Using AH Premium to Predict Related Stock Index with Support Vector Machine

Yiyang Chen*

Chinese Academy of Finance and Development
Central University of Finance and Economics
** Corresponding author: chenyiyang2000@email.cufe.edu.cn*

Abstract
Since the launch of Shanghai-Hong Kong Stock Connect, AH premium of the A share and H share dual-listed companies maintains at a high level. The extra cost for mainland investors to invest in H shares is an important factor. This paper will analyze this phenomenon and apply support vector machine (SVM) to predict related stock index – Hang Seng China Enterprises Index (HSCEI), in order to examine whether investors can invest these dual-listed companies according to the change of AH premium. The forecasting ability of different kinds of SVMs are compared and the results show that when appropriate parameters are selected, the success rate of prediction can reach nearly 56%. Thus investors can invest with reference to changes of AH premium to some extent.

***Keywords**-AH premium; Hang Seng China Enterprises Index (HSCEI); Stock index prediction; Support vector machine*

1. Introduction

Companies in China can issue A-shares in mainland stock exchange and H-shares in Hong Kong stock exchange, but the prices of the same company are always different. Over the past decade, many companies in China sought overseas listings, and now there are over 100 companies both listed in China mainland market and Hong Kong market. The ratio of the A-share price higher than the H-share price is called AH premium.

Analysis shows that the extra costs for mainland investors to invest in H-shares is an important factor of high AH premium. This paper supposes that investors who can freely invest in Hong Kong market will find it possible to make successful investment in H-shares of dual-listed companies according to the change of AH premium and will focus on the feasibility of using AH premium to predict Hang Seng China Enterprises Index (HSCEI), most weight of which is the H-shares of dual-listed companies. Linear and nonlinear support vector machines are applied to make the prediction.

2. Analysis and Literature Review of AH Premium

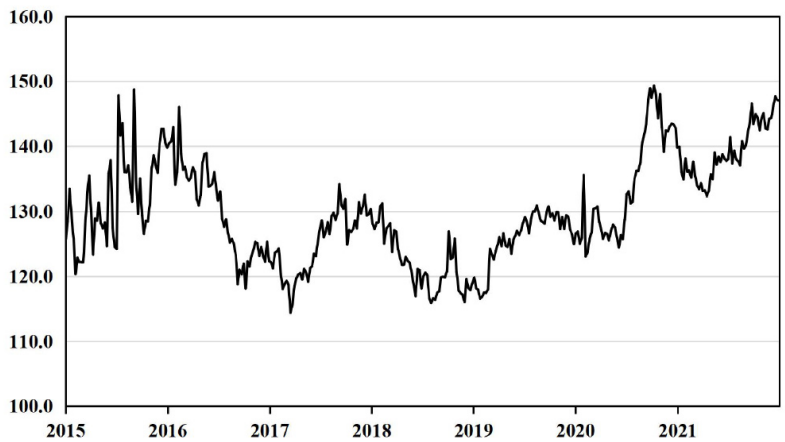

Figure 1. AH Premium Index 2015-2021

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 355–360, 2023.
https://doi.org/10.2991/978-94-6463-042-8_52

AH Premium Index tracks the price difference of the shares of these dual-listed companies and calculates the weighted average premium (or discount) of A-shares over H-shares. After the launch of Shanghai-Hong Kong Stock Connect which allows investors in mainland invest in part of the stocks in Hong Kong market in 2014, the AH Premium Index increased to the average level of 130, which means on average the stocks of dual-listed companies in Hong Kong are 30% more expensive than in mainland. Figure 1 shows the change of the index from 2015 to 2021.

