An Adaptive Model For Stock Market Forecasting Using Modified Genetic Algorithm

Satyakam Behar*, Dr.Anurag Sharma

1* Research Scholar, Dept. of CSE, MATS School of Engineering and Information Technology, Aarang
2Professor, Dept. of CSE, MATS School of Engineering and Information Technology, Aarang

Abstract: This paper proposes a Modified Genetic Algorithm (MGA) in which the local optima problem of Genetic Algorithm (GA) is addressed. The GA uses $\gamma$ distribution for crossover probability due to which GA is sometimes trapped in a local optimum situation, therefore, could not provide globally optimum solution. This problem is solved by using $\beta$ distribution for crossover probability computation in MGA. The $\beta$ distribution does not fall in local optima and therefore provides a global solution to the problem. This paper proposes MGA for optimized feature selection to forecast the stock market movement. Historical stock data of four stock indices (BSE, NITFY50, S&P500, and DJIA) are collected from a financial website which is used to compute the Technical Indicators (TI) which forms the feature set for MGA. K-Nearest Neighbour (KNN) is used as a classification method for the evaluation of MGA. The performance of MGA is compared with the GA and Particle Swarm Optimization (PSO). For evaluation purpose the accuracy, precision, recall, and F-score is taken. The results indicate the superiority of the proposed MGA.

Keyword: Stock Market, Machine Learning (ML), Feature Selection, Genetic Algorithm.

1. Introduction

The stock market plays an essential role in the financial growth of a nation [1]. A stock market is a place where company shares (equity securities) are traded publically to raise money [2]. Forecasting the stock trend is demanding due to the financial benefits and wealth management. However, stock market forecasting is difficult and one cannot always do exact predictions due to the complex behaviour of the financial data. Complexities and nonlinearity exist in the stock market due to the presence of noise [3]. The simplest way to forecast market trends is to analyse the historical pattern of the financial market. Various theories and methods have evolved for market analysis [4]. The conventional way of forecasting is to analyse the company and market growth through technical analysis or fundamental analysis. Technical analysis is best suited for analysing stock performance based on past data. The TI are used for this purpose where TI is computed through the historical stock data. These indicators may be visually represented through graphs and charts so that anyone can easily interpret the market trend by simply looking at these graphs/charts [5]. The ML techniques are gaining popularity due to their
superior performance in every domain and financial time series analysis is not an exception [4].

A wide range of ML methods is available in the literature that is used for financial time series analysis and stock market forecasting [6]. TI play important role in stock market trend analysis. There are a number of TI that can be used along with historical stock data to do predictions. But, the inclusion of all the TI increases the computation time and may reduce the model performance due to redundant and unwanted features. This problem can be solved by an appropriate feature selection technique [7]. Two kinds of feature selection techniques are popular in the literature: filter-based and wrapper-based. The filter method calculates the feature score to identify the importance of the feature and then these scores are used to rank the features. The highest-ranked features are selected whereas lower-ranked features are discarded. Filter methods are simple to implement but feature selection is based on some threshold. If the threshold value is not selected wisely then redundant features may be there in case of lower threshold whereas important features may get discarded in case of higher threshold. Due to this drawback of filter methods, wrapper methods are preferred. The wrapper method follows the stochastic approach in which the optimal feature subset is selected based on fitness value (obtained through objective function) [8,9]. Some of the popular wrapper methods are Genetic Algorithm (GA), Particle Swarm Optimization (PSO), etc. [10,11].

This paper proposes an adaptive stock market forecasting model in which the most important TI are selected through an Evolutionary-based feature selection approach. The key contributions of the paper are:

i. The TI module is designed to generate the feature set.
ii. The Modified GA (MGA) is proposed to select the optimal feature subset to have faster convergence over GA.
iii. The proposed MGA is compared with state-of-the-art methods and statistically analysed.

The remaining paper is organized as: section two provides the literature review; section three provides the methodology and description of the proposed work; section four provides the experimental details and results; section five provide the conclusion and future scope of the work.

2. Literature Review

In a liberalised economy, the stock market is seen as a significant pointer for the strength of the economic system. It looks to be more subject to noneconomic factors such as political damage, terrorist action, and shareholder psychology. According to the literature, several scholars employed a variety of methodologies to simulate stock market indexes [12]. These techniques are based on soft computing like Artificial Neural Network (ANN), Fuzzy Logic, Support Vector Machine (SVM), and evolutionary algorithms (GA, PSO, etc.) [13–17].

