AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 5(5): 4563–4580.
DOI: 10.3934/math.2020293
Received: 09 January 2020
Accepted: 30 April 2020
Published: 22 May 2020

Research article

Stock selection strategy of A-share market based on rotation effect and random forest

Shuai Wang[1], Zhongyan Li[2,*], Jinyun Zhu[2], Zhicen Lin[2] and Meiru Zhong[1]

[1] Department of Finance, Guangdong University of Finance and Economics, Guangzhou City, Guangdong Province, China

[2] Department of statistics and mathematics, Guangdong University of Finance and Economics, Guangzhou City, Guangdong Province, China

* **Correspondence:** Email: 20141230@gdufe.edu.cn; Tel: +136313391161.

Abstract: Due to the random nature of stock market, it is extremely difficult to capture market trends with traditional subjective analysis. Besides, the modeling and forecasting of quantitative investment strategies are not easy. Based on the research of experts and scholars at home and abroad and the rotation effect of large and small styles in China's A-share market, a stock picking strategy combining wheeling effect and random forest is proposed. Firstly, judge the style trend of the A-share market, that is, the relatively strong style in the large and small-sized market. The strategy first judges the trend of the A-share market style, then uses a multi-factor stock selection model through random forest to select stocks among the constituent stocks of the dominant style index, and buys the selected stocks according to the optimal portfolio weights determined by the principle of minimum variance. The empirical results show that the annualized rate of return of the strategy in the eight years from January 1, 2012 to April 1, 2020 is 3.6% higher than that of the single-round strategy, far exceeding the performance of the CSI 300 Index during the same period.

Keywords: quantitative investment; the shift of large cap stocks style and small cap stocks style; random forests; multiple-factor stock selection model; A-share market
Mathematics Subject Classification: 62-XX , 91-XX

1. Introduction

Due to the dynamic nature of stock market volatility, researchers have to formulate different types of stock selection strategies in different periods. The stock selection strategy based on the traditional econometric model cannot dynamically modify the model parameters when the trend of data changed. Therefore, such strategies cannot effectively fit the dynamic trends of the stock market, and it is

difficult to select high-quality stocks in the stock market [1]. With the development of artificial intelligence, big data, cloud computing technology, people gradually know the quantitative investment, which issues trading instructions through quantitative methods and computer programming. Currently, most of quantitative stock selection methods are still based on the simple scoring method and the multi-factor stock selection model of regression method. Therefore, the paper combines the random forest algorithm with strong ability to extract information through data with a multi-factor stock selection model, which makes the stock selection strategy more universal and adaptable.

2. Literature review

In 1987, Merton proposed the rotation effect that the strength and weakness of different types of stocks change with the market. Therefore, domestic and foreign experts have made extensive research about problems that whether there is a style rotation effect in the stock market and how to use the style rotation effect to make a reasonable investment. In foreign studies, Quigley and Sinquefield(2000) [2] studied the relationship between the investment style of institutional investors and their investment income. Froot and Melvyn (2004) [3] studied different types of investment styles from large-cap stocks and small-cap stocks, non-cyclical industries and cyclical industries. In domestic research, Xiao and Wang (2006) [4] empirically tested the momentum of Chinese stock market and pointed out that style momentum reflects the predictability of stock prices. Guo and Wang (2019) [5] used BP neural network to predict the style rotation of the Chinese stock market, and found that the BP neural network model has a better prediction effect on the medium-term style rotation. When judging which stocks have investment value, it is generally necessary to select a factor pool with a high correlation with stock returns as a feature of constructing the multi-factor stock selection model. Mohanram(2005) [6] selected eight indicators from the three aspects of profitability and cash flow performance, growth capacity and book value to market ratio, and based on these indicators to build an index that can affect the establishment of an investment portfolio. Pan (2011) [7] used 9 indicators such as PEG and book-to-market ratio as reference for stock selection. Guerard et al. (2015) [8] combines stock returns, book value, cash flow and sales volume, and momentum in the stock selection model to identify mispriced securities. DeMiguel et al. (2017) [9] based on the investor utility, found that six company characteristics can be used to predict stock returns. Kozak et al. (2019) [10] used PCA to reduce the dimensionality of factors and found that only a few principal component factors are needed to predict the cross-sectional stock returns. In recent years, with the advent of the era of artificial intelligence and the continuous improvement of machine learning algorithms, quantitative technology has gradually become an important technical force in the field of financial investment. The early portfolio investment can be traced back to the "mean variance" model proposed by Markowitz (1952), which has a milestone significance for quantitative investment. After nearly 70 years of development, the technology that combines machine learning algorithms with various stock picking investment theories has become more and more mature. Khaidem and Saha. (2016) [11] used the random forest algorithm to predict stock returns to minimize the risk of stock market investments. Guerard, J.B et al. (2016) [12] generates stock selection models based on robust regression techniques, with significant returns. Phillip (2018) [13] applied neural networks to stock selection strategies and compared it with traditional regression models, drawing the conclusion that deep

learning is more effective for short-term prediction. Dai and Zhu (2020) [14] believe that mixing existing forecasting models can significantly improve prediction performance of stock returns. Dai and Zhou (2020) [15] combine the sum-of-the-parts method, non-negative economic constraint strategy, momentum of return prediction strategy, and three-sigma strategy to improve prediction performance of stock returns. From the existing literature, the existing stock selection strategies are mostly based on simple multi-factor scoring strategies, lacking the processing of high-dimensional and collinear factors. Therefore, this paper proposes a multi-factor stock selection model based on the rotation effect and further introduces the random forest algorithm, so that the model with enhanced data generalization ability can handle high-dimensional data. In the selection of factors, this paper proposes to dynamically select characteristic factors according to the factor contribution of the previous period, which can closely follow the market conditions and the prediction results may be more effective.

