EXTRACTING THE BEST FEATURES FROM MULTI-COMPANY STOCK DATA TO IMPROVE STOCK PRICE PREDICTION

by

GANESH SURESH BONDE

(Under the Direction of Dr. Khaled Rasheed)

ABSTRACT

To predict stock price is always a challenging task. The stock prices are dependent on many factors. Much research has been done in this field but it has been difficult to find the right features, which will help predict the prices with high accuracy. When you try to analyze stock price using historical stock data, it has many attributes (feature set) containing useful as well as redundant features. Thus there is a need to remove the unwanted stock information. In this research we first try to find the best features from the available company data as well as data about similar companies and stock indexes. The best-extracted features are then used to predict stock prices using different machine learning algorithms. Based on the results obtained in the previous experiments, we then implemented two new techniques for predicting stock prices. We used genetic algorithms and evolution strategies. The results obtained using these algorithms were promising. In each case the accuracy obtained was more than 70%. In this research, data of eight companies was used, each having six attributes. Also NASDAQ and S & P 500 data was used for predicting the stock prices.

INDEX WORDS: *Machine learning,stock market, sequential minimal optimization, bagging, genetic algorithm and Evolutionary Strategies.*

EXTRACTING THE BEST FEATURES FROM MULTI-COMPANY STOCK DATA TO IMPROVE STOCK PRICE PREDICTION

by

GANESH SURESH BONDE

B.E, PUNE UNIVERSITY - 2006

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2012

EXTRACTING THE BEST FEATURES FROM MULTI-COMPANY STOCK DATA TO IMPROVE STOCK PRICE PREDICTION

by

GANESH SURESH BONDE

Major Professor:	Dr. Khaled Rasheed
Committee:	Dr. Walter D. Potter
	Dr. Thiab R. Taha

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

ACKNOWLEDGEMENTS

I would like to thank Dr. Khaled Rasheed, for his suggestions and guidance about machine learning and evolutionary computation techniques. I would also like to thank Dr. Rasheed for continuous support throughout this research. It would have not been possible for me to get good results in predicting stock prices without Dr. Rasheed's proper guidance. I also thank Dr. Potter and Dr. Taha for being members of my committee. Above all, I am grateful to the divine force of the universe that helps me in every step I take.

To predict stock prices using machine learning is difficult. I chose this topic and thank Dr. Khaled Rasheed in supporting me to do this research.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS..iv

CHAPTER

1. INTRODUCTION..1

2. EXTRACTING THE BEST FEATURES FOR PREDICTING STOCK PRICES USING MACHINE LEARNING…...6

3. STOCK PRICE PREDICTION USING GENETIC ALGORITHM AND EVOLUTION STRATEGIES …..23

4. SUMMARY AND CONCLUSIONS...43

REFERENCES...46

CHAPTER 1

INTRODUCTION

The prediction of stock prices has always been a challenging task. 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 stock price prediction has become even more difficult than before.

Predicting stock price is a difficult task; hence a good analyst is not someone who is always right in predicting the stock prices but someone who has better accuracy on average, or someone who has better efficiency than others. In the stock market much information is available in short duration of time. Thus a fast response to this information will help us in predicting stock price. However only considering information of only one company in which we are interested may be a bad idea. Hence we should consider companies similar to it or which are in the same domain.

The available information in the stock market is huge. We cannot use all the information available in the market. Thus we should be able to remove the unwanted information from the data and use only the useful data or information to predict the stock price. This will help us in predicting the stock with better accuracy. The data consists of different attributes, which act as feature sets. Thus the data can contain many unwanted features (attributes). Thus there is a need to extract the best features from the information available at that moment. Hence we tried to extract the best features, which will help us in predicting the stock price.

For this research we used eight different companies to predict the highest stock price of each of these companies. For each company six attributes are used from the historical data. 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 are also used. In this research the last five years of historical data is used. This data is downloaded from the Yahoo finance website.

Once the best features are extracted, these features are used to predict the stock price using different machine learning algorithms. The machine learning techniques used in this research are neural network, sequential minimal optimization, bagging, genetic algorithms and evolution strategies.

Chapter 2 of this thesis explains how the best features from each company are extracted. The features selected were based on many different criteria. These are the top 3 companies, previous 3 days, previous 5 days , top 7 attributes, top 10 attributes, Volume + Company, NASDAQ + S&P 500 + company and company alone. Once the best features are extracted, these features are given to different algorithms for predicting the stock price. The different algorithms used are neural network, sequential minimal optimization and a bagging approach in which sequential minimal optimization was used internally. Relative Absolute Error (RAE) is the measure of accuracy used in this case. The RAE is the total relative error, which is normalized by dividing by the total absolute error which is the error of a simple predictor that predicts the average of the actual values. The overall goal to extract the best features is achieved in this part. Then depending on the results and the behavior of the data, the next research objective is identified.

Chapter 3 describes the research in which the genetic algorithm (GA) and evolution strategies (ES) are used to improve the efficiency in predicting the stock price. These algorithms are implemented after analyzing the results obtained in chapter 2. The classification function used is a sigmoid function. The evolutionary search finds the connection weights between each input and the sigmoid function. If the input attribute is of higher importance then automatically higher would be its connections weights. The accuracy is measured on the basis of the number of times the algorithm predicted the stock price correctly. It is found in each case that the accuracy was more than 70%.

Chapter 4 summarizes the research performed in this study and provides conclusions. It also suggests possible future research that could further improve the accuracy in predicting stock prices.