Stock market volatility makes predicting challenging, and new ways to forecasting models are constantly explored. The ML and deep learning (DL) based models are continuously developing by many researchers for predicting stocks. SVM, Neural Network, and Genetic adversarial network (GAN) based ML model was proposed in [18,19] for forecasting the stock
market trend. To analyse the stock another combination of ANN and support vector regression (SVR) model was proposed in [20]. Some of the researchers have proposed [21], a hybrid version method and combines a basic set with ANN for obtaining relative strength index (RSI). The results presented by the researchers are better in hybrid model as compared to ML based model. The ML models are not able to perform well on a volatile or high-dimensional data like stock data, this is also a drawback of ML model.

In addition to DL methods, researchers frequently employed ANN trained models or metaheuristic approaches [22,23]. Das et al. [24] analysed the expected stock price in this perspective, taking into consideration four distinct worldwide indexes such as the FTSE100, BSE Sensex, S&P 500, and NSE Sensex, used TI and statistical measures. In [25] The GA and Firefly algorithms, in conjunction with the evolutionary system, are well thought out for optimising the feature reduction process. The results show that firefly optimised feature reduction utilising an evolutionary scheme applied to the OSELM prediction model outperformed the majority of evaluated models. In [26], author proposed a hybrid ANN model which combined harmony search (HS) and GA to categorize the linking among the TI and the stock exchange for the specific time. Jaya optimization with ELM model is proposed in [27] wherein the model was evaluated and tested on USD to INR, USD to EURO dataset for open stock price. The Adaline Neural Network hybridized with PSO for prediction of HDFC and JSPL open price was tested in [28]. In [29], ANNs and GA with weight optimization model was proposed for prediction of stock index. The adeptness was evaluated for proposed model in [30] which was trained by using GA updated by mutation operator.

Several researchers have been applied different metaheuristic based optimization algorithm for stock market forecasting. The research gap found from the above literature is: there is need to improve the prediction of stock index using optimal feature selection. And none of the researchers are yet achieved better performance using optimization algorithm.

3. Methods and Materials

This section discusses the methods and materials used for the proposed work. Figure 1 illustrates the methodology adopted to forecast the stock market movement.
3.1 Dataset Collection and Pre-processing

First stock market dataset is to be collected. The datasets of BSE, NIFTY50, S&P500, and DJIA is collected from the financial website of Yahoo (Source: https://finance.yahoo.com/). The historical data is collected from 1-Oct-2017 to 30-Sep-2022 which consists of 1230 and 1231 observations for BSE and NIFTY50 respectively whereas 1258 observations for S&P500 and DJIA. Raw stock market data has noise that needs to be handled carefully, therefore data imputation is required. There are a number of ways to impute the missing values. Missing values can be replaced through mean, median, or zero/constant else one can simply drop the rows/columns having missing Values. Next, we have to perform a label assignment to find the stock’s upward/downward movement. Label assignment is to be done by equation 1.

\[
\text{Label} = \begin{cases} 
0, & \text{Close}_n < \text{Close}_{n-1} \\
1, & \text{Close}_n \geq \text{Close}_{n-1} 
\end{cases}
\] (1)

Where, \(\text{Close}_n\) and \(\text{Close}_{n-1}\) is \(n^{th}\) \((n-1)^{th}\) day’s close price.

3.2 TI Generation

TI play important role in analyzing the stock time series and predicting future behaviour. Some of the commonly used TI is Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), On Balance Volume (OBV), Relative Strength Index (RSI), Commodity Channel Index (CCI), Bollinger Band (BB), etc. For the generation of TI, talib library of python is used. Table 1 provides the list of computed TI.
3.2 Modified Genetic Algorithm

The idea behind the development of an evolutionary algorithm (EA) was to provide the solution to optimization problems. EA is apt to solve problems where a huge search space is present with number of possible solutions. The GA is a population-based EA invented by John Holland [31]. The GA’s aim was to study the evolution process so that it could be applied to different applications. GA follows the heuristic approach for searching for solution space through genetics and natural selection. The idea is adopted from the biological process of survival of the fittest. GA evolves a population of chromosomes by crossover and mutation operator where each chromosome contains genes. From the entire population, individuals are chosen by the selection operator based on their fitness. The crossover and mutation operator imitates the
biological processes responsible for introducing diversity to the population. The crossover and mutation are for exploration whereas selection is for exploitation [31]. Following are the parameters of the GA:

- **Population**: A subset of the present generation's solutions is the population. The population is also known as a collection of chromosomes.
- **Fitness Function**: The fitness function is used to evaluate the algorithm where the value of the fitness function indicates the closeness of the solution to the expected outcome.
- **Selection Criteria**: The greater the fit of a chromosome, the greater the likelihood that it will be chosen.
- **Crossover Operator**: The crossover operator couples together two chromosomes (parents) to create a new chromosome (offspring).
- **Mutation Operator**: The mutation operator is used to generate a new chromosome.

