The Nikkei Stock Average Prediction by SVM

Takahide Kaneko and Yumi Asahi(✉)

Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-Ku, Tokyo, Japan
Kanetaka2020@gmail.com

Abstract. The problem of how to extract structures hidden in large amounts of data is called "data mining". Using a support vector machine (SVM), which is one of the data mining methods, I predicted the rise and fall of the Nikkei stock average one day, one week, and one month later. As explanatory variables, we used the historical rate of change in US stock prices and the Nikkei Stock Average. As a result of the analysis, it was possible to stably improve the prediction accuracy of the diary average stock price one day later compared to random prediction. In addition, SHAP was used to analyze whether the explanatory variables were appropriate. As a result, we found that the effect of each explanatory variable on the analysis results differs depending on how the training set and test set are divided. We made it a future task to make stock price predictions using SVMs more concrete and convincing.

Keywords: Support Vector Machine · Stock price prediction · SHAP

1 Introduction

In recent years, computers and the Internet have developed, and a lot of information has been stored on the web. At convenience stores and online shops, customer data such as purchasing history is consolidated every day, and economic data on stock prices and exchange rates can be easily obtained online. These data are originally gathered for some purpose. For example, customer data is the purpose of wanting to increase sales, or in the case of stock price data, the purpose is to predict the price movement of stock prices in the future. However, the data obtained from it is too large, and it is difficult to fulfill its purpose by looking at raw data, and it is necessary to extract the necessary information from large-scale data. The question of how this hidden structure is extracted is called "data mining", and research is underway. One of the data mining methods is the "Support Vector Machine" used in this study. SVM is one of the algorithms commonly used in data analysis sites due to the generalization performance and the size of the applied field. Based on the idea of maximizing the margin, it is mainly used in binary classification issues. It is also possible to apply to multi-class classification and regression problem.

In this study, the purpose of research is to use SVM to predict the Nikkei Stock Average. There is existing research in the prediction of stock prices using SVM. For example, there are research [1], which is classified as a company's stock price "rises" and "drop" after the news article is distributed. Research [1] analyzed the rise and fall

H. Mori and Y. Asahi (Eds.): HCII 2023, LNCS 14015, pp. 211–221, 2023.
https://doi.org/10.1007/978-3-031-35132-7_15

of stock prices a few minutes later. In this study, we focus on the Nikkei Stock Average, not for each company, and predicts the rise or fall of one day, one week, and one month later, rather than a few minutes later. In the research [1], classification was performed when the stock price went up and when it went down. In this study, we will analyze the classification in addition to the case where there is not much change. The reason for analyzing by adding this class classification is that it is common to avoid investing if the stock price does not change much considering the transaction cost, etc., considering the transaction cost. In addition, the explanatory amount is the change rate of the past US stocks and the Nikkei Stock Average. The US economy has a major impact on Japan. In consideration of the impact and the focusing on the past price movements of the Nikkei Stock Average itself, the Nikkei Average will be predicted in the future, and will be treated as explanation. The results obtained by verification clarify the conformity and issues of the prediction by SVM.

2 Method

Here, we describe the basics of SVM, which is one form of supervised learning, and introduce previous research using SVM in the financial field.

2.1 Support Vector Machine (SVM)

Support vector machines (SVM) are one of the machine learning algorithms that are often used in data analysis because of their generalization performance and wide range of applications. Based on an idea called margin maximization, it is mainly used for binary classification problems. Applications to multi-class classification and regression problems are also possible. The features of SVM include that it is difficult to cause overfitting and that it is possible to make highly accurate predictions even with relatively small amounts of data. However, its computational cost is high compared to other machine learning algorithms, making it unsuitable for large datasets.

2.2 Margin Maximization

The n-1 dimensional plane that classifies n-dimensional data is called the separating hyperplane, and the distance to the data (support vector) closest to the separating hyperplane is called the margin. Maximizing this margin is the goal of SVM. A margin that assumes linearly separable data is called a hard margin. A soft margin is a margin that allows for erroneous discrimination on the premise of data that cannot be linearly separable.