3. Research method

3.1. Rotation effect

The rotation of large-cap stock and small-cap stock refers to the phenomenon that the dominant styles in the large-cap stocks and small-cap stocks are alternately replaced. In general, the CSI 300 Index is used as the representative index of large-cap stocks, and the GEM index is the representative index of small-cap stocks. In order to identify the dominant style of the current period, this article chooses an indicator of the relative strength between the large-cap stocks and small-cap stocks, that is, the relative strength index of the index yield: *RSI* (Relative Strength Index), as shown below:

$$RSI_{\mathrm{t}} = \ln \frac{HS\_index_t}{HS\_index_{t-1}} - \ln \frac{GEM\_index_t}{GEM\_index_{t-1}} \tag{3.1}$$

Where $\mathrm{HS\_index_t}$ is the price of the CSI 300 index in day t; $GEM_index_t$is the price of the GEM index in day t; $RSI_t$is the logarithm of the excess return of CSI 300 relative to the GEM in day t; $\mathrm{HS_index_t/HS_index_{t-1}}$is a month yield of the CSI 300 Index, $\mathrm{GEM_index_t/GEM_index_{t-1}}$is a month yield of the GEM index; $RSI$is the difference between the logarithm of a month return rate of the CSI 300 index and the logarithm of a month return rate of the GEM index, being used to judge which style is dominant.

3.2. Multi-factor stock selection

Combined with the discussion of Guerard and Markowitz (2015) [8], the process of multi-factor stock selection is shown in the following Figure 1:

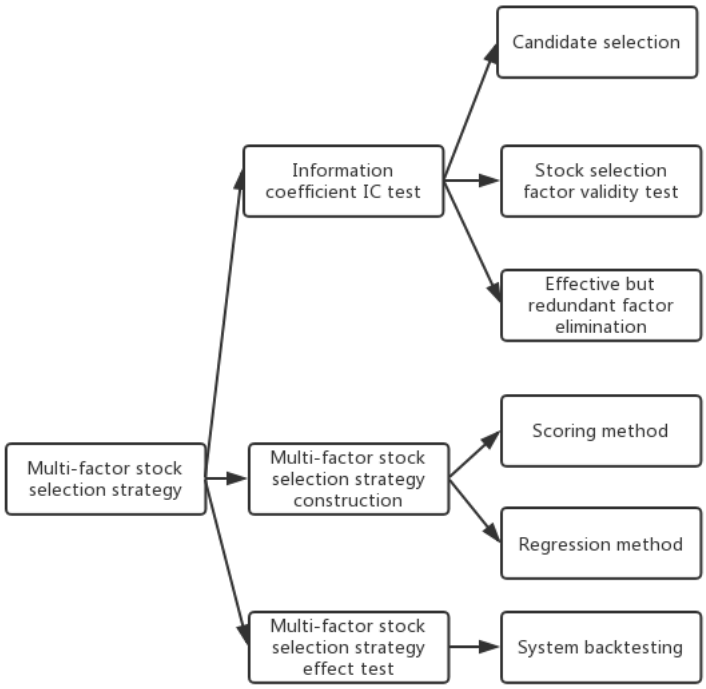

Figure 1. Multi-factor stock selection flow chart.

Multi-factor stock selection strategies generally include single factor information coefficient IC test, multi-factor stock selection strategy construction and multi-factor stock selection strategy effect test. The single factor information coefficient IC test mainly includes the selection of candidate factors, the validity test of stock selection factors for the correlation analysis of the factors and the next yield, and the elimination of the factors with higher correlation between the factors. The effective but redundant factors are eliminated. The construction of multi-factor stock selection strategy includes scoring and regression methods. Here we use similar and regression methods, which use random forest algorithm to predict the next period of profit based on current factors. Finally, the effectiveness of the multi-factor stock selection strategy is tested according to the system back-testing.

3.3. Random forest

3.3.1. Bagging and random forest

Bagging is one of the parallel integrated learning methods. Given a data set containing m samples, first randomly take a sample into the sampling set, and then put the sample back into the initial data set, so that the sample may still be selected at the next sampling. After m random sampling operations performed, a sample set containing m samples is obtained. Some samples in the initial training set appear multiple times in the sample set, and some never appear.

$$H^{oob}(x) = \arg\max_{Y} \sum_{t=1}^{T} I(h_t(x) = y) I(x \notin D_t) \tag{3.2}$$

Where D_t epresents the training sample set actually used by $h_t.H^{oob}(x)$ is the out-of-packet prediction of sample x, that is, only the predictions of the base learners that do not use x training on x.$h_t(x)$ represents a single decision tree classification mode. $y$represents the output variable.I(°) is an indicative function.

The out-of-package estimates for Bagging generalization errors are:

$$\in^{oob} = \frac{1}{|D|} \sum_{(x,y)\in D} I(H^{oob}(x) \neq y) \tag{3.3}$$

Random forest is an extended variant of Bagging. Based on the base learner of decision tree to build Bagging integration, the random forest introduces random attribute selection in the training process of decision tree.

The article assumes that there are a total of d attributes in the sample set. Each decision tree randomly selects a subset of k attributes from the attribute set. In general, take

$$\mathrm{k} = log_2 d \tag{3.4}$$

The schematic diagram of random forest is shown in Figure 2 below:

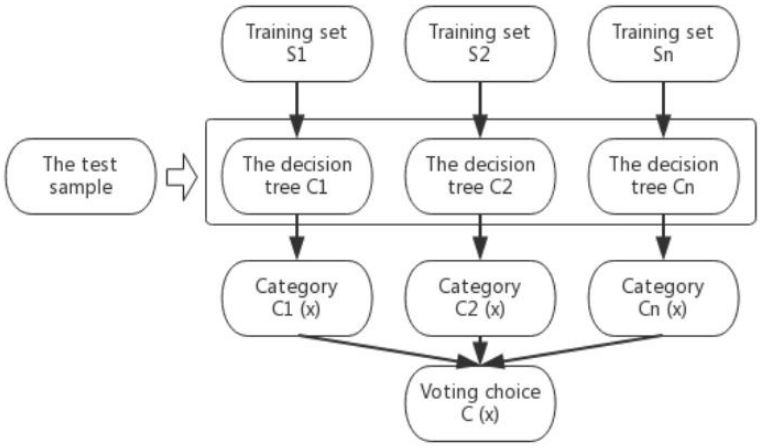

Figure 2. The schematic diagram of random forest.

After each decision tree judges the corresponding category, the learner h_i in the random forest will predict a marker from the category marker set $\{c_1, c_2, \ldots, c_N\}$. The most common combination strategy is to use the voting method, including absolute majority voting, relative majority voting and weighted voting [16].

