Advances in Economics, Business and Management Research, volume 166
Proceedings of the 6th International Conference on Financial Innovation and Economic Development (ICFIED 2021)

Establishment and Analysis of Multi-Factor Stock Selection Model Based on Support Vector Machine in CSI 300 Index Constituent Stocks Market

Changsheng Dou*, Tengzhe Zhao, Ziheng Guo

School of Statistics, Capital University of Economics and Business, Beijing, China, 100070
** Corresponding author, Email: douchangsheng@cueb.edu.cn*

ABSTRACT
This paper uses the SVM (support vector machine) method to model the multi-factor stock selection and conducts research in Chinese Stock Market. The CSI 300 Index accounts for about 60% of the market value of Chinese Stock Market, we uses the principal component analysis for dimensionality reduction, reducing the number of original factors to 13, and the cumulative contribution rate reached 78.5372%, which reduced the complexity of SVM classification. In terms of model building, since the linear SVM method cannot be reasonably classified, this paper uses the radial basic kernel function and then classifies them, and obtains a stock selection model with strong effectiveness, which can beat the benchmark in all test sets. In terms of stock selection, we sort the stocks according to the sample values generated by the prediction of the model, and the reliability of the result obtained was high, which is the innovation of this paper.

Keywords: *Quantitative Stock Selection, SVM, Multi-factor Model, Principal Component Analysis*

1. INTRODUCTION

Quantitative investment is the use of statistical knowledge, quantitative methods and computer information technology to select outstanding securities in the market to establish a portfolio, for high yield and low risk. Because of the development of quantitative technology and continuous improvement of securities market, more and more quantitative investment methods and products come into people's vision. Different from traditional investment, quantitative investment does not use investor experience as the main analysis method to manage investment products, but based on mathematical model and quantitative analysis of much relevant data to get optimal portfolio.

Quantitative investment has developed for decades abroad. In 1952, Markowitz published "portfolio selection: effective diversification of investment", which proposed the mean variance model in the field of investment, and obtained the relationship between the total return and total risk of various risky assets. In 1964, Sharpe et al. proposed the capital asset pricing model (CAPM), which established risk-free return to achieve the pricing effect, and this method is widely used in corporate finance and investment decision-making. In 1973, Black et al. established the option pricing model (OPM), which promoted the development of financial derivatives to a certain extent. In 1976, Ross founded the arbitrage pricing theory (APT). Although it is called arbitrage pricing, it has nothing to do with arbitrage. In fact, it is the promotion of CAPM. It thinks that when the market reaches equilibrium, the return of securities is determined by many factors, and there is an approximate linear relationship, so apt has become the basis of modern quantitative investment. On the basis of CAPM and APT, Fama et al. established Fama-French three factor model in 1992. They found that in the U.S. stock market, only the beta coefficient in CAPM can't explain the difference of different stock returns. The performance of the model becomes better when the price/earnings ratio, market value and book/market ratio are taken as model factors, and a five factor model was established in 1993. On the basis of the original factors, they took the profit level risk and investment level risk as new factors. Based on these excellent research results, more and more scholars introduce machine learning algorithm and apply it to the field of financial quantitative investment. In 2015, Ballings et al. [1] studied the changes of stock profits in European stock market. They took a variety of financial indicators as

factors, and used neural network, SVM, random forest and other algorithms to introduce machine learning into stock prediction. In 2016, thakur et al. [3] combined four different feature selection techniques with SVM to build a new model to predict the financial market, making the performance of the new model better than other single models. In 2018, Ahmadi et al. [4] constructed a hybrid model with SVM, with ICA and GA, respectively to optimize parameters of the models, so as to better predict the stock market. Thakur [2] used machine learning algorithm to predict American stocks in 2018. After dimensionality reduction of data with random forest algorithm, it used SVM to build stock selection model. Kompella [5] implemented the random forest method to predict the stock price in 2019, which proved that the efficiency of random forest for stock price prediction is excellent. Nti et al. [6] found in 2020 that when the data has high noise and high dimension, the over fitting property of SVM is ignored. Therefore, they proposed GASVM, using GA to enhance and the purpose is to predict the stock. In the same year, Nguyen et al. [7] used hidden Markov model (HMM) to score the historical performance of global stocks, and finally realized the choice of stocks. In 2021, Avramov et al. [8] proved that machine learning can identify most stocks with wrong pricing, indicating that it is profitable to use machine learning to select and predict stocks in recent years. In a word, the machine learning algorithm, especially the use of SVM for stock selection and prediction, are good results in foreign countries.