The reasons for high AH premium can be concluded as 3 parts. The first reason is trading cost. The commission charges, stamp taxes and transaction fees of A shares are generally lower than H shares. Particularly, for those mainland investors who invest in H shares through Hong Kong Stock Connect, the dividend payments will shrink due to a higher personal income tax rate. The second reason is liquidity compensation. The trading volume of dual-listed companies is always larger in the A-share market than in the H share market. Additionally, Hong Kong stock investors are relatively mature, which means their investment is more long-term oriented. Also the money from selling H shares through Hong Kong Stock Connect will take 2 more days to be available for investors. Bai et al. studied the price difference between A-shares and H shares of 63 listed companies and found some factors, such as trading liquidity and different risk attitudes of investors, can explain AH premium [1].The third reason is risk compensation. The A shares of the same listed company have rise or fall limits which sometimes can effectively reduce the risk of a sharp fall in stock price, while the H shares do not. Fernald and Rogers found that the exchange rate has a close relationship with the variations of the AH premium [2], which means there is extra exchange rate risk for mainland investors. Furthermore, Chan et al. found that information asymmetry is an important factor contributing to the pricing of the H shares of Chinese dual-listed companies [3]. However, few researches studied the relationship between AH premium and Hang Seng China Enterprises Index (HSCEI).

3. Previous Applications of Support Vector Machines in the Stock Market

3.1. Introduction to Support Vector Machine

Support vector machine (SVM) is a kind of generalized linear classifier [4-5]. Its decision boundary is the maximum-margin hyperplane solved from the training samples. Basic linear SVM means finding a hyperplane serving as the decision boundary to maximize the geometric margin between the data set and the hyperplane in the characteristic space where the input data is located, and the samples are separated into the positive class and the negative class by the boundary. If the data is not separable in the original finite-dimensional space, then nonlinear SVM is needed for classification. The nonlinear SVM is created by applying the kernel trick and mapping the original features into a high-dimensional space, and it allows the algorithm to find the maximum interval hyperplane in the transformed feature space. Commonly used kernel functions include polynomial kernel and Gaussian radial basis function kernel.

3.2. Previous Applications

Support vector machine model is widely applied in stock market prediction. Tay and Cao examined the predictability of financial time-series data using SVMs and compared it with multi-layer back-propagation (BP) neural network [6-7]. Kim applied SVMs to predict the direction of change of Korea composite stock price index (KOSPI) by technical indicators [8]. Huang et al. implemented SVMs to forecast the weekly movement direction of NIKKEI 225 index [9]. Madge S and Bhatt S used daily closing prices for 34 technology stocks to calculate price volatility and momentum and predicted future price changes with SVMs [10]. However, there is few study on AH premium index and its connection with other indexes.

4. Data and Experiment Design

4.1. Data

In order to find an investment target for prediction related to H shares of dual-listed companies, Hang Seng China Enterprises Index (HSCEI) is chosen because most of its components are H shares of those dual-listed companies. This output target is categorized as “0” or “1”. “0” means the index of the next week is lower than this week, and “1” means the index of the next week is higher than this week. Weekly change of AH premium index, RMB exchange rate, Hang Seng Index, China Securities Index 300 Index and Standard & Poor's 500 Index are selected to be the input features after standardization. The data is collected from 2015 to 2021, and the samples between June 2015 and September 2015 and the samples of March 2020 are dropped because there were rapid changes of features in these two periods, and this study does not focus on the prediction of special condition. The total amount of the sample is 340 weeks. Table 1 shows the descriptive statistics of these features before standardization.

TABLE 1. Descriptive Statistics

Feature	Description	Max	Min	Mean	Standard Deviation

AH Premium	Weekly change of AH Premium Index	6.230	-12.440	-0.0049	2.363
HSI	Weekly percentage change of HSI	7.900	-9.491	0.1316	2.433
Exchange rate	Weekly change of RMB/USD Exchange rate	0.1258	-0.123	-0.0003	0.038
CSI300	Weekly percentage change of CSI300 Index	7.5991	-10.081	0.2694	2.710
SP500	Weekly percentage change of S&P500 Index	7.3236	-8.793	0.3184	1.920

