Extracting the best features for predicting stock prices using machine learning

Ganesh Bonde
Institute of Artificial Intelligence
University Of Georgia
Athens,GA-30601
Email: ganesh84@uga.edu

Rasheed Khaled
Institute of Artificial Intelligence
University Of Georgia
Athens,GA-30601
Email: khaled@uga.edu

Abstract:-

Predicting stock price is always a challenging task. In this paper we are trying to predict the next day's highest price for eight different companies individually. For this we are using different feature sets to predict the price. It is observed that the Volume+Company and Nasdaq+S & P 500 +Company sets performed better than any other feature sets used. Also these features were very helpful for predicting stock price using sequential minimal optimization (SMO) and bagging approach. Comparing different methods, the best results were obtained using SMO and bagging.

Keywords:

Machine learning,stock market, sequential minimal optimization, bagging,

I. Introduction

For many years considerable research was devoted to stock market prediction. During the last decade we have relied on various types of intelligent systems to predict stock prices to make trading decisions. Thus numerous models have been depicted to provide the investors with more precise predictions. It has been observed that the stock price of any company does not necessarily depend on the economic situation of the country. It is no more directly linked with the economic development of the country or particular area. Thus the stock price prediction has become even more difficult than before.

These days stock prices are affected by many factors like company related news, political events, natural disasters … etc. The fast data processing of these events with the help of improved technology and communication systems has caused the stock prices to fluctuate very fast. Thus many banks, financial institutions, large-scale investors and stockbrokers have to buy and sell stocks within the shortest possible time. Thus a time span of even a few hours between buying and selling is not unusual.

Kyoung-jae [11] used support vector machines for prediction of stock price index as a time series problem. In this the effect of the value of the upper bound C and the kernel parameter δ^2 in SVM was investigated. It was observed that SVM actually performs better than back propagation and case based reasoning. This is due to the fact that SVM implements the structural risk minimization principle, which leads to better generalization than conventional techniques. Ping-Feng Pai and Chih-Sheng Lin developed a hybrid model, which is a combination of SVM and autoregressive moving average (ARIMA). This actually exploits the individual strengths of both models. Both ARIMA and SVM capture the data characteristics of linear and non-linear domains respectively. This hybrid model performs better when compared with these individual models alone.

Frank Cross [16] tries to find the relationship that could exist between stock price changes on Mondays and Fridays in the stock market. It has been observed that prices on Friday have risen more often than any other day. It has also been observed that on Monday the prices have least often risen compared to other days. Boris Podobnik [17] tries to find cross-correlation between volume change and price change. For the stock prices to changes it takes volume to move the stock price. They found two major empirical results. One is the power law cross-correlation between logarithmic price change and logarithmic volume change and the other is that the logarithmic volume change follows the same cubic law as logarithmic price change.

Many machine-learning techniques are used for predicting different target values [5,6,10]. This could be even to predict stock price. The genetic algorithm has been used for prediction and extraction important features [1,4]. Lot of analysis has been done on what are the factors that affect stock prices and financial market [2,3,8,9]. There are different ways by which stock prices can be predicted. One way is to reduce the complexity by extracting best features or by feature selection [7,13,14]. This approach will help us predict stock prices with better accuracy as the complexity reduces.

The people who invest money in the stock market usually focus only on a particular sector. For example people who want to invest money in Microsoft would not be interested in investing in a chemical industry as they cannot usually have knowledge about two different sectors. Only beginners would be interested in doing something weird like that. Thus the objective of this project is finding the relation between different companies of the same sector so that we can predict stock prices using different machine learning techniques.

II. Feature extraction based prediction

In this project we are trying to predict the highest price of the stocks of a particular company on everyday basis. There are a total of eight companies used for this experiment. These are Adobe, Apple, Google, IBM, Microsoft, Oracle, Sony and Symantec. For each company six different attributes are used. The highest stock prices for next day of these companies will be predicted using different machine learning techniques. For predicting the stock price of each company we are using eight different feature extraction techniques. These eight feature extraction techniques are explained below:-

1) Top 3 companies:-

In this type of feature extraction we are predicting the stock price of each company by finding a relation between different companies. This inter-relation between these companies will be used to predict the stock prices of a particular company in a better way. Each company's data will be individually used to predict each other company's stock price. The top three companies, which can predict a particular company with higher accuracy, will be used together to predict the stock price of a particular company. Along with the top three companies the NASDAQ index and the S&P 500 index would be used in each case.

