IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 5)
Available online at: www.ijariit.com

Multi-classifier ensemble system with dynamic rule based algorithm for stock prediction

Sandeep Sharma
sandeepkalia9960@gmail.com
Rayat Bahra Group of Institutions, Patiala, Punjab

Gurpreet Kaur
gurpreetkhalsa5046@gmail.com
Rayat Bahra Group of Institutions, Patiala, Punjab

ABSTRACT

In monetary markets, it is mutually important and demanding to forecast the daily path of the stock market return. Among the some studies that focus on predicting daily stock market returns, the data mining measures utilized are either unfinished and not efficient, especially when a plethora of features are involved. This paper presents a whole and capable data mining process to anticipate the everyday direction of the S&P 500 Index ETF return based on 60 financial and economic features. To make it more accurate, I study many other techniques but SOP (Self Organizing Map) is a very proficient technique. To attain an accurate stock market prediction, the identification of the effective features is essential. In other words, the representative features of the factors play a key role in prediction efficiency. Technical and fundamental analyses are two indispensable tools in the financial market evaluation. Fundamental analysis can be used to estimate a firm's performance and financial status over a period of time by carefully analyzing the institute's financial statement. Technical analyses (TA), equally, evaluate securities by means of statistics such as past price and volume that are generated by market activities. The major analysis of TA is that it only reflects on the price movement and ignore the fundamental factors related to the company. The multiple classifier ensemble systems (MCS), a single type of machine learning technique, has newly become the focus of a new methodology for obtaining higher accuracy in predictions

Keywords— *Stock market, Machine learning, Self organizing map, Data mining*

1. INTRODUCTION

The future status of companies offering stocks is of great importance to stock market practitioners. According to the efficient market theory, it is impossible to predict prices based on historical stock data. This theory also states that the prediction of the classical criteria of risk and return cannot bring advantages to shareholders. There is abundant evidence in the literature, however, that argues against the efficient nature of the market. A precise prediction of companies' future financial status provides investors with the security to make a confident and profitable investment. To achieve an accurate stock market prediction, the identification of the effective features is crucial. In other words, the representative features of the factors play a key role in prediction efficiency. Technical and fundamental analyses are two essential tools in the financial market evaluation. Fundamental analysis can be used to evaluate a firm's performance and financial status over a period of time by carefully analyzing the institute's financial statement. Technical analysis (TA), equally evaluates securities by means of statistics such as past price and volume that are generated by market activities. The major criticism of TA is that it only considers the price movement and ignores the fundamental factors related to the company. Moreover, TA takes a comparatively

Short-term approach to analyzing the market. Fundamental analysis seeks to find the essential features of stock and market movements. In fact, the logic behind the fundamental analysis is that if a company has a proper fundamental strength, then long-term stock investment in the company will be more secure and stable. Thus, the stocks of these fundamentally strong companies, which are making money, gaining profit and growing their businesses, represent an opportunity for a successful investment. For this reason, in this paper, fundamental analysis is applied in order to determine the fundamental features that decide which company is a good bet for a secure investment. Stock return forecasting is a fascinating endeavor with a long history. From the standpoint of finance practitioners, asset allocation requires real-time forecasts of stock returns and an improved stock return forecast holds the promise of enhancing the investment performance. For an efficient investment, the return consideration is not sufficient. In fact, the risk and return must be considered simultaneously to create an accurate portfolio evaluation. In this paper, the prediction of stock return and risk are implied concurrently based on fundamental features in order to build a more comprehensive model for stock market analysis. Although the statistical approaches such as logistic regression and regression analysis are widely applied to forecast the return and risk of stocks, the results of machine learning approaches are generally superior in comparison to statistical methods. The multiple classifier ensemble systems (MCS), one type of machine learning technique, has recently become the focus of a new methodology for obtaining higher accuracy in predictions.

The rationale is that the optimization of a combination of relatively simpler predictors appears more convenient than optimizing the design of a single complex predictor. In fact, three fundamental issues are effective for establishing a successful MCS model: accuracy of individual classifiers, diversity among classifiers, and the choice of the fusion methods that will be used. The aim of this combination scheme is to gain increased precision with proper single classifiers and eliminate the uncorrelated individual classifier errors, which are the errors made by individual classifiers on various parts of input space [10].

