Stock Market Prediction using Multivariable Neural Network combined with Sentimental Analysis.

Abhijith

Student, Department of Computer Science & Engineering,
Sri Siddhartha Institute of Technology, Tumakuru, Karnataka, India

Key words: Sentimental analysis, Back propagation, FFNN, ANN.

Abstract

Stock market prophecy is the work of trying to establish the future value of a company stock or other financial instrument traded on a barter. The successful prophecy of a stock's future price could acquiesce noteworthy profit. This paper will showcase how to perform stock prediction using Machine Learning algorithms: Linear Regression, Random Forest and Multilayer Perceptron.

I. INTRODUCTION

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Here we are going to discuss the method to predict stock market values using machine learning algorithms. In statistics, linear regression is a linear approach for modeling the relationship between a scalar dependent variable Y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

A multilayer Perceptron (MLP) is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear Perceptron. It can distinguish data that is not linearly separable. The topic focuses is on the task of predicting future values of stock market index. Two indices namely, BSE SENSE and NSE NIFTY are selected for experimental evaluation. The rest of the paper is organized as: Section 2 provides the overview of Literature Survey of Stock Market Prediction, Section 3 describes methodology used in paper. Section 4 shows result analysis. Finally, Section 5 delivers conclusions made through predictions.

II. LITERATURE SURVEY

Kannan, Sekar, Sathik and P. Arumugam in [1] used data mining technology to discover the hidden patterns from the historic data that have probable predictive capability in their investment decisions. The prediction of stock market is challenging task of financial time series predictions.

Jing Tao Yao and chew Lim tan in [2] used artificial neural networks for classification, prediction and recognition. Neural network training is an art. Trading based on neural network outputs, or trading strategy is also an art. Authors discuss a seven-step neural network prediction model building approach in this article. Pre and post data processing/analysis skills, data sampling, training criteria and model recommendation will also be covered in this article.

Tiffany Hui-Kuang and Kun-Huang Huarng in [3] used neural network because of their capabilities in handling nonlinear relationship and also implement a new fuzzy time series model to improve forecasting. The fuzzy relationship is used to forecast the Taiwan stock index. Jigar Patel [7] focuses on the task of predicting future values of stock market index. Two indices namely CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex from Indian stock markets are selected for experimental evaluation. Experiments are based on 10 years of historical data of these two indices. The paper proposes two stage fusion approach involving Support Vector Regression (SVR) in the first stage.

Ching-Hseue cheng, Tai-Liang chen, Liang-Ying Wei in [4] this paper proposed a hybrid forecasting model using multi-technical indicators to predict stock price trends. They used RST algorithm to extract linguistic rules and utilize genetic algorithm to refine the extracted rules to get better forecasting accuracy and stock return.

Fazel Zarandi M.H, Rezaee B, Turksen I.B and Neshat E in [5] used a type-2 fuzzy rule based expert system is developed for stock price analysis. The purposed type-2 fuzzy model applies the technical and fundamental indexes as the input variables. The model used for stock price prediction of an automotive manufactory in Asia. The output membership values were projected onto the input spaces to generate the next membership values of input variables and tuned by genetic algorithm.

Ingoo Han [6] proposes genetic algorithms (GAs) approach to feature discretization and the determination of connection weights for artificial neural networks (ANNs) to predict the stock price index. In this study, GA is employed not only to

improve the learning algorithm, but also to reduce the complexity in feature space.

Time series prediction techniques have been used in many real-world applications such as financial market prediction, electric utility load forecasting , weather and environmental state prediction, and reliability forecasting. Ravi Shankar [8] provides a survey of time series prediction applications using a novel machine learning approach: support vector machines (SVM).

T. Jan [14] surveys machine learning techniques for stock market prediction. He present recent developments in stock market prediction models, and discuss their advantages and disadvantages. In addition, we investigate various global events and their issues on predicting stock markets.

Y. Kara [9] attempted to develop two efficient models and compared their performances in predicting the direction of movement in the daily Istanbul Stock Exchange (ISE) National 100 Index. The models are based on two classification techniques, artificial neural networks (ANN) and support vector machines (SVM).

Bo Qian [10] investigated the predictability of the Dow Jones Industrial Average index to show that not all periods are equally random. He used the Hurst exponent to select a period with great predictability. Some inductive machine-learning classifiers—artificial neural network, decision tree, and *k*-nearest neighbor were then trained with these generated patterns. Through appropriate collaboration of these models, he achieved prediction accuracy up to 65 percent.