CHAPTER 2

EXTRACTING THE BEST FEATURES FOR PREDICTING STOCK PRICES USING MACHINE LEARNING.

Bonde,G and Khaled, R. Submitted to International Conference on Artificial Intelligence, 04/18/2012

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.

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 where δ^2 is the bandwidth of the Gaussian radial basis function. The kernel factor is calculated for high dimensional version to implement non-linear class boundaries. 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 the stock prices on Friday have increased 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]. One of its application where it can be used it 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 study 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 study the data of the previous five days for the company whose highest price we are predicting is used.

4) **Top 7 attributes:-**

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 the class with better accuracy. For this study 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 study also the data of all eight companies and Nasdaq and S & P 500 were used.

6) Volume + Company:-

In this study 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 study 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. QP is a special type of mathematical optimization problem. In this the problem is to optimize a quadratic function of several variables with respect to linear constraints on these variables. 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.

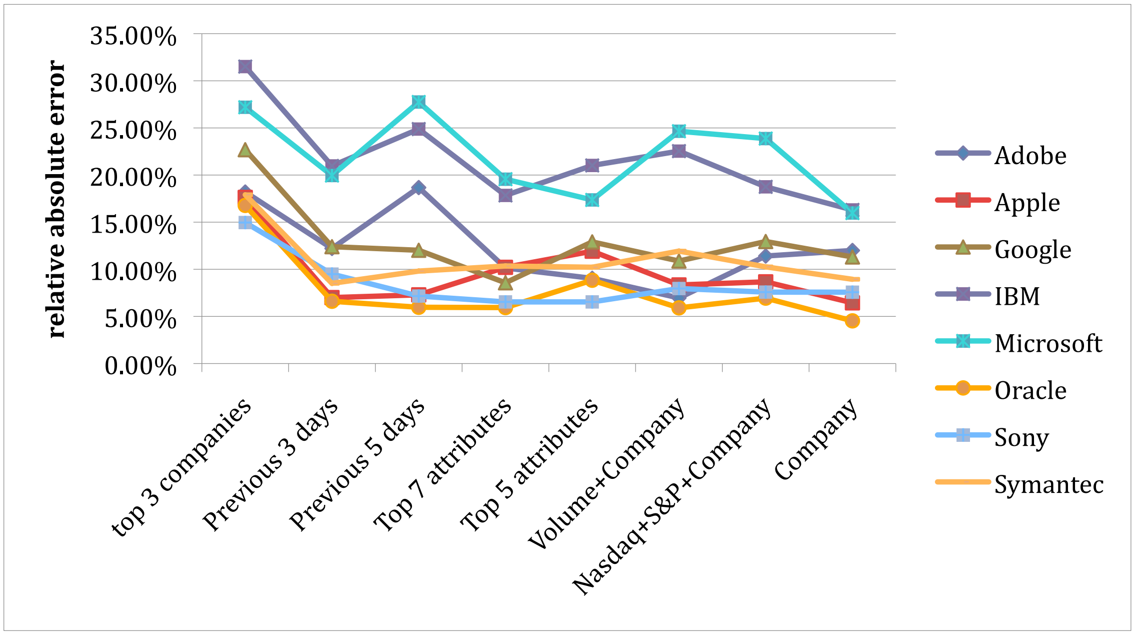

Fig 1:- This figure shows relative absolute error for predicting the highest price for eight companies using neural network. The second dataset is used as testing set in this case

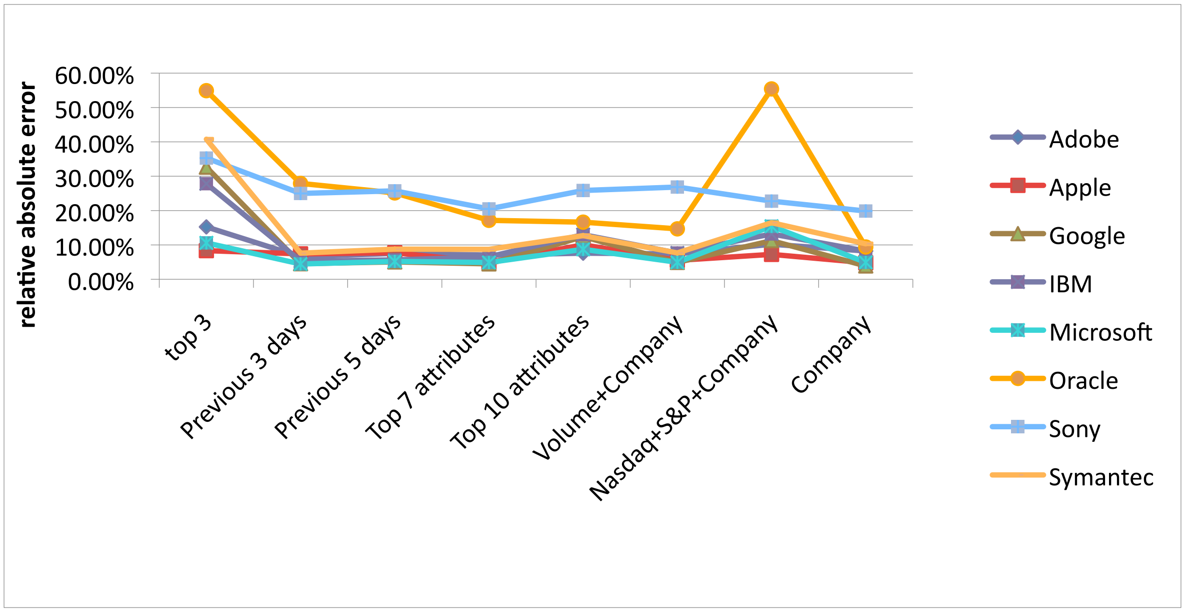

Fig 2:- This figure shows relative absolute error for predicting the highest price for eight companies using neural network. The third dataset is used as testing set in this case.

