Research on Stock Price Prediction Model Based on Sentiment Factor and Multi-Core Bagging Algorithm

Ziqian Wang

School of Science, Tianjin University of Commerce, Tianjin, China, 300134

13363126817@163.com

Abstract. Stock price prediction model is one of the key research points in the quantitative financial field. With the rapid development of data acquisition and storage technology, the data of emotional factors based on nonlinear structure is increasing. How to combine this part of information with common European technology factors to increase the ability of the model is a problem that needs to be solved. Based on this, this paper proposes a method of uncertainty quantification that combines European-and non-European data. Specifically, on the one hand, the Gaussian process regression model is used to capture the unknown relationship between predictive variables and price; on the other hand, the multi-kernel learning technology based on European and non- European kernel functions is used to combine the effective information of European and non- European predictive variables. Moreover, the proposed method can weigh the bias of the prediction model against the variance by using the Bagging algorithm. The simulated data analysis shows that the proposed method not only improves the prediction performance of the European Gaussian process, but also works in the presence of irrelevant predictive variables. Finally, the actual data analysis shows that the proposed method has achieved good prediction results in the task of predicting the rise and fall of stock prices.

Keywords: Emotional Factor, Stock Price Forecast, Gaussian Process, Bagging Algorithm.

1. Introduction

Financial markets, especially the stock market, are highly uncertain and complex. Stock prices are influenced by a variety of factors, including macroeconomic indicators, corporate financial position, industry dynamics, policy changes, and investor sentiment. These factors are intertwined, making stock price fluctuations difficult to predict accurately. Some common prediction methods are using for example linear regression [1], time series analysis (such as ARIMA) [2-5] linear model to establish the link between stock prices and related factors. However, using scenarios is limitations because linear models can only capture linear relationships. In recent years, many scholars have begun to use neural network models with non-linear fitting ability to predict stock prices, and have achieved higher prediction accuracy compared with linear models. For example, Hu Yuwen et al. (2021) proposed an LSTM model combined with PCA and LASSO for dimensionality reduction and feature selection, which improved prediction accuracy [6]. Cao Chaofan et al. (2022) developed the MDT-CNN-LSTM model, utilizing multidirectional delay embedding to enhance prediction precision and timeliness [7]. Liu Yumin et al. (2021) used a method combining RF with SFS for feature extraction, followed by an LSTM model to improve the accuracy of stock price trend prediction [8]. Yu Cilong et al. (2021) utilized BERT to build a financial sentiment prediction system (BERT-FS) that predicts stock price movements from news articles [9]. Kang Ruixue et al. (2023) proposed an LSTM stock price prediction model with self-attention mechanism and multi-source data inputs to improve accuracy [10]; Deng Feiyan et al. (2021) studied the performance differences of LSTM neural networks in predicting short-term price trends using data at different temporal granularities, such as weekly and daily data [11]. These methods provide a powerful tool for stock price forecasting and help investors to make more accurate investment decisions.

In recent years, with the rapid development of data acquisition and storage technologies, the number of some emotional factors based on text data is also increasing rapidly. How to integrate these emotion-like factors into the prediction model to further improve the prediction ability of the model is an important research issue. Some scholars have tried this regard. For example, Guan Jian uses

news features to extract news features, which enriches the input features of the stock prediction model [12]. Wu Bin used a logistic regression model to classify the emotions of stock comments and calculate the daily overall emotion index, as an input feature of stock prediction, which improved the accuracy of the prediction model [13].

In addition, in stock price quantitative trading, the uncertain quantitative output of the forecast model may be more meaningful than in obtaining the forecast value of a stock price. Specifically, it uses Gaussian Process Regression (GPR) to fit nonlinear relationships and quantify uncertainty. Multi-kernel learning characterizes both European and non-European data, ensuring the model's flexibility and comprehensiveness. To address high-dimensional predictors, the Bagging algorithm is employed to reduce the impact of the "curse of dimensionality". This approach improves the model's robustness and reliability, as demonstrated through simulation experiments and real data analysis. Thus, this study contributes to the theoretical understanding of financial market dynamics and offers innovative methods for financial data analysis and prediction.