2) Previous 3 days:-

In this the data of the previous three days for the company whose highest price we are trying to predict is used.

3) Previous 5 days:-

Similarly in this the data of the previous five days for the company whose highest price we are predicting is used.

4) Top 7 attributes:-

In this the top 7 attributes are evaluated using the ReliefFAttributeEval feature selection method. ReliefF algorithm is an extension of Relief. It is not limited to two class problems and is more robust to deal with incomplete and noisy data. The idea of ReliefF is to evaluate partitioning power of attributes according to how well their values distinguish between similar instances. An attribute is given a high score if its values separate similar observations with different class and do not separate similar instances with the same class values. For this the data of all eight companies and NASDAQ index and the S & P 500 index were used.

5) Top 10 attributes:-

In this similarly the top 10 attributes are evaluated using the ReliefFAttributeEval feature selection method. For this also the data of all eight companies and Nasdaq and S & P 500 were used.

6) Volume + Company:-

In this the volume attribute of stock purchased for each of the companies and both stock indexes i.e. NASDAQ index and the S & P 500 indexes are used. Along with this, the data of the company whose highest stock price for next day we are predicting is used for prediction.

7) Nasdaq + S & P + Company :-

In this as the name suggests, we use the NASDAQ index, the S & P 500 index and the data of the company whose highest price we are trying to predict.

8) Company alone:-

In this only the data of the company whose highest price we are predicting is used. So total attributes used in this case are six. These are the opening price, closing price, highest price, lowest price, volume and adjusted closing price.

The different machine learning techniques used for the experiments are briefly explained below:-

1) Neural Network:-

It is inspired from biological neural networks. It consists of interconnected neurons, which process information using a connectionist approach. The network adapts itself according to the information flowing into the network and tries to predict the required data.

2) Sequential Minimal optimization (SMO):-

The sequential minimal optimization solves the QP problem without any extra matrix storages and without using numerical QP optimization steps at all. The SMO decomposes the overall QP problem into QP sub-problems. It is a linear classifier that tries to find the maximum margin i.e. the distance between the classifier and the nearest data points [10,11,12].

3) Bagging using sequential minimal optimization:-

Bagging is a popular re-sampling ensemble method that generates and combines a diversity of classifiers using the same learning algorithm for the base-classifier. The learning algorithm used in this case is sequential minimal optimization. In this a standard training dataset is used which generates new training datasets using sampling. Thus we can learn different models based on the new training datasets generated. These models are combined by averaging the output or by voting to

predict the desired output. In each model the learning algorithm used is sequential minimal optimization.

4) M5P:-

For M5P, a decision-tree induction algorithm is used to build a model tree. To build this tree model divide and conquer approach is used. Secondly, the tree is pruned back from each leaf. To avoid discontinuities between the sub-trees in the model it is smoothened by combining the leaf model prediction with each node along the path back to the root, smoothing it at each of these nodes by combining it with the value predicted by the linear model for that node.

III. Experimental Setup:-

The dataset used for this experiment consists of the stock data for the last five years. Six attributes for each company are used for prediction. These are the Opening price, closing price, highest price, lowest price, volume and adjusted closing price. The values of the NASDAQ and S&P 500 indexes for the last five years are also used. These indexes also have the same six attributes. So there are a total of sixty attributes used for the experiments.

The whole data is divided into three equal sized datasets. These three datasets are sequential. So we train using the first dataset and then use the second dataset for testing. Similarly we train using the second dataset and test using the third dataset.

IV. Results:-

The prices of the stocks were predicted using mainly the four machine-learning techniques mentioned above. The results obtained by these methods are analyzed as given below:-

1) Predicting stock prices using neural network:-

When the neural network is used to predict the highest price for each company, it is observed that the feature extraction from the Company alone performed the best compared with the other feature extraction methods. The results obtained for the different datasets is given in Figures 1 and 2.