- Majority voting

$$H(\mathrm{x}) = \begin{cases} c_j, if \sum_{i=1}^{T} h_i^j(x) > 0.5 \sum_{k=1}^{N} \sum_{i=1}^{T} h_i^k(x) \\ reject, otherwise. \end{cases}; \tag{3.5}$$

If a mark scores more than a half, the forecast is the mark; otherwise, the prediction is rejected.

- Plurality Voting

$$H(\mathrm{x}) = c \arg\max_j \sum_{t=1}^{T} h_i^j(x) \tag{3.6}$$

That is the mark that is predicted to be the most votes. If multiple tags have the highest number of votes at the same time, we randomly select one of them.

- Weighted Voting

$$H(\mathrm{x}) = carg \max_{j} \sum_{t=1}^{T} w_i h_i^j(x) \tag{3.7}$$

w_i is the weight of h_i, usually$w_i \geq 0, \sum_{t=1}^{T} w_i = 1$.

3.3.2. Random forest parameters

Random forest contains many important parameters, here are a few more important parameters such as $n\_estimators$, $max\_features$ and max_depth.

- $N\_estimators$
  The number of trees in the forest. The larger the value of this parameter is, the better the model is. But when the parameter reaches a certain level, the accuracy of the random forest often does not rise but start to fluctuate. The larger the parameter value is, the greater the amount of calculation and memory are required, and the training time will become longer and longer.
- $Max\_features$
  Maximum number of features to consider when dividing trees. Generally speaking, if the number of sample features is not large, we can consider all the number of features when dividing.
- Max_depth
  The maximum depth of the tree. In order to maintain the robustness of the random forest, branches exceeding the maximum depth will be cut off.
- $Min\_samples\_split$
  A node must contain at least the number of training samples. This node is not allowed to branch when the number of samples is insufficient.

3.4. Minimum-variance portfolio

According to Markowitz (1952) [17], the objectives and constraints of the 30-stock minimum variance portfolio are as follows:

$$\begin{aligned} &\min \sum_{i=1}^{30} \sum_{j=1}^{30} X_{\mathrm{i}} X_{\mathrm{j}} \sigma_{ij} \\ &\mathrm{st} \begin{cases} E = \sum_{\mathrm{i}=1}^{30} X_i u_i \\ \sum_{\mathrm{i}=1}^{30} X_i = 1 \\ 0 \leq X_i \leq 1 for \quad i = 1, \ldots, 30 \end{cases} \end{aligned} \tag{3.8}$$

Where X_i,X_j is the investment ratio of the i^{th} and j^{th} stocks, u_i is the average return of the i^{th} stock in one year (252 days), and σ_{ij} is the covariance of the 252-day daily return data of i^{th} and j^{th} stocks.

3.5. Research ideas

The research ideas in this article are based on the rotation of large-cap stock and small-cap stock in China's A-share market, and the change in stock price is based on the fundamental, technical and macro-level economic situation of listed companies. First of all, we will analyze the current situation of whether there is the rotation of large-cap stock and small-cap stock in China's A-share market, and judge the market's style trend by the magnitude of the rise in the previous 3 months of the CSI 300 index (large-cap stock) and the GEM index (small-cap stock). Style as a stock pool. Combining Guerard and Markowitz (2015) to calculate the information coefficient IC for the factor, then we make a correlation analysis between factors and returns, and between factors, and select the growth ability, profitability, quality of return, valuation and market value factors as the alternative factors from the factor bank, which are closely related to return rate. We fix the frequency of shifts, screen the ten most significant characteristic factors through the factor importance in the previous period, use the random forest to predict the stock return rate based on the ten characteristic factors, and then buy the stocks with the highest number of votes after voting through many decision trees, combined with Markowitz (1952 [17], 1959 [18], and 1987 [19]). The number of selected stocks is 30, and the purchase weight is based on the minimum risk of Markowitz effective boundary [19–21]. Finally, a simple stop-loss strategy is added to obtain the Shanghai Composite Index's five-day gain before the entire strategy. If it is less than- 0.1, it will be cleared, and the following stock purchase operations will not be performed.

In this article, the previous stock data of this article is used as training data, using January 1, 2012 to April 1, 2020 as the interval for strategy stock selection. The frequency of position swaps is one month (position adjustment at the end of each month), and the 30 selected stocks are bought based on minimizing risk weights. The initial capital is set to 100,000, the daily buying rate is set to 0.1%, and the selling rate is set to 0.4% (including 0.3% stamp duty).

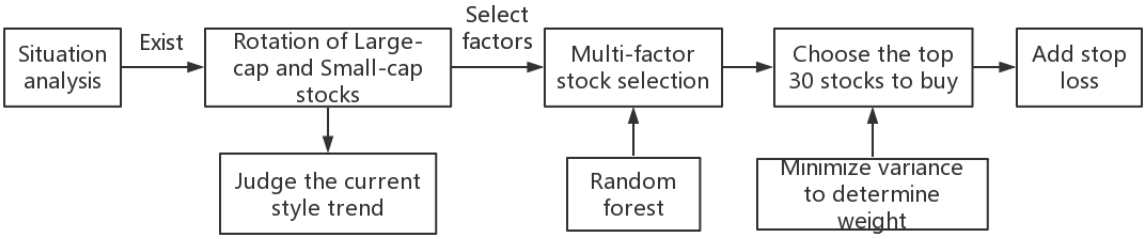

Figure 3. Research ideas.

4. Rotation effect of A-share market

Although it is intuitively obvious that there is a phenomenon on rotation of large-cap stocks and small-cap stocks in China's A-share market, in order to further study the current situation of China's rotation of large-cap stocks and small-cap stocks, this article uses historical data to verify it. If the expected absolute value of the RSI index is too small, the variance is too large, the value has an upward or downward trend, and the value fluctuates around 0, we can judge whether A-share market has the rotation of large-cap stocks and small-cap stocks by that index.

4.1. Graph analysis

From the line chart of RSI, we can see that the index basically fluctuates around 0, and therefore conclude that A-share market has the rotation of large-cap stocks and small-cap stocks.