2.3 Kernel Method

In fact, in most cases, linear separation cannot be performed with the data as it is. Therefore, linear separation may be possible by subjecting the original data to nonlinear transformation to a higher dimension. A machine learning method that performs high-dimensional nonlinear transformation of the feature vectors included in the learning data and identifies the spatial linearity is called the kernel method.

2.4 Kernel Tricks and SVM

The kernel method enables linear separation by transforming the data into a higher dimension. However, there is a fear that calculation will become difficult as the data becomes higher dimensional. Kernel tricks are used there.

A high-dimensional feature vector $\Phi(x^n)(n = 1, 2, \cdots i, \cdots j \cdots N)$ can be obtained by nonlinearly transforming the feature vector $x^n(n = 1, 2, \cdots i, \cdots j \cdots N)$ of the original data. The inner product $x^i \cdot x^j$ of data is required when calculating to solve a quadratic programming problem when executing a linear SVM. Similarly, in SVM in high-dimensional space, the inner product $\Phi\left(x^i\right) \cdot \Phi\left(x^j\right)$ of $\Phi\left(x^i\right)$ and $\Phi\left(x^j\right)$ obtained by transforming data x^i and x^j into high-dimensional space is required. By defining the shape of the inner product $\Phi\left(x^i\right) \cdot \Phi\left(x^j\right)$ after this transformation with the kernel function $K(x^i, x^j)$, the concrete form of $\Phi(x^n)$ is eliminates the need to define. This property of the kernel function is called the kernel trick, and it is possible to prevent the calculation from becoming difficult due to the high dimensionality. Typical kernel functions include the following (1) Polynomial kernel, (2) Gaussian kernel, and (3) Sigmoid kernel. In this study, the Gaussian kernel was used for the analysis.

(1) Polynomial kernel

$$\Phi(x^i) \cdot \Phi(x^j) = K(x^i, x^j) = (x^i \cdot x^j + c)^d$$

(2) Gaussian kernel

$$\Phi(x^i) \cdot \Phi(x^j) = K(x^i, x^j) = exp(-\gamma \|x^i - x^j\|^2)$$

(3) Sigmoid kernel

$$\Phi(x^i) \cdot \Phi(x^j) = K(x^i, x^j) = tanh(cx^i \cdot x^j + \theta)$$

3 Empirical Research

3.1 Usage Data

In this research, we use past Nikkei 225, NY Dow, and S&P500 prices to predict fluctuations in the Nikkei 225 stock price one day, one week, and one month later. Table 1 below shows the source, type, treatment, and period of data used in the analysis.

Table 1. Summary of usage data

Data source	Fact Set
Data type	① Nikkei Stock Average ② NY Dow ③ S&P500
Data handling	Forecast after 1 day Obtain each daily data and calculate the rate of change The rates of change on the previous day and the day before the previous day are used as explanatory variables (feature values) In addition, the daily rate of change of the Nikkei Stock Average is used for class classification Forecast after 1 week Obtain each weekly data and calculate the rate of change The rates of change in the previous week and the week before last are used as explanatory variables (feature values) In addition, the weekly rate of change of the Nikkei Stock Average is used for class classification Forecast after 1 month Obtain each monthly data and calculate the rate of change The rates of change in the previous month and the month before last are used as explanatory variables (feature values) Also, the monthly rate of change of the Nikkei Stock Average is used for class classification
Data period	Forecast after 1 day April to June 2022 Forecast after 1 week September 2020 to August 2022 Forecast after 1 month September 2012 to August 2022

3.2 Analysis Procedure

① Classification according to the price movement of the Japanese stock average

Classify using the rate of change of the acquired Nikkei Stock Average. A method of classifying according to an increase or a decrease and a method of classifying using the average value μ and standard deviation σ of the rate of change during the obtained period are used.

- Classification method 1

(a) When the stock price rises: the stock price change rate is 0 or more
(b) When the stock price falls: the stock price change rate is less than 0

- Classification method 2

(a) When the stock price rises: the rate of change exceeds $\mu + \sigma$.

(b) When the stock price falls: the rate of change is less than $\mu - \sigma$.
(c) Not much change: the rate of change is greater than $\mu + \sigma$ and less than $\mu - \sigma$.

