Short and Long Term Stock Trend Prediction using Decision Tree

Mr. Rupesh A. Kamble
Computer science and engineering
Government College of Engineering
Aurangabad, India.
Kamblerupesh9@gmail.com

Abstract—**This paper presents the results of method designed to predict price trends in the stock market. First objective of this research is to optimize the stock price trend prediction for short term using some oscillators and indicators: Moving Average Convergence Divergence (MACD), the Relative Strength Index (RSI), the Stochastic Oscillator (KDJ) and Bollinger Band (BB). It is observed that using appropriate pre-processing technique and Machine learning model, it is possible to improve accuracy rate of short-term trend prediction. Applying Pre-processing and then using combination of data can yield a better Accuracy rate in Short term Trades, while predicting for Long-term Trend of Stock this Technical indicators are not sufficient. Along with some of this Technical data and Fundamental Data of the company, it is possible to predict Long term stock movement. For Long term Prediction its Debt to Equity, Net profit of pervious 3 year, Promoters holding, Dividend yield and PE ratio is used along with Technical Factors. It is observed that using Fundamental and Technical Data, Long term Stock Prediction is Possible.**

Keywords— MACD, KDJ, RSI,Fundamental data, Random Forest,J48 Decision tree, Bagging.

I. INTRODUCTION

Due to rapid improvement in Computing Performance, many Stock Prediction Techniques are come into picture and methods for fining information from data are gaining more interest[9]. Because large amount of historical stock data is present, it is possible to use data mining to find patterns in stock prices which can be used to predict the future trend of the stock.

A correct data mining approach and correct data processing can yield a better predication .There are 2 Types of stock analysis is there in stock market, those are Technical Analysis and Fundamental analysis. Stock market analyst uses technical analysis for predicting stock prices for short term like 1-2 months stock trend, but along with this technical data, if some fundamental factors are considered, then we can predict the stock value for long term like 6 months to 1 year.

In this paper, some data processing techniques to improve the accuracy of the short and long term analysis is developed. It is concluded that using Technical, Fundamental data and Correct pre-processing , it is possible to predict the stock price for long term , to gain more profit from market.

II. DATA SOURCE , BASIC CONCEPT AND TRADING RULES

A. Data Source

Data from various stocks listed in NSE & BSE is used to build the Training set. More than 3 years of data has been collected as a raw data for pre-processing and training purpose. In analysis closing price of the stock is considered to be a price for both training and testing purpose.

B. Basic Concept and Trading Rules

In stock market analysis mostly used technical indicators are RSI, Bollinger Band, MACD and Stochastic, while in fundamentals analysis information which is used to predict the trend is Net Profit, Dividends, Promoter Holding, debt to equity ratio and PE ratio[7].

MACD is based on moving Averages. For calculating this value, longer exponential moving average (26 day MA) is subtracted from shorter exponential Moving Average (13 day MA)[1].

MACD = 13 day EMA – 26 day EMA
Signal Line = 9 day EMA of MACD line

While calculating EMA for first day SMA is considered.

Stochastic is one of the fast indicator used in analysis [1]. To get values of stochastic oscillator, first RSV is calculated. In stochastic KDJ value is considered to find the short term trend. RSV(Raw Stochastics Value)

$$RSV(n) = \frac{c_t - l_t}{h_n - l_n} \times 100$$

Where l and h stands for high and low prices. K, D and J values are calculated by RSV value.

$$K_{t+1} = \frac{2}{3} \times K_t + \frac{1}{3} \times RSV_n$$
$$D_{t+1} = \frac{2}{3} \times D_t + \frac{1}{3} \times K_{t+1}$$
$$J = 3K - 2D = K + 2(K - D)$$

Stochastic predict the future trend which is for short time frame, that is why it is called as fast indicator.
RSI (Relative Strength Index)is calculated by 14-days Average Gain & Average Loss[10].

$$RSI = 100 - \frac{100}{1 + RS}, \text{ where } RS = \frac{avgGain}{\text{avg}Lost}$$

$$avgGain = (\text{total of gains during past } n \text{ periods}) \div n$$

$$avgLost = (\text{total of losses during past } n \text{ periods}) \div n$$

Most of the trader believes that RSI below 30 is Over-sold zone and RSI above 80 is over brought zone [2]. Over-sold zone can be used as buying opportunity and over-brought zone as selling opportunity.
Bollinger Bands are volatility bands, which changes as volatility increases and decreases. There are 3 bands in Bollinger band - Lower Bollinger band, Middle Bollinger band and Upper Bollinger band[11].