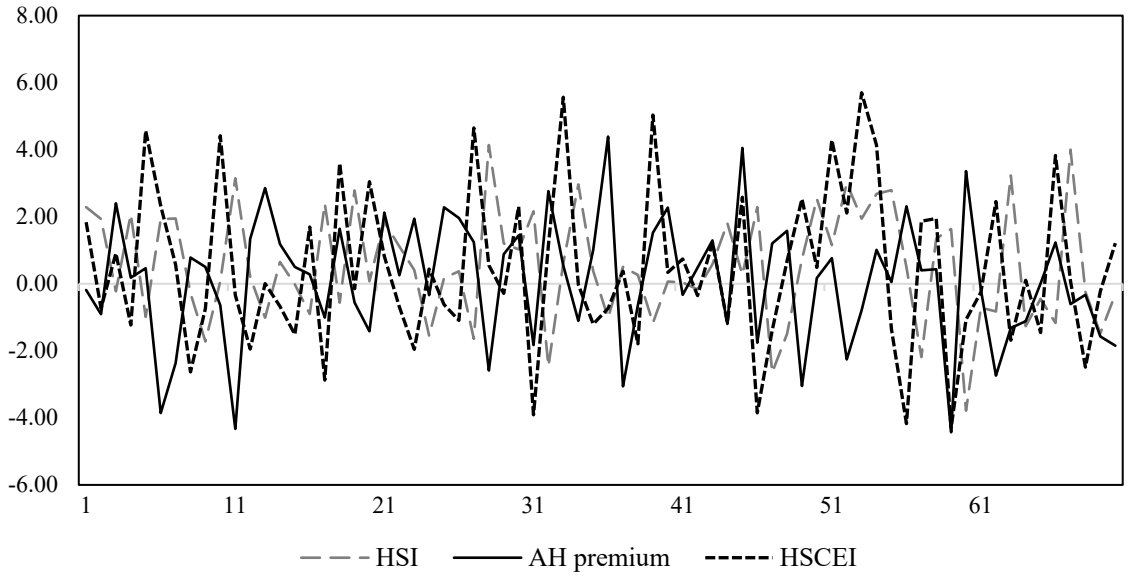

Figure 2. Weekly percentage change of Hang Seng Index (HSI), AH Premium Index and Hang Seng China Enterprises Index next week (HSCEI) (70 samples from January 2017 to May 2018)

As shown in Figure 2, the movement of Hang Seng China Enterprises Index (HSCEI) has relationship with the weekly change of AH Premium Index and Hang Seng Index (HIS). However, the relationship is very complex and explicit formula cannot be derived to describe it. Thus in this paper the support vector machine algorithm will be applied to find the underlying relationship.

4.2. Experiment Design

The target function of soft-boundary Support Vector Machine can be written as [11]

$$\min_{\omega,b,\xi} \ \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{l} \xi_i \qquad (1)$$

subject to

$$y_i(\omega^T \Phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, l \qquad (2)$$

In formula (1), ω is the normal vector of the decision boundary, C is the penalty coefficient for wrong classifications and ξ is the slack variable. In formula (2), x is the input feature, y is the output target, b is the intercept of the decision boundary and $\Phi(x_i)$ is the mapping function when applying kernel trick. In the linear SVM, $\Phi(x_i)$ is just x_i. When applying kernel trick, the kernel function is denoted by $K(\Phi(x), \Phi(z))$.

The function of polynomial kernel can be written as

$$K(x_i, z_i) = (x_i \cdot z_i + r)^d \qquad (3)$$

Particularly, when d=1, r=0, this becomes the linear kernel (the model becomes linear SVM).

The function of Gaussian radial basis function kernel can be written as

$$K(x_i, z_i) = \exp(-\gamma \|x_i - z_i\|^2) \qquad (4)$$

for $\gamma > 0$. Sometimes it is parameterized by $\gamma = \frac{1}{(2\sigma)^2}$.

In this study, three types of SVM will be applied to find the decision boundary for prediction, including linear SVM, nonlinear SVM with Gaussian radial basis function kernel and nonlinear SVM with polynomial kernel. The prediction performance is estimated by the ratio of successful predictions to the total sample size. Also, as the amount of the sample is relatively small, Repeated K-Fold Cross-validation will be used to make the estimation more robust. Repeated K-Fold Cross-validation is a method that effectively utilizes

limited data and the evaluation results can be very close to the model's actual performance on the test set [12]. In this method, the original data is divided into k groups, and each subset data can be used as the test set, while the remaining k-1 subsets of data are used as the training set at one time. So k models can be obtained after one division, and n*k models can be obtained after n times of repeated divisions. These final estimation of the model are the average performance of these n*k models.

