International Journal of Information Science and Computing
Citation: *IJISC*: **6**(1): 09-17, June 2019

INFORMATION SCIENCE AND TECHNOLOGY FOUNDATION

Forecasting Share Prices Using Soft Computing Techniques

G. Vignesh and Ashwathy K. Cherian

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

ABSTRACT

For a long time, there has been the trend of trading of shares. Brokerage firms and dealers buy/sell stocks for clients and companies. Their work is based on knowing how the share price of the company will react in the market. Market/ share price predictions are useful as the investor/broker can attempt to predict the output in order to maximize his dividends or minimize his losses. Using data mining techniques, an attempt is made to estimate a prediction model to help forecast share prices. R and Python will be the tools used to sort, segregate and process the data, and techniques/algorithms such as Genetic Algorithm, ARIMA, Artificial Neural Networks, Linear Regression etc. will be used to forecast results of data. Along with the model data, external factors affecting share prices will also be taken into account. For each of the applied algorithms, their results will be compared and the difference in output with the real time values will be observed and recorded.

Keywords: Ddata mining, genetic algorithms, ARIMA, Artificial Neural Networks, linear regression

The advent of technological advancements and computing methods, has enabled us to perform high level calculations for various problems. Predictive analysis is one area where it has been extensively used. Behaviour of financial markets is one such area where predictive analysis is being extensively applied and has gained monumental prevalence across the world. Selection of shares to invest in is one of the first problems that investors encounter when entering the financial market.

Because large amount of historical data is present, it is possible to use data mining techniques to find patterns in stock prices, which can be used to predict the future trends of the stock.

During the past twenty years, systematic machine trading with artificial intelligence adaptive software has been developed to predict market movements by constructing complex models and sometimes non-linear models. In spite of all recent breakthroughs, predicting market behaviour is a challenging task

due to its complex, dynamic and non-linear nature. The ability to forecast the direction of a share price or an index is very important for various purposes. A few of them being a potential reduction in the risk of investment for the investors and supporting in identifying opportunities for speculators seeking to make profits by investing in stock indexes. Analysis can be primarily done in 2 ways – fundamental and technical. Technical analysis focuses and looks at the variations in the price of the stock. Whereas the fundamental analysis of the market tries to analyse the share market by looking at the cardinal components like the company's news articles, opinion of the analysts, reviewers etc. Fundamental analysis does not seem to be trustworthy because the decision that are made by this way may not have any kind of scientific reasons. The prediction model should not only focus on the prediction accuracy, but the prediction speed as well. Higher accuracy can help people make better decisions and the fast speed of the model helps in simulating more results in a given time.

Various methods/models have been proposed to forecast the market movement with the utilization of ANN, probabilistic models, SVMs, Genetic Algorithm (GA) and other soft computing techniques. In this literature, an attempt will be made to analyse the different techniques from different domains such as ANN, SVM, GA and others. For modelling the system, data from NSE, BSE, Yahoo! Finance or Google Finance may be used, which is then input into the models' algorithms. Apart from the statistical models, the external factors which affect the market behaviour will also be taken into account. The results are then compared and the most accurate and efficient model is found.

Existing Models

Prediction of market behaviour is an arduous task and selecting a suitable model is never an easy task. Emergence of technologies such as big data, high performance computing, etc. have only allowed researchers to create models which yields better results.

Genetic Algorithm is a suitable method that has been applied in different tasks. Often applied to solve global optimization problems, it can also be used to evaluate fitness of a model and help in parameter selection. Selection, crossover and mutation are the three major genetic operations. Assuming that the share market is an evolving system capable of redefining rules, a genetic approach will be well suited to handle the dynamic nature of the problem and predict accurate results. On the basis of the research carried out in[3] which compared the performance of genetic algorithm to a Greedy model and manual calculation, the GA based approach was the more consistent one as compared to the others. However, they only used a single rule to model the GA based model. According to previous researches, the approach and rules selected for any learning model has the capacity to effect the output of the predictive model. This was carried out in[2],[6], and [9], but not on GA approach.