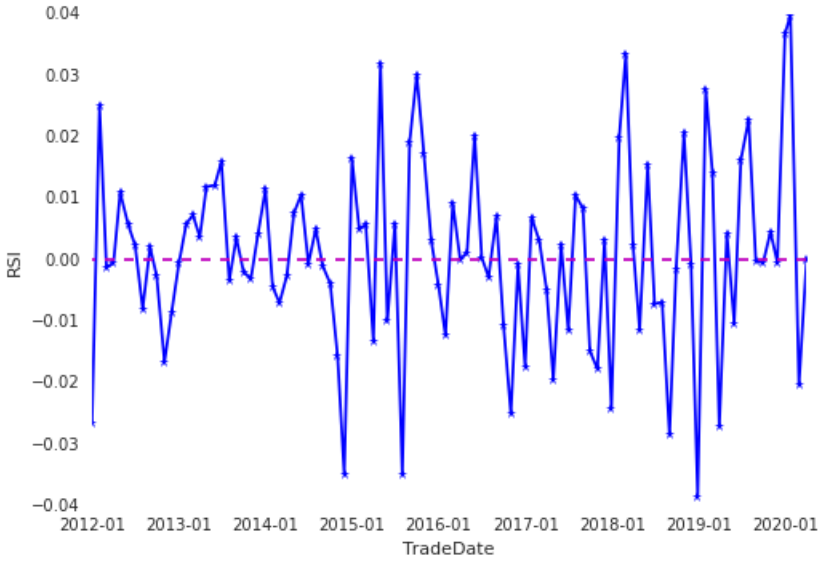

Figure 4. Line chart of RSI.

4.2. Assumes

Due to the following characteristics of financial data:

- The data distribution has the characteristics of "spikes and thick tails";
- Data fluctuation has the characteristics of high fluctuation aggregation.

And this article assumes:

- The average (or median) RSI is equal to 0, indicating that although the market has a rotation effect, the overall return rate tends to 0;
- The larger the standard deviation of the RSI is, the greater the tendency of the market to invest in different styles in different periods is.

In such cases, the standard parametric test is not applicable, so the non-parametric symbol test combined with the standard deviation of descriptive statistics is used to determine whether the market has a rotation effect.

Based on the line chart of RSI, to verify the existence of the rotation of large-cap stocks and small-cap stocks in China's A-share market, we firstly calculate the expected value of the RSI within the interval. The closer to 0 the RSI expectation value is, the more significant the rotation effect is; If the absolute mean of RSI is larger, it indicates that the market bias of a certain style is greater, and the rotation effect is not significant at this time. Secondly, we calculate the standard deviation of the RSI within the interval. The larger the RSI standard deviation is, the greater the market volatility and the

significant rotation effect is; If the RSI standard deviation is smaller, that is, the market volatility is small, there is no rotation basis.

According to the above logic, this article conducts statistical analysis on RSI, the steps of symbol checking are as follows:

- Constructing null hypothesis and alternative hypothesis:
  Null hypothesis $H_0 : M\mathrm{e} = 0$;
  Alternative hypothesis $H_1 : M\mathrm{e} \neq 0$
- Constructing hypothetical statistics Hypothetical statistics $S^+ = \sum_{\mathrm{i}=1}^{\mathrm{n}} Y_i$, $S^- = \sum_{\mathrm{i}=1}^{\mathrm{n}} Z_i$, in the formula:

$$Y_i = I\{\mathrm{x}_i > 0\} = \begin{cases} 1, \mathrm{x}_i > 0 \\ 0, \mathrm{x}_i \leq 0 \end{cases} \tag{4.1}$$

$$Z_i = I\{\mathrm{x}_i \leq 0\} = \begin{cases} 1, \mathrm{x}_i \leq 0 \\ 0, \mathrm{x}_i > 0 \end{cases} \tag{4.2}$$

- Finding the P value

4.3. Statistical analysis

According to the symbol test, the following results are obtained:

Table 1. Statistical analysis of RSI.

Relative Strength Index	Expected Value	Standard Deviation	Maximum	Minimum
RSI	0.0010	0.0153	0.0398	-0.0384

As can be seen from the above table, the mean of RSI is 0.0010, which is close to 0, and the standard deviation is 0.0153. There is a large fluctuation, so it is sufficient to prove that there is the rotation of large-cap stocks and small-cap stocks in China's A-share market.

4.4. Symbol checking

In this article, the median of RSI is tested by non-parametric symbolic test. The test results are as follows:

Table 2. Symbol checklist.

Inspection Type	Statistics S+	Statistics S-	P-values
Sign Test	51	48	0.84

The P-value of the symbol test is greater than 0.05, and the null hypothesis cannot be rejected, so the median of the RSI can be considered close to 0. According to the above statistical analysis results, A-share market has the rotation of large-cap stocks and small-cap stocks.

5. Strategic evidence

5.1. *Large and Small Wheel style rotation strategy*

The specific steps of the rotation strategy of large-cap stocks and small-cap stocks are as follows:

- Get the magnitude of the rise of GEM index and the CSI 300 index in the previous 3 months of the test.
- If the magnitude of the rise of both indexes are less than 0, the clearance is selected to increase the profit and reduce the risk of retracement.
- If the magnitude of the rise of both indexes are not less than 0:
  If the magnitude of the rise of the CSI 300 index is more than that of the GEM index, we select 30 stocks with the maximum market value from the CSI 300 index stocks.
  If the magnitude of the rise of the GEM index is more than that of the CSI 300 index, we select 30 stocks with the minimum market value from the GEM index stocks.
- Obtain the closing price of the first 252 days of the selected 30 stocks to calculate the daily rate of return, determine the weight of 30 stocks according to the minimum variance, and buy and hold according to the weights.
- Add a simple stop-loss strategy: Get the Shanghai Composite Index's five-day gain before the entire strategy, if it is less than -0.1, then the strategy clears the position, and does not perform the following stock purchase operations.

The back-testing results are shown in Table 3 and Figure 5:

Table 3. Multi-disk rotation back-testing information table.

