Research on stock trend prediction based on grey wolf optimized RF and SVM models

Xiangyang Ge

School of Economics and Management, Hefei University, Hefei, 230601,Anhui, China

Abstract. The volatility of the stock market has a profound impact on the stability of the financial market, and then has a significant effect on the overall economic situation, so the accurate prediction of stock trend has always been the focus of research in the financial field. In this paper, the raw minute-level stock data of SSE 50 and CSI 300 in 2021 are normalized, and the historical training and test sets of daily stocks for model training and testing are established based on the datasets. The RF-SVM model based on the gray wolf optimization algorithm is constructed to compare the traditional support vector machine and random forest model, and the relevant performance indexes are calculated. In this study, the RF-SVM model based on gray wolf optimization is highly correlated with the characteristics of stock data, and the prediction accuracy of stock index data is more than 85%. The cross-validation score is over 0.77 and the overall weighted score is as high as 0.98. Compared with the SVM and random forest models under the default parameters, the RF-SVM model based on gray wolf optimization has an average advantage of 25% in performance indicators. The generalization ability is increased by about 26%; The fitting error was reduced by about 51.15%; The overall average score advantage is about 34%. These results prove that the RF-SVM model of gray wolf optimization has superior performance and significant optimization effect in stock prediction, and compared with the traditional grid search optimization algorithm, gray wolf optimization has excellent adaptability and can gradually obtain the global optimal parameter configuration.

Keywords: support vector machine; random forest; parameter optimization; Gray Wolf Optimization;

1. Introduction

In recent decades, with the creation of the self-efficient market hypothesis [1] and the random walk theory [2], and with the profound changes in the financial markets at home and abroad, many researchers have devoted themselves to building various quantitative models, including but not limited to machine learning, reinforcement learning, deep learning and deep reinforcement learning [3], so that accurately grasp the transformation of market states and capture arbitrage opportunities. Tan [4] pointed out that China's stock market does not conform to the efficient market hypothesis, and there are significant differences in the efficiency of information disclosure, regulatory efficiency, and the efficiency of individual investor behavior and trading mechanism, so the slow response of stock prices to information leads to momentum effect and long memory. Gbanador [5] argues that capital markets are considered effective if new information is quickly reflected in stock prices, such as the Nigerian stock market, which is semi-effective in the study. Therefore, many researchers are committed to building various quantitative models aiming to accurately grasp the transformation of market conditions and capture arbitrage opportunities. In terms of random forest algorithm, as a combinatorial classifier algorithm proposed by Leo Breiman [6] in 2001, it has shown strong application potential in the financial field. Pradeep [7] further elaborated on the application scenarios of various machine learning tools, which provide a solid theoretical foundation and practical guidance for the wide application of machine learning in the financial field. The market state judgment framework model constructed by Angelis [8] et al. provides strong support for capturing arbitrage opportunities through the judgment and analysis of market stability and chaos. Gray wolf optimization was proposed by Mirjalili [9] in 2014 as an emerging group optimization algorithm inspired by gray wolf hunting behavior. The algorithm provides a mathematical model for encircling and attacking prey by simulating the stages of tracking, encircling and attacking prey of gray wolves. In the financial field, the application of the gray wolf optimization algorithm helps to achieve more

efficient parameter optimization and model tuning, and provides strong support for accurate decision-making and risk management in the financial market.

2. Method Introduction

2.1. SVM（Support Vector Machine）

Support vector machine (SVM) perform a variety of tasks such as data classification, regression, and outlier detection based on the basic principles of supervised learning. It is considered to be one of the most complete algorithms in the field of machine learning, and in recent decades, SVM theory has been rapidly developed in the study of stock index and stock price change trendsThe results show that SVM can effectively capture the nonlinear dynamic characteristics of the stock market and improve the accuracy of prediction to a certain extent. The core idea of SVM is to find an optimal hyperplane in the feature space that maximizes the separation between two types of data. At this time, different categories of data can be accurately distinguished, so as to achieve data classification. It mainly looks for a plane for a given set of samples and separates the training samples in an optimal way, where y represents the class positive and negative, $D = (x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) y \in -1, +1, \omega^T x + b = 0$,the solid black line is the desired decision surface, which separates the solid and hollow points. The two dashed lines represent the decision plane closest to the two types of points. The sum of the distances from two different kinds of support vectors to the hyperplane being looked for is called the margin. The fundamental logic of support vector level is to find the partition plane with the largest interval, and to find the high-dimensional partition plane in the nonlinearly separable data set. as shown in Fig.1.