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.

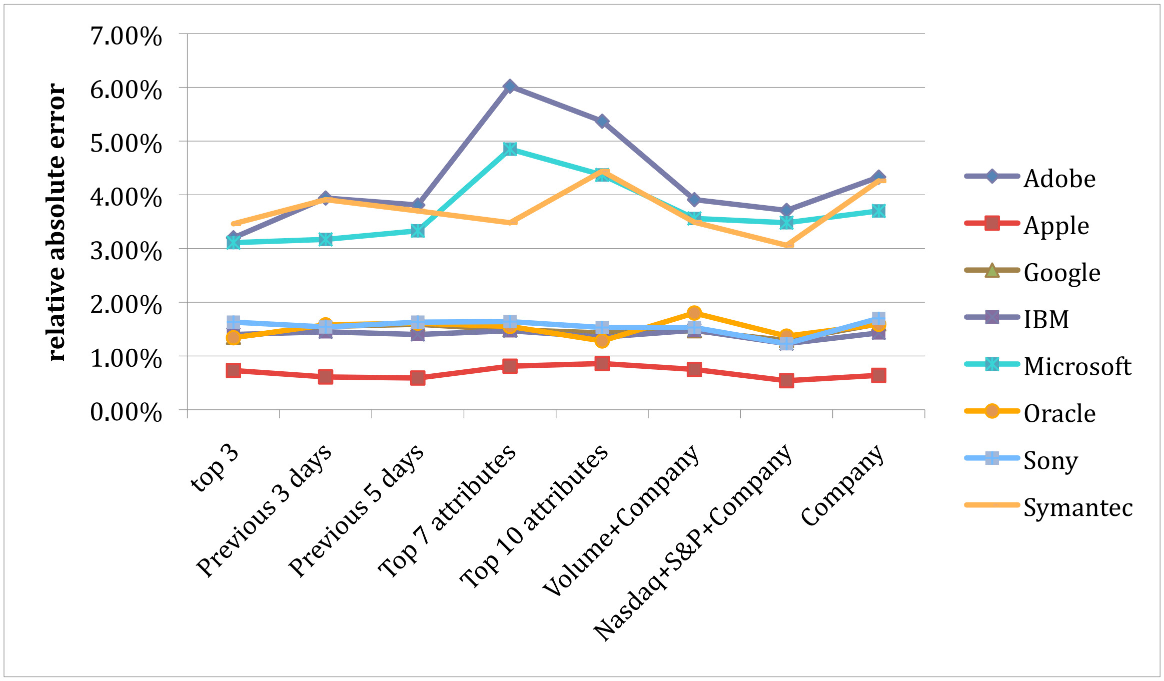

Fig 3:- This figure shows relative absolute error for predicting highest price for eight companies using sequential minimal optimization. The second dataset is used as testing set in this case.

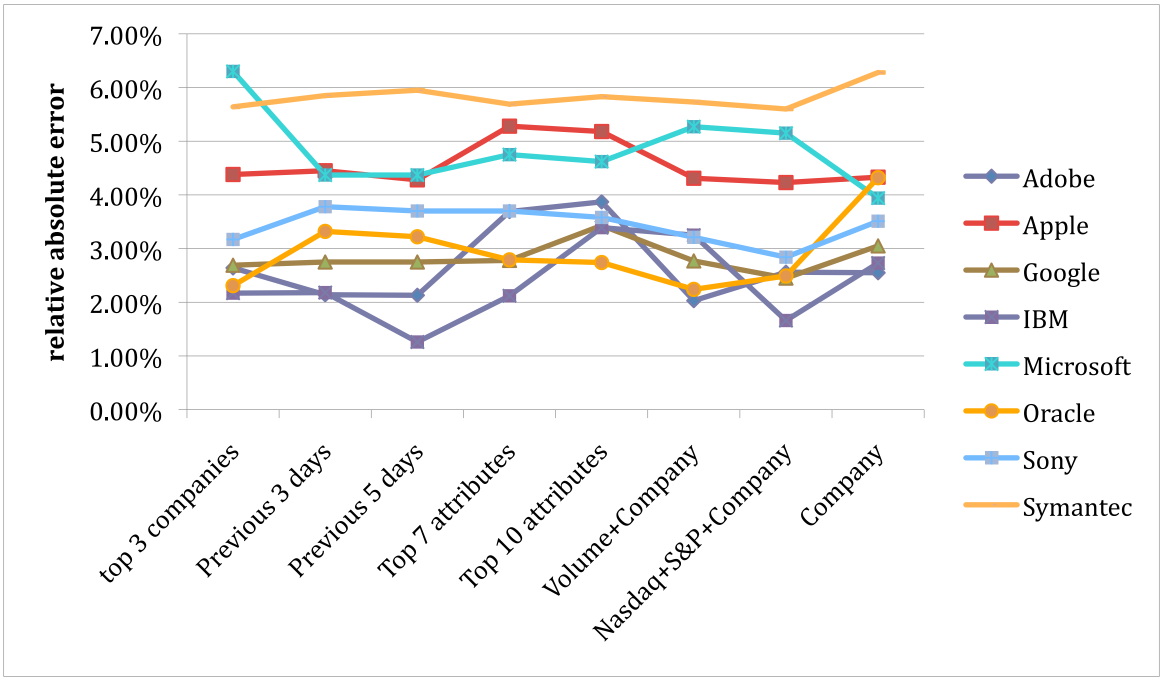

Fig 4:- This figure shows relative absolute error for predicting highest price for eight companies using sequential minimal optimization. The third dataset is used as testing set in this case.