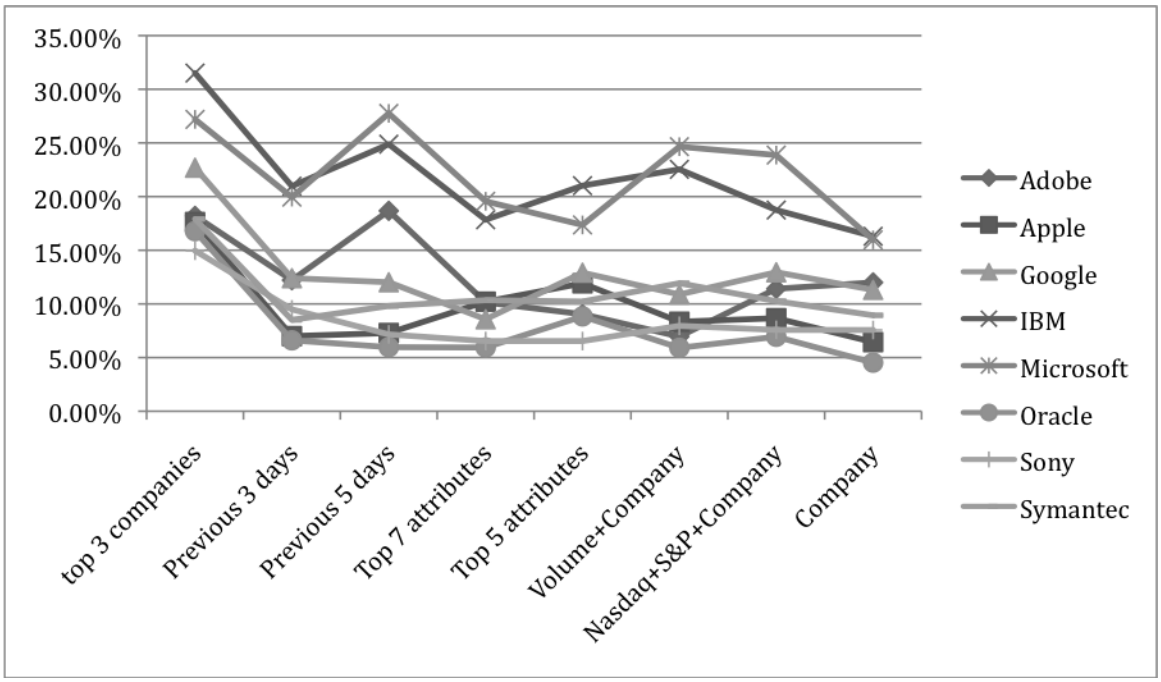

Fig 1:- This figure shows relative absolute error for predicting the highest price for eight companies using neural network. The second dataset is used as testing set in this case

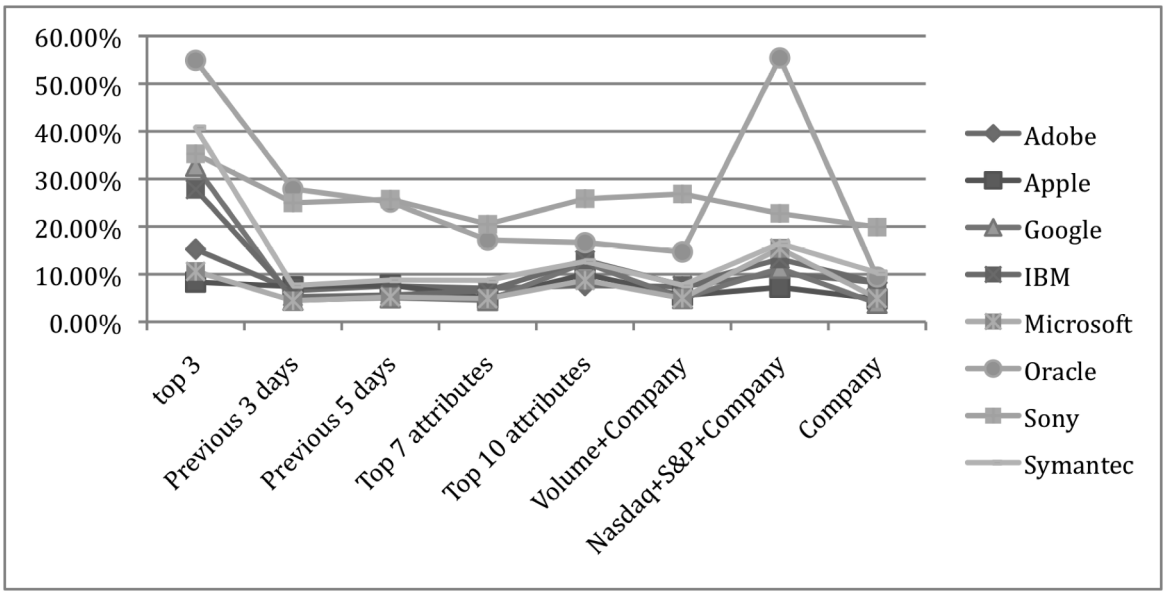

Fig 2:- This figure shows relative absolute error for predicting the highest price for eight companies using neural network. The third dataset is used as testing set in this case.