In this paper, the method of SVM is applied to the field of quantitative investment, and a reasonable multi-factor stock selection model is built. SVM has advantages in optimizing structural risk, nonlinear relationship and data processing in high dimensional space.

(1) China's quantitative investment technology is not yet mature, and its development time is relatively short. Quantitative investment tools are mainly stocks. The model of a multifactor selection can better solve the high-dimensional problems among factors, provide a new idea and direction for quantitative investment in China, and help to select high-quality stocks with full investment potential.

(2) The gap between China's securities market and that of developed countries is still obvious, which is mainly reflected in the slow trading speed, small trading scale and few trading varieties. The establishment of a multi factor selection model will speed up the processing speed of data indicators of listed companies. The reform of non tradable shares and the increasing number of shares in China will undoubtedly make the stock market more open, complex and changeable, and increase the risk. Therefore, it is of practical significance to establish a quantitative model in time for the development of China's securities market.

(3) China has a large base of investors, but the general investment capacity is insufficient. Most investors still choose fundamental analysis and technical analysis for securities selection. In the face of many stocks, investors will not only spend a lot of energy, but also be affected by irrational psychology, which eventually leads to poor investment effect. The quantitative investment method has good discipline and systematicness, overcomes the challenge of behavioral finance, and creates new possibilities for investors to get higher and more stable returns.

The aim of this paper is to get a portfolio better than the market portfolio yield through the model. In order to ensure the timeliness, the model will use the recent data of Shanghai and Shenzhen 300 index. In order to fully consider the relationship between all aspects of the stock index and the rate of return, a variety of indicators related to profitability, solvency, income quality, investor sentiment and so on are used as the original data. Because there is a certain correlation between many stock related indicators, we use principal component to analysis data, and the extracted principal component is used as the factor data needed by the stock selection model. We use kernel to map on the low dimensional space data to the higher dimensional space, adjust the model parameters, and get the optimal hyper-plane which can classify the factor data. The SVM model is established and used to select the component stocks of CSI 300 index in this paper.

2. BASIC THEORY

2.1. Multi-factorial model

Multi-factor model is one of the mature models in the field of quantitative investment. It considers that the return of various securities is affected by several factors. The model usually selects the growth ability, profitability, momentum reversal, scale, management and other factors of the enterprise, carries on the fitting and statistical analysis of the historical data, selects the factors with strong correlation with the yield of securities as the stock selection standard, and buys the securities that meet the standard. In the multi factor model, to determine the sensitivity of the stock return to the changes of these factors, we can establish the regression equation and get the linear relationship between the stock return and the factors:

$$r_p = r_f + \beta_1 F_1 + \beta_2 F_2 + \cdots + \beta_n F_n + \varepsilon \quad (1)$$

Among this, r_p is the yield of the securities under study, r_f is the risks-free rate, β_k is the sensitivity of the security to the change of factor K, and ε is the deviation caused by non systematic risk.

2.2. SVM

SVM is a useful machine learning algorithm of sample data. Different from traditional classification methods, SVM can find the maximum margin hyper-plane in sample space to achieve high-precision classification and minimum structural risk. When dealing with the linear separable problem, the classification results obtained by SVM have the strongest robustness and generalization. When dealing with the linear non-separable problem, so that the linear non-separable problem can be transformed into the linear separable problem. At present, this method is widely used in face recognition, handwritten character recognition, natural language processing and other fields.

2.2.1. linear SVM

Given training sample set $T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}, y_i\in\{-1,+1\}$,

Suppose there is a hyper-plane which can classify the samples and ensure that the nearest sample is farthest from the hyper-plane. The equation of the plane is expressed as:

$$w^T x+b=0 \tag{2}$$

Among this, $w=(w_1,w_2,w_3,\cdots,w_d)$ is a normal vector, b is displacement. Let a point be in the plane $x^{'}$,satisfy $w^T x^{'}=-b$, At the same time, let any point in the sample space x,and the distance from the point to the hyper-plane is:

$$r=\left|\frac{w^T}{\|w\|}(x-x^{'})\right|=\frac{\left|w^T x+b\right|}{\|w\|} \tag{3}$$

