IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

A Comparative Analysis of Different Machine Learning Algorithms Used in Predicting Stock Market Prices

Siddharth Chakravarthy[1], Shraddha Sunil[1]

Department of Electronics and Communication Engineering, R.V. College of Engineering, Visvesvaraya Technological University, Belgaum[1]

Abstract: Stock markets can be defined as a trading platform for the exchange of financial instruments such as debt, equity and derivatives. They work on the principle of price discovery, which is the act of studying the market supply and demand of a commodity and determining its proper price. Essentially when a person is investing in the stock markets, there are two kinds of situations that are prevalent. One is uncertainty and the other is risk. A person would hedge a bet only if the proposition is risky, however if things turn out to be uncertain, he would be averse to investing in the stock. By studying different machine learning algorithms, we analyse the best methods available in predicting the movement of stocks.

Keywords: Neural Networks, Genetic Algorithm, Regression, Decision Tree, SVM.

I. INTRODUCTION

Stocks basically represent a percentage of the company's ownership. The price of the stock is incumbent on the market forces, namely supply and demand. If the demand for the stock is more than the supply, the value of the stock goes up. Similarly, if the supply exceeds the demand, then the value of the stock goes down. A major constraint is that the stock markets are always in flux. In order to assess the risk factor properly and to reduce the losses, an accurate prediction model is required. In this paper, we make a comparison of all the different mathematical tools and ML algorithms that can help reduce the uncertainty in prediction. An accurate prediction would yield higher profit margins for the investor.

II. ALGORITHMS

A. LINEAR REGRESSION

Linear regression focuses on getting the line of best fit. This implies that using existing values in the training data set we can obtain an optimum line which is known as the line of best fit. We can be reasonably certain that the predicted values will lie close to the line of best fit.

The first step involves obtaining scatter plots for the training data. After this, we need to find a straight line that fits through all these point to find the estimated stock values. The predicted errors can be minimised based on the standard deviation as given below.

$$\frac{\Sigma i(x - x')(y - y')}{\Sigma_i (x - x')^2}$$

A Linear regression uses the equation: $y=a_0+a_1x$ which describes a standard linear equation. y=Predicted variable or Dependent variable;

a_0=Error not explained by regression;

a_1=Slope of the regression line

x =Predictor variable or independent variable. x'=mean of set x

y'=mean of set y

A scatter plot of the observations (dependent variable vs independent variable) is first obtained. Then a straight line is drawn to fit through all these points. The values on the line are known as the "Estimated Values". The errors are then minimized based on the estimated values and the actual values.

B. GENETIC ALGORITHMS

Genetic algorithms can be used to predict the values of stocks based on the biological phenomenon of natural selection and natural evolution. It makes use of bio-inspired operators such as mutation, crossover and selection. The first step involves the creation of a "population" of randomly generated strings of 0s and 1s. This population is evaluated at every "generation" based on the fitness function which is used to calculate the fitness of every individual in the population that is being evaluated. The fitness is usually the optimum value of the objective function according to the problem statement.

Once the fitness has been evaluated, the chromosomes that have better chance to reproduce are passed onto the next generation. This is followed by cross-over and mutation. Crossover refers to the process of randomly selecting a cross site and then swapping the genes of the two parent chromosomes along that cross site. After the crossover operation, mutation is performed which involves changing the string by changing relevant bits from 1s to 0s or vice-

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

versa. This is done to generate a variance in the population.
Example of the crossover operation

Chromosome1=10011011 | 10010010
Chromosome2=01101010 | 11010011

Offspring 1=1001101111010011
Offspring 2=0110101010010010

Example of mutation at the 5th and 6th position respectively in 1st and 2nd offspring
Mutated Offspring1= 1001001111010011
Mutated Offspring2= 0110111010010010

C. DECISION TREE
C4.5 is an algorithm for generating a decision tree. Developed by Ross Quinlan, it is an extension of the ID3 decision algorithm. It involves a generation of a decision tree which can be used for classification. C4.5 uses information gain to make a tree of classificatory decisions with respect to a previously known target classification. The information gain can be defined as the reduction in entropy by making a choice as to which attribute to select and at what level. For example, if the user chooses type of connections to discriminate among cases at a given point in its rule construction process, the choice will have some effect on how to tell the classes apart. By taking into consideration the attribute to be taken for discrimination among the cases at a particular node in the tree, we can build a decision tree that allows us to travel from the root node till the decision class at the leaf node by continually examining attributes. The order in which the attributes are chosen depends on the amount of entropy correlated to that given attribute.

D. SUPPORT VECTOR MACHINES
Classification and regression analysis can also be performed by another machine learning algorithm called Support Vector Machine (SVM). A solution space of the given problems is divided into distinct categories (for example, a yes or a no situation). Each data item is a point in an n-dimensional space (where n is the number of features available). Each value of the feature is the value of a particular coordinate. A line or a curve can be used to separate the data in the solution space into different categories in the training set. When new data is added in the test set, the side of the gap on which it falls can be predicted.

E. ARTIFICIAL NEURAL NETWORKS
ANN is a data driven model. The uniqueness of the ANN is its ability to relate the input data to the output data set through a non-linear relationship. The output values can be determined from the input values which have to be transformed using activation functions. The most crucial step involved in this is training a neural network which requires the trainer to perform at the optimum best. Backpropagation algorithm is implemented continuously until the point of occurrence of minimum error. Although being a tedious process, it trains the entire network.