Annualized Rate of Return	**Benchmark CSI 300 Index**	α	β
25.80%	5.80%	21.10%	0.52
Information Ratio	**Maximum Withdrawal**	**Annual Turnover Rate**	**Sharpe Ratio**
0.71	45.1%	1314.21%	0.86

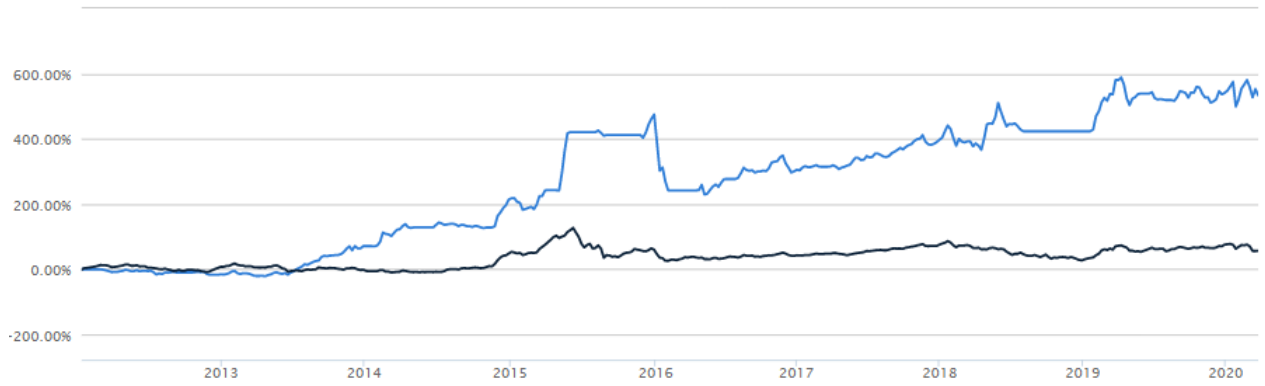

Figure 5. Back-up diagram of the rotation strategy of Large-cap stocks and Small-cap stocks.

5.2. *Combination strategy in the rotation of Large-cap Stocks and Small-cap Stocks and random forest*

5.2.1. The rotation of Large-cap stocks and Small-cap stocks

According to the rotation strategy of large-cap stocks and small-cap stocks, if the current large-cap stocks style dominates, the 60 stocks with the maximum market value will be selected from the CSI 300 index stocks as the forecast stock pool of the random forest. If the current small-cap stocks style dominates, 60 stocks with the minimum market value are selected from the GEM index stocks as the forecast stock pool of the random forest.

5.2.2. Selection of candidate factors

By correlating between the 237 factors in the factor pool from June 1, 2010 to December 31, 2011, and the next return rate and between the factors, we select the factors with larger correlation coefficient with the rate of return and eliminate the factors with greater multicollinearity.

$$r_{x_k y} = \frac{\sum_{i=1}^{ki} (x_{ki} - \bar{x}_k)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{ki} (x_{ki} - \bar{x}_k)^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{5.1}$$

$$r_{x_k x_j} = \frac{\sum_{i=1}^{ki} (x_{ki} - \bar{x}_k)\left(x_{ji} - \bar{x}_{\mathrm{j}}\right)}{\sqrt{\sum_{i=1}^{ki} (x_{ki} - \bar{x}_k)^2 \sum_{i=1}^{ji} \left(x_{ji} - \bar{x}_{\mathrm{j}}\right)^2}} \tag{5.2}$$

$rx_k y$,$rx_j y$ represent the correlation coefficient between factor $\mathrm{x_k}$ and yield y, factor $\mathrm{x_k}$ and factor $\mathrm{x_j}$, respectively. Select the first 40 factors with the largest absolute value of the correlation coefficient, that is $\max top\,|rx_k y|$. If $\left|\mathrm{rx_k x_j}\right| > 0.8$, only select the factor with the largest correlation coefficient with the yield, that is $x_s = \max\left(r_{x_k y}, r_{x_{kj} y}\right)$.

Since there are many factors used for stock analysis in the market, not all factors have an impact on stock returns. According to the previous literature review, this paper combines "Analysis of Quantitative Stock Selection Based on Multi-Factor Model" (Xu, 2017) [22] with "Research on Multi-Factor Quantitative Stock Selection Based on Self-Attention Neural Network" (Zhang, 2020) [23] to select 35 factors in terms of quality, value, growth, momentum, technical indicator, sentiment and basic subject to measure the performance of stocks. Then according to the above steps, a total of 19 factors with high correlation with yield were selected. Then combine these 19 factors with some common growth ability factors, profitability and income quality factors, as well as valuation and market value factors, including 35 factors such as PE, PB, ROA, ROE, etc. The total of 54 factors are used as candidate factors shown in the following table:

Table 4. Factor classification.

Classification	Short name	Factor name
Technical indicators	Mass Index	Mass Index
	MTM	Momentom Index
	EMV6	Ease of Movement Value
	Swing Index	Swing Index
	Chaikin Volatility	Chaikin Volatility
Growth	EGRO	Five-year earnings growth
	Net Asset Grow Rate	Net assets growth rate
	Net Profit Grow Rate	Net profit growth rate
	Operating Revenue Grow Rate	Operating revenue growth rate
	Total Profit Grow Rate	Total profit growth rate
	Operating Profit Grow Rate	Operating profit growth rate
	Financing Cash Grow Rate	Net cash flow growth rate from financing activities
	Total Asset Grow Rate	Total assets growth rate
	OperCash Grow Rate	Net cash flow growth rate from operating activities
	SUOI	Standardized unexpected gross income
Momentum	RC24	Rate of Change
	Aroon	The number of periods that have elapsed since the price reached the highest and lowest recent values
	Aroon Down	Aroon factor intermediate variable
	PLRC6	Price Linear Regression Coefficient
	EARNMOM	Change tendency of net profit in the past eight quarters
Value	PE	Price-earnings ratio
	CTOP	Cash flow to price
	ETOP	Earnings to price
	LFLO	Natural logarithm of float market values
	LCAP	Natural logarithm of total market values
	PB	Price-to-book ratio
	PCF	Price-to-cash-flow ratio
	PS	Price-to-sales ratio
	TA2EV	Assets to enterprise value
Mood indicators	ACD6	Accumulative Distribution
	VOL5	Turnover Rate
	TVSTD6	Turnover Value Standard Deviation
	VDEA	Volume Difference Exponential Average
	VR	Volume Ratio
	WVAD	William's variable accumulation distribution
	BR	Willingness index
	Klinger O scillator	Volume swing indicator
	VSTD10	Volume Standard Deviation
Quality	DEGM	Growth rate of gross income ratio
	ROA	Return on assets
	ROE	Return on equity
	Sale Service Cash To OR	Sale service cash to operating revenues
	Admini Expense Rate	Administrative expense rate
	NOCF To Operating NI	Net cash flow from operating activities Ratio to net income
	Cash To Current Liability	Cash to current liability
	Cash Rate Of Sales	Cash rate of sales
	OperCash In To Current Liability	Cash provided by operations to current liability
	CFO2EV	Cash provided by operations to enterprise value
	NP To TOR	Net profit to total operating revenues
	Operating Profit Ratio	Operating profit ratio
	Net Profit Ratio	Net profit ratio
	Gross Income Ratio	Gross income ratio
Basic Subjects	ASSI	Natural logarithm of total assets
Index for each stock	EPS	Earnings per share