Suppose the equation $y(x_i)=w^T x_i+b$, if $y(x_i)>0$, so $y_i=+1$; when $y(x_i)<0$, then $y_i=-1$, so there is always $y(x_i)\cdot y_i>0$.Make it greater than or equal to 1 by scaling, the optimal hyper-plane problem is:

$$\arg\max_{w,b}\frac{1}{\|w\|}\quad s.t.\ y_i(w^T x_i+b)\geq 1,\ i=1,2,\cdots,n \tag{4}$$

The maximum problem is transformed into the minimum problem, the equation is:

$$\min_{w,b}\frac{1}{2}\|w\|^2\quad s.t.\ y_i(w^T x_i+b)\geq 1,\ i=1,2,\cdots,n \tag{5}$$

Due to the existence of constraints on w and b, after adding Lagrange factor $a_i\geq 0$,The above formula can be rewritten as Lagrange function:

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+\sum_{i=1}^{n}\alpha_i(1-y_i(w^T x_i+b)) \tag{6}$$

Take partial derivatives with w and b respectively, and set its value to 0,to obtain the relation $w=\sum_{i=1}^{n}\alpha_i y_i x_i$ and $0=\sum_{i=1}^{n}\alpha_i y_i$ .Substitute it into Equation (6), and the equation satisfies the convex optimization problem and the KTT condition, so we can write a dual problem:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{7}$$

And the classification model is:

$$f(x)=\sum_{i=1}^{n}\alpha_i y_i(x_i\cdot x)+b \tag{8}$$

2.2.2. Nonlinear SVM

When classifying the linearly indivisible training samples and the optimal partition hyper-plane satisfying the conditions can be found in this space. Let be the eigenvector in the high-dimensional space, then the equation corresponding to dividing the hyper-plane is expressed as:

$$f(x)=w^T\phi(x)+b \tag{9}$$

Similar to the principle of solving linear SVM above, Equation (7) can be rewritten as:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\phi(x_i)^T\phi(x_j)$$
$$s.t.\ \sum_{i=1}^{n}\alpha_i y_i=0,\ \alpha_i\geq 0,\ i=1,2,\cdots,n \tag{10}$$

We solve the question x_i and x_j the inner product of space with the kernel $\kappa(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\phi(x_i)^T\phi(x_j)$

Classification model was obtained by solving Equation

$$f(x)=\sum_{i=1}^{n}\alpha_i y_i\kappa(x,x_i)+b \tag{11}$$

There are several common kernel functions:

（1）Linear kernel function $\kappa(x_i,x_j)=x_i^T x_j$

（2）Polynomial kernel function $\kappa(x_i,x_j)=(x_i^T x_j)^d$

（3）Radial basis kernel function

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

（4）Sigmoid kernel function

$$\kappa(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$$

2.2.3. Soft interval

Soft interval is a tolerance of SVM classification errors, which allows SVM to make errors on some samples, so as to prevent over fitting and reduce the generalization ability of the model. So, the constraints become $y_i(w^T x_i + b) \geq 1 - \xi_i$ after introducing relaxation factor ξ_i. The objective function can be rewritten as:

$$\max_{w,b,\xi_i} \frac{1}{2}\|w\| + C\sum_{i=1}^{n}\xi_i \quad s.t.\ y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,\ i = 1,2,\cdots,n \tag{12}$$

Among this, coefficient C control relaxation, when C is smaller. The error tolerance of the model is higher; when C is bigger. The error tolerance of the model is lower. Introducing Lagrange factor $\alpha_i \geq 0, \mu_i \geq 0$, we can transform Equation (12) to

$$L(w,b,\xi,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y_i(w^T x_i + b)) - \sum_{i=1}^{n}\mu_i\xi_i \tag{13}$$

Though finding the partial derivatives of (13) with respect to w, b, ξ_i, we set the corresponding partial derivatives to 0. By substituting the obtained relation into Equation (13), the original objective function can be transformed into a dual problem:

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \quad s.t.\ \sum_{i=1}^{n}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ i = 1,2,\cdots,n \tag{14}$$

After the kernel function is introduced to process the original training samples, the same classification model as Equation (11) can be obtained.