Artificial Neural Network (ANN) is a model based on the structure of the brain. They are composed of neurons which imitate the behaviour of the brain. These nodes are linked to each other and perform various operations based on data they receive. Each link is assigned a weight. The ability of the ANN to learn is based off the its ability to alter the weights. They are of two types – Feed Forward and FeedbackANNs require large datasets to operate and with the abundance of data present, they are suited for market prediction. An ANN based model was implemented by Hamed, Hussien and Tolba[3], in order to overcome the lack of accuracy and the generalization of the models which existed before. Based on their research, it was concluded that their Kullback Liebler Divergence (KLD) algorithm when applied to

an ANN model performed significantly better as compared to models based on multi-layered perceptron, general feed-forward, time delayed neural networks, and others.

ARIMA (Autoregressive Integrated Moving Average) model is a statistical model. It is stated that prediction can be performed in two ways which includes Artificial intelligence and statistical techniques. It is known for its flexibility and efficiency in the prediction of share markets for a shorter time limit when compared to the artificial neural network techniques, as ARIMA is a time series based model, as compared to neural network based models. They are used on a large scale in financial, economic based and other fields. It must be noted that ARIMA model is only appropriate for a time series that is stationary, i.e. it does not have large change in its values at mediated time intervals.

Triple Exponential Smoothing also known as the *Holt-Winters* method, is one of the many methods or algorithms that can be used to forecast data points in a series, provided that the series is seasonal (repetitive over a some period). HW can said to be special case of ARIMA, but with only three variables. ARIMA has more parameters.

Support Vector Machines (SVM) is supervised training model to classify training data. The aim is to identify the hyper-plane and maximize the margin of the training data by optimally separating the hyper-plane. Support vectors closest to the hyper-plane and at equal distances are identified from both the classes. Value of support vector is evaluated by assigning weights to be learned from the SVM model. Maini and Govinda[12] perform a comparative analysis of SVM and Random Forest models for 1-gram and 2-gram text analysis. They conclude that SVM performs better than Random Forest, but only when the data is in a time series format. They also use a linear and non-linear SVM model and perform a comparison. The non-linear model has a better accuracy as compared to the linear model.

Random Forest is another algorithm which can be used for predicting both short and long term share prices. This algorithm has more accuracy as compared to other decision trees as this is the most advanced decision tree.

Social Media/Sentiment Analysis is another method through which models can be evaluated but they are not useful for obtaining long term results. However, using them as external factors in order to simulate the output from models can be done. Khatri and Srivastava[8] create a predictive model based on sentiment analysis and use it to predict impact of sentiment on prices of various companies. Their result however, was a mean squared average of the different companies they tested on. Thus for individual companies, their difference in expected vs. real values was slightly larger. Rajput and Bobde[12] created a hybrid model based on sentiment analysis and compared the results with a clustering based model. The former performed better than the clustering model, while also leaving scope for improvement by adding more features to the sentiment based hybrid model.

In summary, there are various models which exist but all of them vary in performance and output with the reason being that every model has a different set of parameters that it operates on. Other than this, the nature of the data is another factor which affects the performance of the models, as seasonal data might have higher results in a time series based algorithm and non-seasonal data might perform better in others.

Proposed System

The many existing models do not take into account, the various external factors which affect the financial market. These factors may be social events or political events. There is scope for improving the existing

models by taking into account the factors which might affect the market behavior and integrate the presently known techniques in order to create a better model as compared to the existing technologies.

Thus, a new approach to existing models can be taken, where in by making certain changes the accuracy of existing systems can be improved even further. The model will be built on platforms such as R and Python, along with various externals packages. The aim is to help the user to understand which shares he may be able to invest in based on the predictions of the model, and for an existing investor whether he should buy/sell the current accumulation of shares he has.

Datasets for training and testing the model may be taken from NSE, BSE, Yahoo! Finance or Google Finance. The output will be measured for the years remaining years. The attempt will be to generate outputs on various models and compare the results.

In the proposed system, the aim is to perform predictive analysis of market behavior using statistical and computing methods using tools such as R and Python. The model will be created using genetic algorithm (GA) and the results will be compared with the other available models. The task of share market analysis requires considerable financial acumen, in order to understand how the market works. Creating a predictive model involves certain prior knowledge in computing and statistics.

Genetic Algorithm is heuristic optimization inspired by natural evolution. They can help in optimizing the performance of predictive model by selecting the most relevant features.

