High Frequency Price Duration Prediction of Option Based on XGBoost and GA-BP

Xinzhu Liu

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China.
liuxz@email.cufe.edu.cn

Abstract

In financial high-frequency data, duration data can reflect the transaction intensity and liquidity of financial market, so the prediction of duration is a research hotspot in the financial field in recent years. With the development of machine learning, it is more and more common to apply machine learning method to the research of financial field. This paper screens evaluation indicators via XGBoost and then optimizes the BP neural network with genetic algorithm. Meanwhile, it introduces the duration average high-low spread, duration average absolute return, duration average trading volume, position at time point and trading density as the microstructure variables to build the prediction model of the price duration of Shanghai 50ETF options. It is found that the introduction of genetic algorithm can improve the prediction accuracy and efficiency of BP neural network, and the evaluation indexes MSE, RMSE, and MAE are reduced by 74.14%, 49.41% and 38.82% respectively, with the operation time reduced by 33.25%. This method provides a new research idea for the prediction of high-frequency price duration, and enriches the theoretical research of machine learning method in the field of option high-frequency data prediction.

Keywords

Price Duration; SSE 50ETF Options; XGBoost; GA-BP Prediction Model.

1. Introduction

As a mature basic financial derivative, option has a unique risk management function and can stimulate market innovation. On February 9, 2015, Shanghai Stock Exchange (SSE) officially launched China’s first option product, SSE 50ETF option, which filled a big gap in China’s domestic financial derivatives, promoted the innovation and development of the financial industry, and enhanced the international competitiveness of the capital market. Considering that China’s option market has been developing for just a short time and the option market is still immature, it is necessary to carry out quantitative modeling of the option market. At present, the research on option market is mainly based on low-frequency daily and monthly data. Comparatively, (ultra) high frequency data, measured in units per minute or per second, is closer to the real trading situation of investors than low-frequency daily data, which helps to accurately identify and analyze the market microstructure, and is of great significance to the identification and control of financial risks.

2. Literature Review

High frequency price duration is a very important index of stock options, which reflects the price fluctuation speed and trading intensity. China’s options market starts late, and the existing literature on high-frequency price duration prediction of stock options is relatively limited. However, many scholars have summarized a large number of financial high-frequency data
prediction methods, which are mainly divided into two categories: traditional mathematical statistics method and machine learning method.

Traditional mathematical statistical methods first consider the distribution information of data, and use the distribution information of financial data to establish appropriate time series models, mainly including Auto-Regression and Moving Average (ARMA) Model\cite{1}, Generalized Auto-Regressive Conditional Heteroskedasticity (GARCH)\cite{2}, Auto-regressive Conditional Duration (ACD) Model\cite{3}, etc. This kind of methods has two main characteristics:

First, compared with low-frequency data, high-frequency financial data are often nonlinear and non-stationary, so the traditional time series analysis model cannot be directly used, and needs to be processed by methods such as linear transformation and smooth correction. Liu Xiangli et al. (2015) decomposed the futures return rate by wavelet analysis and predicted the reconstructed sequence using ARMA model, and found that the prediction results were more accurate than the traditional ARMA model\cite{4}. However, because ARIMA model is a linear model, there is a certain lag phenomenon in the prediction. Luo Hua (2017) made improvement on the basis of this, and decomposed the CSI 300 index by wavelet analysis. GM(1,1) model was used to predict the high frequency part after wavelet decomposition, and ARMA was still used for the low frequency part. The results showed that the grey model had a certain adjustment effect on ARMA results\cite{5}.