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.

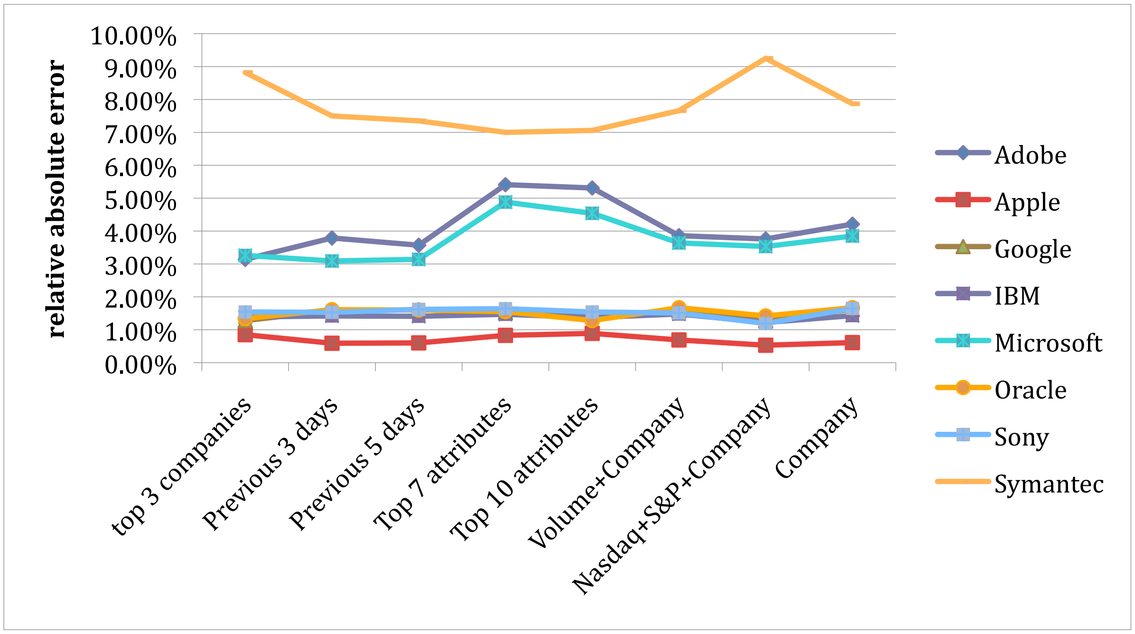

Fig 5:- This figure shows relative absolute error for predicting highest price for eight companies using bagging approach. The second dataset is used as testing set in this case.

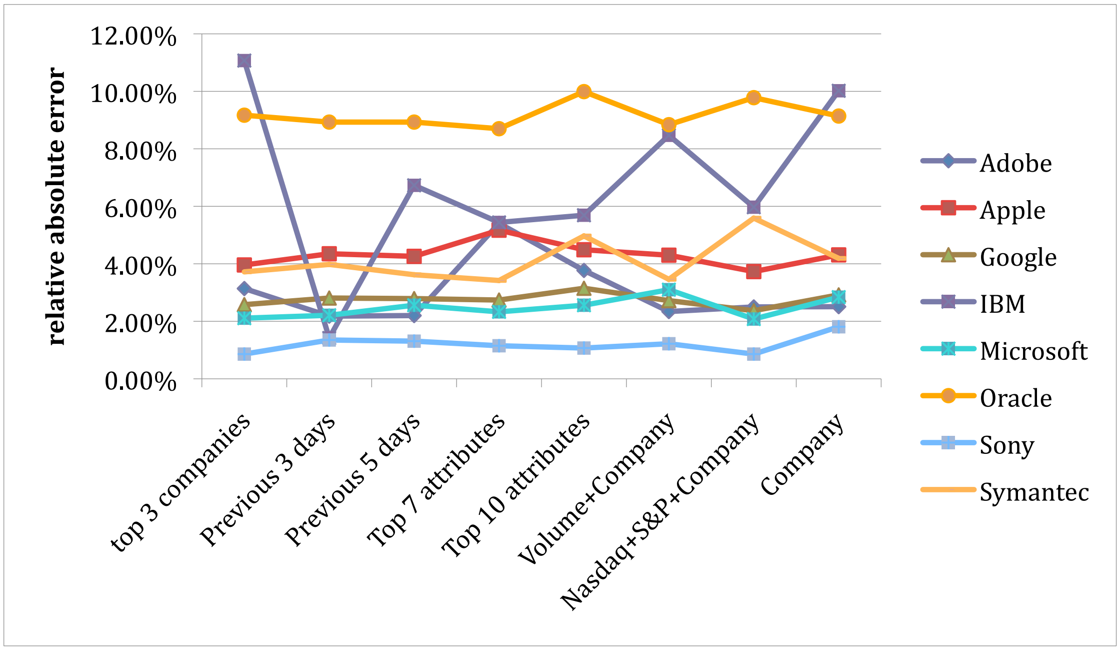

Fig 6:- This figure shows relative absolute error for predicting highest price for eight companies using bagging approach. The third dataset is used as testing set in this case.

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.