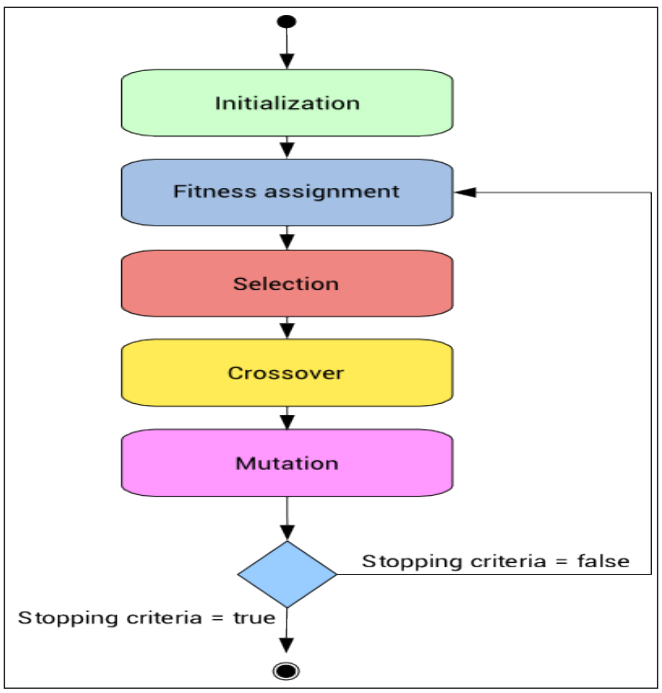

Fig. 1: Process model for generating high diversity children based on input population.

The first step is to initialize the individuals in the population. As it is a stochastic optimization method, the genes of the individuals are initialized at random. Genes are nothing but the features of the population set. The next step is to assess the fitness of the individual. To evaluate fitness, a predictive model with

training data needs to be trained and evaluate its selection error with the selection data. Selection error is nothing but the error of the model on an independent data set not used to create the model. Most used method for fitness assignment is rank based. The individuals are then ranked as per their selection error.

The rank of the individual is multiplied with *k* also called *selective pressure* in order to obtain the fitness value. Individuals with low fitness values will be discarded.

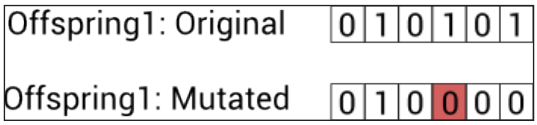

Fig. 2: Mutating the features in an offspring to generate new population

After fitness selection, the individuals need to be chosen that will recombine for the next generation. Only a small group is chosen for recombination, and they are often selected randomly. The crossover function recombines to create new individuals until the population size is the same as the old population. Sometimes, crossover might generate children with the same features as that of the parents, which might cause low diversity. In order to avoid this, mutations are carried out on the offspring. They are done by randomly changing the feature of an offspring by flipping the feature of that individual.

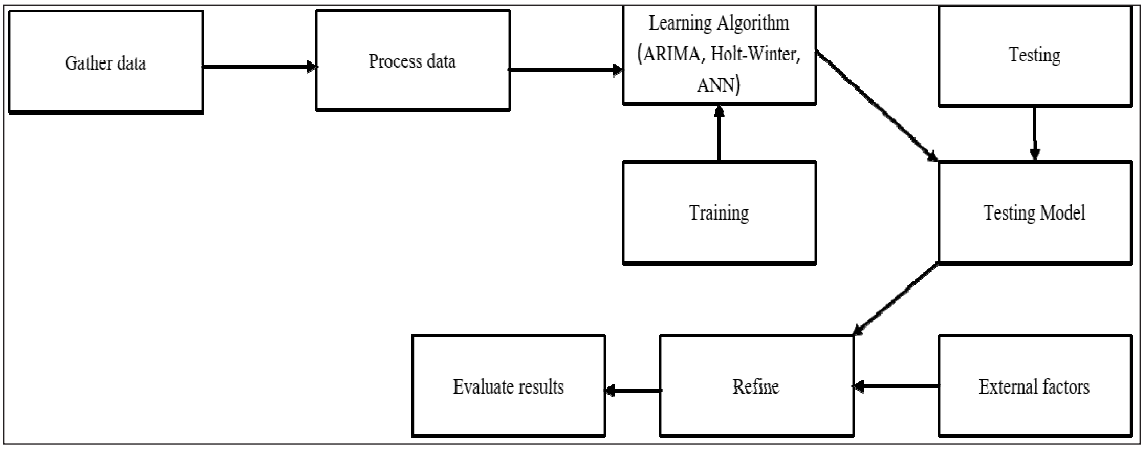

Fig. 3: General learning model for share market prediction

Fitness evaluation, selection, crossover and mutation are done until certain criteria are met. The features will be assigned for the dataset obtained from NSE. The model will deliver high diversity children based on the input data. For other methods, the task is to train a model using data which can be gathered from many sources and can be of varying types. Then, it must be processed and attributes must be selected on which the model must be built.