Second, transactions often occur at unequal intervals, so the traditional time series model cannot be applied. Engle (1997) proposed ACD model for the first time to study the prediction of duration\cite{6}. Since then, many ACD derivation models have been developed based on different conditional duration distribution forms and different error distribution forms. Pacurar (2008)\cite{7} systematically summarized the relevant theories and application conditions of different ACD models. Wang Weiguo et al. (2015) applied different ACD models to China's stock market, and the results showed that the prediction result of asymmetric ACD model was the most stable\cite{8}. Liu Hong (2015) et al., proposed the SEMIFAR-ACD model based on the deterministic trend, differential stationary trend and stationary virtual trend of high-frequency financial data, and found that, compared with the traditional ACD model, this model could better describe the high-frequency financial data\cite{9}. Pan Shuiyang et al. (2017) proposed a trend duration model based on the trend movement characteristics of CSI 300 stock index futures\cite{10}. Danúbia et al. (2020) improved the GBS-ACD model and obtained the GBS-ACD model based on Box-Cox, which has the shape parameters of conditional duration process and the asymmetric effect on impact\cite{11}. Machine learning methods pay more attention to the laws of the data itself, and do not limit the distribution. The distribution of financial data is complex, so the traditional mathematical statistical model is difficult to predict well. Therefore, some scholars introduce machine learning model. Lin Jie (2017) constructed BP and CNN networks to predict futures prices respectively, and the results showed that the prediction effect of BP neural network was better than that of CNN\cite{12}. Huang Qing et al. (2018) simplified the price data into the prediction problem of the direction of price change, and found that the prediction accuracy of XGBoost was about 20% higher than that of BP and SVM when dealing with the dichotomy problem\cite{13}. Hu Bo (2018)\cite{14} divided the high-frequency price series in the gold futures market into three trends: rising, falling and stable. He transformed the price trend into a classification problem and used SVM to make predictions. To sum up, the majority of scholars' studies focus on the optimization of traditional mathematical statistical models, but the introduction of machine learning methods is relatively limited. Even if there are studies, they mainly focus on indicators such as yield rate and price, and the prediction of price duration is in the rise. In addition, the study of high frequency price long period focused on the futures market, with less options market and presents a fragmentation. Therefore, this article uses XGBoost to evaluate the effectiveness and contribution of indicators, then introduces the genetic algorithm to optimize the BP neural
network model, in order to predict the duration of high-frequency price of Shanghai 50ETF options, and compare it with the single BP model results. It is expected to help improve the prediction accuracy and speed, as well as to expand the suitability of machine learning theory in financial data prediction.

3. Model Construction

3.1. Screening Evaluation Indexes

The traditional long-period forecasting model only predicts the long-period series, and does not involve other input variables, which fails to fully use other information in high-frequency data that can reflect the market microstructure.

With the deepening of the research, some scholars have introduced some variables that can reflect the market information into the model to improve the forecasting effect. Liu Xiangli et al. (2012) believed that the influence of position size was weak, and average trading volume and average absolute return rate had obvious negative effects on price duration. Based on this, Wang Shaobin et al. (2014) introduced average trading volume, transaction density and percentage bid-ask spread as microstructural variables, and found that the duration of the result was negatively correlated with average trading volume, trading speed and bid-ask spread.

Based on the above research, this paper hypothesizes that the current price for a long time under the influence of six indicators: price duration of the previous period, average price spread of the previous period, average absolute rate of return of the previous period, average trading volume of the previous period, position at the time point of the previous period, and trading density of the previous period. Besides the price duration itself, the following explanations are given for the other indicators:

3.1.1. Average Price Spread Between High and Low over a Long Period

High spreads often indicate asymmetric information, so high spreads may correspond to short price durations. In order to explore whether the price spread has an effect on the price duration, the average high and low price spread margin in the corresponding price duration $X_i$ is introduced into the model.

3.1.2. Average Absolute Rate of Return in Duration

The rate of return is a direct measure of the return of investors. Therefore, the yield may also have an impact on the price duration of options. The average absolute rate of return in the price duration is defined as: $ayield = \ln(P_t) - \ln(P_{t-1})$

3.1.3. Average Trading Volume in Duration

In previous studies, trading volume often has an impact on trading frequency. Generally speaking, large trading volume may lead to more frequent trading. In order to study whether this kind of effect will appear in the market, the average trading volume $avol$ of the corresponding price duration $X_i$ is introduced into the model.

