A new stock selection model based on Multi-class Support Vector Machine

QiangshengZhang, JingruZhang, ZishengChen, MiaoZhang, Songying Li

Guangdong University of Foreign Studies, School of Finance,
Panyu District, Guangzhou, Guangdong, China

***Abstract:** This paper presents a new stock selection model based on multi-class support vector machine by employing kernel principal component analysis to avoid the risk of speculation and gain the excess return. First, an initial stock pool is choosed according to the industry rotation theory. Then the stock selection index system is constructed based on factor analysis of stock financial indicators and market indicators. Finally, the empirical experiment shows that the proposed stock selection model greatly improves operational efficiency and prediction accuracy for incomplete China's stock market.*

Keywords: Stock selection model, Kernel Principal Component Analysis , Multi-class Support Vector Machine

1. INTRODUCTION

With the economical reform and opening the society has witnessed the rapid development of China's economy and embraced various financial products. However, although people are enthusiastic with the investment portfolios of financial securities, return rate of the majority of the investors is not satisfied due to the impulse and irrational investment. To solve this problem, China has introduced quantitative investment to help both the institutional investors and private individuals to make the investment decision, taking advantage of mathematics and computer science. There are several widespread models such as the neural networks, clustering analysis and so on.

Traditional machine learning algorithms generally base on a hypothesis that the number of training samples tend to be infinite. However, as a matter of fact, the majority of data is small-sample which is high-dimensional and linear inseparable. The invention of support vector machine(SVM) effectively solves the difficulty of pattern recognition and data classification in high dimensional space[1]. Because the parameters of SVM have crucial influence on the calculation speed and classification ability, how to confirm the parameters to guarantee the computational efficiency and classification correctness at the same time becomes a hot issue in the research of SVM. Zhang and Dai[2] proposed that adaptive weighted least squares support vector machine(AWLS-SVM) can improve the weight distribution of exponential distribution and obtain the weights of adaptive sample to eliminate the influence of abnormal data. Phurivit Sangkatsaneea[3] used the two-class SVM to combine features to further improve the classification accuracy. Mohammad[4] proposed a multi-class SVM based on Bayesian theory which had a strong generalization capability, so that the uncertainty problem can be figured out more efficiently. Luo and Cheng[5] used Chebyshev's theorem to set the threshold of the sum of squared errors and compared the sum of squared errors with the threshold to detect the abnormal value.

In terms of stock forecasting, SVM was first applied to stock classification and forecasting in 2001. The results showed that the five-year yield rate of stocks selected from the Australian stock market is much higher than the yield rate of market index[6]. Chien[7] used the stock data of Taiwan Stock Market from 1996 to 2011 as research sample, and established an optimized SVM to predict stock returns. Meryem Ouahilal[8] proved that SVM performed better than back propagation network(BPN) and case-base reasoning(CBR). In addition, when SVM was combined with independent component analysis(ICA) algorithm, this hybrid algorithm improved the accuracy of experimental results. Xue and Yain[9] proposed a new method to locate chart form in financial time series in which the subsequence searching algorithm was viewed as a trainer to train classification of chart form, which achieved significant improvement in terms of speed and accuracy. In China, Jiang[10] and Zhang[11] separately built SVM to predict market trend and return of individual stock. Li[12] established a stock investment value analysis model by virtue of SVM based on fuzzy clustering. Experiments showed that SVM performed much better in classification and prediction than traditional machine learning. Zhang[13] and Lin[14] discussed the application of principal component analysis(PCA) in SVM stock selection model and realized the selection of high quality stocks. Chen[15] employed the heuristic algorithm to reduce the dimension of the data, and then used SVM to select stocks. Furthermore, Chen compared this model with the SVM based on principal component analysis, the results showed that the former performed better in both the classification accuracy rate and the stock selection. Qin[16] believed that the combination of the region labeling method and SVM technology can improve single point labeling method, thus solved the disadvantages of category imbalance and made trading decisions more reasonable. Zhang and Sheng[17] proposed that both SVM and the least squares SVM had obvious effect on the accuracy of price prediction. The difference was that although the operation speed of the latter is faster, the former is more stable. Li and Tan[18] combined wavelet theory with SVM to overcome the curse of dimensionality and over-learning problem to improve anti-noise property. In this paper we will construct a new stock section model to mine the stock financial data and company market data to help investors reduce the risk and gain return. The machine learning and dimensionality reduction algorithm are combined in order to avoid the influence of investors' emotion. The Multi-class Support Vector Machine model is built and empirical tests show that the model has strong learning ability and high prediction accuracy.

