Original Article

Original Article

Grammatical Evolution for Stock Price Prediction: A Time-Series Approach Using MATLAB

Aditya Kusuma Setyanegara✉
Telkom University, Jl. Telekomunikasi No. 1, Terusan Buah Batu, Bojongsoang, Kabupaten Bandung, Jawa Barat 40257, Indonesia.
adityakusuma365@gmail.com✉

Abstract:

Stocks are a popular investment model in contemporary markets, but they carry inherent risks due to price fluctuations. This study aims to improve stock price prediction using Grammatical Evolution, leveraging historical stock price data as input. A comprehensive grammar in Backus-Naur Form notation is developed to generate various predictive models. The proposed approach is evaluated using metrics such as MAPE, demonstrating its effectiveness in forecasting stock prices.

Keywords: Technical Analysis, Grammatical Evolution, Backus-Naur Form

Introduction

*S*tocks are units of value in various financial instruments that represent ownership in a company (Darmadji & Fakhruddin, 2001). To maximize profits and minimize risks, investors typically conduct analysis before making stock transactions. One widely used approach is technical analysis, which predicts stock price movements based on historical market data, particularly price fluctuations and trading volume (Kirkpatrick & Dahlquist, 2006).

In recent years, machine learning and evolutionary computation have been explored to enhance stock price prediction accuracy. One such method is Grammatical Evolution, which has been applied to time series data, including stock prices. Unlike traditional statistical models, Grammatical Evolution offers greater flexibility in discovering both linear and non-linear predictive models. By employing chromosome representation in the form of functions or programs and constructing an extensive grammar in Backus-Naur Form (BNF), this method enables the generation of diverse predictive models (Suyanto, 2008). This study aims to evaluate the effectiveness of Grammatical Evolution in stock price forecasting and compare its performance with established prediction methods.

https://jurnal.usk.ac.id/riwayat

Received: January 6, 2025 – Revised: February 20, 2025- Accepted: March 28, 2025 - Published online: April 7, 2025

Methods

This study aims to develop a system capable of predicting stock prices using Grammatical Evolution (GE). The dataset, obtained from Yahoo Finance, consists of daily and weekly stock prices, divided into two phases: periods 1 to 500 for training and periods 501 to 750 for testing. Before training, the data is normalized using the min-max normalization method to simplify calculations and improve accuracy. During the training phase, the system generates a function with the best MAPE value, derived from the individual with the highest fitness score in each generation. This function is then used to predict stock prices in the testing phase. The Grammatical Evolution process involves several key steps, including population initialization, chromosome translation, the definition of prediction functions using the BNF grammar, individual evaluation, elitism, parent selection via roulette wheel, recombination, mutation, and survivor selection.

Population Initialization

In this stage, chromosomes are defined based on the specified population size, with each chromosome consisting of a number of genes that are randomly generated within a predefined range. The length of each chromosome is determined according to the chosen parameters, and the process involves determining the population size and chromosome length, creating chromosomes of the specified length by generating each gene randomly within the range [1, chromosome length], and repeating this process for the number of individuals specified by the population size.

Chromosome Translation

Chromosome translation involves converting chromosome data, represented by integer values, into functional expressions according to the defined BNF grammar. The process begins by defining the BNF grammar to be used, followed by translating each chromosome into a function that adheres to this grammar. After translation, the validity of each chromosome is verified, and invalid chromosomes—those that do not form valid functions—are regenerated by randomly copying genes from other chromosomes. Finally, any unused genes are trimmed from the chromosome once a valid function is formed.

Defining the Prediction Function Using BNF Grammar

The prediction function in this system is constructed using BNF grammar, which consists of production rules that generate mathematical expressions. The grammar utilized is presented in Table 1.

Table 1. Backus-Naur Form Grammar

	Non-Terminal	Production Rules
1	<expr>	<expr><op><expr> | (<expr><op><expr>) | <pre_op> (<expr>) | <var> | <const>
2	<op>	+ | - | * | / | ^
3	<pre_op>	sin | cos | tan | exp | log
4	<var>	a1 | a2 | a3 | a4 | a5
5	<const>	0.1 | 0.2 | 0.3 | 0.4 | ...|0.9 | 1 | 2 | 3 | 4 | 5

This grammar is used to generate prediction functions for stock prices, incorporating various mathematical operations and trigonometric functions.