3. THE DATA PROCESSING

3.1. Data selection and source

In this paper, all the constituent stocks of CSI 300 Index are taken as the research object, and all the indicators of listed companies are taken as the initial data. The quarterly data of 300 stocks from 2013 to 2017 were selected for the training sample data, including financial information such as profitability, income quality, debt paying ability, etc. At the same time, risk indicators and investor sentiment indicators were introduced. Therefore, the model was more comprehensive and effective.

The selected data are all from Wind financial terminal and GuoTai Junan database, and the authenticity and integrity of the data are high. To ensure the modelling effect, samples with more than half of the missing values and yields more than three times higher than the average will be removed. Other missing values will be supplemented with the average value. This process will be implemented by using the Pandas package in Python software.

3.2. Candidate factor selection

The candidate factors in this paper include stock basic indicators, listed companies' profitability, earnings quality, cash flow, capital structure, debt-paying ability, operating ability, growth ability, as well as risk indicators, investor sentiment indicators and macro data indicators. Specific factor names and meanings are shown in the following table:

Table 1. The candidate factor

type	Name of factor	Factor definition
Basic indicators	Earnings per share EPS	The ratio of after-tax profit to total number of shares
	Book value Per Share BPS	The ratio of shareholders' equity to the total number of shares
	Profit before tax per dividend	The ratio of EBIT to total number of shares
Profitability	ROE	The ratio of net income
	Return on Total Assets Ratio ROA	The ratio of net profit to total assets
	ROIC	The ratio of operating profit and tax to invested capital
	Net profit/gross operating income	The ratio of net profit to gross operating income
Earnings quality	Total net income/profit from operating activities	The ratio of income from operations to total profits

The cash flow	Net cash flow from operations/revenue	The ratio of cash flow from operations to revenue
The capital structure	Asset-liability ratio	The ratio of assets to liabilities
	The rights and interests multiplier	The ratio of total assets to total equity
Debt paying ability	liquidity ratio	The ratio of liquidity assets to current liabilities
	Quick ratio	The ratio of assets to liabilities after removal of inventory
	Cash ratio	The ratio of cash to current liabilities
Operation ability	Inventory turnover	The ratio of cost of goods sold to inventory
	turnover of total capital	Ratio of sales to total assets
	Accounts receivable turnover	The ratio of sales to accounts receivable
Growth ability	Year-over-year growth of underlying earnings per share	The ratio of the difference between the current period and the previous period's earnings per share to the previous period's earnings
	growth rate of net profit	The ratio of the difference between the net profit of the current period and the previous period and the net profit of the previous period
	Year-over-year growth of total assets	The ratio of the difference between the current period and the previous total assets to the period's total assets
Value indicators	PE	The ratio of price to earnings per share
	PBR	The Ratio of price to net asset value per share
Indicators of risk	Beta β	The ratio of the covariance between stock returns and market portfolio returns to the variance of market portfolio returns
Sentiment indicators	Psychological indicator PSY	Investors' psychological expectations for the rise and fall of the stock market
Macroeconomic indicators	Gross domestic product GDP	The core index of national economic accounting

3.3. Principal component analysis

3.3.1. The basic principle of principal component analysis

PCA is a commonly used dimensionality reduction method in machine learning. Its purpose is to transform the original variables that may be correlated into linearly uncorrelated variables. In the mathematical model, the introduction of many influence factors as the model independent variables can be more comprehensive to analyze problems systematically, and the correlation between independent variables will make information have repeat, and too independent variables increases the complexity of the model, not only can reduce the amount of model factors, without change of information covered by the original variable.

Let's say there are n samples and m indicators. After standardization, the initial data is expressed by matrix as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (15)$$

According to Equation (15), the covariance matrix is:

$$C = \frac{1}{m} XX^T = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} x_{1i}^2 & \frac{1}{m}\sum_{i=1}^{m} x_{1i}x_{2i} & \cdots & \frac{1}{m}\sum_{i=1}^{m} x_{1i}x_{ni} \\ \frac{1}{m}\sum_{i=1}^{m} x_{2i}x_{1i} & \frac{1}{m}\sum_{i=1}^{m} x_{2i}^2 & \cdots & \frac{1}{m}\sum_{i=1}^{m} x_{2i}x_{ni} \\ \vdots & \vdots & & \vdots \\ \frac{1}{m}\sum_{i=1}^{m} x_{ni}x_{1i} & \frac{1}{m}\sum_{i=1}^{m} x_{ni}x_{2i} & \cdots & \frac{1}{m}\sum_{i=1}^{m} x_{ni}^2 \end{bmatrix} \quad (16)$$