5.2.3. Filter important feature factors based on factor importance

Each time the position is changed, it is necessary to dynamically select the feature factor with the greater importance of the factor in the current period according to the previous market situation. First, the stocks in the stock pool are ranked by factors, and the top 20% and the bottom 20% stocks are selected to form two combinations; This article divides the average return difference of the two portfolios and the average return difference of the top 20% and the last 20% of all stocks in the stock pool , and the ratio is the factor contribution; The ten important characteristic factors with the largest factor contribution in the previous day for each back-testing day are selected. Each decision tree in the random forest predicts the stock price return rate through ten random factors, and opens (changes positions) to purchase the 30 stocks with the highest votes.

$$c_i = abs\left(\frac{I_{Fi-}I_{Li}}{M_{Fi-}M_{Li}}\right) \tag{5.3}$$

$$x_k = max_{10}\left(c_i\right) \tag{5.4}$$

$$i = 1, 2 \ldots, 6, k = 1, 2, \ldots, 10$$

I_{Fi} is the average rate of return for the top 10% of the factor values of a certain factor of a certain index component, I_{Li} is the average rate of return for the last 10% of the factor values of a certain factor of a certain index component, M_{Fi} is the average rate of return for the top 10% of the factor values of a certain factor in the market. M_{Li} is the average rate of return for the last 10% of the factor values of a certain factor in the market. x_k is the feature factor of the top 10 largest factor contributions selected from 46 candidate factors.

5.2.4. Multi-factor stock selection based on random forest

If the data is missing, assign it a value of 0. Select 40% stocks before and after 60 stocks to eliminate the noise in the middle. Binarize the average return rate of the first five trading days of each stock. If it is greater than the average return rate of all selected stocks in the previous month, it is recorded as +1. If it is less than that, record it as -1. 10 feature factors with maximum factor contribution are used as input variables, and the average rise and fall after binarization are used as output labels. The strategy randomly divides the previous period factor data and output labels at the time of warehouse swap into 90% training set and 10% test set, outputs the training model whose test accuracy reaches a certain standard, and predicts the current rate of return of each constituent stock in the stock pool. Finally, select the 30 stocks with the highest yield prediction value, obtain the closing price of the first 252 days of the selected 30 stocks to calculate the daily yield, determine the weight of 30 stocks according to the minimum variance, and buy and hold according to the weights. Obtain the first five days of the Shanghai Composite Index's gains before the entire strategy. If it is less than -0.1, the strategy will clear the position and do not perform the following stock purchase operations. The overall idea of the strategy is as follows:

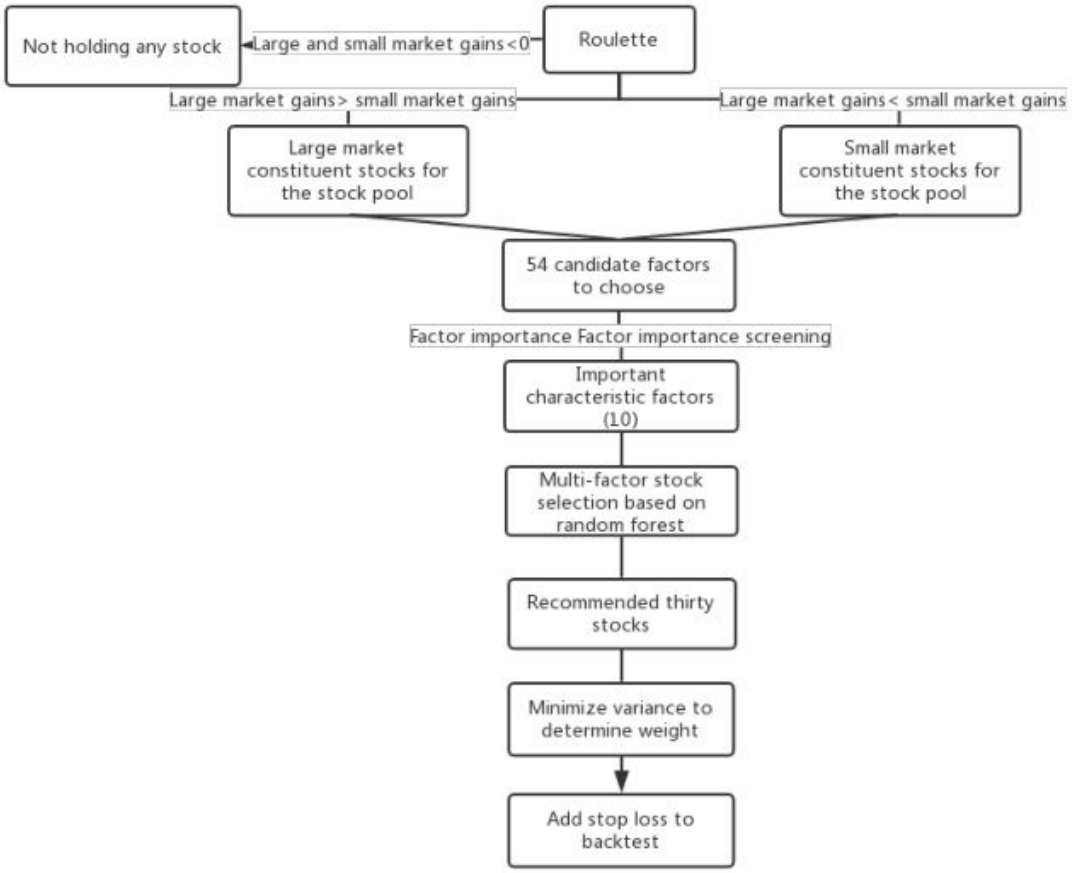

Figure 6. The overall idea of the combination strategy.

