Research on Multi-factor Stock Selection Strategy based on Improved Particle Swarm Support Vector Machine

Canran Xiao [1], Liwei Hou [2], Jun Huang [1]

[1] College of Business Administration, Hunan University, Hunan 410082, China

[2] College of Mechanical and Electrical, Central South University, Hunan 410083, China

Abstract. In recent years, the application of machine learning in quantitative trading has attracted more and more attention. In this paper, a new support vector machine algorithm based on nearest neighbor method and improved discrete particle swarm optimization is proposed. Taking Hu-Shen 300 stocks as the research object, the improved support vector machine and the original support vector machine are used to construct multi-factor stock selection strategies respectively, and the two strategies are compared in the back-test analysis. The results show that the multi-factor stock selection strategy of support vector machine based on nearest neighbor method and improved discrete particle swarm optimization has higher annual return than the original SVM algorithm, and has good execution effect.

Keywords: SVM algorithm, nearest neighbor method, improved discrete particle swarm optimization.

1. Introduction

The rise of quantitative investment can subvert traditional investment. It is an active investment process based on market inefficiency and using computer and mathematical tools to establish investment models to express investment ideas. The development of China's stock market is much shorter than that of Europe and the United States. It is much worse than that of developed countries in terms of market management, market operation and market development. The rapid development of Europe and the United States in financial securities market is related to the adoption of quantitative investment.Therefore, China urgently needs to follow the footsteps of developed countries, explore better valuation theory and explore appropriate investment strategies in combination with China's stock market, which has great potential. This paper hopes to construct a more stable and sensitive stock selection strategy by combining multi-factor theory with support vector machine algorithm.

2. A Multi-factor Stock Selection Strategy based on Improved Particle Swarm Support Vector Machine

2.1 Multi-factor Quantitative Stock Selection Strategy

The most widely used and easy-to-use method of quantitative investment is multi-factor quantitative stock selection strategy. At present, the three-factor model proposed by Fama-French is widely used. Their theory is expressed by formula (1):

$$r = R_f + \beta_1(R_m - R_f) + \beta_2 \cdot SMB + \beta_3 \cdot HML + \alpha \tag{1}$$

SMB refers to the risk factor of company size variable. The difference between the return of portfolio composed of small-capitalized companies and that of large-capitalized companies is expressed by the return of portfolio; HML refers to the risk factor of the market-to-net premium. The difference between the portfolio return of a company with higher book value and that of a company with lower book value is expressed as the difference between the portfolio return of a company with higher book value and that of a company with lower book value;Alpha means excess return. Ideally, the excess return of the portfolio will be explained by three factors, so that alpha should be statistically equal to 0.

The key to the success of multi-factor model in stock selection is that multi-factor model can accurately explain the relationship between factors and returns, but the main problem is that it is difficult to fully explore the impact of each factor on asset returns.Moreover, the deep-seated correlation between factors is often neglected, and some internal relations in reality can not be expressed by simple multiple regression.Therefore, this paper innovatively uses Support Vector Machine (SVM) to construct multi-factor model, and takes the final training model as the basis for stock selection.

2.2 Improved Particle Swarm Optimization SVM

There are many similar data in the quantitative transaction data, and there are many factors involved in the quantitative stock selection, which lead to the investment decision (classification) time becomes longer, and the waste of computing resources.The parameter selection of SVM is closely related to the classification effect. The traditional method is manual selection, but the effect is not ideal, and it wastes time.Therefore, this paper proposes a new SVM algorithm based on Nearest Neighbor (NN) and improved Discrete Particle Swarm Optimization (DPSO) when applying SVM algorithm to quantify investment.This algorithm prunes the training set by using the nearest neighbor method, removes some samples which are not very useful, optimizes feature selection and SVM parameters by using the improved discrete particle swarm optimization algorithm, and can quickly select the appropriate parameters of SVM.

In each iteration of the discrete particle swarm optimization (DPSO), each particle is constantly updated with these two optimal values. After the two optimal values are found, the particle can adjust its speed and position according to formula (2). The velocity and position equations of particles satisfy the following equations:

$$\begin{cases} \mathrm{v}_{\mathrm{id}}(\mathrm{t}+1)=\mathrm{wv}_{\mathrm{id}}(\mathrm{t})+\mathrm{c}_1\mathrm{rand}()(\mathrm{p}_{id}-x_{id}(t))+c_2 rand()(p_{gd}-x_{id}(t)), \\ S(v_{id}(t+1))=\tanh((k+1)v_{id}(t+1)), \\ x_{id}(t+1)=1 \quad if\ rand()<S(v_{id}(t+1)), \\ x_{id}(t+1)=0 \quad else \end{cases} \tag{2}$$

