Stock price prediction using genetic algorithms and evolution strategies

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:-

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.

Keywords:

Machine learning,stock market, genetic algorithm, Eovolutionary Strategies.

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 follows the same cubic law as logarithmic price change.

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 novel and looks very promising.

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 [7,11,12,13,14]. 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 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. The summation of 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 when we do not have detailed information of the input we are trying to predict. This function will classify each input into mainly two classes. 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.

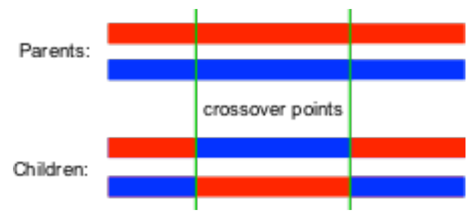

Figure 1. Two point crossover

The creep mutation used works by adding a small value to each gene with probability p. The selection method used to select the population is roulette wheel selection. 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 f_i 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 the selected parent values are averaged to give the child and one of the other parents 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 small amounts of f(x) are randomly calculated using the Gaussian distribution N(ξ, σ). This 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.

$$Xnew = Xold + N(\xi, \sigma) \qquad (4)$$

where ξ=0 and σ=1.

II. 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 representation for the problem is 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

III. 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.

IV. Conclusion and Future Work

The novel 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 can even try to 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