The general model of a linear support vector machine is:

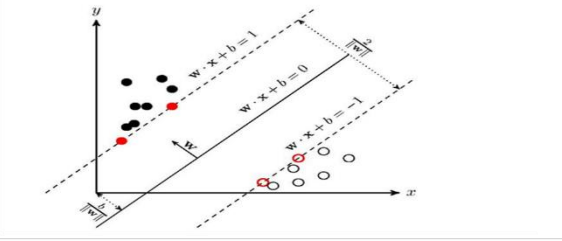

Fig 1. SVM (Support Vector Machine) model

$$max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j,$$
$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, \mathrm{u_i} \geqslant 0, C = u_i + \alpha_i, i = 1,2 \ldots m. \tag{2.1}$$

where the coefficient of the corresponding sample is denoted α_i and the label (1 or -1) of the corresponding sample is denoted y_i and the kernel function is denoted $K(x, y)$and C is the penalty factor.

2.2. RF(Random Forest)

Random forests are a decision tree-based statistical learning method that works by integrating a large number of independently generated decision trees. These decision trees are trained by a self-service sampling technique from a random subset of raw data sets, each based on a different subset of data,as shown in Fig. 2.

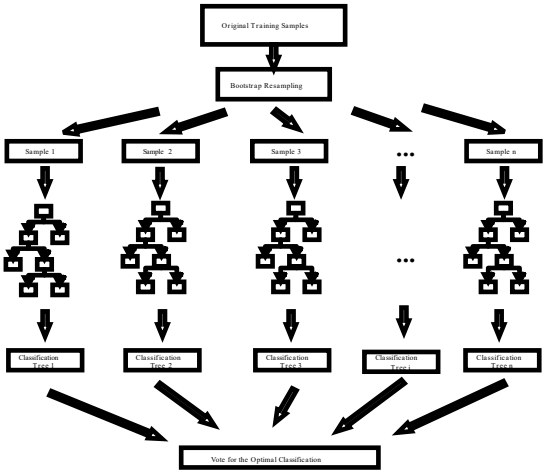

Fig 2. RF (random forest) model

2.3. GWO (Gray Wolf Optimized)

The gray wolf optimization was proposed by Mirjalili and other scholars at Griffith University in Australia in 2014. It draws inspiration from the predatory behavior of gray wolf groups in nature, and cleverly incorporates the gray wolf's leadership hierarchy and hunting mechanics into the algorithm. In the GWO algorithm, the gray wolf group is carefully divided into four roles: alpha (α), beta (β), delta (δ), and omega (ω), which symbolize the leader, sub-leader, third leader, and members of the gray wolf pack, respectively, as shown in Fig.3.

The alpha wolf plays the role of decision-maker, leading key activities such as hunting and resting; Beta Wolf acts as an assistant to assist Alpha Wolf in implementing decisions; Delta wolves, although of a slightly lower status, are responsible for instructing Omega wolves and following the instructions of Alpha and Beta wolves; The Omega Wolf is responsible for carrying out the decisions of the three levels of gray wolves mentioned above, ensuring harmony and coordination within the pack.

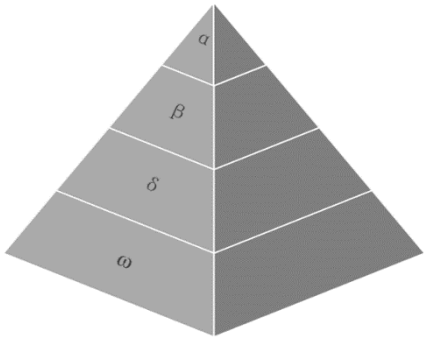

Fig 3. Gray Wolf Optimization Hierarchy Diagram

3. Empirical research

3.1. Data Processing

Preprocess the data before training the model. First of all, choose the SSE 50 Index (000016) and the CSI 300 Index (000300), the SSE 50 Index and the CSI 300 Index are two very important indices in the Chinese stock market, which can not only reflect the overall trend of the stock market but also understand the hot spots and industry performance of the market, so they have important reference value. The data comes from Oriental Fortune Network, which obtains all its minute-level data in 2021, compares its closing price of the day with the closing price of the next day, if it is positive, it is down, and if it is negative, it is up, and this is used as the rise and fall index of the next day, 1 is up, 0 is