Middle Band = 20 day SMA.
Upper Band = 20 day SMA + (20day std deviation of price*2)
Lower Band = 20 day SMA -(20day std deviation of price*2)

Along with this technical oscillator, fundamental data is also considered to predict the stock value in long term, few of them are listed here,
(1) **Net Profit** of last 3 years is considered. Net profit is a actual profit after working expenses.
(2) **Dividend** – Whenever any firm earn profit it distribute some profit to shareholders, It is called as Dividend.
(3) **Promoters Holding** - It is the total number of shares held by the actual owner of the company
(4) **Debt/Equity Ratio** is calculated by dividing a company's total liabilities by its stock holder's equity.
(5) **PE ratio** (Price to earnings ratio) is calculated by current market price of that stock with its earning per share value.

III. MODEL DESIGN

A. Short Term Analysis

Input and Preprocessing - Daily stock prices of different stock from last 5 year are captured and on the basis of its price movement , values for RSI, MACD, Bollinger Bands and Stochastic(KDJ) are prepared. On the basis of those values BUY, SELL or HOLD strategy is made. In stock market it is observed that instead of using MACD and Signal values, more important is the crossover of MACD line over Signal line for BUY decision. In case of stochastic, Fast and Slow stochastic values are considered, but while training they are sampled to make an absolute decision. In Bollinger band, Value of band never matters as analyst always look for the band in which current stock price is lying that is the reason sampling is used as preprocessing technique , so that system can think like a real analyst. RSI value is taken as a numeric value and no sampling can be done as RSI is a volatile Oscillator. This is the Preprocessing Methods which used for preprocessing of data.

Random Forest - After collecting and preprocessing data, Classification algorithm is used to obtain Rules for machine learning. In case of Technical data, it is found that Accuracy of Random Forest model is more, The Problem in other Decision trees is data over fitting, even the Cross-validation Accuracy of that Classification algorithm is more, while working on real time data Accuracy is not good, which is because of data over fitting effect. Random Forests are an ensemble Classification method. In which n- number of trees are constructed while training a data, all this trees are made on the basis of different subsets taken from the original training data. Voting is carried out to decide which class to be selected as a output class. Random Forest is the most advanced decision tree.

Random Forest Algorithm

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

In R language, RWeka and randomForest library are used to create Model from Training database.

```
random <- randomForest(DEC~.,p)
pred.prob <- predict(random,t,type="prob")
```

This two functions are used to find the probablity distribution.

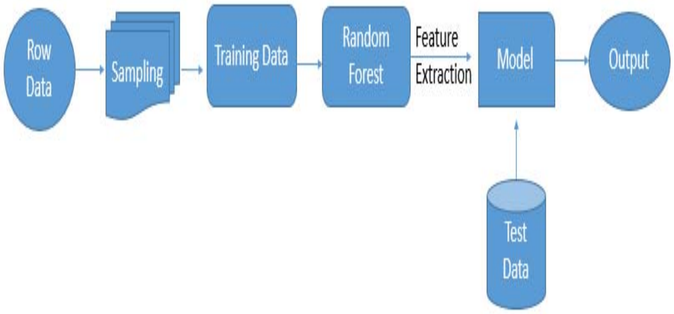


Figure 1. System Architecture for short term

Trading Model - In this paper, Daily trading data is used.Closing price of all the stocks is considered to be Current Market value.All the Trading decisions are made on Current market value of stock. Buy ,Sell and Hold decision can be made on the basis of output of the given model[3][6].If Buy call is Gentarte then that means Stock can show Uptred in Upcoming days, Sell Desicion show Less strength in stock price and hold decision will be considred only, if Buy or sell decision is already taken.Hold decision implies that stock price will be stable for short term. Specific model is bulid to Book

Profit and loss at specific value after decision has been made.If the Buy Decision is Taken then after 10% upmove from the entry price ,Profit booking will be done. In the same way if Stock Price moves down then below 3% Stock Loss is maintained. In this model after Buy Signal,Risk is 3% and Reward is 10%. This model will be used to take action once the Decision is made.