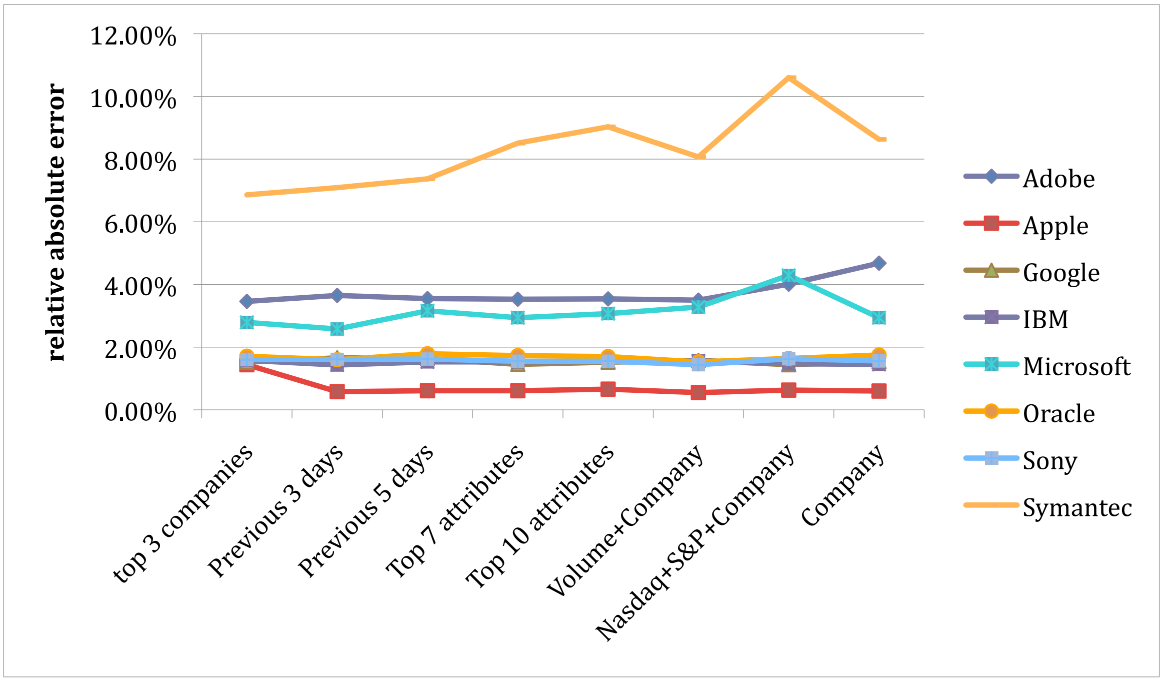

Fig 7:- This figure shows relative absolute error for predicting highest price for eight companies using M5P. The second dataset is used as testing set in this case.

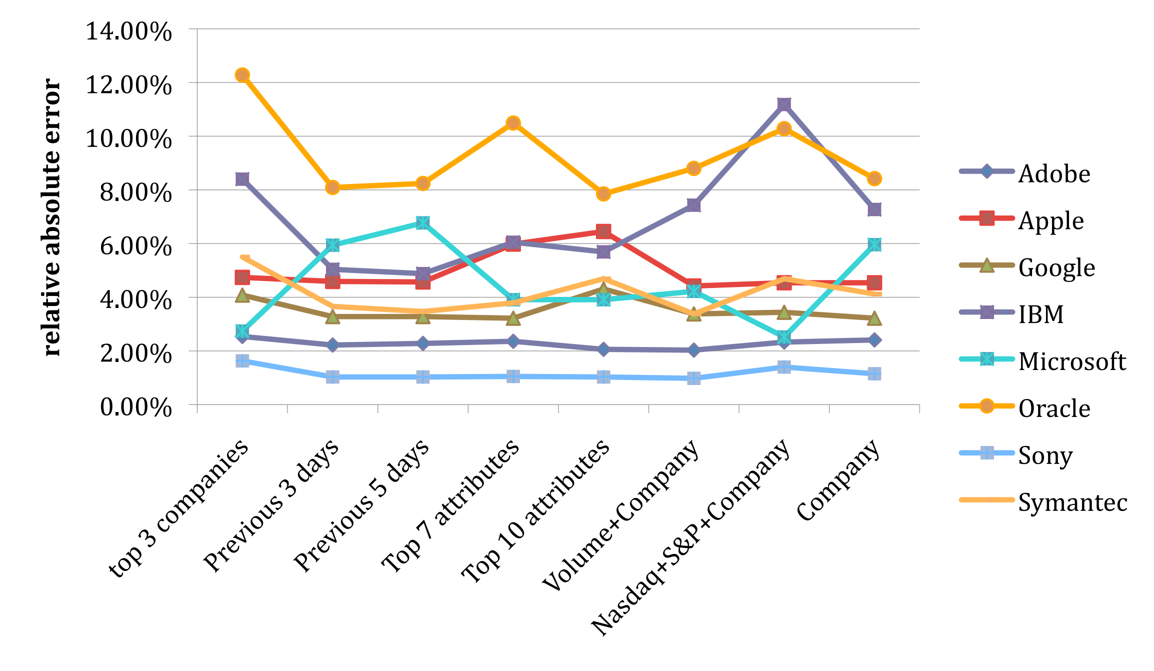

Fig 8:- This figure shows relative absolute error for predicting highest price for eight companies using M5P. The third dataset is used as testing set in this case