After processing the data, the model must be trained using a learning technique using set of data marked as training data. The model will then be trained based upon the input data.

The next step is to test the data based on the learned model. The remaining data marked as testing data, will be input into the learned model to gather the output. After gathering the output, it will be subjected to the measure of external factors which might affect the prices in real time. This will help in the prices

in being closer to real time values. The data will be observed and recorded for comparison.

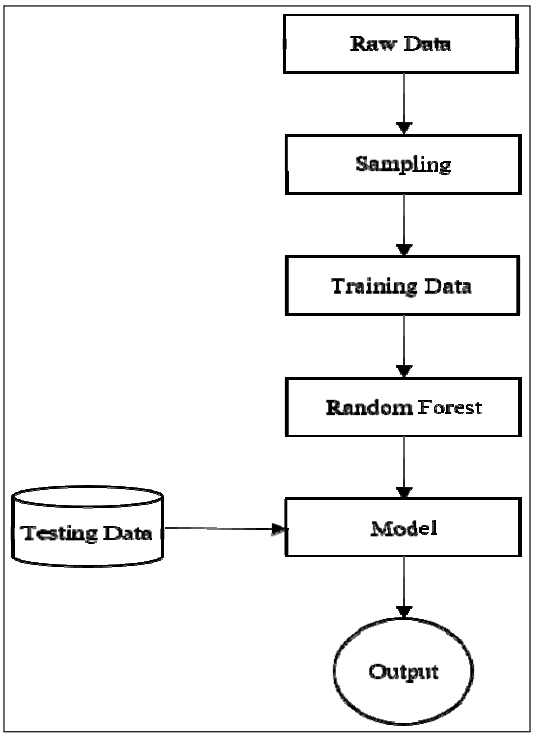

Fig. 4: Model for Random Forest Tree generation and output generate after model creation

Random Forest is better than other decision trees as it takes care of the problem of overfitting. Other than that, it has higher accuracy than other trees and even the cross-validation accuracy of a random forest is more than other decision tress.

The random forest model will be created after collecting and preprocessing data. In this n-number of trees are Fig. 4. Model for Random Forest Tree generation and output generate after model creation constructed after training a data. These trees are created on the basis of different subsets taken from the original training data. Voting is carried out to decide which class is to be selected as output class.

Random Forest can be used to create both short term and long term prediction models. The difference arises on which type of algorithm is used to train the model.

Random Forest Algorithm

1. Create n number of subset using Sample set.
2. Decision Trees are created for each Subset using Information Gain and Entropy.
3. While selecting a node, votes are assigned to each attribute.
4. Node with highest votes is selected as *Key Node*.

After obtaining the training data, the output will be subjected to various external factors which occur in real time in order to infer the difference from real time data.

These factors can be any of the following –

- Market Sentiments
- Company Announcements
- Weather
- Disasters
- Political Events

Factors will be applied to all models in order to get results closer to real time. After obtaining results from all models, they will be compared to calculate the one with least error.

Sentiment analysis is another key factor that is often overlooked as many researchers tend to use it as a separate predictive model. Social events have considerable effect on the financial market. Data gathered from social media can be used to help in obtaining better and more accurate outputs. Social media sites such as Twitter and Facebook have significant impact on the financial market and events such as elections, or major calamities also have a certain effect on financial market.

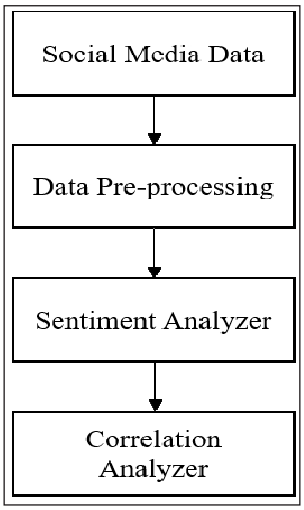

Fig. 5: Sentiment analyzer process model

Using the above process model, the social event which has an impact on market prices can be taken and then factored into the output to obtain a more accurate result.