2. VARIOUS TECHNIQUES OF STOCK PREDICTION

Chen Chen [12]has worked on exploiting social media for stock market prediction with Factorization Machine. Later although the financial news is planned to access market information, there are some demerits for news to predict the stock market. Recently when the micro-blogging service has developed to a popular social media and provides a plethora of real-time messages for numerous users, social media is proposed for the stock market prediction. In that case, the high-dimension of textual feature poses a major challenge. In this study, we propose an original kind of model, Factorization Machine (FM), to predict the trend of the stock market. FM not only improves the impact of high dimensionality but also captures some aspects of basic linguistics. In addition, we shed light on how textual representation influence the prediction and find that FM is constant, which is appropriate to other social media or prediction application normally

Sasan Barak [11] designs a fusion model for returns and risk prediction of stocks in financial market by applying various diversity methods in order to achieve more exact predictions for bearing in mind the simultaneous risk and return prediction of stocks for rising a base classifier selection procedure from candidate procedures by dataset clustering and considering the accuracy of combined classifiers. Developing a wrapper-GA scheme for feature selection and prediction and comparing it with the fusion method.

Xu Feifei and VladoKeelj have worked on Collective Sentiment Mining of Microblogs in 24-Hour Stock Price Movement Prediction. The authors have planned a method for collective sentiment analysis for stock market prediction and examine its ability to predict the change of a stock price for the next day. The planned method is a two-stage process which is based on the latest natural language processing and machine learning algorithms. Their estimation shows the best performance with the SVM approach in sentiment detection, with accuracy rates of 71.84/74.3% for positive and negative sentiment, respectively. The results of sentiment analysis are used in predicting stock price movement (up or down), and we found that users' activity on Stock Twits overnight positively correlate with stock trading on the next business day. The collective sentiments in after-hours have a powerful prediction on the transform of stock price for the next day in 9 out of 15 stocks studied by using the Granger Causality test.

SoujanyaPoria has proposed the use of Sentic patterns for the purpose of the sentiment analysis from the social data. The authors planned the use of dependency-based rules for concept-level sentiment analysis. In this work, the authors have introduced an original paradigm to concept-level sentiment analysis that merges linguistics, common-sense computing, and machine learning for improving the precision of tasks such as polarity detection.

Yassine, Mohamed has worked on the development of a framework for emotion mining from the text in online social networks. This paper presents a new perspective for studying friendship associations and emotions expressed in online social networks where it deals with the nature of these sites and the nature of the language used. It considers Lebanese Facebook users as a case study. The technique adopted is unverified it mostly uses the k-means clustering algorithm.

Vivek Narayanan has worked on a fast and accurate sentiment classification using an enhanced Naive Bayes model. The authors have explored different methods of improving the accuracy of a Naive Bayes classifier for sentiment analysis. They have also experiential that a combination of methods like effective negation handling, word n-grams and feature selection by mutual information results in a considerable improvement in accuracy.

Cambria, Erik et al. [17] has implemented the semantic multidimensional scaling for open-domain sentiment analysis. In this work, the largest existing taxonomy of common knowledge is merged with a natural-language-based semantic network of common-sense knowledge, and multi-dimensional scaling is applied on the resulting knowledge base for open-domain opinion mining and sentiment analysis.

Gun Woodard states that Social Media is at the heart of our communications and are among the most visited places on the Web. Social network services like Facebook, Friendster, Myspace, and Orkut have established themselves as very popular and powerful tools for making and finding friends for identifying other people who have similar interest. Search behavior of Web users reflects the interest of Web users and this also leads to similar profiles. Some research has also been carried so as to identify people who are highly associated with the similar interest of Web search.

Federico also stated that despite much progress in Natural Language Processing (NLP), this field is still a long way from a full Natural Language Understanding (NLU). Basically, NLU requires processing and knowledge that goes beyond parsing and lexical lookup and that is not explicitly conveyed by linguistic elements. Ambiguities and omissions are considered to be a common aspect of human communication. Nowadays the Web sources are more accessible and valuable than ever before, most of the times the truly valuable information is hidden in thousands of textual pages. The transformation into information is therefore strongly linked with their automatic lexical analysis and semantic synthesis.

Baumer et al. also mentioned that social networking Websites create new ways for engaging people belonging to different communities and regions. Social networks facilitate their users to communicate with people exhibiting different moral and social values. These Websites provide a very powerful medium for communication among individuals that leads to mutual learning and sharing of valuable knowledge. The most popular social networking Websites are those that allow people to communicate with each other by joining different communities and groups. Social networking can solve harmonization problems among people that may occur due to geographical distance.