E. Guresan [11] evaluates the effectiveness of neural network models which are known to be dynamic and effective in stock-market predictions. The models analyzed are multi-layer perceptron (MLP), dynamic artificial neural network (DAN2) and the hybrid neural networks which use generalized autoregressive conditional heteroscedasticity (GARCH) to extract new input variables. The comparison for each model is done in two view points: Mean Square Error (MSE) and Mean Absolute Deviate (MAD) using real exchange daily rate values of NASDAQ Stock Exchange index.

B. Nath [12] deals with the application of hybridized soft computing techniques for automated stock market forecasting and trend analysis. We make use of a neural network for one day ahead stock forecasting and a neuro-fuzzy system for analyzing the trend of the predicted stock values.

Cheng-Yi Tesai [13] hybridizes SVR with the self-organizing feature map (SOFM) technique and a filter-based feature selection to reduce the cost of training time and to improve prediction accuracies. The hybrid system conducts the following processes: filter-based feature selection to choose important input attributes; SOFM algorithm to cluster the training samples; and SVR to predict the stock market price index. The proposed model was demonstrated using a real future dataset – Taiwan index futures (FITX) to predict the next day‘s price index.

III. METHODOLOGY

In this paper, first imported package: numpy, pandas and Natural Language Toolkit in Jupyter Notebook. And, read the saved pickled data file in a table. Then, we selected tuples: price and article, and copied them into another table. We added new columns: _compound‘(compound rating of article), _neg‘(negative rating of article),‘ pos‘(positive rating of article) and _neu‘(neutral rating of article). Then apply _SentimentIntensityAnalyzer‘ from Natural Language Toolkit on our new table and calculated sentiment score for each article. Then, we created two data frames for testing data and training data namely, y_test and y_train respectively. Packages _treeintrepreter‘ and _sklearn‘ are imported in Jupyter Notebook. Then, through predict() function of RandomForestRegressor class in package _treeinterpreter‘package, three data frames are created: _prediction‘, _bias‘ and _contribution‘.

The FFNN is a simple and fundamental ANN and was chosen because many other models are based on it. The proposed model will be used to forecast Dow Jones Industrial Average (DJIA) index values and examine how two factors of the FFNN affect the predicted values. By using a FFNN the objective is to investigate how the number of neurons in the network affect the accuracy of the forecasted values, as well as the distribution of the training or learning dataset.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered. With the recent advances in deep learning, the ability of algorithms to analyze text has improved considerably.

First import the package: numpy, pandas and Natural Language Toolkit in Jupyter Notebook. And, read the saved pickled data file in a table.

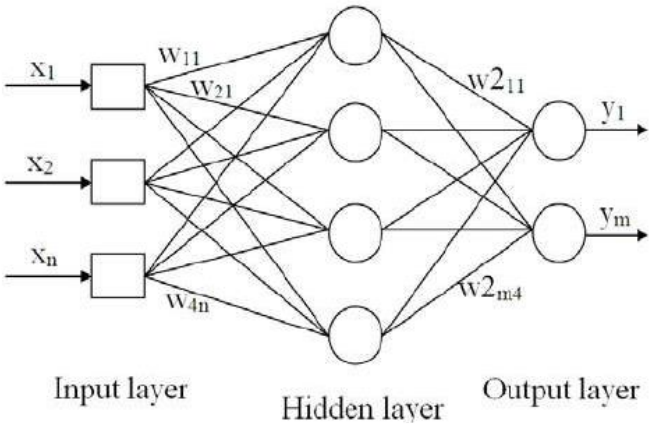


Figure 1. Inside the Neural Network

Then, we selected tuples: price and article, and copied them into another table. We added new columns: _compound‘(compound rating of article), _neg‘(negative rating of article),‘pos‘(positive rating of article) and _neu‘(neutral rating of article). We then applied _Sentiment IntensityAnalyzer‘ from Natural Language Toolkit on our new table and calculated sentiment score for each article. Then, we created two data frames for testing data and training data namely, y_test and y_train respectively. Packages _tree intrepreter‘ and _sklearn‘ are imported in Jupyter Notebook.

Then, through predict() function of Random Forest Regressor class in package_treeinterpreter‘ package, three data frames are created: _prediction‘, _bias‘ and _contribution‘. Another package called _matplotlibs‘ is

imported. A random forest is plotted through pyplot without smoothening. Then, we modify the prices by a constant value so that the predicted prices are close to those actual prices of articles. We add a constant 6117 to all the predicted prices. Then, we apply _Exponential Weighted Mean Average' from pandas to smooth the stock prices. Then, predictions after smoothening are plotted.