down. The data is then cleaned and organized to remove outliers and missing values. Usually from 9:30 a.m. to 11:30 a.m. to 13:00 p.m. to 15:00 p.m. on the same day is the effective trading range, ignoring unstable factors such as extending the closing time, these unexpected data are cut out to ensure the accuracy of analysis. Then we need to normalize the data, so that the numerical range of different features is the same, and the minute-level data is integrated into the daily data for easy calculation and graphing, and the low-frequency data can reflect the overall trend, and the normalization parameter of the data is Micro average. To ensure an accurate assessment of the model's performance, we divided the data. Specifically, we use the top 80% of the dataset as the training set for model learning and fitting. The last 20% of the data is used as a test set to independently evaluate the generalization ability of the model. This partitioning ensures consistency in the data distribution between the training and test sets, thus avoiding bias introduced by improper data partitioning. In this way, we get a standard dataset division (x_train, y_train, x_test, y_test), where x_train and y_train represent the eigenvalues and target values used for training, respectively, while the x_test and y_test represent the eigenvalues and target values used in the actual evaluation, respectively. This division not only helps us to fairly compare the performance of different models, but also provides strong data support for model optimization and tuning. Throughout the experiment, we will strictly follow this data division standard to ensure the reliability and validity of the experimental results. Through such a setting, we will be able to more accurately evaluate the practical application effect of the gray wolf optimization algorithm in the RF-SVM model, and provide a useful reference for the subsequent model optimization and application.

3.2. Model training

In order to more intuitively demonstrate the performance improvement of the optimized model, a set of controlled experiments is carefully designed. The control group was set up according to the principle of control variables, that is, under the same conditions, the default parameters of Support Vector Machine (SVM) and Random Forest (RF) were used for training, and compared with the RF-SVM model based on Gray Wolf Optimization Algorithm (GWO). Eliminate other distractions and focus on evaluating the impact of the gray wolf optimization algorithm on model performance. With this controlled test setup, the differences in the performance of the models on the same dataset before and after optimization were more clearly observed. This can not only provide strong evidence for the improvement of the model, but also provide a useful reference for subsequent model selection and adjustment.

In order to ensure fairness and consistency in the experimental setup of the control group, the widely recognized Radial Basis Function (RBF) was selected as the kernel of the SVM. For RBF kernel functions, set the penalty factor C to 1 and the gamma parameter to 0.1, which are also common choices in the default configuration of the SVM. A similar logic is followed in the parameter setting of the Random Forest (RF), setting the maximum depth (max_depth) to 5 and the number of forest branches (n_estimators), i.e. the number of decision trees, to 20, which are also common configurations for random forests. For the experimental group GWO-RF-SVM, we optimize the model parameters with GWO. In the gray wolf optimization algorithm, the number of wolves N is set to 10, and the accuracy of the random position model parameters is calculated as fitness, which means that the algorithm will search for 10 potential optimal solutions in the solution space at the same time. The maximum number of iterations is set to 100, which ensures that the algorithm has enough time to find the global optimal solution. In the process of parameter optimization, the parameters that can be optimized include C and gamma of the support vector machine, and the max_depth and n_estimators of the random forest. Finally, by comparing the experimental results of the control group and the experimental group, the actual effect of the gray wolf optimization algorithm in improving the performance of the RF-SVM model can be clearly demonstrated.

After the model is constructed, the FIT function is used to fit the x_train and y_train datasets to ensure that the model can fully learn the internal rules and characteristics of the training data. The prediction value (y_pred) was then obtained using x_test as the baseline input for the model prediction.

In order to visually demonstrate the prediction effect, the predicted value was visually compared with the real value (y_test), and the image shown in Fig.4 was plotted.

In Fig.4, the left side shows the trend forecast for the SSE 50 Index, while the right side shows the trend forecast for the CSI 300 Index. The Y-axis label in the chart clearly indicates both the upward and downward states, making it clear that the trend changes at a glance. In the data results of this experiment, node graphs of different colors and shapes were used to distinguish the performance of each model. The red diamond-shaped line represents the real market movement and serves as a benchmark to evaluate the accuracy of the model. The blue squares represent the prediction results of the RF-SVM model optimized by the Gray Wolf Optimization Algorithm (GWO), which shows the excellent performance of the model in fitting and prediction. In addition, the green upper triangle and yellow lower triangle represent the benchmark results of the SVM and RF models under the default parameters, respectively, and provide a reference for the control group. Through this visual display, the performance differences of different models in stock index trend prediction can be intuitively compared, and the effectiveness of the gray wolf optimization algorithm in improving the prediction accuracy of the RF-SVM model can be further verified.