2. SELECTION OF INITIAL STOCK POOL

According to the industry rotation cycle theory, the intrinsic value of a company determines the interest of the portfolio. Therefore, it is important to select well-developed industries as the initial stock pool. The pharmaceutical industry has a loose relation with the macro economy. So China's ongoing economic transition plays an insignificant role in the performance of pharmaceutical industry. China's government has provided massive policy support and subsidy to electronics and information technology industries. As burgeoning industries, the prospects are promising and the development deserves attention from the public. Meanwhile, the machinery industry still plays an important strategic role as the mainstay of the national industry and the demand is growing steadily, no matter domestic or abroad. Based on the above analysis, this paper selects 1548 stocks from the pharmaceutical industry, electronics industry, machinery industry, and information technology industry in the A-share market as the initial stock pool.

Screen the initial stock pool to prepare for the following analysis. Because the indexes and yield rate are special compared with other stocks, so these stocks are supposed to be excluded; then exclude the suspended stocks and stocks cannot be traded in market day. There are four types of financial statements which are respectively first quarterly report, semiannual report, third quarterly report and annual report. Therefore, one quarter is regard as a unit.

3. CONSTRUCTION Of STOCK INDEX SYSTEM

Company's future development trend, continuous operation capability, operational efficiency and revenue, along with the market impact on the company are supposed to be measured comprehensively to build an effective index system. Hence, this paper selects stock indexes from four aspects which are growth indexes, cash flow indexes, technical indexes, and operational indexes. The initial stock indexes system includes total assets growth rate, earnings per share growth rate, net margin growth rate, revenue growth rate, cash flow ratio, net cash flow growth rate from operating activities, shareholder change rate, rate of return on total assets, and total assets turnover rate.

First and foremost, correlation between the stock index and the return of the next quarter is tested. In the first step, stock index should be sorted and divided into five groups. Then calculate the average excess return rate of each group. Secondly, examine the correlation between the stock indexes and the average excess return rate of each group. The correlation test results should higher than 0.75 to pass the test, so net cash flow growth rate from operating activities and the rate of return on total assets are eliminated, all the other stock indexes are tested. Then, after observing the frequency that the yield rate of the fifth group is higher than that of the first group, the average excess return rate of the low-ranking group is subtracted from that factor to form a difference sequence in order to test the validity of the factor. Except for the total assets growth rate, all the other six factors pass the significance test. Additionally, we calculate Spearman rank correlation coefficient between net margin growth rate and earnings per share growth rate is 0.88, which indicates that these two factors are mutually substituted indexes and only one needs be remained. So, we construct the stock index system by five indexes which are total assets turnover rate, revenue growth rate, cash flow ratio, earnings per share growth rate, shareholder change rate.

4. ESTABLISHMENT OF THE STOCK SELECTION MODEL BASED ON SVM

The stock data from the third and fourth quarter of 2016 and the first, second and third quarters of 2017 are selected to train and test the model. For sake of simplifying the complexity of computing process of machine learning, it is necessary to reduce the dimension of the data. If the original data processed by dimension reduction still remain the main information, and the relationship between data remains unchanged or similar, then it is safe to use principal component method to reduce dimension with little side effect. The dimension reduction leads to a reduction of computation as well as efficiency improvement. Therefore, this paper firstly reduces the dimension of the stock indexes system, and then establishes multi-class SVM stock selection models.