B. Long Term Analysis

Input and Preprocessing - In short term analysis, Price history of stock is considered to predict it's future price, and historical data is provided by Google Finance, while when we are predicting the stock price for Long time Horizon, we need to use some core Fundamental factor of that company .while creating a Classification model for long term predication , Fundamental data is also used along with Price movement data[5] .Core Fundamental data includes Debt/equity, Stock and sector's Profit to earn ratio, Net Profit, Share Holding pattern. Fundamental Data can be collected from Company website or BSEindia website. In this paper, Fundamental data is collected from www.bseindia.com, which is highly authenticated.

Same Preprocessing algorithms are used on technical data which are used in short term analysis, and for long term data sampling is done, because accuracy of the proposed model is more on sampled data. Net profit of last 3 years are compared and then data is prepared for Training.PE ratio of the company is compared with PE ratio of that specific sector ,to consider the growth of the firm with stock price. This Pre-processed data is then used to prepare training model.

J48 with Bagging - In case of long term prediction, it is found that Accuracy of J48 with bagging is more. Divide-and-conquer approach is implemented in j48 algorithm.J48 is an an implementation of C4.5 Algorithm which is present in Weka. In J48 algorithm with the training dataset, problem of data over fitting is observed [12]. This can be overcome by using bagging with J48 algorithm [4].

In J48 algorithm, entropy rate for each class is calculated and it's Information Gain is calculated, which is compared with information gain of other classes and class with highest information gain is selected as node. Same procedure is repeated for all the classes to create a decision tree.

H(s) – Entropy is calculated by

$$H(S) = \sum_{x \in X} p(x) \log_2(1/p(x))$$

S – Current data set

X – set of classes in S

P(x)- Proportion of total number of elements in x class to total number of elements in set S.

Information Gain

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

H(s) – Entropy of set s.

T – Subset created from splitting set S by attribute A.

P(t) – proportion of number of elements in t to s.

H(t)-Entropy of subset t.

Improvement from ID3 algorithm

- J48 Algorithm can handle discrete and continuous attributes by splitting the continues values with respect to some threshold values.
- In J48 algorithm missing attributes can be used with '?' sign.
- Tree pruning functionality is present to avoid data overfitting.

Bagging is a technique which is often used with decision trees to handle the effect of data over-fitting. It keeps the bias same and reduces the variance. Bagging is an ensemble technique. In bagging we create multiple bags filled with random data from training set and size of bags are less than training data, then all this bags are used as training set to reduce the variance . It is also called as Bootstrap Aggregation. Bagging is effective because it improves the accuracy of a single model by using multiple copies of it, trained on different sets of data. The real time results showed that J48 with Bagging has better results as compare to J48 alone.

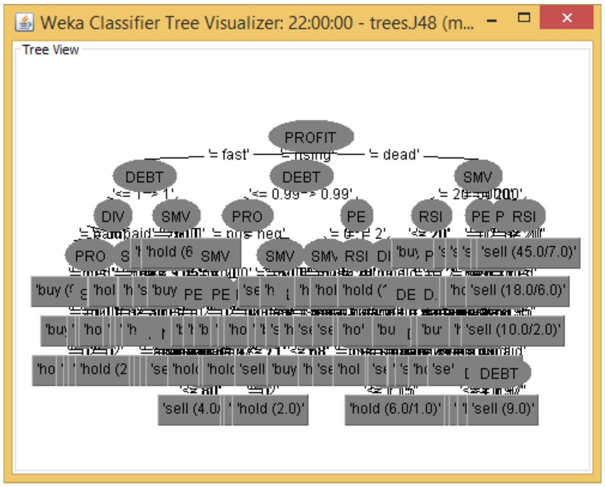


Figure 2. J48 Decision Tree

Rweka Library is used to Perform J48 Algorithm to produce Model .

```
Var1 <- Bagging(DEC~.,data=p,control =
Weka_control(W="J48"))
```

Trading Model - For Long term stock prediction, Long Stop loss and Long Target is considered ,10% Stop loss is used from the Current Stock Price when Buy decision is taken, on the counter part 45% from the Current market price Target is placed[8], and the time period for this will be 1 Year. If stock move 45% up or 10% down in a year then Order will be automatically placed, if no Target or Stop loss reaches in 1 year then sell order will be placed . Even through company is doing well, Stock market Volatility can not be ignored. that is the reason this model is build to book Profits on gain.