Individual Evaluation and Elitism

In the individual evaluation phase, each function generated from the chromosome translation is evaluated by calculating its fitness and MAPE values. The function with the highest fitness value is retained through elitism for each generation. The best individuals, having the highest fitness scores, are preserved to prevent them from being lost in subsequent generations.

Parent Selection Using Roulette Wheel

Parent selection is conducted using the roulette wheel method, where individuals are selected based on their fitness values. The higher the fitness value of an individual, the greater the probability of being selected as a parent for the next generation. The selection process involves summing the fitness values of all individuals in the population and generating a random number within the range [1, total fitness]. The random number determines which individual is selected as a parent. For example, if there are 5 individuals with fitness values {10, 15, 20, 25, 30}, the cumulative fitness values are {10, 25, 45, 70, 100}. If a random number of 40 is generated, the individual with a fitness value between 25 and 45, corresponding to the third individual, is selected as a parent.

Crossover

In the crossover stage, parent pairs are selected to undergo gene exchange. This process utilizes a crossover probability, and two crossover points are randomly selected. For each parent pair, a random number between 0 and 1 is generated. If this value is less than the defined crossover probability, two random crossover points are selected on the chromosome. The chromosome segments between these points are exchanged between the parents, resulting in two offspring.

Mutation and Survivor Selection

Mutation is performed by generating a random number for each gene in the chromosome. If the generated random value is less than the mutation probability, the gene value is replaced with a new random value. After mutation, the survivor selection process is employed to choose the best individuals for the next generation. Survivor selection retains individuals with the highest fitness values, ensuring that only the fittest individuals survive to propagate their genes.

These methods work iteratively to produce the best solutions for stock price prediction, involving selection, recombination, and mutation processes that gradually improve the quality of individuals over time. The combination of these evolutionary strategies allows for robust learning and function optimization, facilitating the effective prediction of stock prices.

Results

During the testing phase, data was obtained from the stock exchange website http://finance.yahoo.com. The data used consists of stock information for PT. Telekomunikasi Indonesia, covering the period from June 2010 to May 2015. The testing process is considered to have reached a solution if the Mean Absolute Percentage Error (MAPE) is less than 20%.

Testing Results of GE Parameter Combinations

The Grammatical Evolution (GE) parameters consist of evolutionary parameters such as population size, number of generations, crossover probability, and mutation probability. The testing process was conducted using various combinations of these parameters. The dataset used includes 500 periods for training and 250 periods for testing. From 30 different parameter combination scenarios, the testing results are presented in Table 2.

Table 2. Testing Results of GE Parameter Combinations.

Population Size	Number of Generations	Pc	Pm	MAPE Training	MAPE Testing
50	50	0,5	0,01	1.4995	0.92988
50	50	0,6	0,01	1.4995	0.92988
50	50	0,7	0,01	1.4995	0.92988
50	50	0,8	0,01	1.4995	0.92988
50	50	0,9	0,01	1.4995	0.92988
50	50	0,5	0,05	1.4995	0.92988
50	50	0,6	0,05	1.4995	0.92988
50	50	0,7	0,05	1.4995	0.92988
50	50	0,8	0,05	1.4995	0.92988
50	50	0,9	0,05	1.4995	0.92988
50	50	0,5	0,1	1.4995	0.92988
50	50	0,6	0,1	1.4995	0.92988
50	50	0,7	0,1	1.4995	0.92988
50	50	0,8	0,1	1.4995	0.92988
50	50	0,9	0,1	1.4995	0.92988
100	50	0,5	0,01	1.4995	0.92988
100	50	0,6	0,01	1.4995	0.92988
100	50	0,7	0,01	1.4995	0.92988
100	50	0,8	0,01	1.4995	0.92988
100	50	0,9	0,01	1.4995	0.92988
100	50	0,5	0,05	1.4995	0.92988
100	50	0,6	0,05	1.4995	0.92988
100	50	0,7	0,05	1.4995	0.92988
100	50	0,8	0,05	1.4995	0.92988
100	50	0,9	0,05	1.4995	0.92988
100	50	0,5	0,1	1.4995	0.92988
100	50	0,6	0,1	1.4995	0.92988
100	50	0,7	0,1	1.4995	0.92988
100	50	0,8	0,1	1.4995	0.92988
100	50	0,9	0,1	1.4995	0.92988