4.1 Kernel Principal Component Analysis

The key of linear principal component analysis(PCA) is to remove noise and reduce dimension based on singular value decomposition(SVD). However, this algorithm implies a hidden condition that data only has the first-order and the second-order structure[19], so it is incompetent to deal with the non-stationary high dimensional data with large quantity of noise and large information overlap ratio. Therefore, it is obviously unreasonable to apply PCA to process stock data, instead, this paper uses the kernel PCA method to perform nonlinear dimension reduction.

The kernel function can exchange nonlinear problems in the input low dimensional space to a linear problem in higher dimensional space by defining a mapping from the low dimension space to the higher one. Therefore, the linear inseparable problem in the low dimensional space is transformed into the linear separable problem in high dimensional space.

Let $X^k \in R^N (k = 1,2,\cdots,l)$ be the original sample which can be mapped by the kernel function $H : R^N \to F, x \to X$ to higher dimensional space $F : H(x_1), H(x_2), \cdots, H(x_l)$ which meets the following prerequisite:

$$\sum_{i=1}^{l} H(x_i) = 0$$

(1)

Then the linear inseparable problem can be transformed into linear separable problem regarding using PCA to reduce the dimension of matrix $\vec{K}=\frac{1}{l}\sum_{j=1}^{l}H(X_j)H(X_j)^T$ in the specific space F [20]. The results of the analysis will be variable as the change of kernel functions. The commonly used kernel functions are as follows:

(1)Linear kernel function: $k(x,x_i)=x\cdot x_i$;

(2)Polynomial kernel function: $K(x,x_i)=[\gamma\cdot(x\cdot x_i)+coef]^d$, where d is the order of the polynomial, and $coef$ is the offset coefficient;

(3)Radial basis function (RBF) kernel function: $K(x,x_i)=\exp\left(-\gamma\cdot\|x-x_i\|^2\right)$, where γ is the width of the kernel function;

(4) Multi-Layer Perception (MLP) kernel function: $K(x,x_i)=\tanh[v(x\cdot x_i)+c]$

Among the four kernel functions, the most widely used is the RBF kernel function. The RBF kernel function has a relatively wide convergence domain which can improve the precision of the calculation. It possesses not only strong generalization ability, but also high computational efficiency. Therefore the RBF kernel function is commonly used in classification field.

4.2 Support Vector Machine Model

The traditional SVM is a binary classifier based on Vapnik-Chervonenkis (VC) Dimension statistics theory which aims to minimize structural risk. Sequential quadratic programming method is used, and the optimal hyperplane divides data into two categories where the distance between categories is maximized and the classification accuracy is ensured. SVM has unique advantages on solving small sample, nonlinear and high-dimensional pattern recognition problems.

Assuming a linear programming sample set is $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}\in(X,Y)$, and $x_i\in X=R$, $y_i\in Y=\{-1,1\}$, hyperplane[21] $H:\omega\cdot x+b=0$ can not only correctly classify the samples, but also guarantee that the distance between samples closet to the hyperplane in different classifications is the farthest, which is as following:

$$\min_{\omega,b}\frac{1}{2}\omega^T\omega$$
$$s.t.y_i[(\omega\cdot x_i+b)]\geq 1,i=1,2,\cdots,n$$

(2)

Lagrangian transformation is performed on the above convex quadratic programming problem:

$$L(\omega,b,\alpha)=\frac{1}{2}\omega^T\omega+\sum_{i=1}^{n}\alpha_i-\sum_{i=1}^{n}\alpha_i y_i\left(\omega^T x_i+b\right)$$

(3)

Solve the above formula can obtain decision function in the following form:

$$f(x)=\operatorname{sgn}\left[\sum_{i=1}^{n}\alpha_i y_i(x_i\cdot x)+b\right]$$

(4)

Most of the problems in reality are linear inseparable problems. A kernel function can map the samples in the input low dimensional space into a higher dimensional space. In other words, it means that use $K(x,x')=(\phi(x)\cdot\phi(x'))$ to replace the dot products $x_i\cdot x_j$ of the low dimensional space[22], so that samples become linear separable in the space. The problem can be transferred into the following quadratic programming problem:

$$\min_{\omega,b,\xi_i}\frac{1}{2}\omega^T\omega+C\sum_{i=1}^{t}\xi_i$$
$$s.t.\begin{cases}y_i\left[\omega^T\phi(x_i)+b\right]\geq 1-\xi_i\\ \xi_i\geq 0,i=1,2,\cdots,n\end{cases}$$