The GA uses γ distribution for the generation of initial population (Pop) using equation (2) and (3).

\[
\text{Pop} = \text{lb} + (\text{ub} - \text{lb}) \times \text{uniform\_rand}()
\]  
\[\text{(2)}\]

Where, \(\text{lb}\) and \(\text{ub}\) are lower bound and upper bound respectively, \(\text{uniform\_rand}\) is uniform random distribution probability calculated by following equation.

\[
p(x) = x^{l-1} \frac{e^{-x/\theta}}{\theta^l \gamma(l)}
\]  
\[\text{(3)}\]

Where, \(l\) is shape, \(\theta\) is scale, and \(\gamma\) is the \(\gamma\) function. The \(\gamma\) distribution is not apt to handle the local optima problem therefore is not suitable for global optimum solution. Therefore, this paper proposed MGA in which \(\beta\) distribution to handle this local optima problem. The equation for the \(\beta\) distribution is computed by following equation.

\[
f(x; a, b) = \frac{1}{B(a, b)} x^{a-1} (1 - x)^{b-1}
\]  
\[\text{(4)}\]

Where \(a\) is probability of success and \(b\) is probability of failure. Figure 2 provides the flowchart for the proposed MGA.

### 3.3 Fitness Evaluation using KNN

Fitness function plays an important role in the selection of feature subset. This paper uses K-Nearest Neighbour (KNN) to evaluate the fitness of chromosomes in population. The KNN is apt to solve the classification problems by computing shortest distance between the testing and training data in the feature space. Let training data for \(n\) number of sample is defined as

\[
d = \{d_1, d_2, ..., d_n\}
\]  
\[\text{(5)}\]
The KNN calculates the Euclidean Distance (ED) for two data points $d_{test}$ and $d_i$ by taking the square root of the sum of the squared difference.

$$ED(d_{test}, d_i) = \sqrt{\sum_{i=1}^{n} (d_{test} - d_i)^2}$$  \hspace{1cm} (6)

In this paper, the KNN uses 5 nearest neighbours and the report classification outcome and error by following equation.

$$\text{argmax} (\text{count}(d_n))$$  \hspace{1cm} (7)

Here, 2 classes \{0, 1\} are used to represent the market movement. In each chromosome, gene value 1 or 0 indicates the selection or rejection of the feature respectively in the feature set index. The chromosome (genome) are bitstrings in the encoded form to represent the features. In each iteration, the individual (combinatorial set of features) in the current population are evaluated and the respective fitness are ranked as per the KNN-based classification error. Individuals with lower fitness have better chance of survival to the next generation or mating pool. The iterations of GA ensure the reduction of error rate and thus picks the individual with best fitness value (here least value is the best one). The fitness value is computed by following equation.

$$\text{fitness} = \frac{\alpha}{N_f} + \exp \left( \frac{-1}{N_f} \right)$$  \hspace{1cm} (8)

Where, $\alpha$ is KNN-based classification error and $N_f$ is cardinality of the selected features. The above equation ensures the learning of GA with minimized error and reduced feature subset.

![Flowchart for MGA](http://www.webology.org)
4. Results and Discussions

This section discusses the experimental outcome of the proposed MGA. The proposed MGA is implemented in the Python 3 with Ryzen 5, 2.10 GHz CPU and 8GB RAM. Table 1 provides the empirically selected parameters of MGA.

Table 1. Selected Parameters of MGA

<table>
<thead>
<tr>
<th>MGA Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Size</td>
<td>100</td>
</tr>
<tr>
<td>Genome Length</td>
<td>100</td>
</tr>
<tr>
<td>Population Type</td>
<td>Bitstrings</td>
</tr>
<tr>
<td>Fitness Function</td>
<td>KNN-based classification error</td>
</tr>
<tr>
<td>Number of Generation</td>
<td>100</td>
</tr>
<tr>
<td>Crossover Rate</td>
<td>0.7</td>
</tr>
<tr>
<td>Crossover Probability</td>
<td>β random distribution</td>
</tr>
<tr>
<td>Mutation Rate</td>
<td>0.01</td>
</tr>
<tr>
<td>Selection Scheme</td>
<td>Roulette Wheel</td>
</tr>
</tbody>
</table>

The MGA evaluated and tested on four datasets (BSE, NIFTY, S&P500, and DJIA). The performance of MGA is compared with the other wrapper feature selection methods (GA and PSO) which is shown in table 2. The MGA attained an average accuracy, precision, recall, F-score as 66.57%, 0.6759, 0.675, and 0.6706 respectively. The MGA achieved the average improvement of 6.96%, 16.87%, 15.91%, and 7.9% in accuracy over GA and PSO for the dataset BSE, NIFTY, S&P500, and DJIA respectively. It is clearly shown that, the MGA outperform the GA and PSO in terms of accuracy, precision, recall, and F-score.