Haghighi, Yazdi et al classifier combination design, it is believed that the success of combinations not only depends on the individual classifier's suitability but also on diversity being inherent among them. In fact, classifiers that are strong in different areas are supposed to be diverse. The entire point of fusing multiple classifiers is to balance the weaknesses of the individual classifiers. This balancing requires classifiers that make errors in different areas of the decision space. Diversity creation methods are generally categorized as explicit and implicit methods.

2012. Nasira, G. M. et. al. [27] has worked on the data mining models for the prediction of the stock prices by classifying the entities related to the certain stock for the prediction of the price. Price prediction helps the farmers and also Government to make an effective decision. Based on the complexity of stock price prediction, making use of the characteristics of neural networks such as self-adapt, self-study, and high fault tolerance, to build up the model of back propagation neural network to predict stock price.

2013. Luo, Chang Shou et. al. [26] has applied the developed solution called SARIMA for the purpose of stock price prediction and analyzed the system on the cucumber prices. The price of stocks is difficult to predict. In order to discover a useful method, this paper fully considers the seasonal variations and uses the seasonal autoregressive integrated moving average model （SARIMA） to calculate the cucumber price

2014. Kaur, Manpreet et. al. [28] has worked on the development of the data mining approach for the price prediction of agriculture crops. In this paper, we will argue about the applications and techniques of Data mining in agriculture. There are different data mining techniques such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) which are used for very latest applications of Data Mining techniques. This paper will consider the difficulty of price prediction of crops. Price Prediction, nowadays, has become a very significant agricultural problem which is to be solved only based on the existing data.

2014. Li, Youzhu et. al. [29] has developed the neural network in the hybrid manner for the short-term prediction of the stock prices. The authors have utilized the HP filter model along with the neural network for the prediction of the stock prices in the trading markets. The linear forecasting model cannot handle nonlinear relationships, while the neural network model alone is not able to deal with both linear and nonlinear patterns at the same time. The linear Hodrick-Prescott (H-P) filter can remove the trend and cyclical components from time series data. We forecast the linear and nonlinear patterns and then combine the two parts linearly to produce a forecast from the original data. This study proposes a structure of a hybrid neural network based on an H-P filter that learns the trend and seasonal patterns separately.

2016. Yoo, Do-il et. al. [24] has worked towards the prediction of the prices based upon the web-based a typical solution. By introducing atypical indexes into the Bayesian structural time series models, we could see that prediction power for stock prices are improved. In other words, it can provide better performances in predicting prices to combine recent Big-Data generated atypical web-search data.

2015. Yang, Lei et. al. [25] has developed the new algorithm called GEP solution for the stock price prediction. In this paper, a new Gene Expression Programming (GEP) algorithm is proposed, which increase “inverted series” and “extract” operator. The new algorithm can well boost the rate of utilization of genes, with convergence speed and solution precision is higher. Taking the Chinese stock price change trend of mooli, scallion as an example, and confer the way to solve the forecasting modeling problem by adopting GEP.

3. RESEARCH GAPS

- The maximum overall accuracy of the existing fusion model of multi-classifier ensemble system (MCS) has been recorded at 88.2%, which can be further improved by optimizing the parameters of classification algorithms using the improved selection of multi-classifier ensemble system to improve the prediction results.
- The existing Fusion of MCS model does not utilize any of the textual information, which primarily involves the news data, social media data, etc. In order to improve the overall accuracy, the proposed model can be improved further using technical features and textual information, in addition to fundamentals features, in order to use more comprehensive features and be able to predict the short term situations of stocks.

4. PROBLEM FORMULATION

The existing model is based upon the multi-classifier ensemble [11], where the classification algorithms of Bagging, Boosting and Adaboost are utilized, for the purpose of stock price prediction. The hybrid classification model in the existing model depends upon the ensemble data without the descriptive analysis of the input data using any of the feature descriptor, which does not let the existing model reduce the effect of independent variables for the high accuracy models.