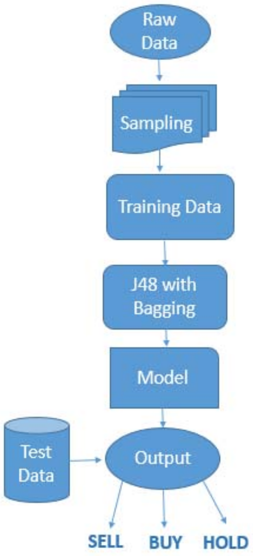
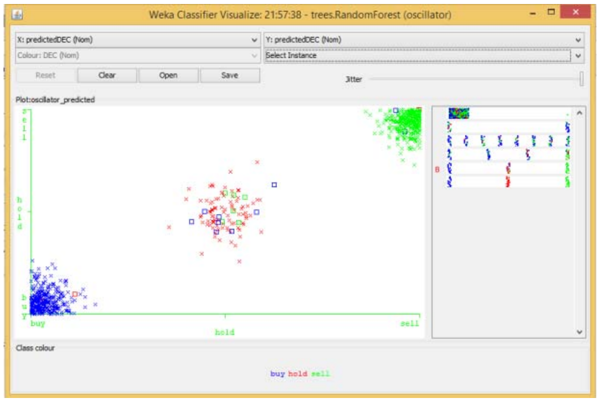


Figure 3. System Architecture for Long term

IV. EXPERIMENTS

Random Forest

In case of random forest ,out of bag(OOB) error rate is used to predict the model accuracy. It uses bootstrap aggregation which sub-samples the training set and applies it to test training data set.When we apply Random forest on training data for short term it is found that OOB estimation of error rate is 15.3% and classification error for Buy,Hold and Sell is 0.11,0.38 and 0.12 respectivly, which shows we can use Buy and Sell signal with high accuracy.


Weka Classifier Visualize: 21:57:38 - trees.RandomForest (oscillator)

X: predictedDEC (Nom)

Y: predictedDEC (Nom)

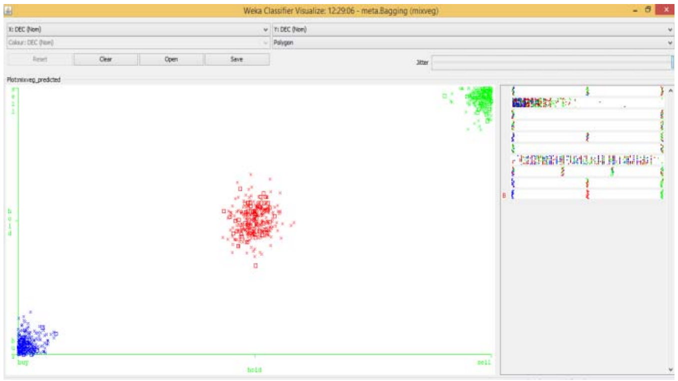
Colour: DEC (Nom)

Select Instance

Plot:oscillator_predicted

Figure 4. OOB estimation on Random Forest

In case of **J48 algorithm** 10-fold cross validation is used to analyze the model accuracy.10-fold cross validation performed on a tarining data set to measure the accuracy of the model. It is also called as Rotation estimation. It is uses as Test model in trainng phase,when we don't have testing data.It gives an insight how the model will genralize to an Independent dataset.After performing 10-fold cross validation on training data for long term analysis. It is found that correctly classified instances are 85.7708 % and Incorrectly classified insatances are 14.2292 % . For buy signal TP- rate was 0.901 and FP-rate was 0.056.