5. RESULTS AND DISCUSSION

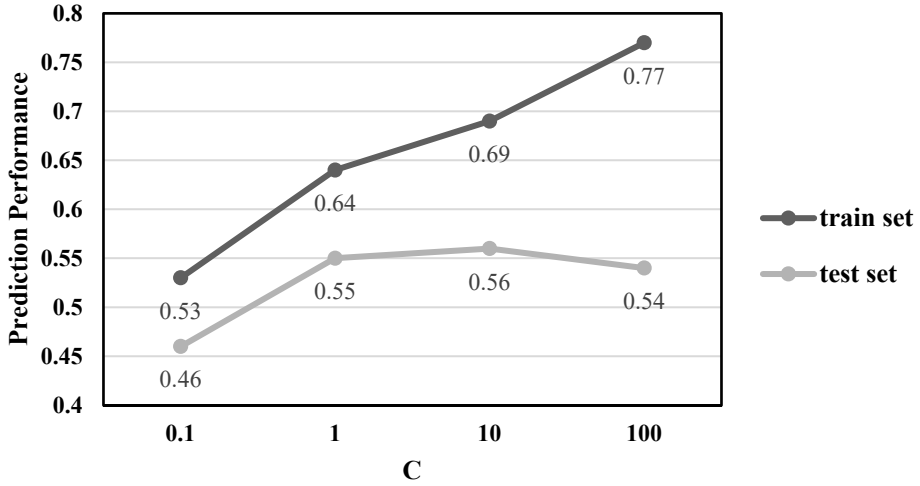

Figure 3. The results of Gaussian radial basis function kernel nonlinear SVMs with different C and fixed gamma 0.1

In this paper, the prediction performance of linear SVM and nonlinear SVM using different kernel functions with different misclassification penalty coefficient C are compared (the gamma of Gaussian radial basis function kernel is set as 0.1 and the degree of the polynomial kernel is set as 3), as is shown in Figure 3. For the linear SVM model, the prediction performance of training data and test data is stable at 0.59 and 0.55, respectively, indicating that changing the penalty coefficient of misclassification C only has a tiny influence on the selection of segmentation hyperplane. For the Gaussian radial basis function kernel model, the prediction performance of the training data increased with the increase of misclassification penalty coefficient, and the prediction performance of the test data will increase in the beginning and then decrease with the increase of penalty coefficient. It shows that when C is small, there is an under-fitting problem, and performance for the forecast of training data and testing data are poor; when C is set at a suitable value, the model performs well on the whole; however, when C is large, there is an over-fitting problem. Although the prediction performance of training data is significantly improved, the prediction performance of test data is poor, as is shown in Figure 3. For the polynomial kernel model, the prediction performance of training data improved with the increase of the misclassification penalty coefficient, similar to Gaussian radial basis function kernel, but the prediction performance of test data was always poor. Therefore, the Gaussian radial basis function kernel models with different parameters are compared next.

TABLE 2. THE RESULTS OF GAUSSIAN RADIAL BASIS FUNCTION KERNEL SVMS WITH VARIOUS PARAMETERS

Value of gamma	Performance of train set	Performance of test set
(a)C=1		
0.01	0.57	0.5
0.1	0.64	0.55
1	0.86	0.52
(b)C=10		
0.01	0.60	0.53
0.1	0.69	0.56
1	0.96	0.49
(c)C=100		
0.01	0.63	0.55
0.1	0.77	0.54
1	1.00	0.49