② Perform supervised learning with kernel SVM

- Learning procedure

(1) Divide the dataset into training set and test set.
(2) Perform grid search using leave-one-out cross-validation on the training set to find appropriate parameters.
(3) Perform SVM learning on the training set with the obtained parameters.
(4) Evaluate the performance using the test dataset.

③ Perform learning multiple times and evaluate the SVM.

The data are randomly classified into training set and test set, and supervised learning is performed multiple times. The following three indicators are used for each analysis.

(1) Accuracy of training set
(2) Accuracy of test set
(3) Average F value

Of these, the average F value is the macro average of each class. By taking the macro-average, it is possible to prevent a situation in which the F value becomes high even when the classification is extremely biased. In Classification Method 1, if the probability of a stock price going up is equal to the probability of a stock price going down, and it is predicted randomly, the hit rate and the average F value will be 0.5, which is the basis for evaluating the accuracy of SVM analysis. Becomes a line. Also, in classification method 2, if we consider the same as in classification method 1, the hit rate and the average F value are both 0.33, which is the baseline for SVM analysis accuracy evaluation.

4 Result

Based on the analysis procedure shown in Sect. 3.2, learning by SVM was performed 100 times for each classification method. Table 2 below summarizes the results of averaging the evaluation indices for each analysis accuracy. As a result, from Table 2, both the correct answer rate and the average F value for the prediction one day later exceeded the baseline. However, the average F value for the predictions after 1 week and 1 month is below the baseline, indicating that the analysis accuracy is poor. Therefore, the proposed method is considered to have a certain degree of predictive power for fluctuations in the Nikkei Stock Average one day later.

Table 2. Analysis accuracy evaluation index results for each classification

		Baseline for Accuracy Evaluation	Accuracy of learning data	Accuracy of evaluation data	Average F value
Classification method 1	1day later	0.5	0.8813	0.6895	**0.6801**
	1week later	0.5	0.7013	0.4929	0.4089
	1month later	0.5	0.7708	0.5858	0.3715
Classification method 2	1 day later	0.33	0.8514	0.5879	**0.4485**
	1week later	0.33	0.7096	0.6805	0.2734
	1month later	0.33	0.7486	0.7338	0.2929

Detailed analysis results after 1 day

Figures 1 and 2 below show the detailed analytical results of stock price forecasts for the next day. From Figs. 1 (left) and 1 (right), for classification method 1 (2 categories), both the hit rate and the F value are consistently above 0.5, suggesting that the analysis accuracy is stable. On the other hand, from Figs. 2 (left) and 2 (right), for classification method 2 (3 categories), the accuracy rate is consistently above 0.33, but the average F value is often below 0.33, so the analysis accuracy is unstable.

Classification method 1 (2 categories)

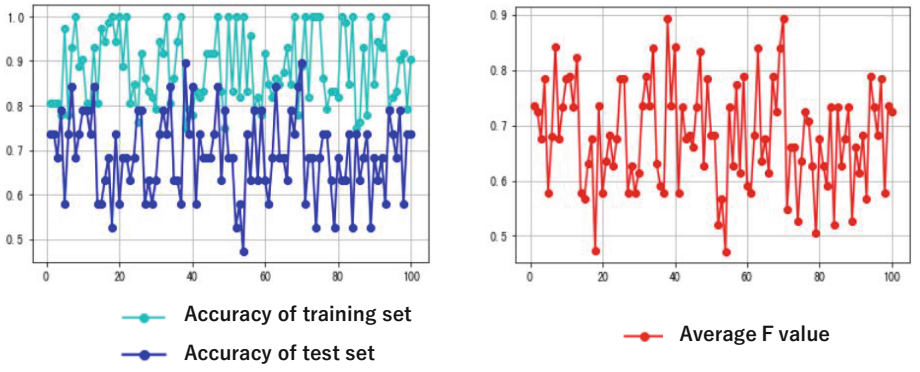

Fig. 1. (left): Accuracy rate for each data when SVM learning is performed 100 times. (right): Each average F value when learning by SVM is performed 100 times

Classification method 2 (3 categories)