Equation (17) is a real symmetric matrix. Hence, let n unit eigenvectors exist, and the constituent matrix can be expressed as $\mathrm{E} = (e_1 \quad e_2 \quad \ldots \quad e_n)$,

We can transform C to a diagonal matrix:

$$\Lambda = E^T CE = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_n \end{bmatrix} \quad (17)$$

Among , $\lambda_1, \lambda_2, ..., \lambda_n$ is the characteristic value. Arrange them in order from large to small, and order their corresponding eigenvectors according to this, and multiply them with the original data matrix to obtain the dimensionality reduction matrix Y.

For reflecting the extraction degree of principal component of the data information, concept of variance contribution rates of principal component will be quoted, and its calculation method is:

$$Z_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{18}$$

Among this, Z_i Represents the proportion of the variance corresponding to the i principal component in the total variance. When Z_i much greater, it shows that the larger the information ratio extracted by the principal component is, the better the extraction effect is.

3.3.2 The experimental process of principal component analysis

We use R Studio software achieve complete the PCA.

（1）Factor data of all samples were read and standardized;

（2）The processed data in (1) is expanded into a covariance matrix;

（3）Use Princ*om* function to find out the principal component and related contribution rate, cumulative contribution rate;

（4）The original data was converted to the data after dimensionality reduction, and the first K principal components were selected.

According to the above steps, the 25 candidate factors selected in this paper are reduced in dimension. Table 2 shows the degree of correlation among some factors (four decimal digits are reserved for all data). It is easy to see that there is obvious correlation among some factors, so it is necessary to carry out principal component analysis on the selected factors.

Table 2. Correlation degree among factors

Correlation degree	EPS	BPS	Profit before tax per dividend	ROE	ROIC	BETA
EPS	1.0000	0.5578	0.5755	0.3300	0.1837	0.0157
BPS	0.5578	1.0000	0.5933	0.1123	0.0459	0.0326
Profit before tax per dividend	0.5755	0.5933	1.0000	0.2041	0.0747	0.0128
ROE	0.3300	0.1123	0.2041	1.0000	0.2703	0.0058
ROIC	0.1837	0.0459	0.0747	0.2703	1.0000	-0.0017
BETA	0.0157	0.0326	0.0128	0.0058	-0.0017	1.0000

From Table 3, Princ*om* function is used to extract the principal components. We select 13 components, because they can get to 78.5372%.

Table 3. Contribution rate

The principal components	contribution	Sum of contribution rate
Comp. 1	11.9459%	11.9459%
Comp. 2	9.8749%	21.8209%
Comp. 3	7.2991%	29.1120%
Comp. 4	6.8739%	35.9939%
Comp. 5	6.3190%	42.3129%
Comp. 6	5.5908%	47.9037%
Comp. 7	5.3417%	53.2453%
Comp. 8	5.1669%	58.4122%
Comp. 9	4.4667%	62.8789%
Comp.10	4.0335%	66.9124%
Comp.11	4.0090%	70.9214%
Comp.12	3.8940%	74.8155%
Comp.13	3.7217%	78.5372%

Therefore, 25 factors are selected, which is related to stock return rate. After PCA dimension reduction, 13 principal components are extracted, and the cumulative contribution rate reached 78.5372%. While ensuring that a large amount of original factor data information could

be retained, the number of factors are reduced, which makes a good preparation in the next section.

4. THE ESTABLISHMENT OF SVM STOCK SELECTION MODEL

4.1. Establish the process of stock selection model

In this chapter, the method of SVM will be used to establish a stock selection model, and the relationship between the 13 principal components extracted in section 3 and the stock return rate will be established. The model takes the sample data of 16 quarters from 2013 to 2016 as the training set, and tests it in the first two quarters of 2017 to test the return effect of the stock portfolio selected by the model. All stocks are held quarterly.

We need to assign y value to each sample, that is, affix the classification label of the sample. Before the establishment of the model, the proportion of 5:5 was used for classification, and the 50% samples with higher return rate were labeled with "1", while the 50% samples with lower return rate were labeled with "-1". The data of some training samples were shown in the following table.