3.1.4. Open Interest at Time Point

Open interest can reflect the market’s interest in the contract. The open interest at the time when the price just changes (that is, the first time of the corresponding price duration $X_i$) is defined as the open interest $oi$ and introduced into the model.

3.1.5. Trading Density

The trading density means the frequency of trading. High trading density often corresponds to drastic fluctuations in the market. This is because the increase in trading volume may be caused by a large number of speculators, which will increase the instability of the market. Transaction density $tint$ is defined as the number of transactions per unit time in the price duration $X_i$. 
In addition, considering that all the above indexes depend on the price duration $X_i$, it is necessary to eliminate the intraday effect of these indicators before using them.

In this paper, the XGBoost algorithm is used to calculate the importance of features to screen the indicators. The XGBoost algorithm is often used for prediction and classification, but it can also be used to evaluate the effectiveness and importance of features. Many scholars have used the XGBoost algorithm to evaluate the effectiveness of indicators, and then follow up the modeling based on the evaluation results\textsuperscript{[17-18]}.

### 3.2. Neural Network Algorithm

As a kind of neural network, BP neural network has great advantages in solving nonlinear problems such as financial high-frequency data prediction. Moreover, BP neural network has high fault tolerance and learning ability, and can well solve the problem of complex influencing factors of high-frequency price duration. Based on the above background, this paper chooses BP neural network to predict the high frequency price duration of options. The specific steps of the algorithm are as follows:

1. Sample data matrix $A_{n \times (p+1)} = (dur_t, dur_{t-1}, o_{i,t-1}, ay_{t-1}, tint_{t-1}, a margin_{t-1}, avol_{t-1})$, where $dur_t$ is the dependent variable.
2. The number of nodes in the input layer is $p$, and the number of nodes in the output layer is $m$. Assume that the number of nodes in the hidden layer is $l$. The weight of neuron $I$ to neuron $j$ is expressed as $w_{ij}$, and the threshold of neuron $j$ is expressed as $\theta_j$. $w_{ij}$ and $\theta_j$ are initialized randomly. The learning rate is $\eta$, and the incentive function is $g(x) = 1/(1 + e^{(-x)})$.
3. Import the preprocessed input vector and expected output into the BP neural network.
   - **Hidden layer output**: $H_i = g(\sum w_{ij} x_i - \theta_i)$, $i = 1,\ldots,n; j = 1,\ldots,l$;
   - **Output layer output**: $O_k = g(\sum w_{ik} H_i - \theta_k)$, $j = 1,\ldots,l; k = 1,\ldots,m$.
4. Update the weight:
   - Update the weight from input layer to hidden layer: $w_{ij} = w_{ij} + \eta H_i(1 - H_i)x_i \sum w_{ik} e_k$,
   - Update the weight from the hidden layer to the output layer: $w_{jk} = w_{jk} + \eta H_j e_k$,
   - where $e_k$ is the training error.
5. Calculate the error: $E = \frac{1}{2} \sum (Y_k - O_k)^2 = \frac{1}{2} \sum e_k^2$. If $E$ is less than the given error value, stop training; otherwise, return to Step (3) and repeat the above step.

### 3.3. BP Model Improved by Genetic Algorithm

Genetic algorithm (GA) is one of the bionic algorithms. BP neural network is improved by GA to get the optimal weight and threshold. Using genetic algorithm to improve BP neural network can not only improve the training speed, improve the accuracy of model prediction, but also overcome the original model’s defect of easily falling into the local extreme value.

