Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 406 – 412

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

A SVM Stock Selection Model within PCA

Huanhuan Yu[a], Rongda Chen[b,*], Guoping Zhang[c]

[a]School of Finance, Zhejiang University of Finance & Economics, Hangzhou, 310018, China
[b]School of Finance, Zhejiang University of Finance & Economics, Hangzhou, 310018, China
[c]School of Economics and International Trade, Zhejiang University of Finance & Economics, Hangzhou, 310018, China

Abstract

In the financial market, well-performing stocks usually have some specific features in financial figures. This paper introduces a machine learning method of support vector machine to construct a stock selection model, which can do the nonlinear classification of stocks. However, the accuracy of SVM classification is very sensitive to the quality of training set. To avoid the direct use of complicated and highly dimensional financial ratios, we bring the principal component analysis (PCA) into SVM model to extract the low-dimensional and efficient feature information, which improves the training accuracy and efficiency as well as preserve the features of initial data. As empirical results show, based on support vector machine, within PCA after norm-standardization, the stock selection model achieves the entire accuracy of 75.4464% in training set and of 61.7925% in test set. Further, the PCA-SVM stock selection model contributes the annual earnings of stock portfolio to outperforming those of A-share index of Shanghai Stock Exchange, significantly.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.

Keywords: machine learning; stock selection; principal components analysis; support vector machine

1. Introduction

Stock has always been one of the most popular investment instruments in financial markets. Investors and researchers are devoting themselves to study out a method that can select accurately the stocks with favorable future

* Corresponding author. Tel.: +860571-85750010; fax: +860571-85212001.
E-mail address: rongdachen@163.com.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.
doi:10.1016/j.procs.2014.05.284

return to be constituents of investment portfolio. Guo and Zhang[1], Kuo et al.[2] and Tsumato et al.[3] develops several method to forecast stock prices or pick qualified ones from large sample. However, some traditional stock selection models usually face challenges when dealing with highly dimensional and nonlinear sample data for the reason that stock selection is a kind of determination with multi objectives and multi restrictions, along with the highly dimensional and huge financial data. The machine learning-based theory, Artificial Neural Network (ANN), can capture the regular patterns hidden behind the complex and high-dimension data through its machine learning [4,5]. Although ANN performs better than traditional methods, it has lots of defects at the same time, such as the difficulty to determine network structures, the problem with local minimum points and the over-fitting. Vapnik [6] proposed a new machine learning-based method called Support Vector Machine (SVM), which can better handle the high-dimension data avoiding the defects of ANN. SVM applies widely in many fields because of its particular advantages. A lot of researches, domestic and abroad, use SVM to predict stock prices or reversal points, as in Yeh et al.[7] and Huang[8]. But it's seldom to establish a stock selection model by SVM, and specifically rare in domestic.

This paper applies SVM into domestic stock market to establish an effective selection model. We treat financial ratios of listed companies in A-share of Shanghai Exchange as original data, and then use the principal components analysis (PCA) to preprocess them. First, we established a stock selection model (PCA-SVM) that recognizes high-return stocks when utilized SVM theory to train the training set. Second, apply PCA-SVM on test set to forecast the high-return stocks in the next year and do a comparison between the forecast and the actual to illustrate effectiveness of the established stock selection model.

2. Principal components analysis (PCA)

Financial ratios of a listed company include earning ability, growth ability, solvency ability and so on. Each ability contains many sub-ratios. If all the ratios were used as inputs in the training set, it would result in redundancy and low efficiency; even decrease the quality of empirical results. New variables can be created through transformation of original variables. Number of variables is less and most information is still retained. These new variables are called principal components.

2.1. Definition of principal components

Principal components can be expressed as follows:

$$\begin{cases} Y_1 = \vec{\alpha}_1^T \cdot \vec{X} = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1n}X_n \\ Y_2 = \vec{\alpha}_2^T \cdot \vec{X} = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2n}X_n \\ \cdots\cdots \\ Y_n = \vec{\alpha}_n^T \cdot \vec{X} = \alpha_{n1}X_1 + \alpha_{n2}X_2 + \cdots + \alpha_{nn}X_n \end{cases}, \tag{1}$$

where X_i is the original variable, Y_i is the principal component and $\vec{\alpha}_i$ is the coefficient vector respectively.

$\vec{\alpha}_i$ can be estimated by maximizing $Var(Y_i)$ with the constraint conditions of $\vec{\alpha}_i^T \cdot \vec{\alpha}_i = 1$ and $Cov(Y_i, Y_j) = \vec{\alpha}_i^T \cdot \sum \cdot \vec{\alpha}_i = 0, j = 1,2,\cdots,i-1$, where $\sum = (\sigma_{ij})_{n\times n}$ is the covariance matrix of $\vec{X}$.