2. Theory and method

2.1. Gaussian Process regression

Gaussian process (GP) regression is based on the Bayesian theory and statistical learning theory developed a new machine learning method is suitable for high dimensional, small samples and nonlinear complex regression problems, and strong generalization ability and neural network, support vector machine is easy to implement, super parametric adaptive acquisition, non-parametric inference is flexible and output has the advantages of probability significance [14]. The core idea of Gaussian process regression is to treat the entire function space as random variables, where any finite set of functions follows a joint normal distribution. Gaussian processes are able to handle high-dimensional data, small sample sets, as well as complex nonlinear relationships, and can give confidence intervals for predictions, which have unique advantages in dealing with uncertainty and risk assessment.

In the GP regression, this paper first define an a priori GP that consists of a mean function $m(x)$ and a covariance function $k(x, x')$. For any two input points $x$and x', their output $f(x)$ and $f(x')$ will follow a joint normal distribution with the mean and covariance determined by the prior GP approach

$$f(x) \sim GP\big(m(x), k(x, x')\big) \tag{1}$$

When the training data $D = \{(x_i, y_i)\}_{i=1}^{n}$ are obtained, this paper can calculate the posterior GP using the Bayes' rule, that is, the GP after considering the data information. The mean and covariance functions of this posterior GP will be used to make predictions for the new input points x_*. The mean and the variance of the predictions are:

$$\mu_*(x_*) = k(x_*, X)[k(X, X) + \sigma_n^2 I]^{-1} y \tag{2}$$

$$\sigma_*^2(x_*) = k(x_*, x_*) - k(x_*, X)[k(X, X) + \sigma_n^2 I]^{-1} k(X, x_*) \tag{3}$$

Predicted mean $\mu_*(x_*)$, provides the best estimate of $f(x_*)$, while the predicted variance $\sigma_*^2(x_*)$ describes the predicted uncertainty. The predictive distribution is a normal distribution who's mean and variance can provide central estimates and uncertainty measures of future price.

X is the matrix of all training input points, y is the vector of all training output points, $k(X, X)$ is the covariance matrix between the training input points, $k(x_*, X)$ is the covariance vector between the test points and all the training input points, $k(x_*, x_*)$ is the covariance of the test point itself, σ_n^2 is the variance of the observation noise, and I is the unit matrix.

2.2. A Gaussian process integration model based on multi-kernel learning and the Bagging algorithm

The Gaussian process integration model based on multi-kernel learning and the Bagging algorithm structure is shown in Figure 1.

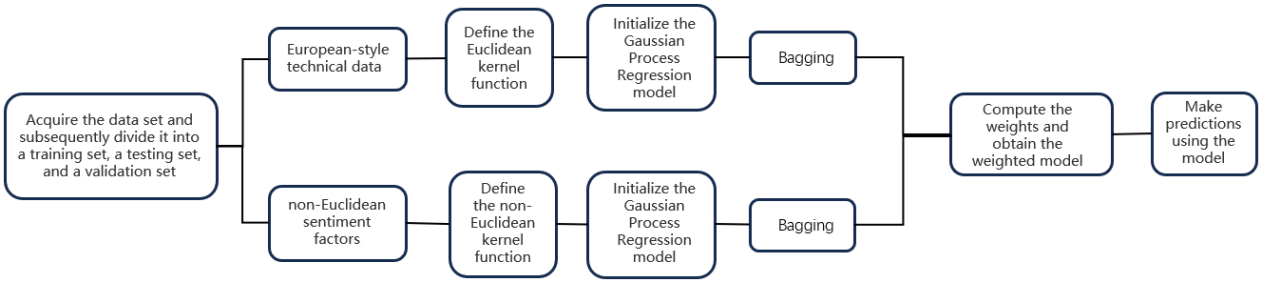

Figure 1. The integration model structure

First, since the Gaussian process supports custom kernel functions, this paper define two kernel functions for different subsets of the training data. For European technical data, this paper use a composite core consisting of constant kernel, radial basis function (RBF) core and white noise core.

$$K_E = \mathrm{C} \times \mathrm{RBF} + K_W \tag{4}$$

For the non- European data subset, this paper choose the cosine similarity kernel, a kernel function suitable for data characterizing the angular similarity between observed values.