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

CHAPTER 3

STOCK PRICE PREDICTION USING THE GENETIC ALGORITHM AND EVOLUTION STRATEGIES

Bonde,G and Khaled, R. Submitted to International Conference on Genetic and Evolutionary Methods, 04/18/2012

Abstract:-

To many, the stock market is a very challenging and interesting field. In this paper we try to predict whether the prices of the stocks are going to increase or decrease on the next day. We are predicting the highest stock price for eight different companies individually. For each company six attributes are used which help us to find whether the prices are going to increase or decrease. The evolutionary techniques used for this experiment are genetic algorithms and evolution strategies. Using these algorithms we are trying to find the connection weight for each attribute, which helps in predicting the highest price of the stock. The input for each attribute is given to a sigmoid function after it is amplified based on its connection weight. The experimental results show that this new way of predicting the stock price is promising. In each case the algorithms were able to predict with an accuracy of at least 70.00%. Since this approach is new any further study in this field can definitely give better results.

I. Introduction

The prediction of stock prices has always been a challenging task. 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 prices prediction has become even more difficult than before.

These days stock prices are affected due to many reasons 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 stock brokers have to buy and sell stocks within the shortest possible time. Thus a time span of even few hours between buying and selling is not unusual.

To invest money in the stock market we need to have an idea whether the prices of stocks are going to increase or decrease on the next day. Thus in this project we are trying to predict whether the highest price of a stock is going to increase or decrease on the next day. In this paper we are trying to predict the price of stocks of eight different companies. For each company we are predicting whether its highest price is increasing or decreasing next day. Thus it is a classification problem with only two classes involved. Thus we have tried to make the problem as simple as possible.

Kyoung-jae Kim and Won Boo Lee [13] developed a feature transformation method using genetic algorithms. This approach reduces the dimensionality of the feature space and removes irrelevant factors involved in stock price prediction. This approach performed better when compared with linear transformation and fuzzification transformation. This GA based transformation looks promising when compared with other feature transformations. Another research done on genetic algorithms (GAs) by Kyoung-jae Kim [4] again to predict stock market is to use a GA not only to improve the learning algorithm, but also to reduce the complexity of the feature space. Thus this approach reduces dimensionality of the feature space and enhances the generalizability of the classifier. Also Ajith Abraham [15] developed a hybrid intelligent system, which consists of a neural network, fuzzy inference systems, approximate reasoning and derivative free optimization techniques. That system also gives promising results but was not compared with any other existing intelligent systems.

Frank Cross [16] tries to find relationships 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] tried to find cross-correlation between volume change and price change. For the stock prices to change, it takes volumes to move the stock price. They found two major empirical results. One is power law cross-correlation between logarithmic price change and logarithmic volume change and the other is that the logarithmic volume change is similar to the same cubic law as logarithmic price change. The logarithmic volume change investigates possible relations between price changes and volume changes.

Abdüsselam Altunkaynak [1] used a genetic algorithm for the prediction of sediment load and discharge. Not many have tried to use only genetic algorithms to predict stock prices. Since the genetic algorithm can perform reasonably well in many cases there has to be a way to predict stock price using GA as well. Hyunchul Ahn [2] suggested that the genetic algorithm can be used to predict in financial bankruptcy. We have also tried to use a similar approach to predict the stock. The method used in this experiment is completely relatively novel and looks very promising.

Many machine-learning techniques are used for predicting different target values [5,6,10]. One of its applications is 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 [7,11,12,13,14]. This approach will help us predict stock prices with better accuracy as the complexity reduces.

In this project the method used for predicting the highest price is relatively novel. We try to find the connection weights of each attribute used for predicting the stock price. There are a total of six attributes used for each company. Hence we use six connection weights, one for each attribute. Each connection weight value defines the contribution given by each attribute in predicting the stock price. For example it could happen that the volume attribute contributes more than other attributes. Thus more importance is given to that attribute. Thus obviously this attribute will have a higher connection weight compared to other attributes. This concept is explained in more detail below.

Feature discretization of each input:-

The main concept in discretization is that we try to normalize each input attribute with respect to each other attribute. Thus we try to find the connection weight for each attribute that decides on the contribution given by that attribute. Each attribute after multiplying by the connection weight is given to a sigmoid function. This function is used to classify the next stock price into increasing or decreasing class.

The sigmoid function in terms of mathematical expression is given below. It is used for classification of the problem into two classes. This function will classify each input into mainly one class predicting the stock price to increase while other class will predict the stock rice to reduce. So it can be used for binary classification problems.

$$P(t) = \frac{1}{1+e^{-t}} \qquad (1)$$

The two evolutionary techniques used for predicting the stock price are given below:-

Genetic Algorithm:-

A genetic algorithm (GA) is a search technique used in computing to find exact or approximate solutions to search and optimization problems. Genetic algorithms are a particular class of evolutionary computation that uses techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover. A genetic algorithm finds the potential solution to a specific problem as a simple chromosome like data structure so as to preserve the critical information.

Its implementation begins with the selection of a population of chromosomes, which is a set of solutions to problems that could occur for a particular scenario. One evaluates its fitness and then does its reproduction to get better solutions with respect to the target problem. The chromosomes, which represent better solutions, are given more chance for reproduction than those which represent poorer solutions. This process continues for a number of generations after which we get the optimal solution.

The operators used for this experiment are two-point crossover and creep mutation. The crossover is a genetic operator used to vary chromosome gene structure where gene information is interchanged between selected parents by selecting two points in the gene structure of each parent.