2.2. Selection of principal components

The covariance matrix of $\vec{X} = (X_1, X_2, \cdots, X_n)^T$, $\sum = (\sigma_{ij})_{n\times n}$, is a symmetric non-negative definite matrix. Therefore it has n characteristic roots $\lambda_1, \lambda_2, \cdots, \lambda_n$, and n characteristic vectors.

Suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ and the orthogonal unit eigenvectors are $\vec{e}_1, \vec{e}_2, \cdots, \vec{e}_n$. The i th principal component of $X_1, X_2, \cdots, X_n$ can be expressed as follows:

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{in}X_n, i = 1,2,\cdots,n \tag{2}$$

with $Var(Y_i) = \vec{e_i}^T \cdot \sum \cdot \vec{e_i} = \lambda_i$ and $Cov(Y_i, Y_j) = \vec{e_i}^T \cdot \sum \cdot \vec{e_i} = 0,\ i \neq j$. The first p principal components' accumulated conribution rate is

$$ACR(p) = \sum_{i=1}^{p} \lambda_i / \sum_{i=1}^{n} \lambda_i \tag{3}$$

which represents the explanation power for original data of the principal components extracted by PCA method. Generally, an ACR of 85% is at least required, or the PCA method would be thought as unsuitable for losing too much original information.

Since the covariance matrix is sensitive to the order of magnitudes of data, we need to standardize the data first. There are two method of standardization in common use:

- Norm-standardization: $X_{ij}^* = (X_{ij} - \overline{X}_j)/s_j$, $\overline{X}_j$ is the mean and s_j is the standard deviation.
- Mean-standardization: $X_{ij}^* = X_{ij}/\overline{X}_j$, $\overline{X}_j$ is the mean.

3. Support vector machine

3.1. Linear classification of SVM

Linear classification of SVM is realized through solving for the optimal separating hyper-plane when the training set is linear separable. If the mingled two classes (C_1, C_2) of a sample can be separated correctly with the linear function (H_0) in a two-dimension plane, this sample is treated as linear separable.

Suppose the training set is $\{(\vec{x_1}, y_1), (\vec{x_2}, y_2), \cdots, (\vec{x_n}, y_n)\}$, where $\vec{x_i}$ is sample information vector ($\vec{x_i}$ is the coordinate vector in a two-dimension plane), $y_i \in Y = \{+1, -1\}$ and +1 represents class C_1 , -1 represents class C_2 . If the linear separating hyper-plane $H_0 : \vec{w}^T \cdot \vec{x} + b = 0$ separates the training set correctly, it is equivalent with the situation: when $y_i = +1$, $\vec{w}^T \cdot \vec{x_i} + b \geq +1$; when $y_i = -1$, $\vec{w}^T \cdot \vec{x_i} + b \leq -1$. If the distance of two data cluster of the sample, D^* , is maximized, this hyper-plane is called the optimal separating hyper-plane in this classification case.

Define $D^* = d_+ + d_-$,

$$d_{\pm} = \min_{i, y=\pm 1} \{ \left| \vec{w}^T \cdot \vec{x_i} + b \right| \Big/ \left\| \vec{w} \right\| \} \tag{4}$$

By substituting $\vec{w}^T \cdot \vec{x} + b = \pm 1$ in (4), we can obtain $D^* = d_+ + d_- = 2/\|\vec{w}\|$ and the problem is transformed to get the $\vec{w}$ minimizing $\|\vec{w}\|$. (b can be calculated by substituting sample points with $\vec{w}$ known)

Additionally, to avoid the situation that distance between the two parallel hyper-planes is maximized while effective classification is not realized, we must pose constraints on this optimization problem as follows:

$$y_i(\vec{w}^T \cdot \vec{x_i} + b) \geq 1 - \xi_i,\ 0 \leq \xi_i \leq 1. \tag{5}$$

ξ_i is the slack variable to tolerate the outliers. And a penalty factor C is also introduced into the objective function to reflect losses for tolerating the outliers. Training a SVM model, i.e. solving the optimization problem, will lead to a quadratic programming problem, as shown in (6).

$$\begin{cases} \max \quad \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j < \vec{x_i}, \vec{x_j} > \\ s.t. \qquad 0 \le \lambda_i \le C, \quad i = 1,2,\cdots,n \\ \qquad \sum_{i=1}^{n}\lambda_i y_i = 0 \end{cases} \tag{6}$$

Suppose λ^* is the solution of (6) and thus the optimal hyper-plane is $\vec{w}^{*T} \cdot \vec{x} + b^* = 0$, where $\vec{w}^* = \sum \lambda_i^* y_i \vec{x_i}$ and b^* can be calculated by the contraints of (5)..