The cosine similarity kernel function is defined as follows:

$$sim(X,Y) = \frac{X \cdot Y^T}{\|X\| \, \|Y\|^T} \tag{5}$$

Where $\|X\|$ and $\|Y\|$ are the norms of X and Y, respectively, i.e.,

$$\|X\| = \sqrt{\sum_{i=1}^{D} X_i^2} \ , \|Y\| = \sqrt{\sum_{i=1}^{D} Y_i^2} \tag{6}$$

Calculating the Kernel Matrix $\boldsymbol{K}$

$$K(X,Y) = \exp\big(\theta \cdot sim(X,Y)\big) \tag{7}$$

Calculating Gradient $\boldsymbol{K_{gradient}}$:

$$\boldsymbol{K_{gradient}} = sim(X,Y) \cdot K(X,Y) \cdot 1_{N \times M \times 1} \tag{8}$$

Where X is an $N \times D$ matrix, where N is the number of samples and D is the number of features. Y is an $M \times D$ matrix, where M is the number of samples in another set. $1_{N \times M \times 1}$ is an $N \times M \times 1$ tensor with all elements equal to 1. Θ (param) is a scalar parameter controlling the scaling of the cosine similarity. K is the kernel matrix, of shape $N \times M$. $\boldsymbol{K_{gradient}}$ is the gradient matrix of the kernel function with respect to the parameter θ, of shape $N \times M \times 1$.

Next, two Gaussian process regression models (GP1 and GP2) were initialized based on the above kernel functions. Subsequently, each Gaussian process regression model was put within the Bagging and obtained $\hat{y}$ by fitting of the Bagging. This ensemble learning strategy can remove the effects of irrelevant variables and reduce the variance of the GP model, thereby improving the accuracy of the prediction and the robustness of the model.

Draw a training set containing k target variables y, a test set of size m, and a validation set of size n. Train both Euclidean and non-Euclidean Bagging models on their respective training sets, $\mathrm{E}_1, \mathrm{E}_2 \ldots \mathrm{E}_\mathrm{k}$ for the Euclidean data and $\mathrm{N}_1, \mathrm{N}_2 \ldots \mathrm{N}_\mathrm{k}$ for the non-Euclidean data, yielding k models $\mathrm{M}_1, \mathrm{M}_2 \ldots \mathrm{M}_\mathrm{k}$. Apply the validation set to the ensemble model $\mathrm{M}_1, \mathrm{M}_2 \ldots \mathrm{M}_\mathrm{k}$, obtaining predictions $\hat{\mathrm{y}}_1, \hat{\mathrm{y}}_2 \ldots \hat{\mathrm{y}}_\mathrm{k}$. Through these predictions, this paper not only obtain the prediction values of each sub-model but also the variances $\varepsilon_1, \varepsilon_2 \ldots \varepsilon_\mathrm{k}$ associated with these prediction values.

Subsequently, to determine the contribution of each submodel in the final ensemble prediction, this paper adopted a weighting strategy: first, record the variance of each submodel on the validation set. Then, according to predict the inverse of the variance of each submodel, in order to more accurately use the collected data, this paper first normalized it, and then took the normalized data values as the weight ω_i of each submodel. Among:

$$\omega_i = \frac{\frac{1}{\varepsilon_i}}{\sum_{i=1}^{k} \frac{1}{\varepsilon_i}} \tag{9}$$

This means that models that predict more stable will receive higher weights. The final predicted value $\hat{y}$ is:

$$\hat{y} = \sum_{i=1}^{k} \omega_i \hat{y}_i \tag{10}$$

Through the weighted average method, the prediction results of each submodels are integrated to form the final prediction output of the integrated model. This process effectively integrates the advantages of multiple submodels and improves the accuracy and reliability of the overall prediction.