Set the number of decision trees of random forest to 100, 500 and 1000 respectively. The backtest results from January 1, 2012 to April 1, 2020 are shown in Table 5:

Table 5. Combination strategy back-testing information table.

Number of decision trees	**annualized return**	**Benchmark annualized return**	α	β	**Information ratio**	**Maximum drawdown**	**Annualized turnover**	**Sharp Rate**
100	17.80%	5.80%	13.10%	0.52	0.45	39.10%	1398.55%	0.54
500	29.40%	5.80%	24.70%	0.53	0.77	44.20%	1439.53%	0.92
1000	24.50%	5.80%	19.80%	0.53	0.64	39.30%	1496.74%	0.75

The effect of combining the random forest of 500 decision and the rotation effect is better, and the back-testing results are shown in Figure 7:

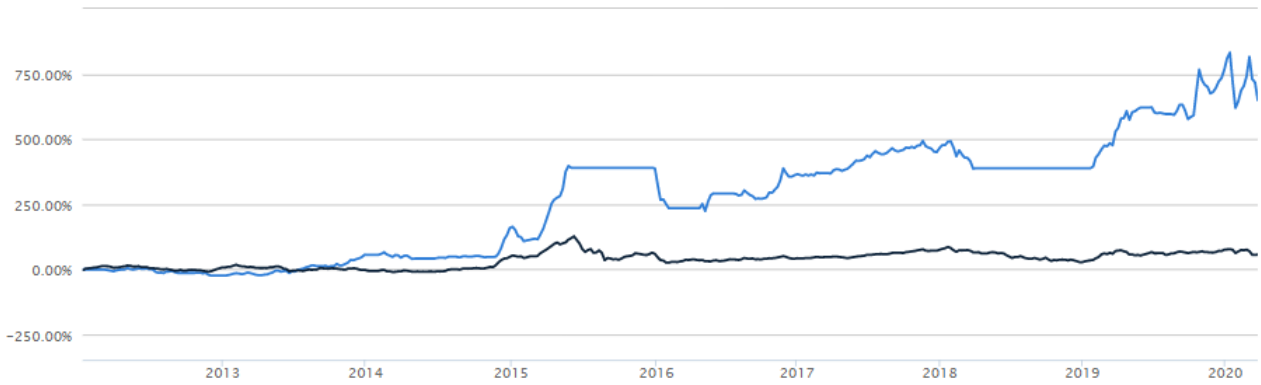

Figure 7. Back-testing chart of combined strategy.

The average accuracy of the random forest on the validation set is 0.66, and the line chart is drawn as shown in Figure 8:

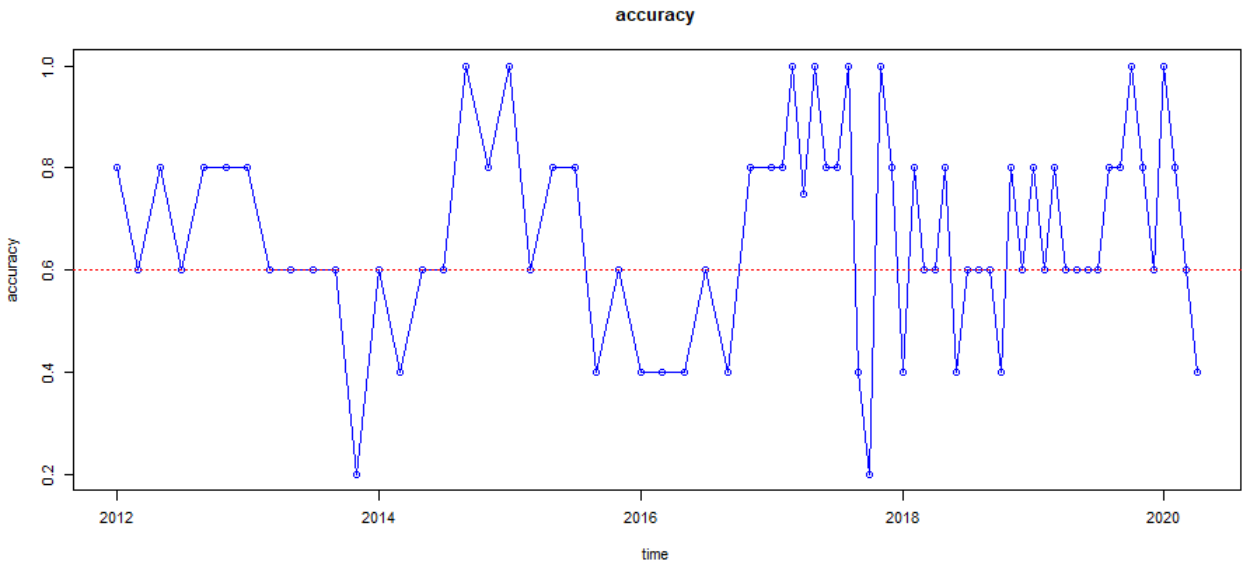

Figure 8. Random forest accuracy.

6. Results analysis

The back-testing index pairs of the two strategies are shown in Table 6:

Table 6. Back-testing table of two strategies.

Strategy type	annualized return	α	β	Sharp Rate	Maximum drawdown	Annualized turnover
Rotation strategy	25.80%	21.10%	0.52	0.86	45.10%	1314.21%
Combination strategy	29.40%	24.70%	0.53	0.92	44.20%	1439.53%

From the comparison of the back-testing results, the annualized return rate of the rotation strategy is 25.80%, and the combined strategy is 29.40%. The combined strategy effect is slightly higher than the rotation strategy effect. The β value of the rotation strategy is 0.52, and the β value of the combination strategy is 0.53, which indicates that the combination strategy is more sensitive to the systemic risk of the stock market. The Sharpe ratio of the rotation strategy is 0.86, and the Sharpe ratio of the combination strategy is 0.92. From the perspective of the excess return of unit risk, the rotation strategy has more advantages than the combination strategy. The maximum drawdown of the rotation strategy is 45.10%, and the maximum drawdown of the combination strategy is 44.20%, which indicates that the combined strategy has better risk control than the rotation strategy.