From the testing results of the 30 parameter variations, the best output function obtained was consistently a1, with a fitness value of 0.66644 during the training process. This optimal function emerged as early as the first generation and at the latest in the seventh generation. The Mean Absolute Percentage Error (MAPE) obtained was 1.4995%, indicating an accuracy of 98.5005% (100% - 1.4995%). Despite being achieved in the early generations, this high accuracy suggests that the GE parameters had minimal impact on changes in both the training and testing MAPE values. Furthermore, the recombination and mutation processes in each generation did not yield significantly better individuals. As a result, all tested GE parameter variations did not substantially improve model performance.

Testing Results of Training and Testing Data Input Scenarios

The testing was conducted using 16 different data scenarios, employing consistent Grammatical Evolution (GE) parameters to ensure reliable performance across the scenarios. The parameters used were a population size of 50, a total of 50 generations, a crossover probability of 0.8, and a mutation probability of 0.01. The Mean Absolute Percentage Error (MAPE) values were computed for both the training and testing phases, providing insights into the model's accuracy across different data input scenarios.

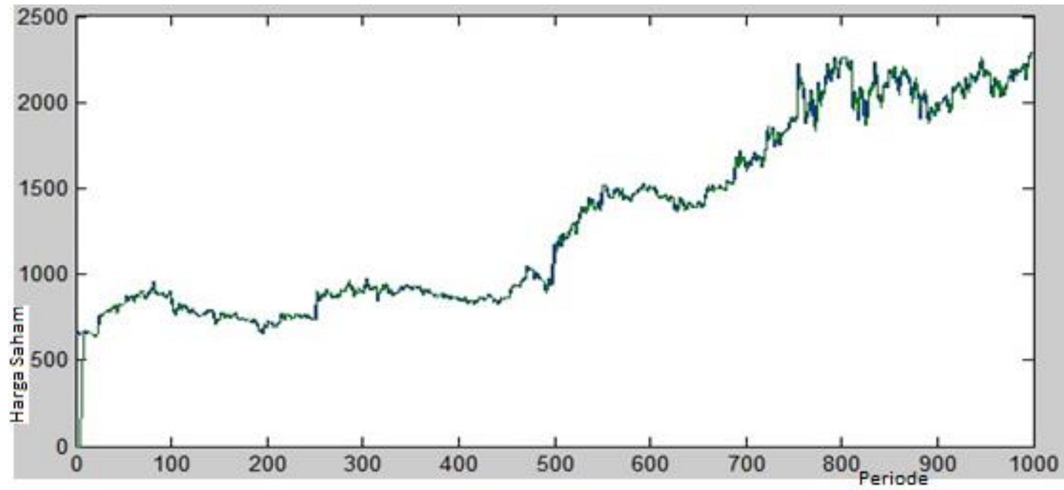
Table 3. Testing Results of Training and Testing Data Scenarios.

Training	Testing	MAPE Training	Mape Testing	Rata-rata Training & Testing	Rata Aktual Training (Rp)	Rata Aktual Testing (Rp)
1000	1000	1.7051	1.3041	1.5046	12.1634	23.284
750	1000	1.717	1.3041	1.51055	11.4713	23.284
500	1000	1.4073	1.3041	1.3557	10.5277	23.284
250	1000	1.3978	1.3041	1.35095	11.0698	23.284
1000	750	1.6125	1.3058	1.45915	11.6904	26.9363
750	750	1.3718	1.3058	1.3388	11.137	26.9363
500	750	1.3491	1.3058	1.32745	11.713	26.9363
250	750	1.3025	1.3058	1.30415	12.3686	26.9363
1000	500	1.3523	1.2887	1.3205	13.4777	29.687
750	500	1.3307	1.2887	1.3097	14.646	29.687
500	500	1.2984	1.2887	1.29355	16.4518	29.687
250	500	1.2976	1.2887	1.29315	20.5862	29.687
1000	250	1.4229	0.92988	1.17639	19.8442	24.9878
750	250	1.4323	0.92988	1.18109	22.7884	24.9878
500	250	1.4995	0.92988	1.21469	28.0351	24.9878
250	250	1.6628	0.92988	1.29634	34.7136	24.9878

From the results presented in Table 3, the best performance was achieved with a training dataset of 1000 periods and a testing dataset of 250 periods, yielding an average MAPE of 1.17639% for both training and testing. The variations in MAPE values were influenced by the differences in stock price patterns across datasets. The best training MAPE was achieved using data from the period June 25, 2012, to June 10, 2013, with a MAPE value of 1.2976%. Some test results also indicated that larger datasets tended to produce worse MAPE values for both training and testing compared to smaller datasets.