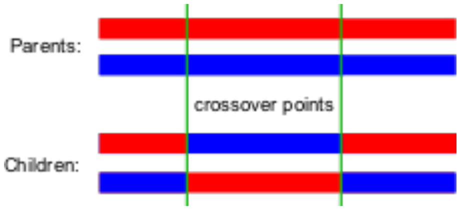

Figure 1. Two point crossover

The creep mutation used works by adding a small value to each gene with probability p. The population is selected using roulette wheel selection method. In this method, the fitness assigned to each individual is used for the selection process. This fitness is used to associate a probability selection with each individual. This can be given as below:-

$$P_i = \frac{f_i}{\sum_{j=1}^{N} f_j} \qquad (2)$$

Where *fi* is the fitness of the ith individual and N is the population size.

Evolution Strategies:-

The evolution strategy (ES) is also an idea inspired by concepts of adaptation and evolution. This type of algorithm is mainly used for continuous parameter optimization. The representation of the gene is vector. The intermediate recombination technique is used in this algorithm. In this study, the selected parent values are averaged to give the child and one of the parent is selected randomly so that two individual can go to the next generation.

The algorithm for evolutionary strategies is given below:

1. Randomly create an initial population of individuals.

2. From the current population generate offspring by applying a reproduction operator (described below).

3. Determine the fitness of each individual.

4. Select the fittest individuals for survival. Discard the other individuals.

5. Proceed to step 2 unless the number of generations have been exhausted.

In this experiment we are using a (μ , λ)-ES strategy in which the parents (candidate solutions) produce offspring (new solutions) by mutating one or more problem parameters. Offspring compete for survival; only the best (i.e., those with the highest fitness) will survive to reproduce in the next generation. If done properly, the population will evolve towards increasingly better regions of the search space by means of reproduction and survival of the fittest.

The mutation technique used is based on a Gaussian distribution requiring mainly two parameters the mean ξ and the standard deviation σ. In this study, small amounts of f(x) are randomly calculated using the Gaussian distribution N(ξ, σ) where ξ is the mean and σ is the standard deviation . This probability distribution function is given as

$$f(x)= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (3)$$

The new value of x is calculated as the sum of previous gene value and some small random value calculated using the above equation.

Xnew = Xold + N(ξ, σ) (4)

where ξ=0 and σ=1.

I. Experimental Setup

Dataset used:

The dataset used for this experiment consists of data for the last five years. A total of six attributes for each company are used for prediction. These are opening price, closing price, highest price, lowest price, volume and adjusted closing price. The eight companies used for this experiment are Adobe, Apple, Google, IBM, Microsoft, Oracle, Sony and Symantec.

Two datasets are used for the experiment. One training dataset is used for finding the connection weights for each attribute used. We used another testing dataset so that we can verify the result. Thus we can check if over fitting is occurring or not. The results obtained actually showed that no over fitting occurred.

The problem is represented using floating point so each connection weight used for a particular attribute is a floating point number. The fitness used in this problem is the number of times the connection weights result in predicting stock price correctly. So if it was able to predict the stock price correctly in 500 data points, then its fitness is 500. There are a total of 620 data entries for each dataset, which we need to predict. We first use the training dataset to find the exact connection weight for each attribute and then using these connection weights we try to predict the testing data. The different parameter settings for each algorithm are given below:

The parameter settings for the Genetic algorithm are:-

No.	**Parameters**	**Values**
1	Population Size:	100
2	Crossover Probability:	0.5
3	Mutation Probability:	0.013
4	Selection:	Roulette Wheel
5	Stopping Criteria:	1000 generations

Chart 1: Parameter settings for the genetic algorithm

The parameter settings for the Evolution strategy algorithm are given below:-

No	Parameters	Values
1	Population Size with (μ , λ)-ES strategy	20-100
2	Crossover Probability:	0.6
3	Mutation Probability:	0.015
4	Selection:	Roulette Wheel used only for initial population.
5	Stopping Criteria:	1000 generations

Chart 2: Parameter settings for the evolutionary strategies

I. Results

Tables 1 and 2 show the optimal connection weights used for predicting stock price in each algorithm. Table 3 shows the best fitness values evaluated for each company. Table 4 shows the accuracy of the algorithm to predict the highest price. The connection weights are calculated using the training dataset and is tested on the testing dataset. This protects against any over-fitting occurring in the model. From the results shown in Table 3 and 4 it can be seen that over-fitting is not occurring. The fitness also indicates the number of times it actually predicted the stock price correctly. The total number of entries present in each set is 620.

It can be seen from Table 4 that we were able to predict the stock price with considerable accuracy. The search space for this problem is very large. This is because the connection weight can range from zero to even a million or more. Since we have restriction on space search we have kept the upper end to be 1000 only for floating representation.

From table 4 it can be seen that the connection weight evaluated for each attribute do not get over-fitted. In fact in some cases the accuracy for prediction is higher for testing data than training data. The highest accuracy obtained using the genetic algorithm is 73.87% and using the evolutionary strategies is 71.77%.

Company	Open price	Closing price	Highest price	Lowest price	Volume	Adjusted closing price
Adobe	995.0	10.0	27.0	83.0	929.0	38.0
Apple	98.0	12.0	85.0	18.0	30.0	17.0
Google	89.0	12.0	18.0	15.0	87.0	21.0
IBM	87.0	5.0	39.0	44.0	71.0	23.0
Microsoft	1212.0	135.0	223.0	138.0	218.0	148.0
Oracle	963.0	1.0	24.0	18.0	989.0	28.0
Sony	921.0	7.0	54.0	37.0	975.0	38.0
Symantec	976.0	8.0	23.0	18.0	55.0	2.0

Table 1: Connection weights for each company using the genetic algorithm.

