Financial Trading System using Combination of Textual and Numerical Data

Shital N. Dange
Computer Science Department, Walchand Institute of Technology, Solapur, India

Rajesh V. Argiddi
Assistant Prof. Computer Science Department, Walchand Institute of Technology, Solapur, India

S. S. Apte
Professor and Head of Department Of Computer Science and Engineering, Walchand Institute of Technology, Solapur, India

ABSTRACT

There is large amount of financial data that are generated and evaluated at a high speed. These financial data is coming continuously, changing with time and may be unpredictable. Therefore there is a critical need for automated approaches to effective and efficient utilization of large amount of data to support companies and individuals for decision-making. Data mining techniques can be used to uncover hidden patterns, to discover the behavior of the stock market, to find out the trends in financial markets and so on. For predicting stock trends and making financial trading decisions, a new model is presented. It is based on combination of data and text mining techniques which takes the textual contents of time-stamped web documents along with numerical time series data and performs the future prediction. By using this model, we will show that the accuracy of result will be improved.

Keywords

Data mining, pre-processing, feature extraction.

1. INTRODUCTION

Over the past decades, many attempts have been made at understanding and predicting the future using data mining methods. Among them, to forecast the price movements in stock market and making the decision is considered as major challenge. However, most methods suffer from serious drawback and therefore results are hard to understand and producing inaccurate predictions. Therefore, predicting stock price movements is difficult. Data mining techniques are able to detect future trends and behaviors in financial markets.

The Efficient Market Hypothesis (EMH), as stated by Fama ([2], [3], [4]), assumes that stock prices fully reflect all their relevant information at any given point in time and that everyone has some degree of access to the information. In 'Random Walks ([4])' theory stock prediction is considered to be impossible, where stock prices are changes randomly. Most financial specialists try to use the time gap of the markets adjustments to new information for making their own predictions. They do this by combining both technical and fundamental analysis strategies. In technical analysis, it performs the prediction based on past price, while in fundamental analysis; it is based on real economy factors, such as trading volume, organizational changes in the company, etc [1]. Therefore stock market data or financial news articles can be used to get data required by these two strategies. The conventional approach to modeling stock market returns is to model the univariate time-series with autoregressive (AR) and moving average (MA) models. The autoregressive conditional heteroskedasticity (ARCH) class of models was originally introduced by (Engle, 1982) and has become a core part of empirical finance. Recently, Engle [9] and Bollereslev [10] provided a new very powerful tool for the modeling of financial data in general and stock market returns in particular. The new process suggested by Engle and Bollereslev [11] is different from earlier conventional time series models in that, instead of making the assumption that the variances are constant they allow the conditional variances to change over time as functions of past errors. These models are deterministic in the sense that they attempt to use mathematical equations to describe the process that generates the time-series. A disadvantage of these models lays in the assumption that trader or financial analyst needs to determine the appropriate number of lags and sometimes the successful analysis is based on the experience of analyzing the enormous variety of time series econometrical models. The advantage of these models lays in their ultimate interpretability.

Different similarity queries on time-series have been introduces ([12], [13]). Mining different queries from huge time-series data is one of the important issues for researchers. In useful data mining techniques like classification and clustering, to handle time-series data is one of the stimulating research issues. Given a set of cases with class labels as training set, classification is to build a model (classifier) to predict future data objects for which the class label is unknown. Classification is one kind of data mining technique to identify essential features of different classes based on a set of training data and then classify unseen instances into the appropriate classes. Decision trees [5] have been found very effective for classification of huge and frequently modifiable databases e.g., Stock Market, Shopping Mall etc**.** Decision trees are analytical tools used to discover rules and relationships by systematically breaking down and subdividing the information contained in data set.

Now a day's more and more important and commercially valuable information becomes available on the World Wide Web. Also financial services companies are making their products increasingly available on the Web. There are various types of financial information sources on the Web. Internet provides almost all possible information on the stock market worldwide through various useful websites as Google finance, Financial Times, Yahoo finance and many more. The reliability of the information depends on the reputation and the quality of the source sites. The news in the web pages is also related with the time in the publisher country.