In order to improve the speed of optimization, this paper proposes a new improved discrete algorithm, which uses interval trisection instead of random probability in reference [1].It is found that this method can shorten the optimization time and ensure that there is little difference in the time of each experiment.If a random number satisfies rand()<0.33,then x_{id}(new)=x_{id}(old).Otherwise, if the random number satisfies rand()≥0.33 and vid < 0.66, there is x_{id}(new)=p_{id}.For other cases, x_{id}(new)=p_{gd} can be used.rand() is a random number in the range of (0,1), pid is the optimal value of the particle itself, pgd is the global optimal value, and vid is the velocity, which is a uniformly distributed random number in an interval in the range of (0,1). A balance should be maintained between local search ability and global search ability.

In SVM, the relaxation factor ξ represents the wrong expectation in the process of sample data classification.Both penalty parameter C and relaxation factor need to be optimized.In order to simplify the calculation, the parameters σ, C and ξ are recorded as t1, t2 and t3 respectively, where σ is the kernel parameter.The feature set has n feature vectors.The feature selection set is denoted as $F=\{f_1,f_2,...,f_n\}$,and $f_i=1,i=1,...,n$ represents that the ith feature is selected. $f_i=0,i=1,...,n$ represents that the ith feature is not selected. Combining the above parameters, the training model of SVM can be constructed. The performance E of SVM is the objective function to optimize the training model of SVM. It can be expressed as equation (3) and used as the adaptive function of the improved discrete particle swarm optimization.

$$\begin{cases} \max E(T,F) \\ T=\{t1,t2,t3\},t_i>0,i=1,...,l \\ F=\{f_1,f_2,...,f_n\},f_i=\{0,1\},i=1,...,n \end{cases} \tag{3}$$

The kernel function chosen in this paper is the radial basis function:

$$k(x_i,x_j) = \exp(-(x_i - x_j)^2/(2\sigma^2)) \quad (4)$$

The improved particle swarm optimization SVM algorithm is as follows:

Firstly, the nearest neighbor (NN) method is used to prune the training set. Find the nearest sample points of each sample and determine whether they belong to one class or not. If so, keep the sample. Otherwise, remove the sample from the training set.

Initialize the particle swarm. The parameters such as t1, t2 and t3 should be each particle involved in the optimization. Each parameter takes its value within its limits.

Calculating Particle Fitness Value.

Compare the current fitness of a particle with its historical maximum. If the maximum value of the particle itself is less than the current fitness value, it is replaced. Otherwise, the fitness value remains unchanged. The maximum fitness in particle swarm optimization is selected as the global maximum. Updating the position and velocity of particle swarm according to the improved discrete PSO algorithm equation.

Determine whether the termination condition is satisfied: if it is, stop iteration; otherwise, return to step c.

3. Empirical Analysis

3.1 Stock Factor Selection

In the strategy, the constituent stocks of Shanghai and Shenzhen 300 are traded, and 11 stock factors are selected according to four categories, as shown in Table 1.

Table 1. Stock Factor Selection

Index	Factor	Illustration
Profit indexes	P/E ratio	Share price/earnings per share
	Price-to-sales Ratio	Share price/sales per share
	Earnings per share	-
Growth indexes	Net profit growth rate	-
	Net asset growth rate	(Net assets at the end of the period-Net assets at the beginning of the period)/Net assets at the beginning of the period
Solvency indexes	Current ratio	Current assets/current liabilities
	Equity ratio	Total liabilities/total owner's equity
Technical indexes	Logarithmic Stock Market Value	-
	5-day average turnover rate	Weekly turnover/total negotiable shares
	AR	By comparing the highest and lowest opening prices on the same day, the market sentiment is reflected through the position of opening prices in the stock price in a certain period of time
	BR	Based on the closing price of the previous day, the fluctuation of the day's market is shown by figures, which are used to predict the trend of stock movements

3.2 Data Preprocessing

In this paper, Z standardization method is used for data preprocessing. The formulas are as follows:

$$z_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j} \tag{5}$$

Among them, $\overline{X}_j$ is the mean of the j-th feature and S_j is the standard deviation of the j-th sequence.By this method, we can eliminate the dimension of different factors and make them obey the expectation of 0 and the standard deviation of 1.

3.3 Strategic Establishment

This paper establishes multi-factor stock selection strategy based on Improved PSO SVM. The related algorithms have been introduced in the previous paper, so the following will be the specific strategy establishment process.