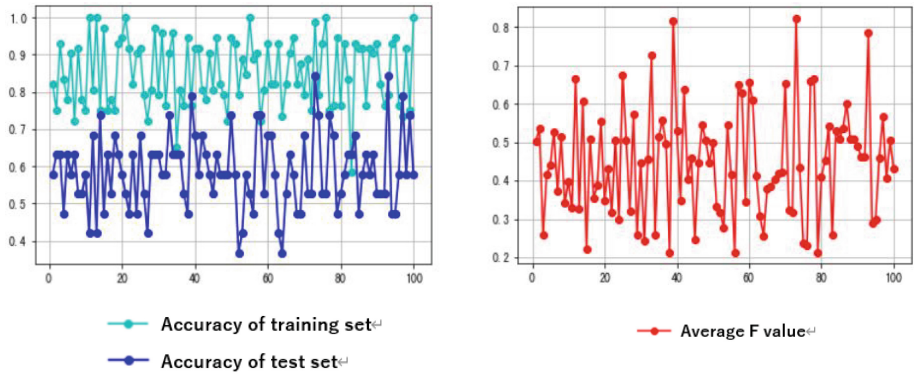

Fig. 2. (left): Accuracy rate for each data when SVM learning is performed 100 times. (right): Each average F value when learning by SVM is performed 100 times

Detailed analysis results after 1 week

Figures 3 and 4 below show the detailed analysis results of stock price forecasts for the next week. From Fig. 3, we can see that the hit rate for classification method 2 is consistently above 0.5, but from Fig. 4, we can see that the average F value is low, suggesting that the classification is excessively biased.

Classification method 1 (2 categories)

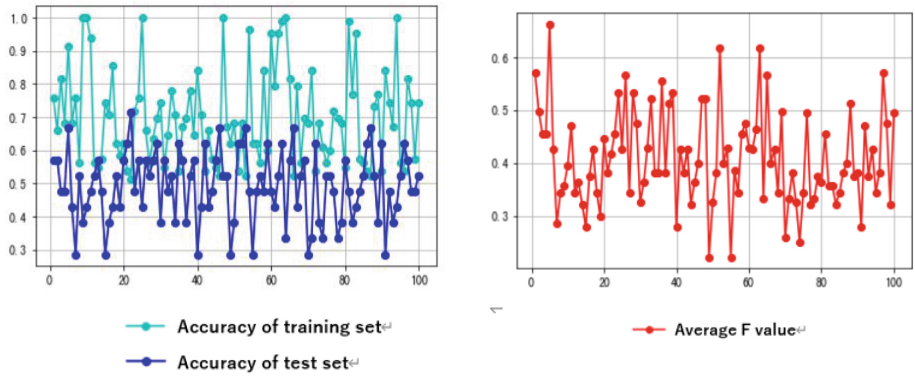

Fig. 3. (left): Accuracy rate for each data when SVM learning is performed 100 times. (right): Each average F value when learning by SVM is performed 100 times

Classification method 2 (3 categories)

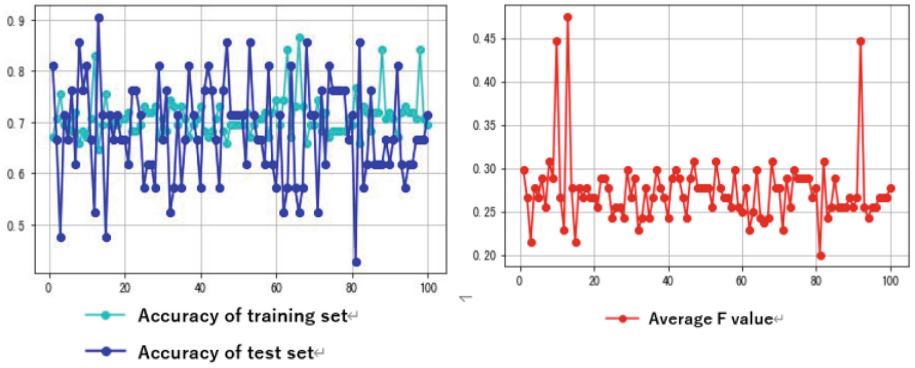

Fig. 4. (left): Accuracy rate for each data when SVM learning is performed 100 times. (right): Each average F value when learning by SVM is performed 100 times

Detailed analysis results after 1 month

Figures 5 and 6 below show the detailed analysis results of the stock price forecast one month later. From Fig. 5, we can see that the hit rate for classification method 2 is consistently above 0.5, but from Fig. 6, we can see that the average F value is low, suggesting that the classification is excessively biased.

