose = 0

Repeat for nine consecutive days ending today

If (TC > YC)

UpClose = (Upclose + TC)

Else if (TC < YC)

DownClose = (Down Close + TC)

End if

$$RSI=100-\left[\frac{100}{\left(1+\left[\frac{upclose}{downclose}\right]\right)}\right]$$

E. Bollinger Bands

It is a technical indicator which creates two bands i.e. upper band and lower band around a moving average and are based on the standard deviation of the price. If the volatility is high then the bands will wide and when there is little volatility then the bands will narrow.

The Upper and Lower Bands are calculated as

$$stdDev = \sum_{i=1}^{N}(price(i) - MA(N))^2$$

$$Upperband = MA+D\sqrt{\sum_{i=1}^{N}\left[\frac{(price(i)-MA)^2}{N}\right]}$$

$$Lowerband = MA-D\sqrt{\sum_{i=1}^{N}\left[\frac{(price(i)-MA)^2}{N}\right]}$$

Advantages

- Using the above methods prediction was correct at least 50% of the time.

Limitations

- Above methods performed well on half of the stocks and not so well on the other half of the stocks.

Debashish Das and Mohammad Shorif Uddin [14] introduced data miming and neural network technique for prediction for stock market. Data analysis tools were used to predict future trends and behavior which helping organizations in active business solutions for knowledge driven decisions. Intelligent data analysis tools produce a database to search for hidden patterns, finding projecting information that may be missed due to beyond experts prediction.

Data mining technique steps are as follows.

1. Analysis of survey data
2. Explanatory simulation
3. Analytical modeling
4. Identify patterns and rules
5. Acquisition summary

Due to ability in dealing with fuzzy, uncertain and insufficient data which may fluctuate rapidly in very short period of time the neural network was used. In this computer units connected together such that each neuron can transmit and receive signals from each other. The framework of neural network is as follows.

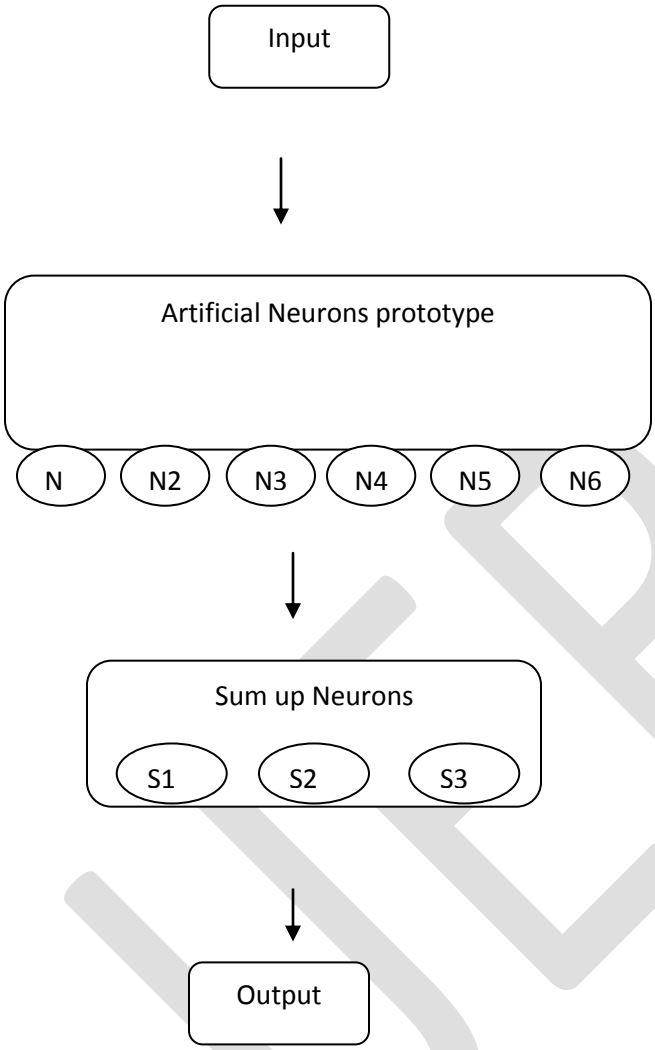

Figure2: Neural network model architecture

Advantages

- Neural network gives high computation speed.
- The neural network technique is fault tolerance.
- The data mining technique discover useful patterns from a dataset which are used for prediction of market.

Limitations

- Handling of time series data in neural networks is a very complicated.
- For the prediction of stock market with data mining technique requires high volumes of data for training.

III. CONCLUSION

In this paper we see the different technique for prediction of stock market. Stock market prediction is the activity of interpreting the future state of the share prices. It gives the guidance for proper investment. Most stock market prediction technique such as data mining technique, neural network technique, different technical trading indicators are depends on the financial news approaches and success of these technique is also depends on proper information extraction about stock market. Another approach such as stock market news extraction which uses different clssifire such as viewr pro tool, naïve bayes classifier, Term Frequency and Inverse Document Frequency for providing the market event model. In this paper, prototypes that used for prediction of stock market. also focused.

IV. ACKNOWLEDGMENT

I express great many thanks 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] Wijnand Nuij, Viorel Milea, Frederik Hogenboom, Flavius Frasincar, and Uzay Kaymak ,An Automated Framework for Incorporating News into Stock Trading Strategies. IEEE transactions on knowledge and data engineering, VOL. 26, NO. 4, APRIL 2014

[2] J. Borsje, F. Hogenboom, and F. Frasincar, Semi-Automatic Financial Events Discovery Based on Lexico-Semantic Patterns, Int'l J. Web Eng. and Technology, vol. 6, no. 2, pp. 115-140, 2010.

[3] 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..

[4] 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

[5] Cho, V.: Knowledge Discovery from Distributed and Textual Data. Dissertation Hong Kong University of Science and Technology. Hong Kong 1999.

[6] Cho, V.; Wüthrich, B.; Zhang, J.: Text Processing for Classification. In: Journal of Computational Intelligence in Finance 7 (1999) 2, pp. 6-22.

[7] Wüthrich, B.; Cho, V.; Leung, S.; Peramunetilleke, D.; Sankaran, K.: Zhang, J.; Lam, W.: Daily Prediction of Major Stock Indices from Textual WWW Data. In: Proceedings 4th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining. New York 1998, S. 364-368.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] F. Allen and R. Karjalainen, Using Genetic Algorithms to Find Technical Trading Rules, J. Economics, vol. 51, no. 2, pp. 245-271, 1999.

[13] K. Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam, Financial Stock Market Forecast using Data Mining Techniques, Preceding of the International MultiConferance of engineers and Computer Scientists 2010Vol I, IMECS 2010, March 17-19,2010, Hong kong

[14] Debashish Das and Mohammad Shorif Uddin, Data mining and Neural network techniques in stock market prediction: a methodological review, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.4, No.1, January 2013