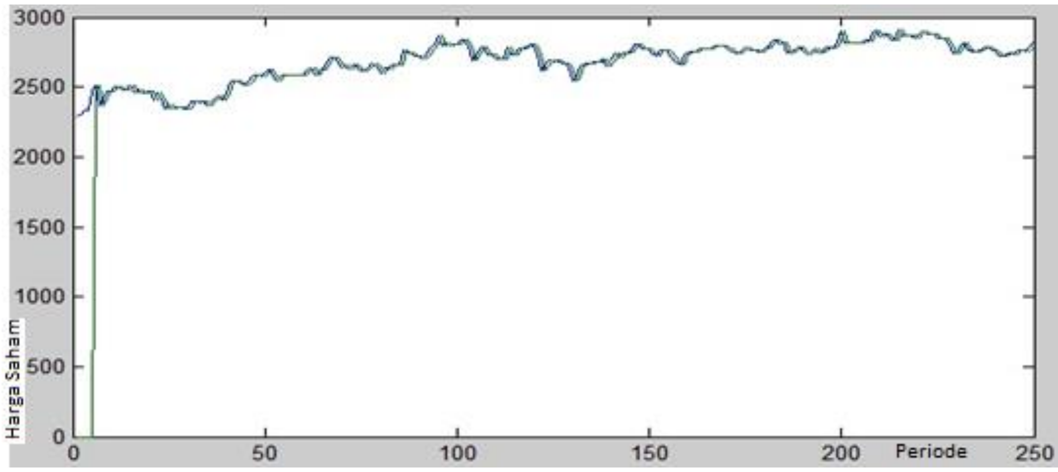
Forecasting Test Results Up to H+10

The testing aimed to evaluate the Grammatical Evolution (GE) system's accuracy in predicting stock prices for H+10, using forecasted data from H+1. The experiment used an input training dataset of 1000 periods and an input testing dataset of 250 periods, with the GE system configured with parameters including a population size of 50, a total of 50 generations, a crossover probability of 0.8, and a mutation probability of 0.01. These settings were selected to optimize the evolutionary process and improve the model's predictive performance. Figures 1, 2, and 3 show the graphs of the training, testing, and prediction phases, respectively.



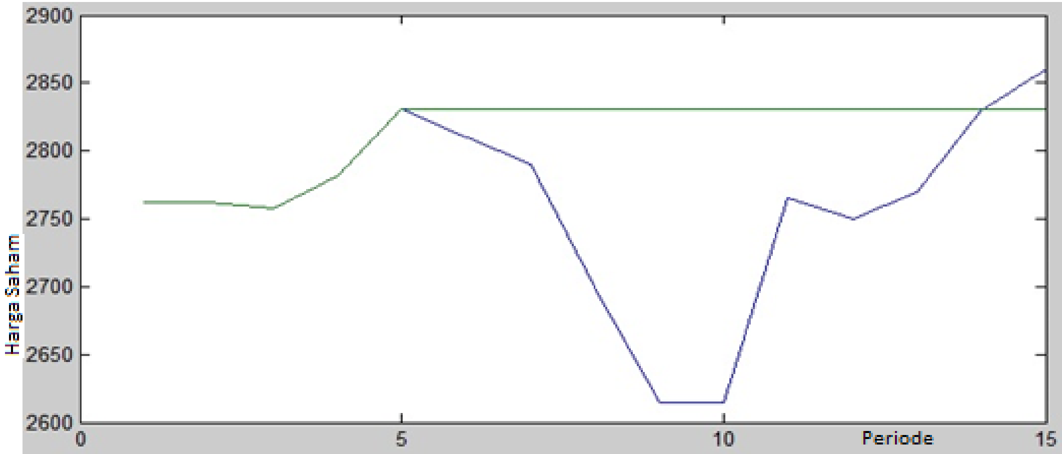
Source: Output from Matlab

Figure 1. Training Phase Graph (Blue: Actual Data, Green: Predicted Data).