Table 4. Data of some training samples

Stock name	Time	The label	Comp.1	Comp.2	Comp.3	...
China petroleum& chemical corporation	In the first quarter of 2013	1	-1.1826	0.3321	-0.0429	...
China Life Insurance	In the first quarter of 2013	1	-0.5666	1.0035	0.3477	...
PORT OF NINGBO	In the first quarter of 2013	1	-0.2150	-0.3011	-0.3951	...
Sichuan for energy	In the first quarter of 2013	1	-0.4207	0.9370	-0.5867	...
Century huatong	In the first quarter of 2013	1	0.6433	0.6255	-0.2506	...
Shandong Gold	In the first quarter of 2013	-1	-1.0387	-1.8648	0.4695	...
Fuyao Group	In the first quarter of 2013	-1	0.0704	0.4974	-0.1127	...
Yutong	In the first quarter of 2013	-1	0.1626	0.7718	0.3731	...
Huaxia happiness	In the first quarter of 2013	-1	-0.8743	1.7714	0.2546	...
Nanshan aluminum	In the first quarter of 2013	-1	0.5305	-0.5338	0.0061	...
...	...	...	...	...	...	...

Although after principal component analysis to reduce the original 25 factor to 13 factors, but in SVM classification model, independent variable quantity still more, sample data is complex, in low dimensional space cannot achieve good classification effect for binary classification, while radial basis kernel function has better effect in dealing with complex nonlinear problems, faster running speed and higher classification accuracy, so the radial basic kernel will be used.

In addition, there are two parameters C and γ that need to be selected in the model. C is the penalty which indicates the degree of relaxation of the model. When C is small, the tolerance of the model to

misclassification is high. When C is large, the tolerance of the model to misclassification. Different values of γ will change the complexity of sample data subspace and affect the classification error to a certain extent. A pair of appropriate $(C,\ \gamma)$ can not only avoid the occurrence of over-fitting, but also further improve the accuracy of classification. We choose $C=2,\gamma=0.5$, the model training and testing will be implemented by using the E1071 package in R Studio.

4.2. Stock selection and inspection

After the establishment of the stock selection model according to the above steps, use the predict function in R Studio to predict the test set, and each sample will get a value, namely $y(Comp.1,Comp.2,...,Comp.13)$. When the value is greater than 0, the corresponding sample is classified as "1". When the value is less than 0, the corresponding sample is classified as "-1". The larger the absolute value, the more reliable the classification. In this paper, we hope to select stocks with high yield and relative stability to enter the stock pool, so we rank the generated values from the largest to the smallest. The stocks corresponding to the higher ranking values are not only classified into the "1" category by the model, but also have the highest reliability. We select the top 30 stocks to build the portfolio. In the first and second quarters of 2017, the specific stock selection situation is shown in the table below.

Table 5. Stock selection

First quarter of 2017		The second quarter of 2017	
Bank of Nanjing	Hengtong Optoelectronics	Songcheng Performance Development	ChinaNetCenter
China galaxy	Supor	Zhifei biology	Tongrentang
Europee home	Stone based information	Huiding Technology	COSL
Meinian health	Shaanxi coal industry	Dongfang Yuhong	Minmetals Resources
Sichuan investment energy	Minmetals Resources	White Cloud Mountain	Watson biology
AVIC Electronics	Guanglianda	Yan'an Bikang	Bank of Beijing
Hengtong Optoelectronics	Shanghai Airport	Shaanxi coal industry	The Oriental Pearl
Guodian power	Aier ophthalmology	Haitong Securities	Hengyi petrochemical
Xinwei communication	Hang Seng Electronics	Chenguang stationery	Midea Group
Yanghe Co., Ltd	Meijin energy	Youngor	Notoginseng and entertainment
Fosun Pharma	Shanghai Laishi	Guiyang bank	Guodian power
Hengrui medicine	Bohai leasing	Lujiazui	Yunnan baiyao
Zhaoyi innovation	Anson trust	Stone based information	Aier ophthalmology
Huadian International	Ziguang Co., Ltd	Europee home	Dragon Python Baili
Shanghai Lingang	ChuanHua Zhilian	Luzhou Laojiao	Aerospace Information

According to the above Table 5, the equal weight of the stock is selected, the equal weight return of the portfolio is calculated, and the return of CSI 300 index in the same period is used as the benchmark to test the effectiveness of the model. In the first quarter of 2017, the result was 4.4097%, the sum was 12.7550%, and the excess return was 8.3453%; in the second quarter of 2017, the result was 6.0981%, the sum was 7.3878%, and the excess return was 1.2897%. In both quarters, the portfolio outperformed the CSI 300 index, and performed well in the first quarter of 2017. In addition, the comprehensive accuracy rate of the top 30 "1" stocks selected by the model in the two quarters is 61.6667%.