The use of robust and flexible feature descriptor, such as Self Organized Maps (SOM), for a description of the features from the input training and testing data, can further recover the overall performance of the proposed model. The multiple classification algorithms of Bagging, Naïve Bayes and Support vector machine (SVM) will be used in the proposed multi-classifier ensemble. For the purpose of result fusion by the multi-classifier ensemble, the dynamic rule-based greedy algorithm (DRGA) can be used for the preparation of final classification results. Because we can’t represent multiple price predictions for a stock or stock market index as the final result, hence we need to apply feature amalgamation or fitness based selection for the final prediction representative. The DRGA is applied upon the multiple results obtained from the multi-classifier ensemble. The DRGA algorithm will be designed for the amalgamation by selection of the prospective results of the multi-classifier ensemble model, which will represent the final figure for the estimation or prediction of the stock prices. The proposed model would also incorporate the business news analysis in order to predict the stock market trends in the short-term, which will be processed using the N-gram analysis based feature extraction, polarity weights based classification. As per our study and assessments, the customization of the existing approach for the prediction of risk and return in a particular industry or investigating the accuracy of the procedure using data from other popular stock markets, such as the Indian stock markets (Bombay Stock Exchange (BSE) or National Stock Exchange (NSE)), which may result in new dimensions for this procedure.

5. METHODOLOGY

At the first stage, a detailed literature study would be conducted on the stock prediction algorithm and data classification methods; and to know their advantages and disadvantages. Literature study will lead us toward refining the structure of the proposed security solution design to overcome the shortcomings of the existing schemes while keeping their advantages intact in order to build a robust system. Afterward, the proposed solution will be implemented in PYTHON simulator with all essential input and output parameters. Then the implementation will undergo a thorough performance analysis and detailed comparison with the existing models. The input data acquisition is done on the historical data of the stock market for the last 2-10 years. The historical data of stock markets contains the readings of the stock market listings on all of the days of trading, which contains the starting and closing value of each day accounted in the historical data. During the implementation, the proposed model would be designed using the self-organized maps (SOM) with support vector machine (SVM) for the purpose of historical data processing, which is generally utilized for the long-term prediction. After the classification, the intensity of the individual stocks is inquired and calculated, which prepares and classifies the stock trend and stock price handling. The Self Organized Maps (SOM) is the operation performed on the given data stream obtained after the stock market data analysis. The structural equation modeling is the stock to flexibility and variability relationship and the stock price and its relation building, which predicts the stock trend and its expected price. The short-term predictive analysis is usually done on the basis of the business news analysis using the natural language processing (specifically textual processing models) for the purpose of news polarity analysis. The news polarity analysis gives the short-term trends for the particular stocks.

6. PERFORMANCE ASSESSMENT PARAMETERS

- Precision
- Recall
- Overall accuracy
- F1-error

7. CONCLUSION AND FUTURE SCOPE

The stock markets play the vital role in the national and international economies and are volatile in nature. The stock markets work on the basis of the market indices, which are presented as the overall presenter for the variations in the prices and sales based details of the stocks enlisted on the stock markets. The change in the indices or the specific stocks in the markets can be predicted in the long-term and short-term domains. The long-term domain based predictions are usually based upon the historical data assessment, which gives the nearly concrete results, which works on the basis of quarter over quarter (QOQ), Year over a year (YOY) etc. like analytical paradigms. The short-term assessment is generally made on the basis of the current business or other news related to the stocks, sectors or the other related issues, which carries the influential capacity in the price variations in the unlisted stocks. In the proposed model, we are working on the prediction of the stocks on the basis of historical data and market news.

8. REFERENCES

[1] R.Cervello-Royo, F. Guijarro and K. Michniuk, "Stock market trading rule based on pattern recognition and technical analysis: forecasting DJIA index with intraday data," Expert System Application, vol. 42, 2015, pp. 5963-5975.

[2] C.-F. Huang, "A hybrid stock selection model using genetic algorithms and support vector regression," Appl. Soft Comput., vol. 12, 2012, pp. 807-818.

[3] S. Barak, J. H. Dahooie and T. Tichy, "Wrapper ANFIS-ICA method to do stock market timing and feature selection on the basis of Japanese candlestick," Expert System Application, vol. 42, 2015, pp. 9221-9235.

[4] D.E Rapach and G. Zhou, "Forecasting Stock returns," Handbook of Economic Forecasting, vol. 2, 2013, pp. 328-383.

[5] S. Barak, M. Abessi and M. Modarres, "Fuzzy turnover rate chance constraints portfolio model," European Journal of Operational Research, vol. 228, 2013, pp. 141-147.

[6] S. H. Cheng, "A hybrid predicting stock return model based on Bayesian network and decision tree," in Modern Advances in Applied Intelligence, Springer International Publishing, Taiwan, 2014.

[7] S.H. Cheng, "A hybrid predicting stock return model based on logistic stepwise regression and CART algorithm," in Intelligent Information and Database Systems, Springer International Publishing, Taiwan, 2015.