Company	Open price	Closing price	Highest price	Lowest price	Volume	Adjusted closing price
Adobe	804.0	36.0	767.0	18.0	601.0	727.0
Apple	309.0	20.0	116.0	8.0	158.0	111.0
Google	890.0	15.0	27.0	46.0	43.0	830.0
IBM	247.0	23.0	35.0	8.0	907.0	72.0
Microsoft	285.0	5.0	70.0	42.0	24.0	183.0
Oracle	842.0	1.0	769.0	7.0	103.0	281.0
Sony	856.0	9.0	861.0	44.0	854.0	42.0
Symantec	778.0	13.0	161.0	302.0	938.0	23.0

Table 2: Connection weights for each company using the evolutionary strategy.

Company	Fitness Value Using GA		Fitness using Evolutionary Strategy	
	Training data	Testing data	Training data	Testing data
Adobe	447	454	450	434
Apple	457	439	460	445
Google	465	430	462	435
IBM	438	439	452	442
Microsoft	467	436	472	440
Oracle	445	452	434	444
Sony	412	431	421	441
Symantec	440	458	431	439

Table 3: The best fitness calculated for each company.

\

Company	Fitness Value Using GA		Fitness using Evolutionary Strategy	
	Training data	Testing data	Training data	Testing data
Adobe	72.09%	73.22%	72.58%	70.00%
Apple	73.70%	70.80%	74.19%	71.77%
Google	75.00%	69.35%	74.51%	70.16%
IBM	70.64%	70.80%	72.90%	71.29%
Microsoft	75.32%	70.32%	76.12%	70.96%
Oracle	71.77%	72.90%	70.00%	71.61%
Sony	66.45%	69.51%	67.90%	71.11%
Symantec	70.96%	73.87%	69.51%	70.80%

Table 4: The accuracy with which the stock price was predicted for each company.

I. Conclusion and Future Work

The new method of predicting stock prices using the genetic algorithm and evolutionary strategies looks promising. It was found that the genetic algorithm and evolution strategies have performed almost evenly. The best accuracy found using the genetic algorithm was 73.87% and using evolutionary strategies was 71.77%. The genetic algorithm was able to predict better than the evolutionary strategies in five cases. The evolutionary strategy reached an accuracy of 70% or better in all cases.

We used two different datasets for predicting the stock prices. The first one acts as training set and the other acts as testing set. This division is required so that we can test if over-fitting is occurring or not. The results show that over-fitting has not occurred.

There are many aspects we can consider in the future. We need to include more attributes to predict stock prices. The six attributes used are very similar to each other hence we need more attributes, which are not similar but affect the prices.

We can try different activation functions for classification. Thus instead of using the sigmoid function we can use some other function.

This method can be compared with other popular algorithms used for stock price prediction such as neural networks and support vector machines.

Future Work:-

The evolutionary algorithms used for this experiment looks very promising. Therefore, further research is required in this field. We must use attributes of other companies to predict the prices to check whether they help in predicting the prices. Thus we can use only those company's data, which will help in predicting the data in a better way. There is a high chance that the accuracy for prediction will be above 80.0% if we used other companies' data also instead of using just individual company's data.

Since the results obtained are above 70.0% in every case then we can test the performance on real time data as well. This will give us an idea whether only historical data is good enough to predict data or not. If not, then we need to find the factors other than historical data which affect the prices. This information can also be fed to the algorithms we used for this experiment. There is a high chance that the accuracy will increase.

The companies used in this experiment were big companies. We can check the performance of those algorithms on small size companies as well.

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

CHAPTER 4

SUMMARY AND CONCLUSIONS

The goal of this research was to extract the best features from the available data and try to predict stock prices with better accuracy. We used eight different companies to predict the highest price of each company. We also used all the company's data for predicting the stock prices. We found that information of other companies also affects the stock price of a particular company. It was not useful for us to use only the company's data alone. When we used volume information of other companies, the results obtained were better than previous ones.

We came to know that NASDAQ and S & P 500 were also useful features that helped in predicting the stock prices. It can be seen that many combinations of available information of stock market data can be tried. We have a huge amount of information available in the market. We have to logically prioritize the data to be used. For example we can use countries currency value, which changes on daily basis. We can also use petrol, diesel or similar prices which might affect the stock price.

Thus there are many things that can be done in the future in this field. We need information or news which would be useful to predict the price. We can also extract the news announced on the news websites. This information may be crucial but to extract only the wanted information would be a challenging task. Natural Language Processing is probably the way to go about it. This is also a big field, which can be explored. At last the financial data and the news can be combined together to form a hybrid model. This hybrid model will probably give better results.

In the first part of the research we realized that the extracted features improve the efficiency in predicting stock price. Thus we implemented the genetic algorithm and the evolutionary strategies which would first find the importance of each attribute and assign a connection weight to it. If a particular attribute is of higher importance, we get its connection weight to be higher. Due to this kind of implementation we were able to achieve 70% accuracy in each case. This is a novel method and more research is required to be done in this field.

There are many aspects we can consider in the future. We need to include more attributes to predict stock prices. The six attributes used are very similar to each other hence we need more attributes, which are not similar but affect the prices.

We can try different activation functions for classification. Thus instead of using the sigmoid function we can use some other function.

Last, the research done on feature extraction for stock price has given us positive results and we need to explore this field more so that we can get better accuracy than present.

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