2) **Predicting stock prices using sequential minimal optimization (SMO):-**

When sequential minimal optimization is used to predict the highest price for each company, it is observed that the feature extraction methods Company + Volume and Company + NASDAQ +S & P has performed the best when compared with other feature extraction method. In this case these extraction techniques performed better than Company alone which was not the case for neural network. The other two are previous 3 days and previous 5 days. The results obtained for the different datasets are given in Figures 3 and 4.

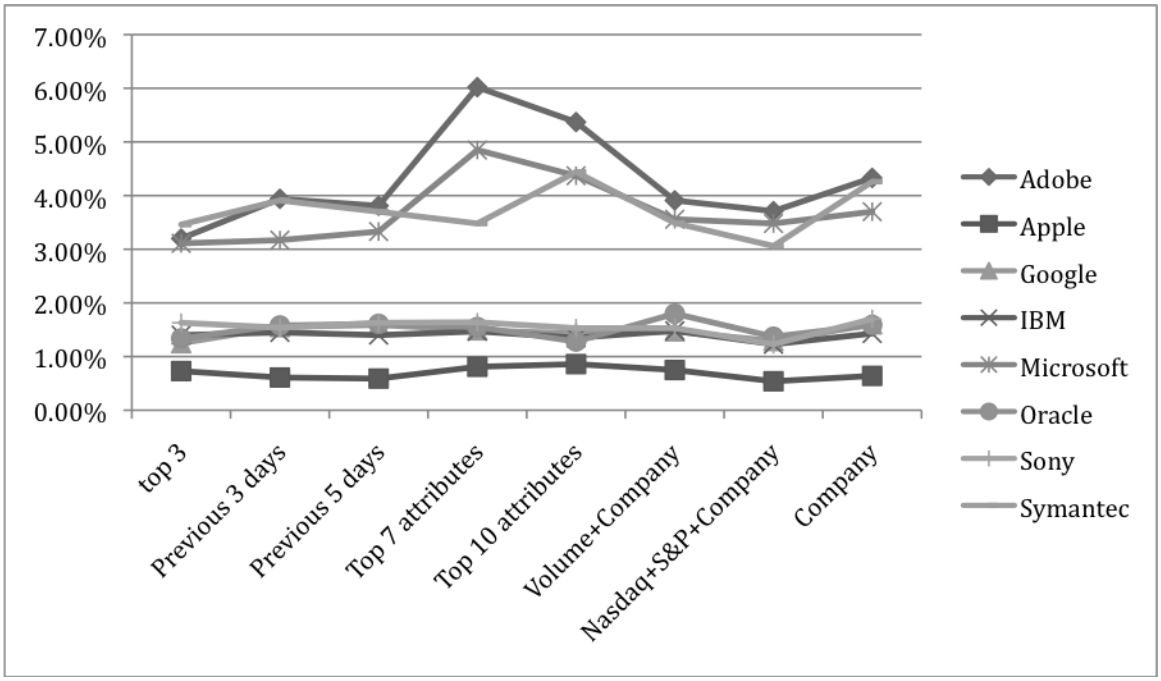

Fig 3:- This figure shows relative absolute error for predicting highest price for eight companies using sequential minimal optimization. The second dataset is used as testing set in this case.