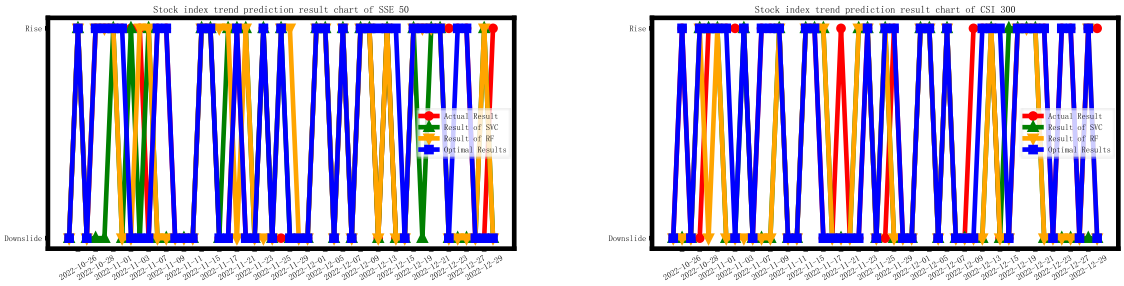

Fig 4. Trend forecast chart of each model stock index

In the default SVM and random forest models, the AUC areas are 0.64 and 0.76 for the SSE 50 Index, and 0.76 and 0.60 for the CSI 300 Index, respectively. However, when we use the RF-SVM model optimized by the Gray Wolf Optimization Algorithm (GWO) and add likelihood estimation to calculate the prediction probability, the AUC value increases significantly, reaching 0.91 and 0.92, respectively. This shows that the GWO-RF-SVM model has the most accurate performance in predicting the SSE 50 Index and the CSI 300 Index, and has achieved the expected results.

3.3. Model evaluation

In order to evaluate the performance of the model more comprehensively, we summarized the experimental data, calculated the performance indicators such as accuracy, precision, recall, and F1 score, and evaluated the validation error, training error, and cross-validation score to further verify the generalization ability and stability of the model. These aggregated data will be presented in Table 1 to provide a comprehensive basis for evaluating the performance of the model. It is worth noting that a custom weighted calculation method is designed here, and the total score is weighted for accuracy, precision, recall, F1 score, cross-validation, validation error, and training error. The main reasons for choosing such a weight allocation are as follows: First, accuracy is the most basic and important evaluation index in the classification problem, which directly reflects the proportion of samples correctly classified by the model. Therefore, we have assigned it the highest weight of 0.3 to highlight its importance. Secondly, precision, recall and F1 score are supplementary indicators to evaluate the performance of classifiers. Precision measures the proportion of samples that the model predicts to be positive that are truly positive, and recall measures the proportion of all true positives that are correctly predicted by the model. The F1 score is a blended average of precision and recall, taking into account the performance of both. To balance the importance of these three indicators, they are each assigned a weight of 0.2. Furthermore, cross-validation is an effective way to evaluate the generalization ability of a model, and through multiple training and validation, the performance of the model on unseen data can be estimated. Therefore, it is assigned the same weight of 0.3 as the

accuracy. Finally, the validation error and training error reflect the performance of the model on the validation set and the training set, respectively. While these two metrics are important in the model tuning process, they focus more on how well the model fits than on the ability to generalize. Therefore, they are each assigned a lower weight of -0.1, which means that these two indicators will act as a certain penalty when calculating the total score to prevent the model from overfitting. With this custom weighted calculation, we can evaluate the performance of the model more holistically, while emphasizing the importance of accuracy, precision, recall, F1 score, and cross-validation, as well as appropriate penalties for validation errors and training errors. This weight allocation helps to understand the performance of the model in different aspects more clearly, and provides strong guidance for model optimization.

Table 1. Model data summary

	Prediction results of the SSE 50			Prediction results of the CSI 300		
	RF-SVM	SVM	RF	RF-SVM	SVM	RF
Accuracy	0.8571	0.7143	0.6735	0.8571	0.6327	0.6531
Precision	0.8583	0.7167	0.6667	0.8586	0.8182	0.8333
Recall	0.8571	0.4583	0.6667	0.8571	0.3600	0.4000
F1 score	0.8574	0.6111	0.6667	0.8575	0.5000	0.5405
Cross validation score	0.7768	0.7143	0.5306	0.7718	0.5714	0.6122
Training error	0.1429	0.2252	0.1126	0.1429	0.2789	0.0950
Test error	0.1347	0.2645	0.4343	0.1244	0.3760	0.4429
Weighted Score	0.9850	0.7665	0.7202	0.9852	0.6242	0.6734