Then, we modify the prices by a constant value so that the predicted prices are close to those actual prices of articles. We add a constant 6117 to all the predicted prices. Then, we apply _Exponential Weighted Mean Average' from pandas to smooth the stock prices. Then, predictions after smoothening are plotted.

In ANN, BP network models are common to engineer. So called a BP network model, which is the feed-forward artificial neural network structure and a back-propagation algorithm (BP). It has proved that BP network model with three-layer is satisfied for the forecasting and simulating as a general approximator. Thus, a three layer BP network model trained by Levenberg-Marquardt optimization algorithm is chosen for this study.

In Figure 2, three-layered feed forward neural networks (FFNNs), which have been usually used in forecasting hydrologic time series, provide a general framework for representing nonlinear functional mapping between a set of input variables and the output. Three-layered FFNNs are based on a linear combination of the input variables, which are transformed by a

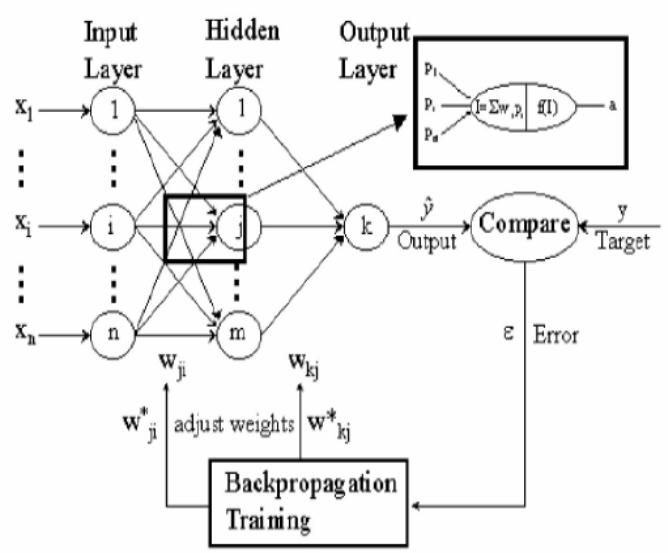


Figure 2. A three layered FFNN with BP training algorithm

nonlinear activation function. In the Figure 2, *i*, *j* and *k* denote input layer, hidden layer and output layer neurons, respectively and *w* is the applied weight of the neuron. The term "feed-forward" means that a neuron connection only exists from a neuron in the input layer to other neurons in the hidden layer or from a neuron in the hidden layer to neurons in the output layer and the neurons within a layer are not interconnected to each other. The explicit expression for an output value of FFNNs is given by

$$\hat{y}_k = f_o\left[\sum_{j=1}^{m} W_{kj} f_h\left(\sum_{i=1}^{n} W_{ji} x_i + W_{jo}\right) + W_{ko}\right]$$

where *wji* is a weight in the hidden layer connecting the *i*-th neuron in the input layer and the *j*-th neuron in the hidden layer, *wjo* is the bias for the *j*-th hidden neuron, *fh*is the activation function of the hidden neuron, *wkj* is a weight in the output layer connecting the *j*-th neuron in the hidden layer and the *k*-th neuron in the output layer, *wko* is the bias for the *k*-th output neuron, *fo* is the activation function for the output neuron, *xi* is *i*-th input variable for input layer and $\hat{y}$, *y* are computed and observed output variables. The *n* and *m* respectively are the number of neurons in input and hidden layers.

The weights are different in the hidden and output layers, and their values can be changed during the process of network training. According to the mentioned concept a computational code has been developed by the author in order to do any related hydrological modeling.

IV. ANALYSIS

The transition from mathematics to code or vice versa can be made with the aid of a few rules. They are listed here for future reference. To change from mathematics notation to MATLAB notation, the user needs to:

- Change superscripts to cell array indices.
- Change subscripts to parentheses indices.
- Change parentheses indices to a second cell array index.
- Change mathematics operators to MATLAB operators and toolbox functions.

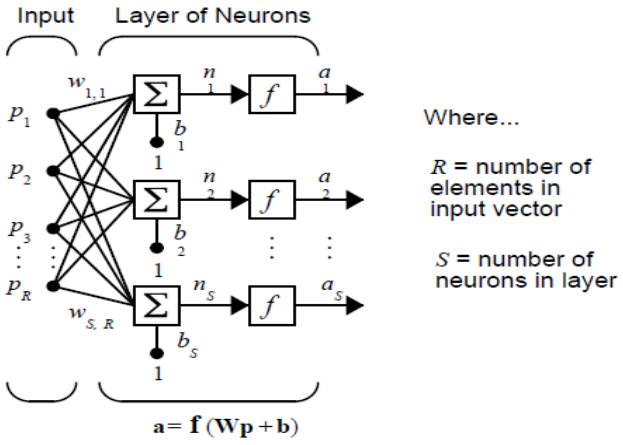
The following equations illustrate the notation used,

p1->p{1}

$$n = w_{1,1}p_1 + w_{1,2}p_2 + \ldots + w_{1,R}p_R + b$$

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\ & & & \\ w_{S,1} & w_{S,2} & \cdots & w_{S,R} \end{bmatrix}$$