(5)

C is a penalty function. The larger the C value is, the severer the penalty for classifying errors is. The dual form of the above formula is:

$$\min_{\alpha}\frac{1}{2}\alpha^TQ\alpha+\frac{1}{4C}-e^T\alpha$$
$$s.t.\begin{cases}y^T\alpha=0\\ 0\leq\alpha_i\leq C,i=1,2\cdots,n\end{cases}$$

(6)

The decision function is obtained to classify the samples as following:

$$f(x)=\operatorname{sgn}\left[\sum_{i=1}^{n}\alpha_i y_i K(x_i\cdot x)+b\right]$$

(7)

Because built-in SVM model in MATLAB can only perform binary pattern recognition, which means that the samples can only be identified as "0" or "1". Unfortunately, this model is not suitable for the analysis of this paper. Therefore, this paper uses LibSVM, software which can solve the multi-pattern recognition problem quickly and efficiently.

5. EMPIRICAL TEST OF STOCK SELECTION MODEL BASED ON SVM

The experimental data of stock selection model is divided into training samples and test samples. The training samples are the third, fourth quarter of 2016 and the first and second quarters of 2017, and the test sample is the third quarter of 2017.

Utilizing the kernel PCA to reduce the dimension of the stock index system we first select the four principal components whose contribution of variance is 20.376%, 20.027%, 20.015%, 19.995% respectively. Then we set up the classification grade based on stock's return rate where the stocks are divided into sixteen categories. Since the performance of the SVM is not affected by the missing data, even some stocks are in lack of the kernel principal component, SVM can still classify them. So the number of training samples is 935. Stocks are classified as follows:

Table 1: Stock Classification Grades

1st	2nd	3rd	4th
<-40%	-40~-30%	-30~-25%	-25~-20%
5th	**6th**	**7th**	**8th**
-20~-15%	-15~-10%	-10~-5%	-5%~0
9th	**10th**	**11th**	**12th**
0~5%	5~10%	10~15%	15~20%
13th	**14th**	**15th**	**16th**
20~25%	25~30%	30~40%	>40%

In this paper, the libsvmtrain() function in the LibSVM package is used to implement stock classification. The function libsvm_options() is used to control several parameters such as kernel function type, kernel function parameter, weights, penalty function and so on. The selection of these parameters determines the performance and prediction accuracy of the SVM. After several empirical tests, the best generalization ability of the classifier is obtained by taking the cross-check factor as thirteen. Since this is a complex linear inseparable stock classification problem with high dimensions, to guarantee the classification precision and operation speed, RBF kernel function is the first choice.

After kernel function is determined, kernel function parameter σ, penalty parameter C and weight parameter ω should be optimized to balance the complexity of machine learning and generalization ability. Among them, σ significantly affects the sensitivity of SVM to samples. If σ is too large, the operation complexity of classifier will increase and the speed will decrease. If σ is too small, the classifier will overreact to minor changes in the input sample, leading to failure of classification. C adjusts the balance between the generalization ability of classifier and the classification accuracy rate. If C is too large, for one hand it leads to the improvement of classification accuracy, for other hand it reduces the probability to apply the model in other problems. ω is the weight of penalty factors. Since the traditional SVM pursues maximization of the overall classification accuracy, the classification accuracy of small sample has been ignored. Therefore, it is bound to present the phenomenon that the classification accuracy of the large sample is high while the accuracy rate of small sample is low. In order to solve this problem, different weight of the penalty factors are assigned to different classes, that is, the penalty factors assigned to small sample is larger than that of the large sample.