Table 1 shows the prediction performance of Gaussian radial basis function kernel SVM with different parameters. Parameter gamma is correspondent to γ in the kernel function mentioned above. It reflects the effective width of a single sample. When gamma is small, the effective width of each sample is large and the model complexity is low, which means the under-fitting problem is more possible to happen. When gamma is close to 0, the decision boundary of the Gaussian radial basis function kernel SVM is close to the linear SVM. When gamma is large, the effective width of each sample is smaller and the model complexity is high, which means the over-fitting problem is more possible to happen. The table shows that there is a strong

relationship between prediction performance and the parameters chosen. When the misclassification penalty coefficient C is relatively small, small value of gamma will make the under-fitting problem even worse, resulting in low accuracy in prediction of both training data and test data. Meanwhile, a large and suitable value of gamma will resolve the under-fitting problem and generate a better result. On the other hand, when the penalty coefficient C is relatively large, large value of gamma will make the over-fitting problem even worse, resulting in high accuracy (close to 100%) in prediction of training data but low accuracy in prediction of test data. Similarly, a small and suitable value of gamma will resolve the over-fitting problem and generate a better result. Among the parameter combinations used in the table, the model performed best when C=10 and gamma =0.1, with prediction accuracy of 69% for the train set and 56% for the test set.

6. Conclusion

This paper introduces the AH Premium index and analyzes the reasons for the AH premium. One of the most important reasons is that mainland investors need to pay extra cost when buying Hong Kong stocks through the Hong Kong Stock Connect. Therefore, this paper attempts to test whether investors who can trade directly in the Hong Kong market can follow the changes in the AH premium index and invest in the H shares of the dual-listed companies in Hong Kong successfully.

In this paper, the SVMs are used to predict the rise and fall of Hang Seng China Enterprise Index, and three kernel functions including Linear kernel, Gaussian kernel and Polynomial kernel are selected to build linear SVM and nonlinear SVM models. In addition, this paper also discusses the influence of parameter changes in SVMs on the prediction results of training set and test set. In general, when appropriate parameters are selected, the success rate of test set prediction of linear SVM and nonlinear SVM using Gaussian radial basis function kernel (corresponding to function (4) in part IV) can reach nearly 56%. Therefore, investors can make better investment in H shares (the stocks traded in the Hong Kong stock market) of these dual listed companies with the guidance of AH premium index.

Future researches can be done in two ways. The first is to find more effective input features to improve the prediction performance of SVM. The second is to verify the rise and fall of a specifically designed portfolio of H shares of dual listed companies.

References

[1] Bai, Yu, W. M. Tang, and K. F. C. Yiu. "Analysis of Price Differences Between A and H Shares." Asia-Pacific Financial Markets 26.4 (2019): 529-552.

[2] Fernald, John, and John H. Rogers. "Puzzles in the Chinese stock market." Review of Economics and Statistics 84.3 (2002): 416-432.

[3] Chan, Kalok, Albert J. Menkveld, and Zhishu Yang. "Information asymmetry and asset prices: Evidence from the China foreign share discount." The Journal of Finance 63.1 (2008): 159-196.

[4] Vapnik, Vladimir N. "An overview of statistical learning theory." IEEE transactions on neural networks 10.5 (1999): 988-999.

[5] Cristianini, Nello, and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[6] Tay, Francis EH, and Lijuan Cao. "Application of support vector machines in financial time series forecasting." omega 29.4 (2001): 309-317.

[7] Tay, Francis EH, and L. J. Cao. "Modified support vector machines in financial time series forecasting." Neurocomputing 48.1-4 (2002): 847-861.

[8] Kim, Kyoung-jae. "Financial time series forecasting using support vector machines." Neurocomputing 55.1-2 (2003): 307-319.

[9] Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine." Computers & operations research 32.10 (2005): 2513-2522.

[10] Madge, Saahil, and Swati Bhatt. "Predicting stock price direction using support vector machines." Independent work report spring 45 (2015). .

[11] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM transactions on intelligent systems and technology (TIST) 2.3 (2011): 1-27.

[12] Refaeilzadeh, Payam, Lei Tang, and Huan Liu. "Cross-validation." Encyclopedia of database systems 5 (2009): 532-538.