All these source of information contain global and regional political and economic news, as well as recommendations from financial analysts. This is the kind of information that moves bond, stock and currency markets in Asia. This rich variety of information and news make it an attractive resource from which to mine knowledge. Techniques are presented enabling to predict the movements of major stock market indices from up-to date textual financial analysis and research

information. Therefore, exploiting textual information especially in addition to numeric time series data increases the quality of the input. Hence improved predictions are expected from this kind of input. There is a variety of prediction techniques used for stock market analysis. Statistical techniques and regression analysis [8] provides quantitative forecasts. Technical analysis helps to visualize and anticipate the future trend of the stock market. Technical analysis only makes use of quantifiable information. But there are also immeasurable factors such as "general political news" which largely affect the world's stock markets.

In this project, we apply data mining techniques on Indian Stock market. Because Indian Stock market is now well developed and with the larger data sets. Our work is based on finding movements in the stock markets and making the decisions. There are large numbers of attributes that can be used to classify texts in the news articles. This attributes are mainly the words that can represents positive, negative or neutral meanings to indicate the possibility of the direction of the stock movement. After having prepared both numerical and textual data and trends are assigned to stock's prices data mining algorithms like decision tree is applied for making future prediction and making market action recommendation.

2. METHODOLOGY

The block diagram of our system is shown below in fig [1]. It consists of following major steps:

1. Data Collection
2. Pre-processing
3. Generate Prediction Model

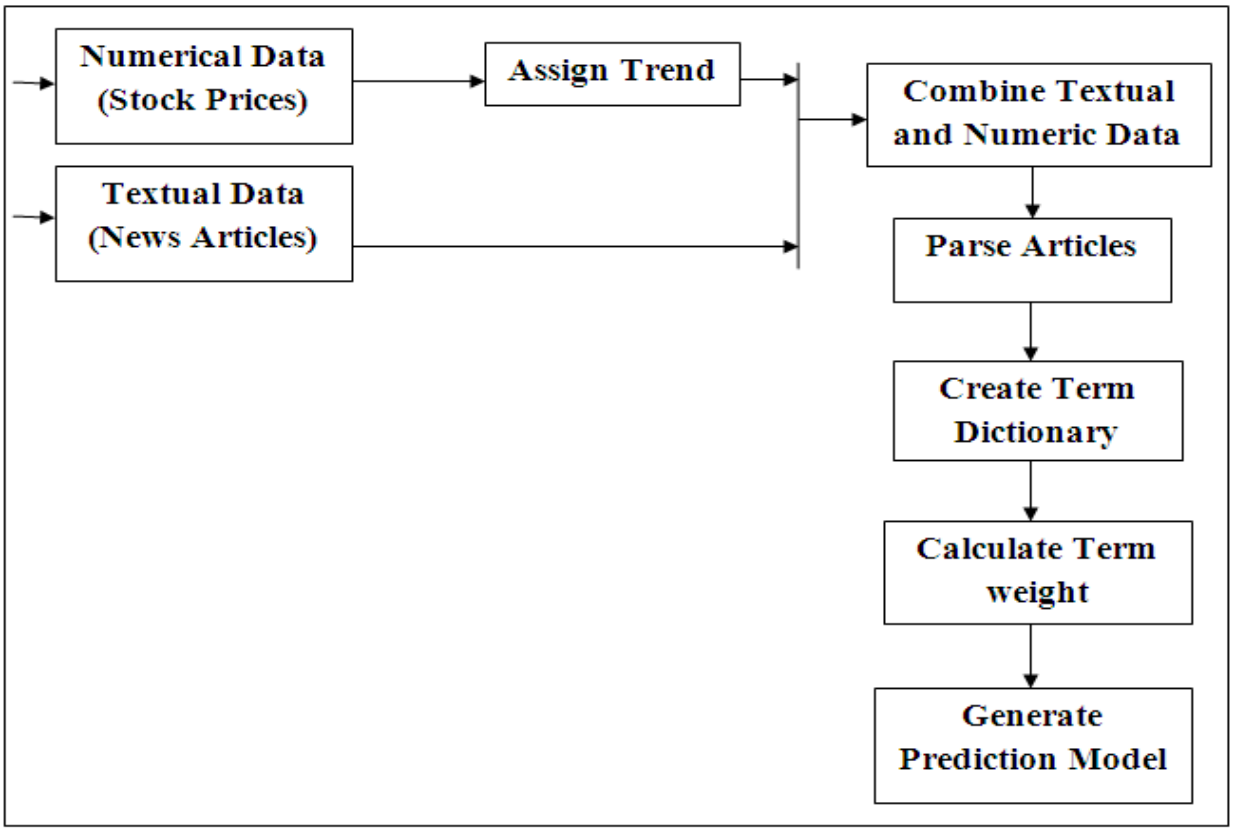

Figure 1. Block Diagram of Prediction Model with combination of Textual and Numeric Data

The detailed description of each step is given below:

2.1 Data Collection

Data Collection is the first step which includes numeric data as well as textual data. Numeric data is history of stock prices, while its related textual data is financial news articles. These data are collected from web sites like http://www.moneycontrol.com and http://www.nseindia.com. The data is easily available from above sites; therefore we have chosen these websites. The textual data is collected once a day, and for numeric data closing price of stock data is selected.

2.2 Pre-processing

Once data is collected, it is necessary to pre-process it. For numeric data, depending on its closing price we assign trend like up, down, expected, slight-up, and slight-down [1]. Example of Historical data of Reliance is shown in fig [2] below:

For textual data, retrieving particular term and then assigning weight to that term is done which is explained in detail below. After having collected both numeric data and textual data, and assigned trends to numeric data, we have to combine both data. Therefore stock prices and its related news article are grouped into single file which we will use for prediction.

2.2.1 Feature Extraction:

Feature extraction is one of the most important steps of our prediction model and it is used for textual data. It includes parsing of news articles and creation of term dictionary.

When news articles are collected, the first step is to extract key word and key phrases from predefined Window of Influence [1]. Depending on it, TermDictionary is automatically created.

We collect news articles and it is first parsed by using open source *HTML parser*. Html parser is a Java library used to parse HTML page.

Date	Open Pric	High Price	Low Price	Last Trade	Close Price	Total Trac	Turnover
23-May-12	685.2	689.95	679.55	689	686.85	2583348	17722.9
22-May-12	704.6	704.6	688.75	690	691.1	2938047	20449.93
21-May-12	686.5	700.7	686	697	695.5	2334127	16249.99
18-May-12	676.9	696.4	675	689	688.55	2363960	16223.61
17-May-12	680.25	689.85	677.8	686.8	685.15	3560488	24379.02
16-May-12	675	682.8	673.05	676.15	676.1	3101760	20996.93
15-May-12	679	688	674.1	682.3	681.65	3696346	25185.65
14-May-12	698	702.8	678.55	681	681.3	4081094	28132.72
11-May-12	693	703	688.3	698	697.2	3042510	21193.2
10-May-12	697.25	708.4	690.55	694.5	694.15	3391819	23738.84
09-May-12	702	708.3	691.05	697.6	695.1	3670660	25597.49
08-May-12	718	718.9	699	708.65	708.35	3885415	27512.14
07-May-12	715	722	705.1	715.5	715.7	4591050	32716.3
04-May-12	733.4	740	722	727	726.45	3778468	27606.71
03-May-12	738.5	744.45	735.15	740.75	738.85	1897162	14038.24
02-May-12	747.9	751.4	740	744	743.55	2861902	21310.84
30-Apr-12	737.5	747.9	737.5	745	745.1	2776770	20669.1

Figure 2. Example of Historical data of Reliance

The htmlparser attempts to balance opening tags with ending tags to present the structure of the page, htmllexer simply splits out nodes. nextNode(), nodeIterator(), elementAt(), etc methods are used to parse it. *Standford-postagger* is used to extract keywords. It is a piece of software which is used as Extractor. With the help of extractor, we analyze all collected web articles related to specific stock.