Table 6. Income statement

Rate of return	First quarter of 2017	Second quarter of 2017
The CSI 300 Index	4.4097%	6.0981%
Portfolio	12.7550%	7.3878%
The excess return on the portfolio	8.3453%	1.2897%

From Table 6, after determining the use of radial basis kernel function and selecting appropriate parameters, the SVM stock selection model is established, and good results are obtained on the test set. In the first quarter of 2017, the excess return was 8.3453%, and in the second quarter of 2017, the excess return was 1.2897%, both beating the market.

5. CONCLUSION

(1) After using PCA for dimension reduction, the model of SVM is established and tested in the empirical analysis.

1. With respect to factor selection, this paper selects the multidimensional indexes, which not only include profitability, earnings quality, capital structure, solvency, growth ability and other financial data index, PSY investor sentiment index and GDP, at the same time takes into account the individual, the macro economy and the impact of investors intend to return, made more comprehensive factor data. At the same time, we use PCA for dimensionality reduction, reducing the number of original factors to 13, and the cumulative contribution rate reached 78.5372%, which reduced the complexity of SVM classification.

2. In terms of model building, since the linear SVM method cannot be reasonably classified, this paper uses the radial basic kernel function and then classifies them, and obtains a stock selection model with strong effectiveness, which can beat the benchmark in all test sets.

3. In terms of stock selection, this paper sorted the stocks according to the sample values generated by the prediction of the model, and the reliability of the result obtained was high. Compared with the researches of other relevant scholars, this point is the innovation of this paper.

(2) Due to the limited ability and time of the author, there are still many deficiencies in the SVM stock selection model established in this paper, which are mainly reflected in the following aspects:

1. Stock selection model in this paper cannot achieve good stock selection effect when the market is abnormal. China's stock market experienced a severe bear market from the second half of 2017 to 2018, and the portfolio obtained by using this model cannot achieve good return effect. The accuracy of model classification also needs to be further strengthened.

2. The time lag isn’t taken into account in the paper. The release of the company's financial data has a certain lag period, which will affect investors' stock selection.

3. The effect of the stock selection model in other markets and other investment periods needs to be further studied. The stocks in this paper only consider the constituent stocks of the CSI 300 Index, and do not test on the A stock market with more stocks, or on the daily and monthly position strategies.

ACKNOWLEDGMENTS

Dou was supported by the Special Fund for Fundamental Scientific Research of the Beijing Colleges in CUEB, Top young talents of Beijing Gaochuang project, CUEB's Fund Project for reserved discipline leader.

REFERENCES

[1] M. Ballings, D. V. Den Poel, Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications, 42(20)(2015), 7046-7056.

[2] M, Thakur, D. Kumar, A hybrid financial trading support system using multi-category classifiers and random forest, Applied Soft Computing, 67 (2018), 337-349.

[3] D. Kumar, S.S. Meghwani, M. Thakur, Proximal SVM based hybrid prediction models for trend forecasting in financial markets, Journal of Computational Science, 17(2016), 1-13.

[4] E. Ahmadi, M. Jasemi, L. Monplaisir, M.A. Nabavi, A. Mahmoodi, P.A. Jam, New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the SVM and Heuristic Algorithms of Imperialist Competition and Genetic, Expert Systems with Applications, 94 (2018), 21-31.

[5] S. Kompella, K. Chakravarthy Chilukuri, Stock Market Prediction Using Machine Learning Methods. International Journal of Computer Engineering and Technology, 10(3) (2019), 20-30.

[6] I. K. Nti, A. F. Adekoya, B. A. Weyori, Efficient Stock-Market Prediction Using Ensemble SVM, Open Computer Science, 10(1) (2020), 153-163.

[7] N. Nguyen, D. Nguyen, Global Stock Selection with Hidden Markov Model, Risks, 9 (2021), 9, 1-18.

[8] D. Avramov, S. Cheng, L. Metzker, Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability, (2021).