In this paper, the minimizing of cross-check error is taken as the target, and grid search is adopted. The initial parameters and search step size are set in the following table:

Table 2: Initial parameters and search step size

Parameter	Initial Setting	Search Step Length
Kernel Function Parameter σ	(0.005,0.5)	0.005
Penalty Parameter C	(1,10)	0.5
Weight Parameter ω	(0.002,0.2)	0.001

The optimal parameters of the kernel function obtained through search are as follows: the kernel function parameter σ is 0.01, the penalty parameter C is 5, and the weight parameter ω is assigned as follow: (0.134, 0.173, 0.067, 0.083, 0.025, 0.010, 0.004, 0.014, 0.015, 0.018, 0.031, 0.057, 0.035, 0.149, 0.172).

The actual classification and predicted classification results are shown as follows: taking the first class as an example, in the actual classification, there are 13 first class stocks, while there are 9 first class stocks in the forecast classification. A total number of 5 stocks are included both in the actual classification as well as in the predicted classification, so the classification accuracy is 38.46%. A total number of 8 stocks are actually in the first class, but are wrongly classified into the other classes in the predicted classification. So the classification error rate was 61.54%. The overall prediction accuracy is 83.64%. It can be observed intuitively from the chart that the correct rate of classification shows a trend of increasing firstly and then decreasing. The classification accuracy in 6th class to 12th class remains higher than 80%, indicating that SVM has higher accuracy in the prediction of the stocks whose rate of return are from -15% to 20%, and the accuracy is significantly reduced when return rate of stocks is too high or too low.

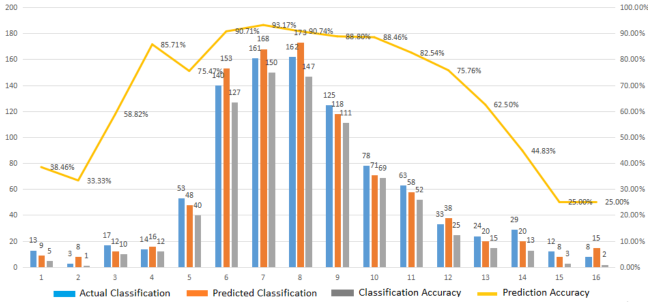

Figure 1 The accuracy rate of prediction and the performance of SVM

Compared with the traditional machine learning methods such as neural network which supposes that sample size tend to be infinite, the SVM performs prominently in small samples, so this paper designs a comparative trial. The empirical test is carried out in pharmaceutical industry, electronics industry, machinery industry, and information technology industry respectively. In addition, all stocks whose predicted rate of return is higher than 0, in other words, whose predicted classification is above or equal to 9th class, are regarded as "buying stocks". Compare the average return rate of “buying stocks” with that of certain industry to investigate the performance of SVM in the specific industry.

Table 3: Earnings Per Share (EPS) of All Stocks and Selected Stocks

Industry	**Pharmaceutical Industry**	**Electronics Industry**
Classification Accuracy	83.37%	80.21%
Total Number of Stocks	126	40
The Number of Stocks (Rate of Return>0)	49	13
The number of Selected Stocks	56	11
Return rate of All Stocks	-0.0099	-0.0180
Return Rate of Selected Stocks	0.0569	0.0601
Industry	**Machinery Industry**	**Information Technology Industry**
Classification Accuracy	82.59%	80.82%
Total Number of Stocks	477	292
The Number of Stocks (Rate of Return>0)	146	100
The number of Selected Stocks	133	109
Return rate of All Stocks	-0.0227	-0.0184
Return Rate of Selected Stocks	0.0623	0.0591

The results of empirical tests show that the classification accuracy rate of the four industries can reach 80%, and a high excess rate of return can be obtained. The results of the comparative trial shows that the multi-class SVM stock selection model has strong generalization ability and is suitable to be applied in individual industries. Therefore, the SVM model has outstanding and stable performance in stock selection.

6. CONCLUSION

This paper employs kernel principal component analysis to extract principal components of financial indexes as input variables in the multi-class SVM model to establish the stock selection model which can classify the stocks according to predicted quarterly return rate. The industry rotation theory is used to select initial stock pool. The selection of stocks is transformed into a common prediction and classification problems. The results show that multi-class SVM is a effective model to circumvent risk and increase return for investors.

Acknowledgement

This paper is supported by the Natural Science Foundation of Guangdong Province, China under Grant 2017A030313435.