The specific steps are as follows:

1. Population initialization. Set the population size as $P$, and the initial population $W=(W_1, W_2, \ldots, W_p)^T$ can be obtained by randomly generating $P$ individuals. Each individual in the population corresponds to a set of weights, equivalent to a chromosome. In order to get the high precision weight, floating-point coding is used to encode it. Each individual is a floating-point string containing the initial weight value and threshold value of a BP neural network.
2. Calculate the fitness value of the population to find the optimal individual. According to the initial weights and thresholds of BP neural network obtained by individuals, BP neural network was trained with training data for prediction, and output results are obtained. Adaptation values were calculated according to the output results:
\[ f_i = \frac{k}{|y_i - \hat{y}_i|}, \quad i = 1, 2, \ldots, P, \]

where \( \hat{y}_i \) is the predicted value of output; \( y_j \) is the expected output value; \( P \) is population size; \( k \) as the coefficient.

(3) Roulette is used to select each chromosome, and the selection probability of each individual \( i \) was \( p_i = \frac{f_i}{\sum_{i=1}^{P} f_i} \), \( i = 1, 2, \ldots, P \)

(4) Cross calculation of the chromosomes selected from the previous generation. The cross-operation method of the No. \( s \) gene \( w_s \) and the No. \( t \) gene \( w_t \) is as follows:

\[ w_s = \alpha w_s + (1 - \alpha) w_t; \quad w_t = (1 - \alpha) w_s + \alpha w_t, \]

where \( \alpha \) is the random number among \([0,1]\).

(5) Variation operation. Variation refers to the random change of the value of some genes of an individual in the population with a small probability, so as to maintain the diversity of the population and realize local random search.

(6) Judge whether the evolution has finished. If not, go back to Step (2) and repeat the above steps.

Based on the above steps, the optimal individual obtained is put into the BP network as the initial weight value and threshold value for training. The specific training process is shown in Figure 1:

---

**Figure 1. Flow Chart of BP Network Optimization by Genetic Algorithm**
4. Simulation Experiment and Analysis

4.1. Data Source and Description

This paper selects the transaction data of the September contract of Shanghai 50ETF option from August 1, 2020 to August 31, 2020 as the research object, with a total of 21 complete trading days and 7,215 transaction data obtained (data source: Monopoly Data Center).

The following calculation of price duration is based on the transaction prices of 7215 transactions in the original data set, where $X_i = T_i - T_{i-1}$ represents the time interval between the occurrence of two events, or duration. In this research, an event is a change in the transaction price. By calculation, a total of 5405 valid price duration data are obtained.

![Figure 2. Logarithmic Price Duration Sequence of Options](image)

As shown in Fig. 2, the vertical axis represents the value of price duration (in seconds). It can be seen that the price duration sequence fluctuates around 3 with continuous rise and decline, indicating that the price duration has aggregation and periodicity with a period of about 190, which is also in line with the intraday effect of high-frequency data.

In order to ensure that no outliers appear in the forecast results of price duration, $\ln(x_i)$ is obtained via logarithmic processing. The autocorrelation graph of the logarithmic price duration series is examined, and the lag order is 580 (three days). As can be seen from Figure 2, the autocorrelation graph of the logarithmic price duration series is trailing and shows obvious aggregation. In addition, it can also be seen from Figure 2 that the fluctuation of the autocorrelation function presents periodicity.

![Figure 3. Autocorrelation Graph of Option Logarithmic Price Duration](image)
The duration can be decomposed into the deterministic part and the random part. This paper fits the deterministic part by the method of linear spline function, and eliminates the random part, to get the intraday trend of the price duration. In Figure 3, the horizontal axis represents the time (in seconds) from 0 am of the day, and the vertical axis represents the price duration. The points in the figure represent the deterministic part of the high-frequency price duration.

![Figure 3. Intraday Trend of Price Duration](image)

The trading time of SSE 50ETF on each trading day is 9:30-11:30 and 13:00-15:00. As shown in Figure 4, there is an intraday effect in the duration of option price. The sequence first decreases and then increases in the morning, and the duration at the closing time in the morning is greater than that at the opening time. The duration at the afternoon first rises to the maximum, and then fluctuates. In a word, the price duration changes show a “V” shape in the morning and a “N” shape in the afternoon. The price duration just before the opening and closing of the day is shorter, and the price duration before and after lunch break is longer.