For each new training set, Term Dictionary is created. Term Dictionary is collection of words which we define as termCount. For each word, we calculate a score. But only K highest words will be used from term dictionary. Score is calculated by using following equation [1]

Score= (1/2) *(termFrequency/termCount) +1/2 (P/L * B/L)

Where:

termFrequency :the no. of occurrence of word in time frame

termCount: total no. of words in term Dictionary

P: number of days to the last occurrence of word

L: time frame

B: time window between the first and last occurrence of word

2.2.2 Calculate Term Weight:

In order to distinguish the importance of each feature, the weight of each feature has to be re-calculated. Therefore for each word in term dictionary, we need to assign weight whose value range from zero to one. This weight is calculated by dividing number of occurrence of word during particular period by the occurrence of that word in the term dictionary. [1]

2.3 Generate Prediction Model

After all preprocessing, classification algorithm of weka (Waikato Environment for Knowledge Analysis) is applied on it. Classification also called supervised learning, the training data are accompanied by labels indicating the class of the observations, and the new data (testing set) are classified based on the training set [7]. Decision tree induction is one of the methods using the classification approach. C4.5 algorithm which was developed by Quinlan [6] is used that generate associations between different features and different trend.

For this model, we have collected reliance stock data and its related news article. An arff file is created which is given to weka. Pre-processing is done on collected data as given in above steps. After preprocessing J48 algorithm is applied which gives results in terms of accuracy is shown below in figure3:

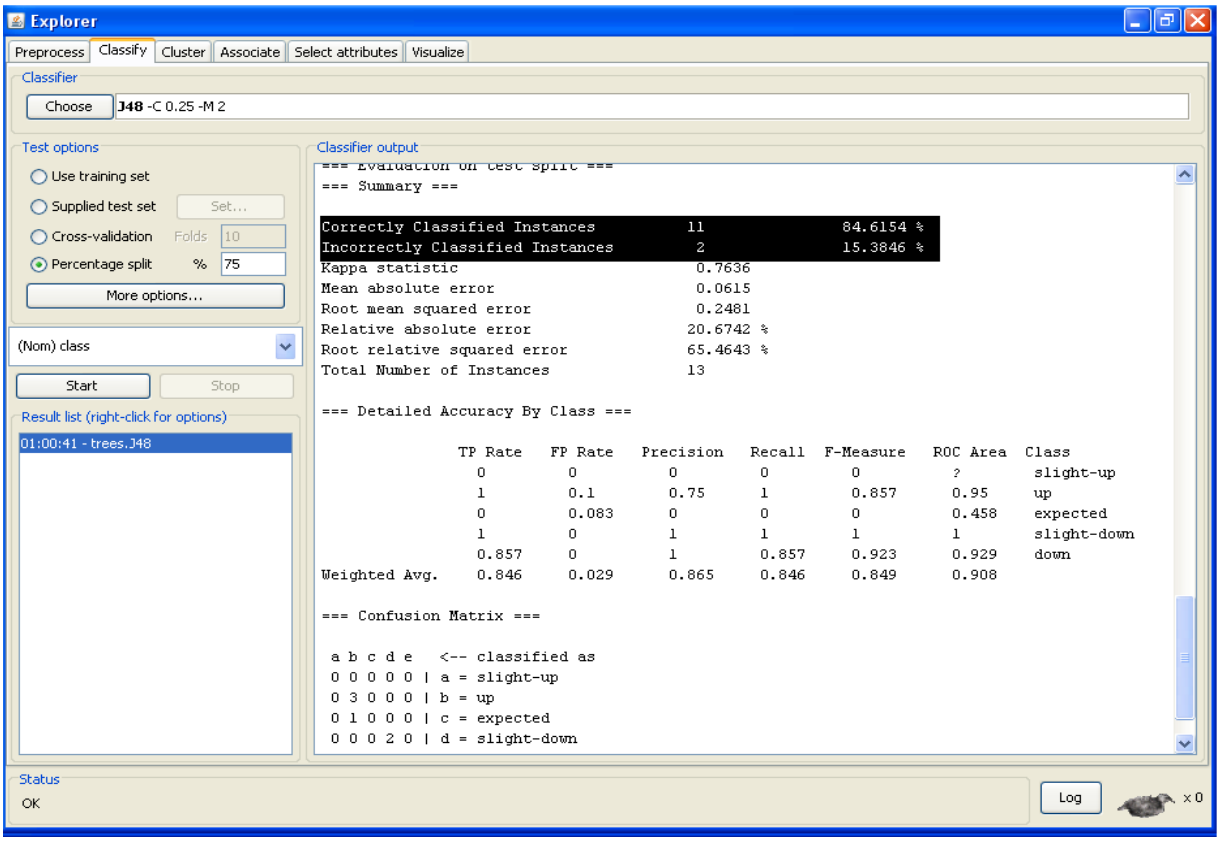

Figure 3.shows the prediction accuracy based on textual and numeric data.

There are 84.62% correctly classified instances and 15.38% incorrectly classified instances.

When we use only numeric data, it shows that there are only 64% correctly classified instances (see fig.4) .Therefore to improve the accuracy we have to combine numeric data along with textual data. From this we can say that, when we both textual and numeric data; the accuracy of result is improved.