III. LITERATURE SURVEY

- Farhad Soleimanian Gharehchopogh, Tahmineh Haddadi Bonab and Seyyed Reza Khaze in [1] have written about using linear regression to predict S&P 500 index behaviour. The paper explores the relationship between trading volume as the dependent variable and average price per share as the independent variable of the regression equation. The data is divided into two parts of training and testing. The training set is used for analysis while the testing data is used to calculate the accuracy of the algorithm. The accuracy is found to be 63.6%. The authors conclude by saying that for more comprehensive results, K-means can be used.
- Ganesh Bonde and Rasheed Khaled in [2] have used the following six metrics for the prediction of stock market prices. They are Opening price, closing price, Highest Price, Lowest Price, Trading Volume, Adjusted Closing Price. The fitness computed here corresponds to the number of times a correct prediction has been made. The highest accuracy found is quite reasonable and is found to be 73.87%. It can be seen that the metrics chosen have a good degree of correlation amongst them. So, if the accuracy is to be improved, then more dissimilar parameters have to be chosen.
- R Lakshman Naik, D. Ramesh, B. Manjula, Dr. A. Govardhan in [3] propose genetic algorithm as a solution to the stock market price prediction problem by defining a set of rules that will yield the maximum profit. They use five conditions to determine whether or not the person should buy or sell.
- S.S Panigrahi Dr. J. K. Mantri in [4] state that the decision tree is remarkable in terms of predicting the stock market prices. In this paper, an accuracy of 90.8213% is shown in terms of predicting stock market prices. It has also been shown that the accuracy value is superior to that of using a normal tree. The reason for the good accuracy values is because in the confusion matrix that has been developed, the false positive rate is lower.
- Binoy B Nair, V P Mohandas, N R Sakthivel in [5] show how rough sets are used for predicting the stock market prices. Such sets are useful in dealing with incomplete or imperfect knowledge. They create a lower bound approximation for those elements that are known with a degree of certainty and the upper-bound approximation that has elements whose knowledge is incomplete. The accuracy is found to be 90.22% which is better than the accuracy of Naive Bayes prediction method which has an accuracy of 72.36%.
- Qasem A. Al-Radaideh, Adel Abu Assaf, Eman Alnagi in [6] have developed a model via the CRISP-DM. WEKA software has been used to simulate Decision

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

tree id3 and c4.5. The dataset that has been used is Amman Stock Exchange. The two evaluation methods used are K-fold cross validation and percentage split method. As the accuracy isn't very good (44-55%) the author's analyse the reason for such a low accuracy. The reason they came up with is that the price of a stock may be affected by other factors such as news of the company and the company's performance in the market.

- In [7], a new method of using SVM with financial statement analysis for prediction of stocks is proposed. The experimental results show that there is a higher degree of accuracy using this new approach than by just using SVM because financial indices are used as experimental parameters. The experimental results are much more reliable, non-volatile and valid.
- In [8] stock market prices are predicted using ANN. ANN is a popular method for identifying hidden patterns suitable for stock market prediction. The whole process is divided into two modules. For the first model Back propagation algorithm is used while for the second module the Multilayer feedforward network is used.
- In [9], it was stated that ANN is suitable for prediction of stock market prices because of the large amount of data that's involved and also because no linear model can be used to describe the stock market, and also there is a large set of interacting input series which is needed to explain the stance of the stock price which suits Neural Network.

REFERENCES

[1] Farhad Soleimanian Gharehchopogh1 , Tahmineh Haddadi Bonab2 and Seyyed Reza, "A linear regression approach to prediction of stock market trading volume", International Journal of Managing Value and Supply Chains (IJMVSC) Vol.4, No. 3, September 2013
[2] Ganesh Bonde, Rasheed Khaled, "Stock price prediction using genetic algorithms and evolution strategies"
[3] D. Ramesh1 , B. Manjula1 , Dr. A. Govardhan, "Prediction of Stock Market Index Using Genetic Algorithm R. Lakshman Naik1" , Computer Engineering and intelligent systems, Vol 3, No 7, 2012
[4] SS Panigrahi, Dr J.K Mantri, "A text based decision tree model for stock market forecasting", International Conference on Green Computing and internet of things
[5] Binoy B.Nair, V.P Mohandas, N.R. Sakthivel, "A decision tree-rough set hybrid system for stock market trend prediction", International Journal of Computer Applications, Volume 6- No 9, September 2010
[6] Qasem A. Al-Radaideh, Adel Abu Assaf, Eman Alnagi, "Predicting Stock Prices using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013)
[7] Shuo Han, Rung-Ching Chen, "Using SVM with financial statement analysis for prediction of stocks", Volume 7, Issue 4, 2007
[8] Zabir Haider Khan, Tasnim Sharmin Alin, Md.Akter Hussain, "Price prediction of Share Market using Artificial Neural Network (ANN)", International Journal of Computer Application(0975-887), Volume 22, No.2, May 2011
[9] Neelima Budhani, Dr. CK Jha, Sandeep K Budhani, "Prediction of stock market using ANN", 2014 IEEE