Weka Classifier Visualize: 12:29:06 - meta.Bagging (mixveg)

X: DEC (Nom)

Y: DEC (Nom)

Colour: DEC (Nom)

Polygon

Plot:mixveg_predicted

Figure 5. 10-fold Cross validation result on J48 with bagging

Experimental setup

To analyze real time accuracy of model 1000 stocks are applied to given system for short term and long term analysis. For short term analysis whenever is buy signal is generated buy order is placed with 10% Target and 3% stop loss . While in case of fundamental analysis whenever buy signal is generated buy order is placed and Target will be 45%, stop loss will be 10% and Timeframe for this will be 1 year.

Experimental Resuts

Around 1000 Stocks are applied to find the accuracy of the given short term model. It is observed that around 563 stocks has given Buy signal, then 10% target and 3% stop loss is applied. It is found that 376 stocks hit target price, 91 stocks hit stop loss price, and 96 stocks neither hit target price nor stop loss price. Which proves that accuracy of the model is 66.8% for Buy signal.

For analysis of result of long term model, again 1000 stocks are tested. Technical and Fundamentals values of this stocks are provided to find the results. It is observed that in 273 stocks Buy signal is generated , then 45% target and 10% stop loss is provided according to experimental setup. Out of 273 , 207 Stocks has reached the target price, 47 stocks didn't touched either stop loss or target. But significantly only 19 stocks hit stop loss price. So it makes sense that if someone puts his money on the basis of this model chances of making loss is only 16% for short term and only 7% for long term, and reward is much higher than risk.

978-1-5386-2745-7/17/$31.00 ©2017 IEEE

V. CONCLUSION

A. Finding

In this paper, we have proposed a stock recommendation system using Random Forest and J48 Algorithm. Instead of using raw data, sampled data is used to predict future trend ,but sampled data cause the effect of data overfitting which is further reduced by pruning tree , Random forest and Bagging technique. It is observed that in real time scenario this model yield good result with less risk.

B. Further Research

Future researchers may include more methods for finding the best model for predicting stock prices. Twitter **sentiment** analysis technique can be used to perform voting system along with this model, which can avoid the complete dependence on historical stock data. **News based system** can also be integrated to get the fresh news along with previous historical results of the specific firm to avoid losses in market. Instead of always booking profit on 10% rise and Selling on 3% fall, a new **Trading model** can be build on the basis of probability distribution of the decision based on the model output.

ACKNOWLEDGEMENT

The author would like to thank Computer department, GECA for their support and guidance in building Short and Long Term Stock Trend Prediction System.

REFERENCES

[1]. Mingyuan WU, Xiaotian DIAO, Technical Analysis of Three Stock Oscillators Testing MACD, RSI and KDJ Rules in SH & SZ Stock Markets, 2015 4th International Conference on Computer Science and Network Technology (ICCSNT 2015)

[2]. Sabaithip Boonpeng and Piyasak Jeatrakul. Decision Support System for Investing in Stock Market by using OAA-Neural Network, 8th International Conference on Advanced Computational Intelligence Chiang Mai, Thailand; February 14-16, 2016

[3] Chai Chee Yong and Shakirah Mohd Taib, Designing a Decision Support System Model for Stock Investment Strategy , WCECS 2009, October 20-22, 2009, San Francisco, USA

[4] Indriana Hidayah , Adhistya Erna P., Monica Agustami Kristy, Application of J48 and Bagging for Classification of Vertebral Column Pathologies. 2014 International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia.

[5] Carol Hargreaves, Yi Hao . Does the use of Technical & Fundamental Analysis improve Stock Choice? : A Data Mining Approach applied to the Australian Stock Market .

[6] Yibu Ma, The Research of Stock Predictive Model based on the Combination of CART and DBSCAN, 2013 Ninth International Conference on Computational Intelligence and Security.

[7] Wilder, J. Welles. New concepts in technical trading system.Trend Reseach ,1978.

[8] Wilder, J. Welles. New concepts in technical trading systems. Trend Research, 1978.

[9] Kannan K S, Sekar P S, Sathik M M, et al. Financial stock market forecast using data mining techniques[C]//Proceedings of the International Multi conference of Engineers and computer scientists. 2010, 1.

[10] Murphy, John J. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications, Penguin, 1999.

[11] http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:bollinger_bands

[12] K R Pradeep; N C Na K R Pradeep; N C Naveen veen, Predictive analysis of diabetes using J48 algorithm of classification techniques, 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I).