3.2. Nonlinear classification of SVM

Linear classification of SVM we talked about in the prior section can be only applied when sample is linear separable. In this section, an improved nonlinear SVM method is proposed to solve the complicated and high-dimensional financial ratios.

A kernel function φ is very important here because it can map the original date into high-dimensional space H, i.e. $\varphi : R^n \to H; \vec{x} \to \varphi(\vec{x})$, which can let the data can be linear separable in H. Then an optimal separating hyper-plane discussed in prior section can be obtained to do the classification.

Suppose the training set is $\{(\vec{x_1}, y_1), (\vec{x_2}, y_2), \cdots, (\vec{x_n}, y_n)\}$, $\vec{x_i}$ is the highly dimensional information vector of the sample and $y_i \in Y = \{1, -1\}$. A quadratic programming similar with (8) is obtained through mapping φ:

$$\begin{cases} \max \quad \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j < \varphi(\vec{x_i}), \varphi(\vec{x_j}) > \\ s.t. \qquad 0 \le \lambda_i \le C, \quad i = 1,2,\cdots,n \\ \qquad \sum_{i=1}^{n}\lambda_i y_i = 0 \end{cases} \tag{7}$$

To solve (7), $\varphi : R^n \to H; \vec{x} \to \varphi(\vec{x})$ is needed to know, so we choose Gauss radial based kernel function (RBF) to get the inner product value as $k(x, y) = < \varphi(x), \varphi(y) >$ directly without searching for the complex φ.

4. Data selection

Table 1. Financial ratios and sample stocks information

Sample stock	Earnings ability *A*	Activity ratio *B*	Shareholder return *C*
2009, 677 stocks 2010, 679 stocks	EBIT a_1 ROA a_2 ROE a_3	Turnover of accounts receivable b_1 Turnover of inventory b_2 Turnover of current assets b_3	EPS c_1 Price-to-book ratio c_2 Common stock profitability c_3 P/CF c_4
Cash ratios *D*	**Growth ratios *E***	**Risk level *F***	**Solvency ratios *G***
EBIT-to-Cash ratio d_1 Cash-to-Assets ratio d_2 Operating ratio d_3	Growth of total assets e	Financial leverage f_1 Operating leverage f_2	Quick ratio g_1 Debt-to-Asset ratio g_2 EBIT/Interest ratio g_3 EBIT/Fixed charge ratio g_4

This paper selects 7 categories of financial ratios of companies in A-share Shanghai Stock Exchange from their annual reports of 2009 and 2010. The detailed financial indexes chosen are shown in Table 1. Our objective is to

separate the high-return stocks from the low ones according to their features hidden inside the financial ratios, thus it is necessary to label each stock with the return characteristic. After statistical analysis, all the companies have announced their annual report before 1th/May in 2009 and 2010. Therefore we label the stock as +1 if its return ranks the first 25% of all the sample stocks, i.e. $y_i = 1$ and $y_i = -1$ for the rest stocks. Labels of a part of sample are presented in Table 2.

5. Stock selection of model and analysis

5.1. Extraction of training set based on PCA method

Financial ratios of 677 stocks in 2009 are the original data. We apply PCA to extract the principal components satisfying the condition of $ACR \geq 85\%$. Since our sample is large, if we apply PCA on all of the ratios of 677 stocks directly, we would lose the local information and the effect of dimension reduction is also smaller. Thus we do PCA extraction one time for every 40 sample stocks. The training set is in Table 2.

Table 2. Training set of SVM nonlinear classification (part of 677 stocks)

Stock code	Earnings ability	Activity ratios	Shareholder return	Cash ratios	Growth ratios	Risk levels	Solvency ratios	y
PCA with norm-standardization								
600069	-1.6114	-0.9830	-0.4337	-1.0664	-0.4253	0.7874	0.1431	1
600070	0.5249	-0.3005	-0.8563	-0.5438	-0.0903	-0.1103	0.0136	-1
600071	2.1843	0.1875	-1.5191	1.1364	-0.6570	-1.7170	0.7624	1
PCA with mean-standardization								
600069	0.8222	-1.3006	0.8049	1.0620	-0.9571	0.3681	1.8768	1
600070	4.6133	1.0647	-0.3712	-1.1497	0.8309	1.6046	1.5020	-1
600071	7.0948	1.1286	-0.7982	0.2286	0.2485	-0.2133	2.0515	1

5.2. SVM stock selection model and analysis

The total scores obtained in the prior section combined with return labels of sample stocks constitute the complete training set of SVM. By applying the nonlinear classification of SVM introduced in section 3 on the training set, we can obtain the optimal separating hyper-plane. If we use this hyper-plane on test set, stocks in test set can be classified into the high-return part and the low-return part. It can be seen as a prediction of stocks' future return characteristic. The accuracy of classification and prediction is presented in Table 3.