Classification method 1 (2 categories)

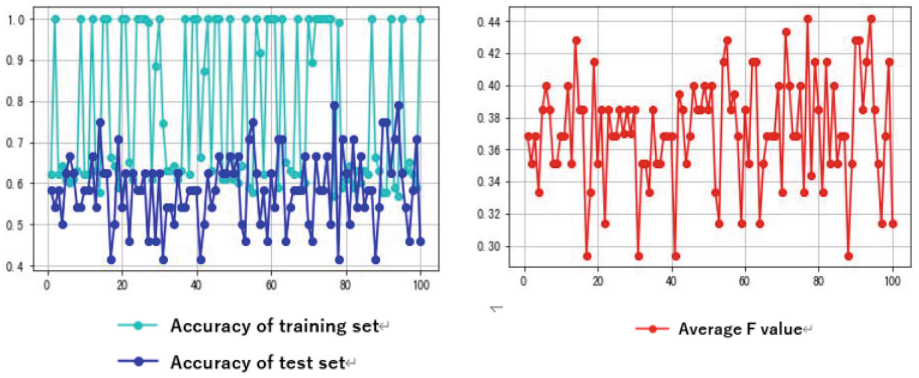

Fig. 5. (left): Accuracy rate for each data when SVM learning is performed 100 times. (right): Each average F value when learning by SVM is performed 100 times

Classification method 2 (3 categories)

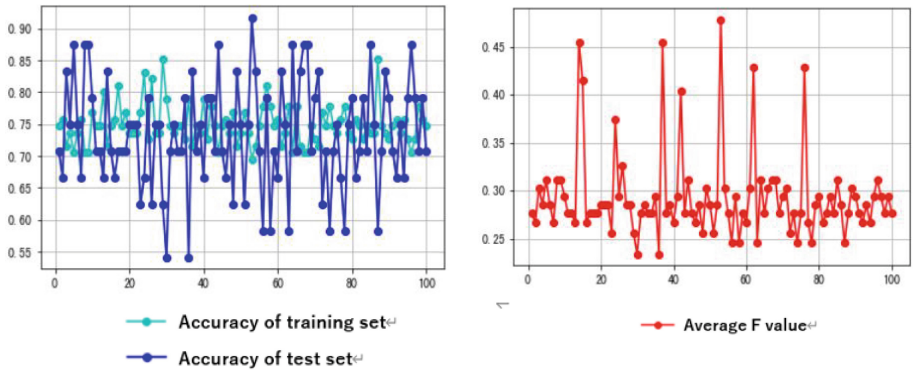

Fig. 6. (left): Accuracy rate for each data when SVM learning is performed 100 times. (right): Each average F value when learning by SVM is performed 100 times

5 Discussion

In this study, we used past Nikkei 225, Dow Jones Industrial Average, and S&P500 prices as explanatory variables to predict fluctuations in the Nikkei Stock Average. To apply this method in practice, it is necessary to specifically grasp how much each explanatory variable affects the results. Here, we used Shapley additive explanations (SHAP) as an indicator.

5.1 SHAP

SHAP is an abbreviation for Shapley additive explanations. A method of calculating how much each feature value contributes to the prediction result of a model based on the concept of the Shapley value proposed in the field of game theory. It is possible to visualize the effect of increasing or decreasing the value of the feature amount.

5.2 Shapley Value

The Shapley value was proposed in a field called cooperative game theory. Cooperative game theory considers how to distribute rewards according to each player's degree of contribution in a game in which multiple players cooperate to clear the game. At that time, the contribution of each player is obtained as a Shapley value.

5.3 Overview of Calculation of SHAP Value

By recognizing the player as a feature and the reward as a predictor, we can apply the Shapley value concept to machine learning, but there are differences in details. In

SHAP, "how much each feature value affects the predicted value" is measured by how much each feature value raises or lowers the predicted value from the average. With the Shapley value, the reward is 0 if no one plays the game. Both directional contributions are output. The problem when applying the Shapley value concept to machine learning is how to obtain the predicted value when "some features are missing". A method called KernelSHAP uses marginalization, which fixes the "present" features and takes the average of the predicted values for the "absent" features.