A one-layer network with *R* input elements and *S* neurons follows. In this network, each element of the input vector **p** is connected to each neuron input through the weight matrix **W**. The *i*th neuron has a summer that gathers its weighted inputs and bias to form its own scalar output *n(i)*. The various *n(i)* taken together form an *S*-element net input vector **n**. Finally, the neuron layer outputs form a column vector **a**. Note that it is common for the number of inputs to a layer to be different from the number of neurons (i.e., *R* ¦ *S*). A layer is not constrained to have the number of its inputs equal to the number of its neurons.



V. CONCLUSION

Stock market prediction is important factor in finance. It is considered to be dynamic in nature. The paper presented how to predict stock values based on the data of NY Times of 10 years using Machine Learning algorithms: Logistic Regression, Random Forest and Multilayer Perceptron(MLP). We also concluded that MLP is better than the other two algorithms because, within a certain range, the difference between actual price and predicted price is quite small as compared to those in Logistic Regression and Random Forest. Also, Random Forest is better than Logistic Regression, but inferior to MLP, in predicting stock values.

ACKNOWLEDGEMENT

First and foremost, I wish to express my profound gratitude to (your guide & HOD names) for giving me the opportunity to carry out the project . My heartfelt thanks for their invaluable guidance, immense help, support & useful suggestions throughout the course of my project work.

REFERENCES

[1] K. Senthamarai Kannan, P. Sailapathi Sekar, M. Mohamed Sathik and P. Arumugam, ―Financial stock market forecast using data mining Techniques", Proceedings of the international multiconference of engineers and computer scientists, 2010.

[2] JingTao YAO, Chew Lim TAN, ―Guidelines for Financial Prediction with Artificial neural networks―.

[3] Tiffany Hui-Kuang yu, Kun-Huang Huarng, ―A Neural network-based fuzzy time series model to improve forecasting‖, Elsevier, 2010, pp: 3366-3372..

[4] Ching-Hsue Cheng, Tai-Liang Chen, Liang-Ying Wei, ― A hybrid model based on rough set theory and genetic algorithms for stock price forecasting‖, Pages. 1610-1629, 2010.

[5] M. H. Fazel Zarandi, B. Rezaee, I. B. Turksen and E.Neshat, ―A type-2 fuzzy rule-based experts system model for stock price analysis‖, Expert systems with Applications, Pages. 139-154, 2009.

[6] Kyoung-jae Kim, Ingoo Han, ―Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index‖,Expert Systems with Applications, Vol. 19, Issue 2, Pages. 125-132, 2000.

[7] J. Patel, S. Shah, P. Thakkar, K. Kotecha, ―Predicting stock market index using fusion of machine learning techniques‖, Expert Systems with Applications, Vol. 42, Issue 4, Pages. 2162-2172, March 2015.

[8] R. Shankar, N.I. Sapankevych, ―Time Series Prediction Using Support Vector Machines: A Survey‖, IEEE Computational Intelligence Magazine, Vol. 4, Issue 2, Pages 2162-2172, March 2009.

[9] Y. Kara, M. A. Boyacioglu, Ö. K. Baykan, ―Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange‖, Expert Systems with Applications, Vol. 38, Issue 5, Pages 5311-5319, May 2011.

[10] Bo Qian, K. Rasheed, ―Stock market prediction with multiple classifiers‖, Applied Intelligence, Vol. 26, Issue 1, Pages 25-33, February 2007.

[11] E. Guresan, G. Kayakutlu, T. U. Diam,, ―Using artificial neural network models in stock market index prediction‖, Expert Systems with Applications, Vol. 38, Issue 8, Pages 10389-10397, August 2011.

[12] A. Abraham, B. Nath, P. K. Mahanti,, ―Hybrid Intelligent Systems for Stock Market Analysis‖, Computational Science - ICCS 2001, Pages. 337-347.

[13] Cheng-Yi Tesai, Cheng-Lung Huang, ―A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting‖, Expert Systems with Applications, Vol. 36, Issue 2, Pages 1529-1539, March 2009.

[14] P. D. Yoo, M. H. Kim, T. Jan, ―Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation‖, IEEE Computational Intelligence for Modelling, November 2005.