REFERENCE

[1] Cherkassky, V. "The Nature Of Statistical Learning Theory." Technometrics 38.4(2002):409-409.

[2] Zhao, Chao, et al. "Soft sensor modeling for penicillin fermentation process based on adaptive weighted least squares support vector machine." Journal of Nanjing University of Science & Technology (2017).

[3] Sangkatsanee, Phurivit, N. Wattanapongsakorn, and C. Charnsripinyo. "Practical real-time intrusion detection using machine learning approaches." Computer Communications 34.18(2011):2227-2235.

[4] Sadooghi, Mohammad Saleh, and S. E. Khadem. "Improving one class support vector machine novelty detection scheme using nonlinear features." Pattern Recognition 83(2018).

[5] Luo, Songrong, and J.Cheng. "VPMCD based novelty detection method on and its application to fault identification for local characteristic-scale decomposition." Cluster Computing 20.25(2017):1-11.

[6] Fan, A., and M. Palaniswami. "Stock selection using support vector machines." International Joint Conference on Neural Networks, 2001. Proceedings. IJCNN IEEE, 2001:1793-1798 vol.3.

[7] Huang, Chien Feng. "A hybrid stock selection model using genetic algorithms and support vector regression." Applied Soft Computing Journal 12.2(2012):807-818.

[8] Ouahilal, Meryem, et al. "A novel hybrid model based on Hodrick–Prescott filter and support vector regression algorithm for optimizing stock market price prediction." Journal of Big Data 4.1(2017):31.

[9] Gong, Xueyuan, et al. "Financial time series pattern matching with extended UCR Suite and Support Vector Machine." Expert Systems with Applications 55.C(2016):284-296.

[10] Jiang Hanqiao. " Research on Securities Investment Decision Making Based on Support Vector Machine[D]." Wuhan University,2005

[11] Zhang Chenxi, Zhang Yanping, Zhang Yingchun et al. " Stock Prediction Based On Support Vector Machine[J]." Computer Technology and Development 16.6(2006):35-37.

[12] Li Yunfei, Hui Xiaofeng. " Research on Stock Investment Value Classification Model Based on Support Vector Machine [J]." Chinese Soft Science 1(2008):135-140.

[13] Zhang Yuchuan, Zhang Zuoquan, Huang Zhen. " Application of Support Vector Machines in Selecting High-Quality Stocks [J]." Statistics and Decision-Making 4(2008):163-165.

[14] Quan Lin, Jiang Xiuzhen, Zhao Junhe, etc. " Research on Stock Selection Model Based on SVM Classification Algorithm [J]." Journal of Shanghai JiaoTong University 9(2009):1412-1416.

[15] Chen Rongda, Yu Huanhuan. “Support Vector Machine Stock Selection Model Based on Heuristic Algorithm [J].” System Engineering 2(2014).

[16] Qin Lu, Li Xuwei. "Study on Cost Sensitive Support Vector Machines Based on Regional Marker Method in Stock Prediction." Journal of Sichuan University2(2018).

[17] Zhang Jiankuan, Yan Shengping. "Prediction of the Stock Price Fluctuation Based on Support Vector Machine." Journal of Beijing University of Information Technology32.3(2017):41-44.

[18] Li kun, Tan Mengyu. "Stock Prediction Based on Wavelet Support Vector Machine Regression." Statistics and decision(2014):32-36.

[19] Chen Bin, Lu Congde, Liu Guangding. "Time Domain Airborne Electromagnetic De-Noising Method Based on Kernel Principal Component Analysis[J]." Journal of Geophysics 57.1(2014):295-302.

[20] Dong Hao et al. "Short - Term Power Load Forecasting Based on Kernel Principal Component Analysis and Limit Learning Machine[J]." Journal of Electronic Measurement and Instrumentation 1.2018,188-193

[21] Junhua Chen. ˝Research on stock selection model based on multi-classification support vector machine [D]. ˝ Huazhong university of science and technology, 2010.

[22] Fang Xu. ˝The research of high quality stock selection based on support vector machine[D]. ˝ Chongqing jiaotong university, 2013