Stock price prediction using genetic algorithms And evolution strategies

Sonal Sable
MIT BHOPAL
11sonalsable@gmail.com

Ankita Porwal
ASS. PRO. (CSE)MIT BHOPAL

Upendra Singh
upendrasingh49@gmail.com

ABSTRACT:-

Stock market is a very challenging and an interesting field. In this paper, we are trying to predict the target prices of the stocks for the short term. We are predicting the target priceof script individually for eight different scripts. For each script, six attributes are used which help us to find, whether the prices are going up or down. The evolutionary techniques used for this experiment are the genetic algorithms and evolution strategies. By using these algorithms, we are trying to find the connection weight for each attribute, which helps us in predicting the target price of the stock. An input for each attribute is given to a sigmoid function after it is amplified based on its connection weight. The experimental results show that usingthis approach, predicting the stock price is promising. In each case, the algorithms were able to predict with an accuracy of at least 70.00.

Keywords:

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

I. INTRODUCTION

The prediction of the stock prices has always been a challenging task. It has been observed, that the stock price of any company does not necessarily depends on the economic conditions of the country. It is not directly linked with the economic development of the country or particular area anymore. Thus, the stock prices prediction has become even more difficult than before.

These days the stock prices are affected due to many reasons like company fundamentals and related news, political events, natural disasters etc. The fast data processing of these events with the help of an improved technology and the 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 the stocks within the shortest period of time. Thus, a time span of even few hours between the buying and selling is not unusual.

To invest the money in the stock market, we need to have an idea whether the prices of the 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 the stock is going to increase or decrease on the next day. In this paper, we are trying to predict the price of the stock 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 involving only two classes. 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 the genetic algorithms. This approach reduces the dimensionality of the feature space and removes the irrelevant factors involved in the stock price prediction. This approach performed better when compared with the linear transformation and fuzzification transformation. This GA based transformation looks promising when compared with the other transformations. Another research done on the genetic algorithms (GAs) by Kyoung-jae Kim [4] again to predict the stock market is to use GA, not only to improve the learning algorithm, but also to reduce the complexity of the feature space. Thus, this approach reduces the dimensionality of the feature space and enhances the generalizability of classifier. Ajith Abraham [15] developed a hybrid intelligent system, which consists of neural network, fuzzy inference systems, approximate reasoning and derivative free optimization techniques. The system also gives the promising results but was not compared with any other existing intelligent systems.

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

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

Many machine-learning techniques are used for predicting the different target values [5,6,10]. This could be even to predict stock price. The genetic algorithm has been used for the prediction and extraction of important features [1,4]. Lots of analysis has been done on what are the factors that affect the stock prices and financial market [2,3,8,9]. There are different ways by which the stock prices can be predicted. One way is to reduce the complexity by extracting the best features or by the feature selection [7,11,12,13,14]. This approach will help us to predict the stock prices with the better accuracy as the complexity reduces.

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

Feature discretization of each input:-

Main concept in the discretization is that, we try to normalize each input attribute with respect to other attributes. Thus, we try to find the connection weight for each attribute that decides on contribution given by that attribute. After multiplying by the connection weight, the summation of each attribute is given to a sigmoid function. This function is used to classify the next stock price into an increasing or a decreasing class.

The sigmoid function in terms of the mathematical expression is given below. It is used when we do not have the detailed information of an input we are trying to predict. This function will classify each input mainly into two classes. So, it can be used for the 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 which is used in computing to find an exact or approximate solution to search and optimization of the problems. Genetic algorithms are aparticularclass of evolutionary computationthat uses the techniquesinspiredby an evolutionary biology such as inheritance, mutation, selection, and crossover. A genetic algorithm finds a 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 the population of chromosomes which is a set of solutions to the problems that could occur for a particular scenario. One evaluates its fitness and does its reproduction to get the better solutions with respect to the targeted problem. The chromosomes which represent the better solutionswill begiven more chance for the reproduction than those which represent the poor solutions. This process continues for a number of generations after which we get an optimal solution.

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

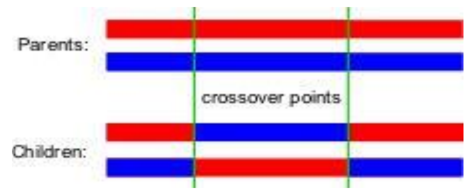

Figure 1. Two point crossover

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

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

The algorithm for the evolutionary strategies is givenbelow:-

1. Randomly create an initial population of individuals.

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

3. Determine fitness of each individual.

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

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

In this experiment, we are using a (μ , λ)-ES strategy in which the parents (candidate solutions) produce offspring (new solutions) by mutating the one or more problem parameters. An offspring compete for the 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 the means of reproduction and survival of the fittest.

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

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

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

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

where ξ=0 and σ=1.

II. Experimental Setup

Dataset used:

The dataset used for this experiment consists of the data for the last five years. A total of six attributes for each company are used for the prediction. These are an opening price, closing price, highest price, lowest price, volume and an 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 used attribute. We usedanother testing dataset, so that we can verify the results. Thus, we can check if over fitting is occurring or not. The results obtained actually showed that no over fitting is occurred.

The representation for the problem is a 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 the times the connection weights result in correctly predicting the stock price. So, if it was able to predict the stock price correctly in 500 data points, then its fitness is 500. A total of 620 data entries are there for each dataset, which we need to predict. Firstly, we used 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. Thedifferent parameter settings for each algorithm are as given below:-The parameter settings for the Genetic algorithm are:-

Tab 1: Parameter settings for the genetic algorithm

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

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

Tab 2: Parameter settings for the evolutionary strategies

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

. Table 6 shows the accuracy of the algorithm for predicting the highest price. The connection weights are calculated using the training dataset and is tested on a testing dataset. This protects against any over-fitting occurring in the model. From the results shown in Table 5 and 6 it can be seen that over-fitting is not occurring. The fitness also indicates the number of the times it actually predicted the stock price correctly. The total number of the entries present in each set is 620.

It can be seen from the Table 6 that we were able to predict the stock price with the 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 restrictions on the space search, we have kept the upper end to be 1000 only for the floating representation.

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

Table 3: Connection weights for each company using the genetic algorithm.

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 4: Connection weights for each company using the evolutionary strategy.

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 5: 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	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 6: The accuracy with which the stock price was predicted 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%

IV. Conclusion and Future Work

The novel method for predicting the stock prices using the genetic algorithm and evolutionary strategies looks very promising. It was found that genetic algorithm and evolution strategies have performed almost evenly. The best accuracy found using 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 at an accuracy of 70% or better in all the cases.

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

There are many aspects that we can consider in the future. We need to include more attributes to predict the 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 the classification. Thus, instead of using the sigmoid function we can use some other functions.

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

Future Work:-

The evolutionary algorithms used for this experiment looks promising. Therefore, the further research is required in this field. We can even try to use the attributes of the other companies to predict the prices to check whether they help in predicting the prices or not. 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 the prediction will beabove 80.0% if we usethe other companies' data instead of using just an individual company's data.

Since, the results obtained are above 70.0% in every case, then we can test the performance on the real time data as well. This will give us an idea whether only historical data is good enough to predict the data or not. If not, then we need to find the factors other than the historical data which affects 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.

In this experiment, the companies used were the big companies. We can check the performance of those algorithms on the 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-IChou, 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 andApplications (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