N-Gram and Word2Vec may be used to select the features. The correlation analysis is done by checking for words that match those related to the data set. The emotions due to media are labelled as positive and negative. Features extracted from above data are fed to a classifier and trained using random forest algorithm.

Based on work done in[13], it can be established that there is a strong correlation between sentiments and share prices.

By classifying emotions into various categories, they were able to correlate the impact of social media (Twitter) to the share prices of a company. Their output was only valid for the next day share prices for the company as they analyzed tweets for only 3 days. However, by using social media platform used by traders and stock enthusiasts (Stocktwits), a better model can be created. Also, from research carried out in[5] using high performance decision maker, an experimental model to better predict impact of sentiments and events may be done and can be further used to improve the overall accuracy of the model.

CONCLUSION

Predictions of share market prices was done using techniques such as GA, ARIMA, Holt-Winter, etc. Instead of just relying on model output data, the various external factors which affect the market prices are also taken into account as external factors to make a more accurate prediction compared to real-time values. The variation in output values may occur as the parameters of the performance might vary for the models.

REFERENCES

1. Wang, C. and Lin, Y. 2015. *"The Prediction System for Data Analysis of Stock Market by Using Genetic Algorithm,"* 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, pp. 1721-1725.

2. Wang, F., Zhang, Y., Xiao, H., Kuang, L. and Lai, Y. 2015. "*Enhancing Stock Price Prediction with a Hybrid Approach based Extreme Learning Model,"* 2015 IEEE 15th International Conference on Data Mining Workshops, IEEE, pp. 1568-1570.

3. Lin, I., Cao, L., Wang, J. and Zhang, C. 2004. *"The Applications of Genetic Algorithms in Stock Market Data Mining Optimisation,",* https://www.researchgate.net/publication/252682655_The_Applications_of_Genetic_Algorithms_in_Stock_Market_Data_Mining_Optimisation, [Accessed 29-Oct-2018]

4. Hamed, I.M., Hussien, A.S. and Tolba, M.F. 2011. "*An Intelligent Model for Stock Market Prediction,*" 2011 IEEE, pp. 105-110.

5. Dormehl, L. 2018. *"What is an artificial neural network? Here's everything you need to know," Digital Trends, 13-Sep-2018.* [Online]. Available: https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/. [Accessed: 27-Oct-2018].

6. Ponnam, L.T., Rao, V.S., Srinivas, K. and Raavi, V. 2016. *A comparative study on techniques used for prediction of stock market," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT).*

7. Srinivasan, N. and Laskhmi, C. 2017. *"Forecasting Stock Price Using Soft Computing Techniques," 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM),* IEEE, 2017, pp. 158-161.

8. Momoh, O. 2018. *"Stock Analysis," Investopedia, 05-Aug-2018. [Online]. Available:https://www.investopedia.com/terms/s/stock-analysis.asp.* [Accessed: 02-Nov-2018]

9. Kamble, R.A. 2017. "Short and Long Term Stock Trend Prediction using Decision Tree," *International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1371-1375.

10. Nivetha, R.Y. and Dhaya, C. 2017. *"Developing a Prediction Model for Stock Analysis," 2017 International Conference on Technical Advancements in Computers and Communications.*

11. Khatri, S.K. and Srivastava, A. 2016. *"Using Sentimental Analysis in Prediction of Stock Market Investment," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions),* Sep. 7-9, 2016, AIIT, Amity University Uttar Pradesh, Noida, India, pp. 566-569.

12. Maini, S.S. and Govinda, K. 2017.*"Stock market prediction using data mining techniques," 2017 International Conference on Intelligent Sustainable Systems (ICISS).*

13. Tiwari, S., Bharadwaj, A. and Gupta, S. 2017. *"Stock Price Prediction Using Data Analytics," IEEE, 2017.*

14. Xing, T., Sun, Y., Wang, Q. and Yu, G. 2013.*"The Analysis and Prediction of Stock Price," 2013 IEEE International Conference on Granular Computing (GrC).*

15. Rajput, V. and Bobde, S. 2016. *"Stock Market Prediction Using Hybrid Approach," International Conference on Computing, Communication and Automation (ICCCA2016).*

16. Pagolu, V.S., Challa, K.N.R., Panda, G. and Majhi, B. 2016. *"Sentiment Analysis of Twitter Data for Predicting Stock Market Movements,"* International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016, IEEE, pp. 1345-1350.