Comparing the strategy of this article with the quantitative investment funds in the market, the three-year annualized return rate comparison is shown in Table 7:

Table 7. Back-testing table of two strategies.

Strategy / fund company name	annualized return	Strategy / fund company name	annualized return
Rotation strategy	25.80%	**Cathay Juxin Value Advantage Hybrid Fund A**	21.97%
Combination strategy	29.40%	**Changsheng Medical Industry Quantitative Allocation of Stock Fund**	21.72%
Bocom Alpha Core Hybrid Fund	27.07%	**Cathay Juxin Value Advantage Hybrid Fund C**	21.12%
Agricultural Bank of China Strategic Hybrid Fund	23.10%	**Everbright Industry Rotation Hybrid Fund**	19.31%

It can be seen that the rotation strategy and the combination strategy of this article can outperform the quantitative investment funds in the market, indicating that the strategy of this article has the value of in-depth research.

7. Conclusions

The quantitative stock selection strategy for the market periodic rotation effect and random forest constructed in this paper is improved on the basis of the traditional style stock selection and multi-factor stock selection strategy. The random forest algorithm in machine learning, to some extent overcome the subjectivity, blindness and speculation of investment. It is a rational investment with significant returns. The selection of characteristic factors in the random forest changes with market conditions, and the frequency of changing positions is appropriate. It keeps up with the market while ensuring transaction costs and avoids the lag of stock market information. The average prediction accuracy of random forest is 0.67, and empirical results show that the strategy based on the rotation effect and the random forest algorithm has a good performance. Therefore, the strategy of this paper has certain efficiency in stock selection of China's A-share market and has certain reference value.

However, due to the limited level of research, the feature factor selection in this paper only roughly selected 54 important factors based on the analysis of correlation and collinearity among more than 200 stock market indicators. The random forest algorithm also does not use the rolling prediction method and loses some accuracy. In terms of risk control, the stop loss strategy is relatively simple and needs further research and optimization.

Acknowledgments

Research project of Humanities and Social Sciences in general colleges and universities of Guangdong Province (2015WQNCX038); National University Student Innovation Training Project (201910592003); Guangdong provincial department of education young creative talents project (2016WQNCX046).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. H. Yuan, *Overview of the application of data mining models in stock market prediction*, Chin. Collect. Econ., **33** (2017), 66–67.
2. G. Quigley, R. Sinquefield, *Performance of UK Equity Unit Trusts*, J. Asset Manage., **1** (2000), 72–92.
3. K. Froot, M. Teo, *Style investing and institutional investors*, J. Financ. Quant. Anal., **43** (2008), 883–906.
4. J. Xiao, Y. X. Wang, W. Z. Chen, et al. *Empirical Research on style level momentum strategy in the China stock market*, Finance Econ., **03** (2006), 23–29.
5. Y. R. Guo, X. L. Wang, *Rotational forecast of stock market over and under style based on BP neural network*, Comput. Simulat., **36** (2019), 239–242.
6. P. S. Mohanram, *Separating winners from losers among low Book-to-Market stocks using financial statement analysis*, Rev. Account. Stud., **10** (2005), 133–170.
7. F. Pan, *Multi-factor Stock Selection Model Based on Effective Factors*, Shenzhen: Anxin Securities Research Center, (2011), 1–21.
8. J. B. Guerard Jr., H. M. Markowitz, G. L. Xu, et al. *Earnings forecasting in a global stock selection model and efficient portfolio construction and management*, Int. J. Forecasting, **31** (2015), 550–560.
9. V. DeMiguel, A. Martin-Utrera, F. J. Nogales, et al. *A Transaction-Cost perspective on the multitude of firm characteristics*, LBS Working Paper, 2017.
10. S. Kozak, S. Nagel, S. Santosh, et al. *Shrinking the Cross- section*, J. Financ. Econ., **135** (2020), 271–292.
11. L. Khaidem, S. Saha, S. R. Dey, et al. *Predicting the direction of stock market prices using random forest*, Appl. Math. Financ., (2016), 1–20.
12. J. B. Guerard, R. A. Gillam, H. Markowitz, et al. *Data mining corrections testing in Chinese stocks*, Interfaces, **48** (2018), 108–120.
13. H. Phillip, *Using autoregressive modelling and machine learning for stock market prediction and trading: ICICT 2018*, Adv. Intel. Sys. Comput., **2** (2018), 767–774.
14. Z. F. Dai, H. Zhu, *Stock return predictability from a mixed model perspective*, J Pacific-Basin Financ. J., (2020), 101267.
15. Z. F. Dai, H. T. Zhou, *Prediction of stock returns: Sum-of-the-Parts method and economic constraint method*, Sustainability, **12** (2020), 541.
16. Z. H. Zhou, *Machine Learning*, Tsinghua University Press Bei Jing, 2016.
17. H. M. Markowitz, *Portfolio selection*, J. Financ., **7** (1952), 77–91.

18. H. M. Markowitz, *Portfolio selection: Efficient diversification of investment*, John Wiley and Sons, New York, 1959.
19. H. M. Markowitz, *Mean-Variance analysis in portfolio choice and capital markets*, Basil Blackwell, London, 1987.
20. J. B. Guerard, H. M. Markowitz, G. Xu, et al. *Earnings forecasting in a global stock selection model and efficient portfolio construction and management*, Int. J. Forecasting, **31** (2015), 550–560.
21. H. M. Markowitz, *Risk-Return analysis: The theory and practice of rational investing*, McGraw-Hill, New York, 2013.
22. J. Z. Xu, *Analysis of quantitative stock selection based on multi-factor model*, Financ. Theor. Explorat., **3** (2017), 30–38.
23. H. Zhang, H. L. Shen, Y. C. Liu, et al. Research on Multi-factor Quantitative Stock Picking Problem Based on Self-Attention Neural Network. Mathematical Statistics and Management, 2020. Available from: `https://doi.org/10.13860/j.cnki.sltj.20200403-001`.

AIMS Press