Table 3. Accuracy of SVM nonlinear classification

	Method used	Mean-standardization PCA-SVM	Norm-standardization PCA-SVM
Training	Whole accuracy a	88.6905%	75.4464%
	Accuracy of +1 a	100%	58.5366%
	Accuracy of -1 a	85.0394%	80.9055%
Test	Whole accuracy b	69.1943%	61.7925%
	Accuracy of +1 b	10.1266%	24.5283%
	Accuracy of -1 b	88.8421%	74.2138%

Training and testing of SVM proceed with Livsvm 3.1 in Matlab. To achieve the best generalization ability, the optimal penalty factor C and the coefficient σ in RBF is determined by Grid Searching method.

By observing Table 3, we can find that the accuracy of mean-standardization PCA-SVM for label +1 in training set is 100%. However, the accuracy of the same label in test set is only 10.1266%. It is the over-fitting phenomenon that too many support vectors were used to explain the training set, which could has a good classification effect on training set while a bad effect on predictions. The accuracy of norm-standardization PCA-SVM is obviously better.

For further analysis, we construct an equal weighted portfolios with stocks selected by PCA-SVM and do a comparison between the accumulated return (ACR) gained by this model and the A-share index of Shanghai Stock Exchange. The comparison is presented in Fig.1. It manifests that PCA-SVM has higher accumulated return over the A-share index, which means SVM classification method is accurate and highly efficient when dealing with complex and highly dimensional data.

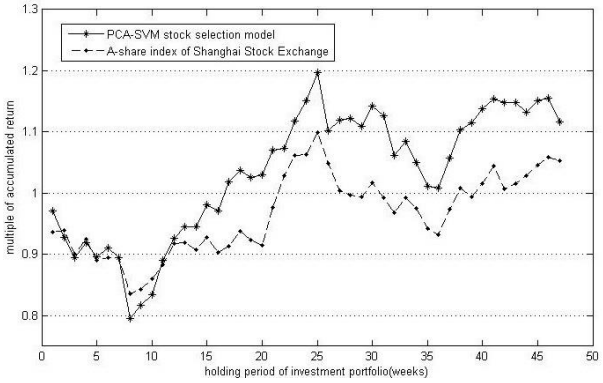

Fig.1. Comparison between PCA-SVM and A-share index of Shanghai Stock Exchange

6. Conclusions

Support Vector Machine is commonly used to train the time-series data of stocks for price forecasting. In this paper, SVM is employed to generate an optimal separating hyper-plane in high-dimensional space based on the training set. To increase the accuracy and efficiency of SVM classification model, we apply PCA to process the original data. Finally, the empirical result has suggested that the return of stocks selected by PCA-SVM is apparently superior to A-share index.

Information features of financial ratios of companies vary with their industries. We believe that the quality of training set can be improved if we apply PCA on each industry separately. Additionally, it is quite meaningful for achieving higher returns if stocks could have different weights according to their risk-return characteristics when portfolios are constructed.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 71171176).

References

1. Ming Guo, Yuan-Biao Zhang. A Stock Selection Model Based on Analytic Hierarchy Process. Factor Analysis and TOPSIS//*The International Conference on Computer and Communication Technologies in Agriculture Engineerin*. 2010. p. 466-469.
2. Kuo R.J., Chen C.H.& Hwang Y.C. A Intelligent Stock Trading Decision Support System Through Integration of Genetic Algorithm based Fuzzy Neural Network and Artificial Neural Network. *Fuzzy Sets and Systems*. 2001; 118: 21-45.
3. Tsumato S., Slowinski S., Komorowsk J. & Grzymala-Busse J.W. Lecturenotes in Artificial Intelligence. *The fourth international conference on rough sets and current trends in computing*. 2004.

4. E.L. de Faria, Marcelo P. Albuquerque, J.L. Gonzalez, J.T.P. Cavalcante, Marcio P. Albuquerque. Predicting the Brazilian Stock Market Through Neural Networks and Adaptive Exponential Smoothing Methods. *Expert Systems with Application*. 2009; 36:12506-12509.
5. Yudong Zhang, Lenan Wu. Stock Market Prediction of S&P 500 via Combination of Improved BCO Approach and BP Neural Network. *Expert Systems with Applications*. 2009; 36: 8849-8854.
6. Vladimir N. Vapnik. Statistical Learning Theory. *Publishing House of Electronics Industry*. 2004.
7. Chi-Yuan Yeh, Chi-Wei Huang, Shie-Jue Lee. A multiple-kernel support vector regression approach for stock market price forecasting. *Expert Systems with Applications*.2011; 38: 2177-2186.
8. Pengpeng Huang. Prediction of the Turnover Points in Stock Trend Based on Support Vector Machine. *College of Software, Fudan University*. 2010.