From the tabular data provided in this paper, it is clear that after the evaluation of the custom weighted scoring model designed by us, the RF-SVM model adjusted by the gray wolf optimization algorithm shows significant advantages, whether it is for the SSE 50 index or the CSI 300 index. Compared with the Support Vector Machine (SVM) model and the Random Forest (RF) model under the default parameters, the performance of the RF-SVM model optimized by Gray Wolf has been significantly improved. This result not only verifies the effectiveness of the gray wolf optimization algorithm in parameter tuning, but also further highlights the potential of the RF-SVM model in financial market forecasting. This finding is of great significance for improving the accuracy and reliability of financial forecasting models, and provides a more accurate and scientific basis for investors in the decision-making process.

4. Conclusion

Machine learning-based stock trend forecasting has become a high-profile area of research with great promise and potential. In this study, the RF-SVM model is tuned by introducing the gray wolf optimization algorithm, and a training model suitable for different stock index datasets is successfully constructed. Experimental results show that the model shows excellent performance in stock trend prediction, and its accuracy is far better than that of machine learning models using default parameters. In particular, on the SSE 50 and CSI 300 indices, the model achieves an average advantage of about 24.1% and 33.6%, respectively, through the non-average weighting method.

1. Performance indicators: The accuracy, precision, recall, and F1 score performance indicators in machine learning were used to evaluate the model, and the average index data of the three models were 0.8575, 0.6324, and 0.6376 from Table 1. Therefore, the RF-SVM model optimized by gray wolf is about 26% better than the support vector machine, about 25% better than the random forest, and the average advantage is about 25.5%, which means that the optimized cross model is more accurate and efficient in terms of model accuracy.

2. According to Table 1, the average 5-fold average cross-validation score of the RF-SVM model in the SSE 50 and CSI 300 datasets is about 0.774, while the support vector machine and random forest are about 0.664 and 0.571, respectively, with an individual advantage of 16.5% and 35.5%, and

an average advantage of about 26%. The generalization ability represents the performance of the model on the unfitted training set, and the generalization error of the RF-SVM model optimized by gray wolf is relatively small.

3. Error shrinkage: Finally, the error comparison, these two indicators play a penalty role to prevent the model from overfitting, the RF-SVM model optimized by gray wolf has an average error reduction of 0.1425 compared with the original model, and the overall error has decreased by 51.15%. This achievement is quite significant in this study, and the bottoming error indicates that the model fits well, and there is neither underfitting that leads to a decrease in accuracy, nor overfitting that causes the model to be overly complex.

In summary, the RF-SVM model optimized by gray wolf performs the best in this stock index trend prediction, with an average advantage of about 34% in the three model classification studies, which has high potential and can be further improved. In addition, optimization algorithms and cross-models can be improved to improve performance.

Reference

[1] KINGSTONE N, YUDHVIR S. Alternatives to the efficient market hypothesis: an overview [J]. Journal of Capital Markets Studies, 2023, 7(2): 111-124.

[2] TOMáŠ M, MARTINA K, LUDVíK F, et al. Backtesting the evaluation of Value-at-Risk methods for exchange rates [J]. Studies in Economics and Finance, 2023, 40(1): 175-191.

[3] KUMAR I P, BALAKESAVAREDDY P, KUMAR S A. Stock price prediction methodology using random forest algorithm and support vector machine [J]. Materials Today: Proceedings, 2022, 56(P4): 1776-1782.

[4] ZHENGXUN T, YAO F, HONG C, et al. Stock prices' long memory in China and the United States [J]. International Journal of Emerging Markets, 2022, 17(5): 1292-1314.

[5] GBANADOR M A. An Empirical Test for Semi-strong form Efficient Market Hypothesis of the Nigeria Stock Market [J]. Asian Journal of Economics, Business and Accounting, 2021: 42-53.

[6] BREIMAN L. Randomizing Outputs to Increase Prediction Accuracy [J]. Machine Learning, 2000, 40(3): 229-242.

[7] PRADEEP S. Fundamentals and Methods of Machine and Deep Learning:Algorithms, Tools and Applications [M]. John Wiley & Sons, Inc.

[8] ANGELIS L D, PAAS L J. A dynamic analysis of stock markets using a hidden Markov model [J]. Journal of Applied Statistics, 2013, 40(8): 1682-1700.

[9] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey Wolf Optimizer [J]. Advances in Engineering Software, 2014, 69: 46-61.