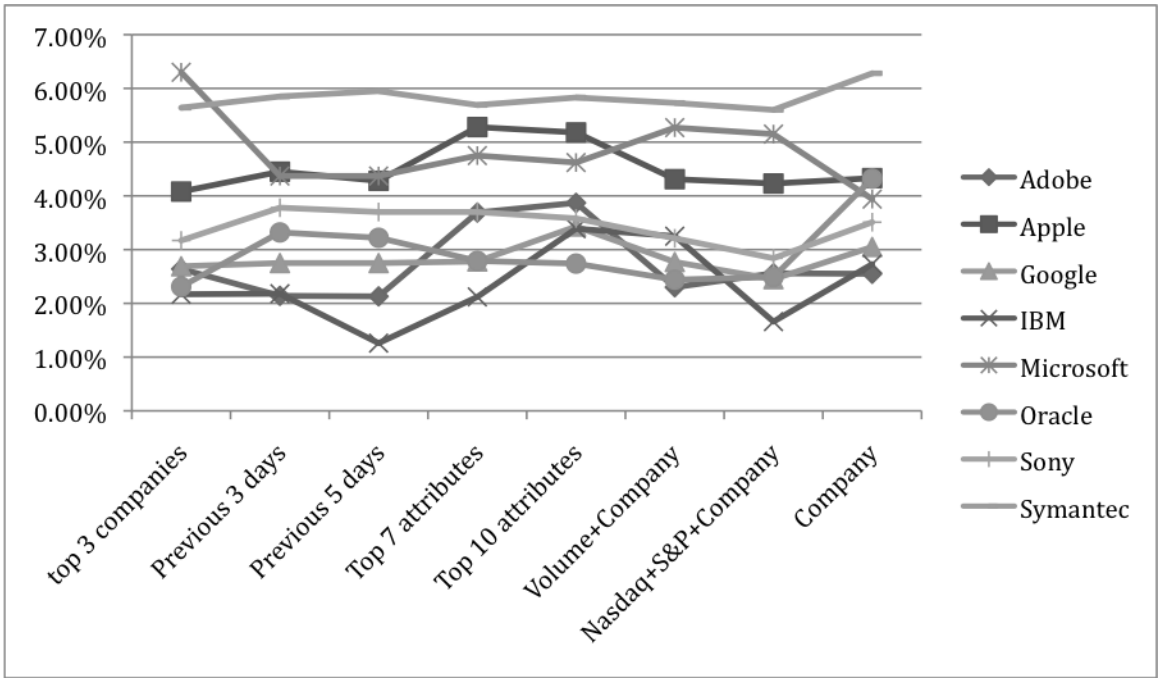

Fig 4:- This figure shows relative absolute error for predicting highest price for eight companies using sequential minimal optimization. The third dataset is used as testing set in this case.

3) **Predicting stock prices using bagging:-**

When bagging is used to predict the highest price for each company, it is observed that the feature extraction methods Company + Volume and Company + Nasdaq +S & P performed the best when compared with other feature extraction methods. Similar results were observed when we tried to predict the stock market using sequential minimal optimization. The results obtained for different datasets are given in Figures 5 and 6. The SMO algorithm was used internally to predict the stock price of each company.

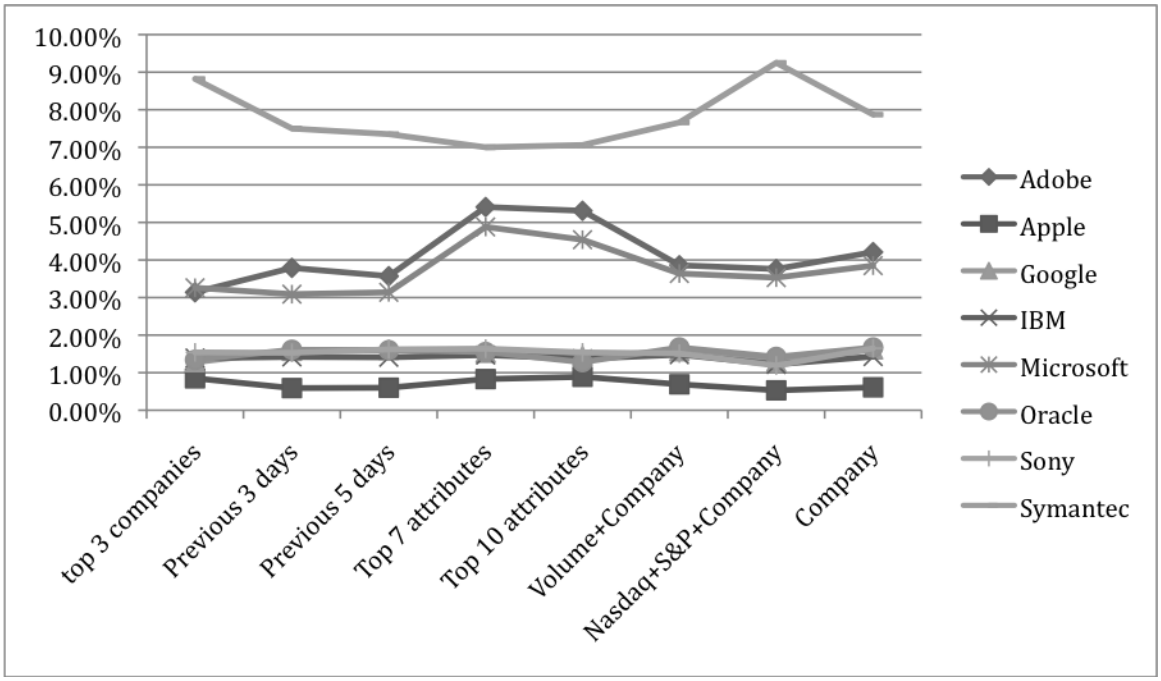

Fig 5:- This figure shows relative absolute error for predicting highest price for eight companies using bagging approach. The second dataset is used as testing set in this case.