The training set selected in this paper is the Shanghai and Shenzhen 300 component stocks on May 31, 2018. The label is the monthly return rate in June, 2018. Because the monthly return rate is continuous, this paper makes the positive sample whose return rate is more than 0.15, labeled as 1, and the negative sample whose return rate is less than 0.15, labeled as -1. These data are used to train the model, and then the factor value of July 1 is used to predict whether stocks should be bought in July. After 20 trading days, we use the new factor data to determine which stocks should be bought and which stocks should be sold. Since the length of data we used to train the model was one month, our warehousing frequency was 20 trading days (about one month).

240 stocks were randomly selected from 300 stocks to train the support vector machine model, and then the remaining 60 stocks were used to score the model. This score was based on the correct number of samples classified to the total number of samples, i.e. 60. By training the model, the final score was 92.5%.

Then we started trading on July 1, 2018 and ended trading on February 1, 2018. We use this model to predict stocks and get the stocks we need to buy. If the stocks we have held before are on the list of stocks we buy, we will continue to hold them and sell them if not. For stocks that need to be bought, we sell them at the market value of 90% of the currently available funds, and adjust positions every 20 trading days.

3.4 Policy Implementation

Fig. 1 shows the cumulative rate of return of the strategy from July 2018 to February 2019. It can be seen that the improved SVM strategy is higher than the original SVM strategy, and the original SVM strategy is slightly higher than the performance benchmark (Shen-Hu 300).

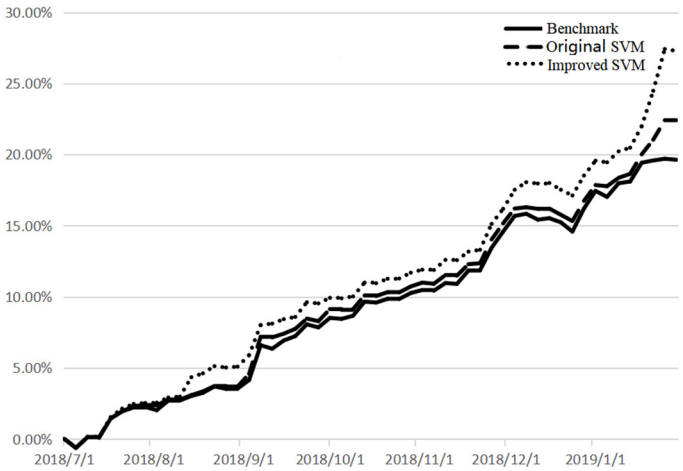

Figure 1. Accumulated Return Rate

The annual rate of return of the improved SVM strategy is 44.6%, and the Alpha of it is 17.2%. Alpha represents the excess return beyond the market, which shows that the excess return of this strategy is still very good. Beta is 0.97, which measures the systemic risk of the portfolio. The beta of this strategy is close to 1, which shows that the systemic risk is equal to the risk of the whole market. Sharp ratio represents the difference between the return rate of the strategy and the risk-free interest rate divided by volatility. The Sharp ratio of this strategy is 3.37, which is a relatively high value, indicating that the implementation effect of the strategy is very good. The information ratio reflects the overall effect of strategy prediction. The information ratio of 2.23 is quite good. The lower the maximum back-test, the better. The 5.8% maximum back-test is relatively low. As shown in the figure, there is little difference between the strategy return curve and the benchmark return curve when the strategy is first implemented, and the difference between the strategy and the benchmark becomes more and more obvious with the passage of time.

4. Summary

The test results show that the quantified investment strategy based on improved particle swarm optimization SVM algorithm is effective. The yield of the quantitative investment strategy used in this paper is stable higher than the performance reference benchmark, and its maximum return is only 5.8%. This shows that the risk of the strategy is relatively small, which shows that the quantitative investment strategy can improve the portfolio return and control the portfolio risk well. Quantitative investment strategy avoids people's emotional interference, which also reflects to a certain extent the reduction of risk caused by avoiding human factors in investment strategy.

References

[1]. Shen Q, Jiang J H, Jiao C X, et al. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists[J]. European Journal of Pharmaceutical Sciences, 2004, 22(2-3):145-152.

[2]. Dong Zhi -qiang,Liu Yong -nian&Wei Li -hua.Fault detection of automatic car equipment based on AdaBoost and SVM combination[J].Electronics World,2018,(2):17-19.

[3]. Fama E. the Behavior of Stock Market prices [J]. Journal of Business,2010(05):35-105.

[4]. Chou R Y, et al. Volatility persistence and stock valuations: Some empirical evidence using GARCH [J]. Journal of Applied Econometrics,2008(09):279-294.

[5]. MassoudMetghalchi, Yung-Ho. Technical Analysis of the Tai-waneseStock Market[J]. International Journal of Economics and Finance,2012,4 (1):90-102.