Source: Output from Matlab

Figure 2. Testing Phase Graph (Blue: Actual Data, Green: Predicted Data).



Source: Output from Matlab

Figure 3. Prediction Phase 3 Graph (Blue: Actual Data, Green: Predicted Data).

The prediction process up to H+10 resulted in a MAPE value of 3.2011%. Despite this, the function derived from the training process (a1) produced a relatively low MAPE, with an average prediction error of Rp. 85.86. In the prediction graph for H+10, the stock values generated by the training-derived function tended to remain constant, indicating a certain stability in the model's predictions.

Conclusion

Based on the research and experimental results, stock price prediction using Grammatical Evolution (GE) shows that the best function obtained is a1 or a function similar to a1, where the stock price on day H-1 has a dominant influence on future price predictions. Changes in GE parameters do not significantly impact the MAPE value if the system has already achieved high accuracy in the early generations. Even with a relatively high recombination and mutation probability, the evolutionary process does not produce a better solution than previous generations. However, variations in input data scenarios affect the MAPE values during both training and testing phases. In predictions up to H+10, the system achieved a MAPE of 3.2011%, with an average prediction error of Rp 85.86, demonstrating the model's effectiveness in forecasting stock prices.

Suggestion

In future research, the development of a stock price prediction system based on Grammatical Evolution can be enhanced by integrating various time series models such as ARIMA, Prophet, Exponential Smoothing, or other similar approaches into the Backus-Naur Form (BNF). This study can focus on designing a more adaptive grammar, enabling the optimal combination of multiple prediction models with dynamic weighting based on market conditions. Additionally, the exploration of parameter optimization techniques, such as evolutionary strategies or Bayesian optimization, can be implemented to improve predictive accuracy. The system's performance evaluation can be conducted by comparing prediction results against historical data using metrics such as RMSE, MAPE, and R^2. Through this approach, the system is expected to generate more accurate stock price predictions that are adaptable to dynamic market patterns.

References

Brealey, A. R., Myers, S. C., & Marcus, A. J. (2007). *Fundamentals of corporate finance*

(5th ed.). McGraw-Hill.

Dalton, M. J. (2001). *How the stock market works* (3rd ed.). NYIF.

Darmadji, T., & Fakhruddin, H. M. (2001). *Pasar modal di Indonesia*. Salemba Empat.

Fauzan, R., Harsono, K., Meisandy, R. P., Barokah, M., & Muhaimin, M. I. (2024). Optimising human resource management as an effort to improve employee performance through digital attendance. *Riwayat: Educational Journal of History and Humanities, 7*(1), 16–25.

Frankel, J. A., & Froot, K. A. (1990). Chartists, fundamentalists, and trading in the foreign exchange market. *The American Economic Review, 80*(2), 181–185.

Gunaidi, A. A. (2010). *The shortcut of MATLAB programming*. Informatika.

Imran, M. F. (2024). Criminological examination of physical and psychological violence committed against children in the school environment. *Riwayat: Educational Journal of History and Humanities, 7*(1), 41–47.

Kirkpatrick, C. D., & Dahlquist, J. R. (2006). *Technical analysis: The complete resource for financial market technicians*. Financial Times Press.

Leung, J., Kern, K., & Dawson, J. (2007). *Genetic algorithms and evolution strategies* [Presentation slides].

Osler, K. (2000). Support for resistance: Technical analysis and intraday exchange rates. *FRBNY Economic Policy Review, July*.

Romulo, C. S., & Dalimunthe, Z. (2024). Effect of related party transaction and tax haven utilization on tax avoidance moderated by country-by-country reporting. *Riwayat: Educational Journal of History and Humanities, 7*(1), 26–40.

Sugiharto, A. (2006). *Pemrograman GUI dengan MATLAB*. Andi.

Suyanto. (2008). *Evolutionary computation: Komputasi berbasis evolusi dan genetika*. Informatika.

Taylor, M. P., & Allen, H. (1992). The use of technical analysis in the foreign exchange market. *Journal of International Money and Finance, 11*(3), 304–314.

Widhiantika, B. A. (2010). *Peramalan harga dinar di Indonesia menggunakan grammatical evolution* [Undergraduate thesis, Institut Teknologi Telkom].