5.4 Implementation Results of SHAP

We used SHAP to visualize the impact of each feature value (Table 3) on stock price prediction one day later.

Table 3. Feature value

Feature0	NY Dow 2 days ago
Feature1	NY Dow 1 day ago
Feature2	SP500 2 days ago
Feature3	SP500 1 day ago
Feature4	Nikkei average 2 days ago
Feature5	Nikkei average 1 day ago

Figure 7 below shows the results of May stock price prediction using the April dataset as training set, and Fig. 7 shows the results of June stock price prediction using the May dataset as training set. The horizontal axis represents the SHAP value. In addition, "High" and "Low" written on the right side indicate that "High" indicates an increase in stock prices and "Low" indicates a decrease in stock prices. As a result of analysis using SHAP, we found that the impact of each explanatory variable on the analysis results differs depending on how the training set and test set are divided. It is a future task to make stock price prediction using SVM more concrete and convincing.

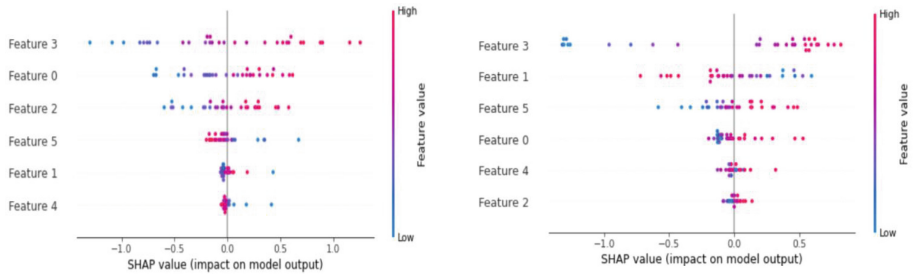

Fig. 7. (left): Result of May stock price prediction from April data set. (right): Result of June stock price prediction from May data set

6 Conclusion

In this research, based on the past rate of change of US stocks and the Nikkei Stock Average, we used SVM, which is one of supervised learning, to predict the Nikkei Stock Average one day, one week, and one month later. We verified its practicality. As a result, we confirmed that it is possible to improve the accuracy of stock price prediction for the next day compared to random prediction. In the two-class classification of "when the stock price rises" and "when the stock price falls", a high hit rate and average F-value were stably calculated. In addition, in the 3-class classification of "when stock prices go up", "when stock prices go down", and "not much change", there were variations in the average F value, but overall, the hit rate and average F value was high. On the other hand, it became clear that the proposed method is not practical for predicting stock prices one week and one month later. A high hit rate was calculated, but a low average F value was calculated. This suggested that the analysis was overly biased.

In addition, we conducted an analysis using SHAP to determine whether the explanatory variables were appropriate for the one-day stock price prediction, which had high prediction accuracy. As a result, we found that the effect of each explanatory variable on the analysis results differs depending on how the training data and evaluation data are divided. We made it a future task to make stock price predictions using SVMs more concrete and convincing.

References

1. Yusuke, I., Danushka, B., Hitoshi, I.: Using news articles of foreign exchange to predict stock prices by SVMs. SIG-FIN-012-09
2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
3. Wen, F., Xiao, J., He, Z., Gong, X.: Stock price prediction based on SSA and SVM. Procedia Comput. Sci. **31**, 625–631 (2014)
4. Tanaka, K., Nakagawa, H.: Proposal of SVM method for determining corporate ratings and validation of effectiveness by comparison with sequential logit model. Trans. Oper. Res. Soc. Jpn. **57**, 92–111 (2014)
5. Lahmiri, S.: A comparison of PNN and SVM for stock market trend prediction using economic and technical information. Int. J. Comput. Appl. **29**(3), 0975–8887 (2011)
6. Akaho, S.: Kaneru tahennryou kaiseki (Kernel multivariate analysis). Iwanami-Shotenn, Japan (2008)