### 4.2. Selection of Input Nodes

The algorithm is used to screen the indexes and calculate the importance of each index. Set the threshold value to 0.05, that is, the indicators whose importances are greater than 0.05 are regarded as effective.

In the experiment, the XGBoost algorithm calculates the importance of features through gini impure, adopts the gbtree for machine learning, sets the maximum depth of the tree to 6, iterations to 500, and the learning rate to 0.1.

<table>
<thead>
<tr>
<th>Index Name</th>
<th>Index Interpretation</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$oi$</td>
<td>the open interest at the time point of the last duration</td>
<td>0.276</td>
</tr>
<tr>
<td>$ayield$</td>
<td>The average absolute rate of return in the last duration</td>
<td>0.238</td>
</tr>
<tr>
<td>$tint$</td>
<td>The transaction density of the last duration</td>
<td>0.181</td>
</tr>
<tr>
<td>$amargin$</td>
<td>The average price difference between high and low in the last duration</td>
<td>0.139</td>
</tr>
<tr>
<td>$dur$</td>
<td>The last duration</td>
<td>0.113</td>
</tr>
<tr>
<td>$avol$</td>
<td>The average trading volume in the last duration</td>
<td>0.053</td>
</tr>
</tbody>
</table>

The results are shown in Table 1, which shows that the above selected indicators are effective, and the open interest at the time point has the greatest impact. The open interest reflects the market’s interest in the contract. In contrast, the impact of average trading volume is the smallest, and trading volume can reflect the market activity of a contract.
4.3. Empirical Research

In this paper, the simulation model is implemented by MATLAB, and the data is from the data set processed in the previous sections. The dependent variable is the current price duration, and the independent variable is the duration of the previous duration, the average high-low price spread of the previous duration, the average absolute rate of return of the previous duration, and the average trading volume of the previous duration. The first 5234 items of the training set are used for model training, 171 data of the last afternoon are predicted, and the predicted value is compared with the real value. Considering the limitations of different evaluation indexes in practical application, it is difficult for a single evaluation index to comprehensively and comprehensively measure the model training results. Therefore, this paper selects three indexes as model evaluation indexes to evaluate the prediction effect, which are mean-square error (MSE), root mean square error (RMSE), and mean absolute error (MAE). As the most commonly used evaluation index, and MSE, RMSE, and MAE directly reflect the error between the fitting value and the actual value, where \( y_i \) is the actual value of the price duration, \( \hat{y}_i \) is the forecast value of price duration. The smaller the two indexes are, the better the fitting effect of the model is.

\[
\text{MSE} = \frac{1}{k} \sum_{i=1}^{k} (y_i - \hat{y}_i)^2
\]

\[
\text{RMSE} = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (y_i - \hat{y}_i)^2}
\]

\[
\text{MAE} = \frac{1}{k} \sum_{i=1}^{k} |y_i - \hat{y}_i|
\]

Next, the six indexes after XGBoost screening and normalization are substituted into the model as input variables, and the current price duration is used as output variables to train the BP neural network model and the GA-BP model respectively. The prediction results are shown in Table 2. Among them, the population algebra of genetic algorithm is set as 400, the population size is 5000, the crossover probability is 1, and the mutation probability is 0.03. At the same time, the range of hidden layer nodes is 4-12, the maximum number of iterations is 500, and the training error is 0.02.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
<th>RMSE</th>
<th>MAE</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>6.420547</td>
<td>2.533880</td>
<td>27.54314</td>
<td>4.27</td>
</tr>
<tr>
<td>GA – BP</td>
<td>1.659567</td>
<td>1.288242</td>
<td>16.85475</td>
<td>2.85</td>
</tr>
</tbody>
</table>

It can be seen from Table 2 that the BP neural network model with genetic algorithm is the best in all indexes, indicating that the prediction effect of GA-BP model is better than that of single BP model, which improves the prediction accuracy and calculation efficiency. At the same time, in order to verify the rationality of the model, this paper also uses the data of different years and different quarters for comprehensive comparison, the results show that GA-BP model has better prediction effect.