Finally, input the test set $E_1 \ldots E_n, N_1 \ldots N_n$ into the ensemble model $M_1, M_2 \ldots M_k$ to predict the new feature space $\hat{y}_1, \hat{y}_2 \ldots \hat{y}_k$. Utilize these predictions through Equation (10) to generate the final predicted values $\hat{y}$ on the test set.

3. Result

3.1. Data collection and preprocessing

From Oriental wealth network obtained A shares from 2017 to 2022 monthly stock closing price data, and obtained the relative strength index (RSI), smooth average (MACD), rising direction line (PDI), trend index (DMI), energy tide (OBV), the index (CCI) European technical indicators, and closed fund average discount (DCEF), average initial public offering yield (RIPO), new accounts (NA), market turnover (TURN) last month, consumer confidence index (CCI) non-European sentiment factor data. This paper used the pandas library in Python for data processing by scaling the stock close divided by 1000 so that it distribution distributed between 1 and 10.

To simulate the influence of European and non-European features in stock price prediction, this paper constructed a set of mixed feature data. Specifically, this paper first generate three European eigenvectors, and at the same time, considering the complexity and nonlinearity of financial markets, this paper introduce a 256-dimensional feature word vector to simulate sentiment factors in the non-European feature space.

To incorporate these two classes of features into a unified prediction framework, this paper perform a simple mathematical transformation of the European features to simulate the correlation of the eigenvectors with the stock price. Furthermore, this paper calculated the cosine similarity between non- European features as a measure of their interrelations. Finally, this paper add a random vector designed to simulate the stochastic volatility factors in the market, resulting in the final sequence of simulated stock prices.

Subsequently, this paper set optimizers corresponding to European and non-European features in the prediction model to explore the impact of different optimization strategies on model performance. Through a series of experiments, this paper evaluated the effect of different numbers of optimizers on the predictive accuracy of the model, and adjusted the number of European-and non-European model optimizers to observe the effect of the model prediction.

The experimental results show that when the number of optimizers for both the European and non-European models is set to 10, the prediction effect of the model reaches the best state. Based on our experiments, this paper determined that using 10 optimizers achieves the best performance for the stock price prediction model when handling both European and non-European features. Considering the diversity of stock predictions, this paper try two different data segmentation strategies——

external expansion and interpolation, to construct the training set and test set. In order to deeply explore the performance of the model under different conditions, this paper designed three prediction scenarios to test the sensitivity of the model to the number of features and its generalization ability:

Case 1: Using the Complete Feature Set

Define the feature set as $E = \{E_1, E_2 \ldots E_6\}$ and $N = \{N_1, N_2 \ldots N_6\}$, The model is trained and tested using the complete feature set.

Case 2: Reducing the Number of Feature Vectors

Extract a subset of feature vectors $E' = \{E_1, E_2 \ldots E_i\}$ and $N' = \{N_1, N_2 \ldots N_i\}$, where $i < 6$ and $E' \in E, N' \in N$, The model is trained and tested using a subset of feature vectors.

Case 3: Reducing the Number of Feature Vectors Only in the Training Set

For the training dataset, use a feature subset $E'' = \{E_1, E_2 \ldots E_j\}$ and $N'' = \{N_1, N_2 \ldots N_j\}$, where $j < 6$ and $E'' \in E$, $N'' \in N$,Train the model using a subset of feature vectors (E'' and N'') and test it on the complete feature set $E = \{E_1, E_2 \ldots E_6\}$ and $N = \{N_1, N_2 \ldots N_6\}$.

To assess the adaptation and generalization of the model when the features are limited during the training phase. In order to ensure the robustness of the model results, this paper compare it with the traditional Gaussian process prediction model: the European and non-European data are directly stitched together as the independent variable of our model ($X = \{E_1 \ldots E_6, N_1 \ldots N_6\}$, $X' = \{E_1 \ldots E_i, N_1 \ldots N_i\}$ and $X = \{E_1 \ldots E_j, N_1 \ldots N_j\}$) and the substitution kernel function is

$$K_T = C \times BF \tag{11}$$

Fit predictions in the Gaussian regression model.