Table 2. Comparative Analysis

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSE</td>
<td>GA</td>
<td>60.98</td>
<td>0.6127</td>
<td>0.5958</td>
<td>0.6292</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>52.75</td>
<td>0.5234</td>
<td>0.5376</td>
<td>0.5175</td>
</tr>
<tr>
<td></td>
<td>MGA</td>
<td>63.82</td>
<td>0.6491</td>
<td>0.6399</td>
<td>0.6387</td>
</tr>
<tr>
<td>NIFTY</td>
<td>GA</td>
<td>43.21</td>
<td>0.4247</td>
<td>0.4328</td>
<td>0.4318</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>50.18</td>
<td>0.5121</td>
<td>0.5287</td>
<td>0.5163</td>
</tr>
<tr>
<td></td>
<td>MGA</td>
<td>63.56</td>
<td>0.6417</td>
<td>0.6429</td>
<td>0.6382</td>
</tr>
<tr>
<td>S&amp;P500</td>
<td>GA</td>
<td>52.69</td>
<td>0.5269</td>
<td>0.5266</td>
<td>0.5281</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>51.74</td>
<td>0.5038</td>
<td>0.5160</td>
<td>0.5183</td>
</tr>
<tr>
<td></td>
<td>MGA</td>
<td>68.26</td>
<td>0.6867</td>
<td>0.6989</td>
<td>0.6872</td>
</tr>
<tr>
<td>DJIA</td>
<td>GA</td>
<td>63.09</td>
<td>0.6245</td>
<td>0.6136</td>
<td>0.6256</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>62.37</td>
<td>0.6243</td>
<td>0.6363</td>
<td>0.6255</td>
</tr>
<tr>
<td></td>
<td>MGA</td>
<td>70.63</td>
<td>0.7264</td>
<td>0.7183</td>
<td>0.7229</td>
</tr>
</tbody>
</table>

Figure 3 shows the convergence graph of GA and MGA for fitness values evaluated on all four datasets for 100 iterations. In which, X-axis denotes number of iterations and Y-axis denotes
the fitness value. In this paper, fitness function (equation 8) used is for minimization, therefore, the graph convergences towards the minimum value.

![Figure 3. Convergence Graph](http://www.webology.org)
When comparing GA with MGA, it is observed from the convergence graph (figure 3) that MGA gives faster convergence and provides stagnant performance after initial number of iterations. The significance of result is checked through some statistical test. Therefore, this paper used Wilcoxon signed rank test to analyse the performance (table 3). For Wilcoxon test, average accuracy, precision, recall, and F-score is passed to get the statistical significance. The level of significance ($\alpha$) is taken as 0.05. In table 3, ‘+’ signifies the win of the MGA and ‘=’ signifies the tie of MGA. The results illustrate the superiority of the proposed MGA over the original GA and PSO as well.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSE</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
</tr>
<tr>
<td>NIFTY</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
</tr>
<tr>
<td>S&amp;P500</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
</tr>
<tr>
<td>DJIA</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
<td>0.05(+)</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper proposes MGA for selecting optimal feature set for stock market data. The TI were computed to generate the feature set because the stock trend analysis can be better analysed through the TI than the historical data. Although, all the information within the TI are not highly correlated with the stock market movement. Therefore, we come up with the solution by proposing MGA for optimal feature subset selection. The paper proposed a modification in GA to overcome the local optima problem of GA. The proposed MGA improved the crossover probability function by replacing the $\gamma$ distribution by $\beta$ distribution. The MGA was tested on the four datasets (BSE, NIFTY, S&P500, DJIA). The results of MGA were compared with the GA and PSO to validate the model performance. The MGA attained an average improvement of 11.91% in terms of accuracy. The MGA’s performance was statistically analysed through Wilcoxon test which also shows the superiority of MGA. The MGA provided maximum accuracy of 70.63% for DJIA dataset which is still not very good. Accuracy plays important role in stock market prediction due to the associated financial benefits. Therefore, the work can be extended to attain the more accuracy by improvising MGA with mutation operator.

References


[29] X. Pan, J. Zhang and K.Y. Szeto, Application of mutation only genetic algorithm for the extraction of investment strategy in financial time series, in 2005 International Conference on Neural Networks and Brain, 3 (2005), pp. 1682–1686.