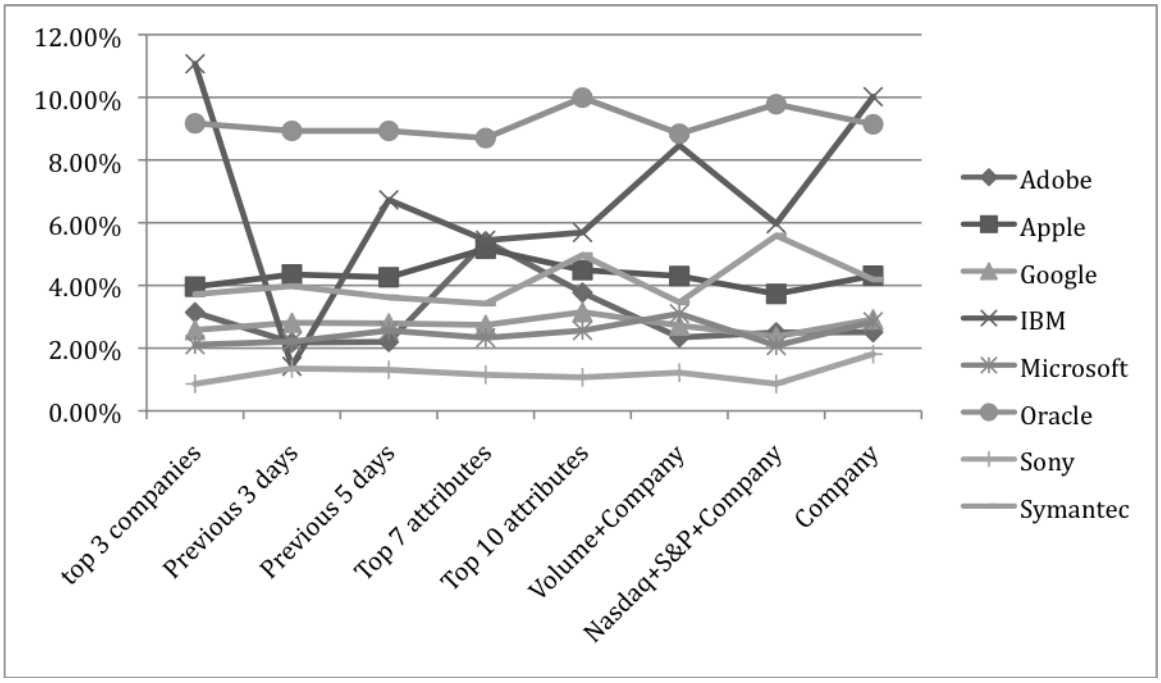

Fig 6:- This figure shows relative absolute error for predicting highest price for eight companies using bagging approach. The third dataset is used as testing set in this case.

4) **Predicting stock prices using M5P:-**

When M5P is used to predict the highest price for each company, it is observed that the feature extraction methods "Company + Volume" and "Previous 3 days" performed the best when compared with other feature extraction methods. The third best feature extraction method is Company alone. The results obtained for the different datasets are given in Figures 7 and 8.