In addition, multiple repeated experiments were performed for each prediction scenario, with 30,50 and 100 repeated experiments. The results of each experiment, that is, Mean Squared Error (MSE) between the predicted value and the actual stock price, were compared with the traditional Gaussian process prediction model. It can not only quantify the prediction accuracy of the model under different feature configurations, but also can deeply understand the stability and reliability of the model.

3.2. Comparison of the results

This paper randomly sampled data from the training, test and test sets from the real dataset for analysis. Moreover, considering that stock data is also related to time, for time series data, this paper should consider the influence of time. In order to retain the original time order, this paper adopt the interpolated training set and test set division method. Earlier data were used for training and later data for testing to reflect the model performance when processing future data.

This paper first consider case 1, keeping the training set and prediction set unchanged, using the external and interpolated data segmentation methods, respectively, divide the training set and test set, and make predictions on the real data.

From Table 1, this paper can obtain the values and variances of the two prediction methods.

Table 1. Case 1 comparative analysis of integrated model and Gaussian process

Type of experiment	Extended MSE	External expansion SD	Plug-in MSE	Built-in SD
30 times of the ensemble model	0.0362	0.0299	0.2446	0.0885
30 times in the GP regression model	0.5153	1.7169	3.3447	1.7352
50 times of the ensemble model	0.0393	0.0297	0.2712	0.0944
50 GP regression models	0.3790	1.3502	3.2180	1.3441
And 100 times of the ensemble model	0.0433	0.0353	0.2383	0.0971
100 GP regression models	0.3041	1.0627	3.0279	0.0000

The results in Table 2 were obtained under the condition where the feature vectors of the two training sets and the prediction sets are also used to make predictions on the real data.

Table 2. Case 2 Comparative analysis of integrated model and Gaussian process

Type of experiment	Extended MSE	External expansion SD	Plug-in MSE	Built-in SD
30 times of the ensemble model	0.0358	0.0265	0.2516	0.0796
30 times in the GP regression model	0.9824	2.4912	2.9911	1.8020
50 times of the ensemble model	0.0422	0.0420	0.2642	0.0732
50 GP regression models	0.3223	0.1831	2.6621	0.0000
And 100 times of the ensemble model	0.0420	0.0342	0.2781	0.0833
100 GP regression models	0.4097	0.2438	2.7608	0.9870

The results in Table 3 were obtained under the condition where only the feature vectors of the two training sets were removed, while keeping the test set features unchanged.

Table 3. Comparative analysis of three integrated model and Gaussian process

Type of experiment	Extended MSE	External expansion SD	Plug-in MSE	Built-in SD
30 times of the ensemble model	0.0380	0.0362	0.1178	9.5402
30 times in the GP regression model	0.3237	0.2737	2.8712	1.7909
50 times of the ensemble model	0.0311	0.0271	0.1334	0.0873
50 GP regression models	0.5754	1.4402	2.8712	0.0000
And 100 times of the ensemble model	0.0301	0.0233	0.1136	0.0962
100 GP regression models	0.5421	1.3322	3.2577	1.9027

Through comparison, it is found that the prediction accuracy of our model is better than that of the traditional Gaussian process, and our model is more stable.

It can be seen from the table that the proposed method shows significant performance advantages compared with the traditional Gaussian process. The main reasons are as follows: first, this paper effectively integrate the market sentiment factors; second, this paper adopt the non- European cosine similarity function to show the interaction between these factors, and reduce the deviation and variance of the prediction results through the multi-core Bagging algorithm. In contrast, traditional Gaussian processes ignore the impact of market sentiment on stock price fluctuations. By comparing Table 1, Table 2, and Table 3 horizontally, this paper can also find that changes in the number of feature factors do not significantly affect the model's predictions, indicating that our model has robustness. Moreover, when dealing with the emotion factors in the non- European feature space, this paper adopted a cosine similarity function to measure the correlation between different factors. This complex link between non-European metric performance sentiment factors and stock prices further enhances the sensitivity and prediction ability of the model to market sentiment fluctuations. In summary, the proposed method not only improves the stability of the model and reduces the risk of overfitting, but also increases its robustness to noise and outliers.