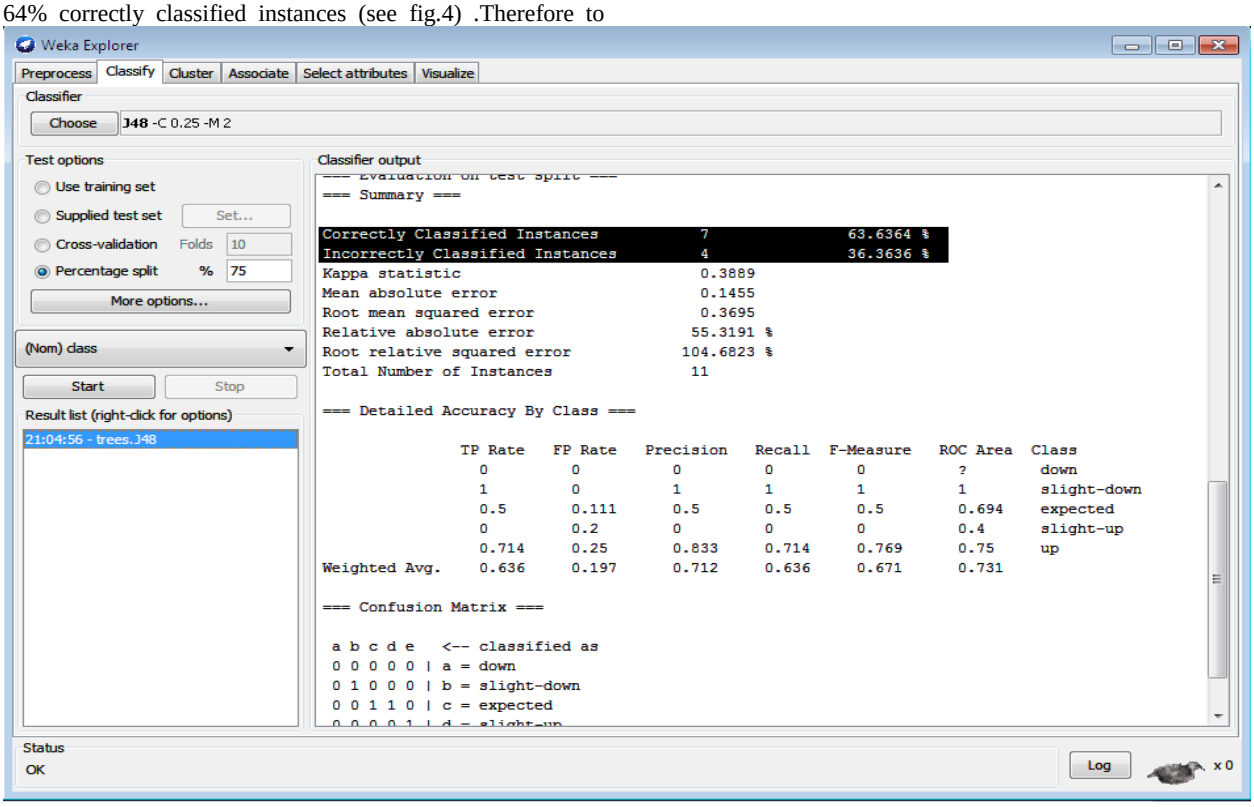

Figure 4.shows the prediction accuracy based on numeric data.

3. CONCLUSION

This paper presents a new model for predicting the market direction more accurately when stocks data and textual data are correlated to each other with a lag of K number of trading days. The main tasks such as data collection, pre-processing of the data are presented. And a decision tree induction algorithm is used so that an automated data mining process can be programmed. Thus we conclude that the results that produced using this model are more accurate. As mentioned, the accuracy of results hits 84%.

As more and more information becomes available on the Web, further research may include other input sources like different stock markets which can affects on prediction and prove to be of higher quality.

4. REFERENCES

[1] Gil Rachlin, Mark Last, Dima Alberg and Abraham Kandel, 2007. ADMIRAL: A Data Mining Based Financial Trading System, Proceeding of the 2007 IEEE Symposium on Computational Intelligence and Data Mining(CIDM 2007).

[2] E.F. Fama, 1970. Efficient Capital Markets: A Review of Theory and Empirical Work, Journal of Finance, 25 (May 1970): 383-417.

[3] E.F. Fama, 1991.Efficient Capital Markets: II, Journal of Finance, 46 (December 1991): 1575-1617.

[4] E.F. Fama, 1995 .Random Walks in Stock Market Prices, Financial Analysts Journal, September/ October 1965 (reprinted in January-February 1995).

[5] Quinlan, J. R. 1986. Induction of Decision Tree. *Machine Leaning*, Vol, pp.81-106.

[6] J.R. Quinlan, 1993 C4.5: Programs for Machine Learning, Morgan Kaufman Publishers Inc., San Francisco.

[7] Luk Chi Wa, Analyzing Stock Quotes using Data Mining Technique, The University of Hong Kong

[8] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, W. Lam, 1998.Daily Stock Market Forecast from Textual Web Data, In IEEE International Conference on Systems, Man. and Cybernetics, Volume: 3, Page(s): 2720 -2725.

[9] R.Engle and T. Bollerslev, 1986 .Modelling the Persistence in Conditional Variances, Econometric Reviews, 5, 81 -87.

[10] T. Bollerslev, 1986. Generalized Autoregressive Conditional Heteroscedasticity, Journal of Econometrics, 31, 307-327.

[11] R. Engle, 1982 .Autoregressive Conditional Hetroscedasticity with Estimates of the Variance of United Kingdom Infation, Econometrica, 50(4), 987-1007.

[12] Agrawal, R., Lin, K.-I., Sawhney, H.S., and Shim, K. (1995a). Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. *Proc. 21st Int'l Conf. Very Large Data Bases (VLDB '95)*, Sept. 1995, pp. 490±501

[13] Agrawal, R., Psaila, G., Wimmers, E.L., and Zait, M. (1995b).Querying Shapes of Histories, Proc. 21st Int'l Conf. Very Large Data Bases (VLDB '95), Sept. 1995, pp. 502±514.