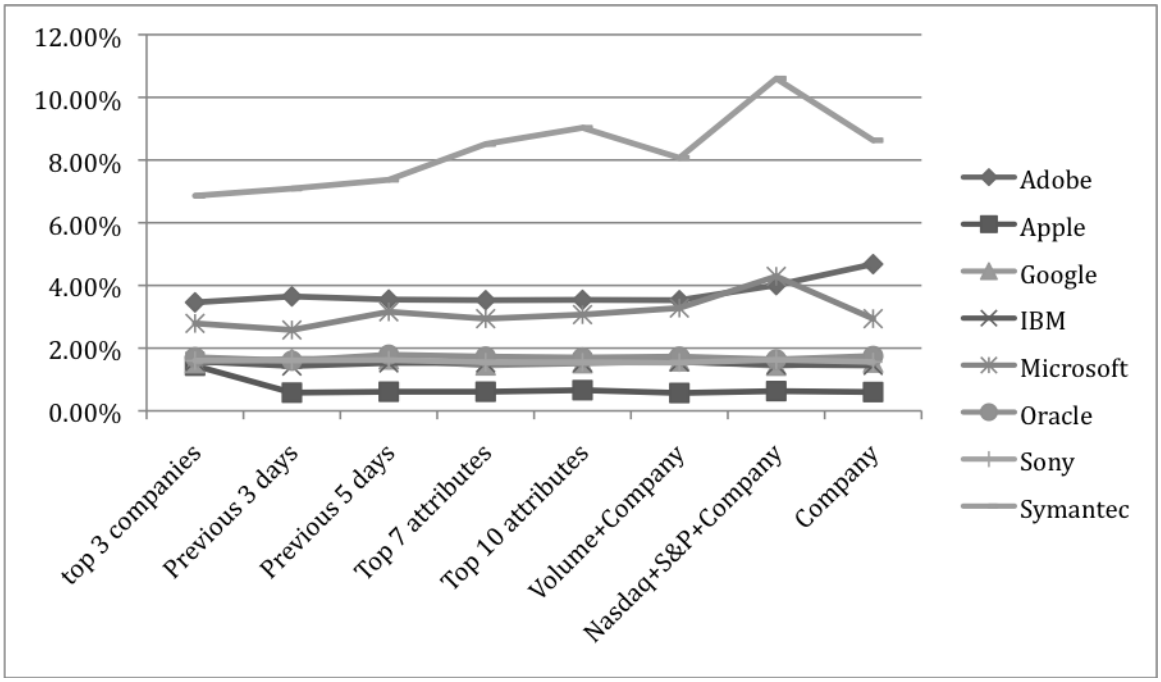

Fig 7:- This figure shows relative absolute error for predicting highest price for eight companies using M5P. The second dataset is used as testing set in this case.

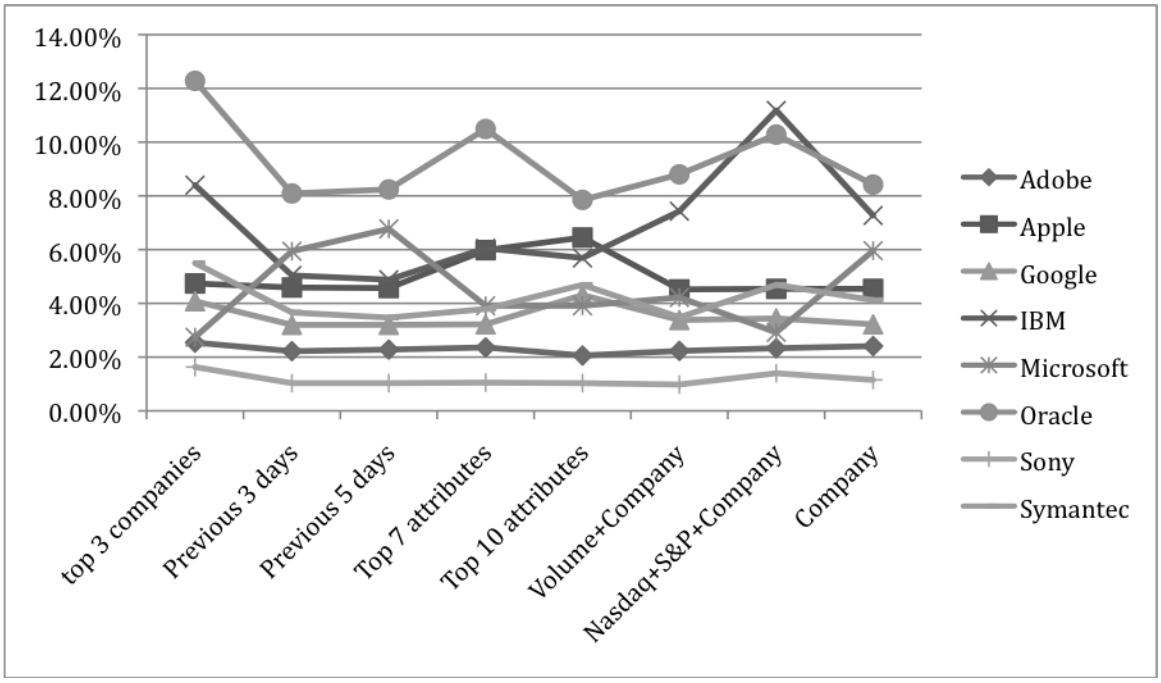

Fig 8:- This figure shows relative absolute error for predicting highest price for eight companies using M5P. The third dataset is used as testing set in this case.