4. Conclusion

This paper proposes a stock price model based on sentiment factor and multi-core Bagging algorithm, in order to improve the accuracy of financial market forecast by integrating the advantages of European and non-European data. Gaussian Process regression was used to explore the link between the characteristic factors and stock prices. To further improve the model accuracy, so that the stock prediction effect is not affected by irrelevant factors, this paper introduced the Bagging

algorithm. Through experimental verification, our method makes substantial progress in prediction accuracy, and maintains good adaptability and robustness.

For future research directions, on the one hand, consider using the Bayesian Model Averaging (BMA) method to weight each submodel. For example, BMA is a Bayesian statistical method, which assigns weights according to the posterior probability of the model, and the model with higher posterior probability will get greater weight. On the other hand, try to apply the model of stock value prediction to other fields, such as commodity futures market, because commodity prices are affected by various factors such as supply and demand, weather, policy and, similar to the stock market, the futures market also needs to predict the price trend. Stock forecasting models can be adjusted to handle these variables, predicting fluctuations in commodity prices. In the foreign exchange market, stock forecasting models can be used to analyze the impact of economic indicators, political events and monetary policy on exchange rates. By applying this model to multiple scenarios, the universality of this method can be tested.

References

[1] Ma Jingjing. Regression and classification study in stock price trend prediction [J]. Computer Knowledge and Technology, 2024, 20 (12): 12-14+23.

[2] Han Ying, Zhang Dong, Sun Kaiqiang, et al. New model of stock prediction combining long and short-time memory networks and width learning [J]. Operations Research and Management, 2023, 32 (08): 187-192.

[3] Guo Gaiwen & Wang Shihan. Stock price forecast based on grey theory and ARIMA model [J]. Journal of Henan Institute of Education (Natural Science Edition), 2023, 32 (02): 22-27.

[4] Huang Xingdan. Master's thesis on stock price index forecast research based on time series and deep learning model [D]. Shandong Industry and Business School, 2023.

[5] Huang Mengting. Empirical study on stock price prediction based on ARIMA model [J]. Neijiang Technology, 2023, 44 (03): 61-62.

[6] Hu Yuwen. Stock prediction based on optimized LSTM model [J]. Computer Science, 2021, 48 (S1): 151-157.

[7] Cao Chaofan, Luo Zenan, Xie Jiaxin, et al. Research on stock price prediction using the MDT-CNN-LSTM model [J]. Computer Engineering and Applications, 2022, 58 (05): 280-286.

[8] Liu Yumin, Li Yang, & Zhao Zheyun. Component stock price trend prediction based on feature selection RF-LSTM model [J]. Statistics and Decision, 2021, 37 (01): 157-160.

[9] Yu Cilong, Shi Zhenyu, Xie Yunhao, et al. Sentiment analysis and stock price movement prediction system based on natural language processing [J]. Systems Engineering, 2021, 39 (05): 114-123.

[10] Kang Ruixue, Niu Baoning, Li Xian, et al. LSTM stock price prediction with self-attention mechanism and multi-source data inputs [J]. Journal of Chinese Mini-Micro Computer Systems, 2023, 44 (02): 326-333.

[11] Deng Feiyan, Cen Shaoqi, Zhong Fengqi, et al. Short-term price trend prediction based on LSTM neural network [J]. Computer Systems & Applications, 2021, 30 (04): 187-192.

[12] Wu Bin. Master's thesis on stock prediction research, which combines multiple head self-attention mechanism and news public opinion characteristics [D]. Jiangxi University of Finance and Economics, 2022.

[13] He Zhikun, Liu Guangbin, Zhao Xijing, et al. Review of the Gaussian process regression methods [J]. Control and Decision, 2013, 28 (08): 1121-1129+1137.

[14] Wu Bin. Master's thesis on stock prediction research, which combines multiple head self-attention mechanism and news public opinion characteristics [D]. Jiangxi University of Finance and Economics, 2022.