[8] S. Barak and M. Modarres, "Developing an approach to evaluate stocks by forecasting effective features with data mining methods," Expert System Application, vol. 42, 2015, pp. 1325-1339.

[9] S. Barak and S.S. Sadegh, "Forecasting energy consumption using ensemble ARIMA-ANFIS algorithm," International Journal of Electrical Power Energy System, vol. 82, 2016, pp. 92-104.

[10] C.-F. Tsai, Y.-F. Hsu and D.C. Yen, "A comparative study of classifier ensembles for bankruptcy prediction," Appl. Soft Computing, vol. 24, 2014, pp. 977-984.

[11] Sasan Barak, Azadeh Arjmand and Sergio Ortobelli, "Fusion of multiple diverse predictors in the stock market," Springer, vol. 36, 2016, pp. 90-102.

[12] Chen Chen, Wu Dongxing, Hou Chunyan and Yuan Xiajie, "Exploiting Social Media For Stock Market Prediction With Factorization Machine," in Web Intelligence (WT) and Intelligent Agent Technologies (IAT), Poland, 2014.

[13] Xu Feifei and Vlado Keelj, "Collective Sentiment Mining of Microblogs in 24-Hour Stock Price Movement Prediction," in Business Informatics (CBI), Geneva, Switzerland, 2014.

[14] Soujanya Poria, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," Knowledge-Based Systems, vol. 69, 2014, pp. 45-63.

[15] Yassine, Mohamed and Hazem Hajj, "A Framework for Emotion Mining from Text in Online Social Networks," in Data Mining Workshops (ICDMW), Sydney, Austraila, 2010.

[16] Vivek Narayanan, Ishan Arora, and Arjun Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," Springer Berlin Heidelberg, vol. 8206, 2013, pp. 194-201.

[17] Cambria, Erik, Yangqiu Song, Haixun Wang, and Newton Howard, "Semantic multidimensional scaling for open-domain sentiment analysis," IEEE Intelligent Systems, vol. 29, no. 2, 2014, pp. 44-51.

[18] Gun Woopark, SooJin Lee and SangHoon Lee, "To Enhance Web Search Based on Topic Sensitive_Social Relationship Ranking Algorithm in Social Networks," in IEEE/WIC/ACM Web Intelligence and Intelligent Agent Technology, Milan, Italy, 2009.

[19] Federico Neri, Paolo Geraci and Furio Camillo, "Monitoring the Web Sentiment, The Italian Prime Minister's Case," in IEEE Advances in Social Network Analysis and mining, Denmark, 2010.

[20] Baumer E. P. S., Sinclair J and Tomlinson B., "America is like Metamucil: fostering critical and creative thinking about metaphor in political blogs," in Human Factor In Computing Systems ACM, Atlanta, USA, 2010.

[21] M. S, Haghighi, A. Vahedian and H. S. Yazdi, "Creating and measuring diversity in multiple classifier systems using support vector data description," Appl. Soft Computing, vol. 11, 2011, pp. 4931-4942.

[22] Yoo, Do-il. "Vegetable Price Prediction Using Atypical Web-Search Data." In 2016 Annual Meeting, July 31-August 2, 2016, Boston, Massachusetts, no. 236211. Agricultural and Applied Economics Association, 2016.

[23] Yang, Lei, Kangshun Li, Wensheng Zhang, and Yalong Kong. "A New GEP Algorithm and Its Applications in Vegetable Price Forecasting Modeling Problems." In International Symposium on Intelligence Computation and Applications, pp. 139-149. Springer Singapore, 2015.

[24] Luo, Chang Shou, Li Ying Zhou, and Qing Feng Wei. "Application of SARIMA model in cucumber price forecast." In Applied Mechanics and Materials, vol. 373, pp. 1686-1690. Trans Tech Publications, 2013.

[25] Nasira, G. M., and N. Hemageetha. "Vegetable price prediction using data mining classification technique." In Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on, pp. 99-102. IEEE, 2012.

[26] Kaur, Manpreet, Heena Gulati, and Harish Kundra. "Data Mining in Agriculture on Crop Price Prediction: Techniques and Applications." International Journal of Computer Applications 99, no. 12 (2014): 1-3.

[27] Li, Youzhu, Chongguang Li, and Mingyang Zheng. "A Hybrid Neural Network and HP Filter Model for Short-Term Vegetable Price Forecasting." Mathematical Problems in Engineering 2014 (2014).