V. Conclusion:-

It can be observed from Figures 1 through 8 that the best machine learning techniques for predicting the stock price are sequential minimal optimization and bagging using SMO. Using these methods the best features extracted to predict stock prices are "Volume + Company" and "Nasdaq + S & P +Company". Thus when the volume attributes of all eight companies are used along with individual data of the company whose price we are trying to predict will represent "Volume +Company". Similarly the whole data for Nasdaq, S & P 500 and individual companies data will represent "Nasdaq + S & P +Company".

Generally neural networks perform well but in this case the performance is not satisfactory. Also the results obtained using neural networks do not match the trends of the remaining learning techniques. Hence proper tuning of the different parameters is required so that neural networks may perform well like the other three learning algorithms.

REFERENCES:

[1] Abdüsselam Altunkaynak, Sediment load prediction by genetic algorithms Advances in Engineering Software, Volume 40, Issue 9, September 2009, Pages 928–934

[2] Hyunchul Ahn , Kyoung-jae Kim[b]. Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, Applied Soft Computing,Volume 9, Issue 2, March 2009, Pages 599–607

[3] Po-Chang Ko, Ping-Chen Lin. An evolution-based approach with modularized evaluations to forecast financial distress, Knowledge-Based Systems,Volume 19, Issue 1, March 2006, Pages 84–91

[4] 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, 2000.

[5] Chung-I Chou, You-ling Chu and Sai-Ping Li . Evolutionary Strategy for Political Districting Problem Using Genetic Algorithm, Lecture Notes in Computer Science, 2007, Volume 4490/2007, 1163-1166.

[6] Guangwen Li, Qiuling Jia, Jingping Shi , The Identification of Unmanned Helicopter Based on Improved Evolutionary Strategy, Intelligent Computation Technology and Automation, 2009. ICICTA '09. Second International Conference on, 205-208

[7] Chih-Fong Tsai , Yu-Chieh Hsiao . Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches, Decision Support Systems, Volume 50, Issue 1, December 2010, Pages 258–269.

[8] Xiaodong Li, Chao Wang, Jiawei Dong, Feng Wang, Xiaotie Deng, Shanfeng Zhu. Improving stock market prediction by integrating both market news and stock prices

[9] F. Mokhatab Rafiei, Manzari, S. Bostanian, Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence, Expert Systems with Applications**,** Volume 38, Issue 8, August 2011, Pages 10210–10217

[10] George S. Atsalakis, Kimon P. Valavanis . Surveying stock market forecasting techniques – Part II: Soft computing methods, Expert Systems with Applications**,** Volume 36, Issue 3, Part 2, April 2009, Pages 5932–5941

[11] Kyoung-jae Kim. Financial time series forecasting using support vector machines, *Neurocomputing*, Volume 55, Issues 1-2 (September 2003), Pages 307-319.

[12] Ping-Feng Pai, Chih-sheng Lin. A hybrid ARIMA and support vector machines model in stock price forecasting, Omega **,**Volume 33, Issue 6, December 2005, Pages 497–505.

[13] Kyoung-jae Kim, Won Boo Lee. Stock market prediction using artificial neural networks with optimal feature transformation. Neural Computing and Applications (2004), Volume: 13, Issue: 3, Publisher: Citeseer, Pages: 255-260

[14] 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, Volume 19, Issue 2, August 2000, Pages 125–132.

[15] Ajith Abraham, Baikunth Nath and P. K. Mahanti. Hybrid intelligent systems for stock market analysis. Proceedings of the International Conference on Computational Science Part 2, Pages 337-345.

[16] Frank Cross. The behavior of stock prices on Fridays and Mondays. Financial Analyst Journal Vol. 29 No. 6, pages 67-69.

[17] Boris Podobnik, Davor Horvatic, Alexander M. Peterson and Eugene Stanley. Cross-correlations between volume change and price change. Proceedings of the National Academy of Sciences of the United States of America, Vol. 106, No. 52, pp. 22079-22084, December 2009