In order to explore the influence of the sample size of the training set on the accuracy and efficiency of the model fitting, the training set and the test set are divided according to different proportions, and the ratios of the training set and the test set are set to 24, 15, 10, 7, 5 and 2 respectively. The data sets divided by different proportions are substituted into the model, and MSE, RMSE, MAE and time (s) are calculated. The results are shown in Figure 5.
As shown in Figure 5, the horizontal axis represents the ratio of the training set to the test set, and the vertical axis is the evaluation index of accuracy and efficiency. From the perspective of prediction accuracy, with the increase of the proportion of training set, the prediction error MSE, RMSE and MAE show a significant downward trend, in which MAE has the largest decline and RMSE is the smallest. From the perspective of prediction efficiency, with the increase of the sample size of the training set, the time consumed by the model operation also fluctuates.

To sum up, this paper uses GA-BP neural network model to predict the price duration of 50ETF option in Shanghai Stock Exchange. After 300 generations of searching, the algorithm proves that the average fitness of chromosome tends to be stable, and the prediction error of the algorithm is less than 10%. Through the above comparative analysis, the stability of BP neural network model optimized by genetic algorithm is better than that of BP neural network model. This shows that the price duration forecasting model is effective.

4.4. Result Analysis
Through the above data analysis and empirical research, the following conclusions can be drawn.
(1) Through the observation of high frequency price duration series, autocorrelation chart and intraday trend, it is found that there is obvious intraday effect in the price duration of Shanghai 50ETF options, which presents “V” type in the morning and “N” type in the afternoon. The results show that the trading frequency is higher in the opening and closing time, and the price fluctuation is larger, so the price risk is also larger.
(2) Through XGBoost calculation of feature importance, it is found that the six indicators selected in this paper have an impact on the price duration. Among them, the position has the greatest impact, and the average trading volume has the least impact.
(3) The results show that GA-BP model can well predict the high frequency price duration series of SSE 50ETF, and the prediction effect is better than single BP model. In the whole process of using genetic algorithm to optimize the traditional BP neural network model, it not only improves the performance of the model, but also overcomes the shortcomings of the original model, such as low prediction accuracy, slow training speed and easy to fall into local extremum.
(4) The research shows that the sample size of training set will affect the accuracy and efficiency of prediction. If higher prediction accuracy is aimed, large sample training should be used; the operation speed is focused on, the training sample size needs to be reduced appropriately.
5. **Summary**

With the rapid development of financial market, more and more economists pay attention to the price duration prediction of options. The price duration of options is not only regulated by the market, but also has many factors that are difficult to control. In view of the characteristics of BP algorithm, such as slow training speed and easily falling into local optimum, this paper introduces genetic algorithm to optimize BP network, and selects Shanghai Stock Exchange 50ETF option price for simulation training. The experiment shows that the improved model overcomes the shortcomings of traditional model and improves the accuracy and efficiency of prediction.

This research provides a new idea for the prediction of option market price duration. It shows that GA-BP model can well predict the high-frequency price duration series of SSE 50ETF, and the prediction accuracy is better than a single BP neural network model, which further proves the applicability of machine learning method in the field of financial prediction.

Financial market data is complex and changeable, especially for high-frequency data, it needs to consider more data features when forecasting, so machine learning method has advantages. In the future, more independent variables can be introduced to improve the prediction accuracy. In addition, although machine learning method has good prediction effect, it weakens the explanatory power of independent variables. Therefore, in the future, the economic explanation of machine learning method for financial data prediction can be further studied.

**References**


