Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Gaussian Process Regression Model Based on Cross-Validation Weighted Bagging Algorithm and Its Applications

Yongqi An[1,*], Yizhe Feng[2]

[1]School of Mathematics and Big Data, Dezhou University, Dezhou, China, 253023

[2]School of Mathematics and Big Data, Jining University, Qufu, China, 273100

* Corresponding Author Email: 15965219561@163.com

Abstract. As a non-parametric regression technique widely applied across various disciplines, Gaussian Process Regression (GPR) faces certain challenges in the selection of the kernel function. To address this issue, this paper innovatively adopts a model aggregation strategy to dynamically determine the kernel function, rather than selecting a single model, thereby enhancing the model's adaptability and predictive power. Specifically, on one hand, the proposed method integrates Gaussian processes with different kernel function representations using a bagging algorithm. On the other hand, considering the varying importance of each sub-model, this paper employs cross-validation for weighted processing to balance the overfitting and underfitting issues of the predictive model. Finally, through extensive simulation experiments and real data analysis, the results demonstrate that the proposed method significantly improves the predictive capability compared to some classic predictive models, possesses better generalization performance, and can be effectively applied in various application scenarios.

Keywords: Gaussian Process Regression; Kernel Function Selection; Model Averaging; Bagging Algorithm; Cross-Validation; Weighted Processing.

1. Introduction

Gaussian Process Regression (GPR), a Bayesian non-parametric method, offers a highly flexible solution for complex regression tasks. By defining a probability distribution within the function space, GPR naturally incorporates data noise and model uncertainty, thereby providing confidence intervals for predictions, which is crucial for tasks involving uncertainty quantification. GPR is applied in various fields, including stock price forecasting in finance, and the assessment of corn nutritional value, processing characteristics, storage stability, and meat quality in agriculture and food industries. The performance of Gaussian process regression models is highly dependent on the choice of the kernel function. The selection of the kernel function not only affects the model's adaptability but also the key to predictive accuracy and the width of confidence intervals. Given that different datasets and problems may require different kernel functions, the selection of the kernel function naturally becomes a challenging issue. Some scholars have conducted research on this.

For example, in a pioneering study, Wang and He [1] introduced an innovative approach to evaluate the selection of kernel functions by examining the proximity of incorrectly classified instances to the decision boundary delineated by support vectors, which serves as an indicator of misclassification severity. They further applied a rank sum test to this premise. Another insightful research by Liang, Chen, Deng, et al. [2] explored kernel function selection through a methodology inspired by fractal theory, quantifying data complexity via information entropy and leveraging the fractal dimension to dissect the sample distribution, thereby facilitating kernel function choice. Zhong [3] offered a novel perspective, who conceptualized the smallest hypersphere encapsulating the entire sample space, utilizing this metric to discern the spatial distribution characteristics and subsequently guiding kernel function selection in alignment with the kernel's geometric attributes. In a groundbreaking work, Gao and Jia [4] harnessed the Gram-Schmidt Orthogonalization technique and the Hilbert-Schmidt Independence Criterion to devise a novel strategy for kernel function selection. Their technique commences by employing the GSO to mitigate redundancy among various kernel functions,

subsequently applying the HSIC to gauge the likeness of each kernel to an exemplary one, culminating in the identification of the kernel exhibiting superior discriminative capabilities. However, in the model selection process, the uncertainty of the model is often neglected, which may lead to poor estimation effects. The estimation of a single model may pose risks, including a reduction in accuracy due to model selection bias, and an inability to fully utilize the large amount of information that has been discarded. Therefore, we draw on the method of model averaging [5], by combining multiple models and averaging the model estimation results with weighted averages. The model averaging method generally does not regard a selected model as the true data generation process, but considers all models with reasonable weights, providing a safeguard mechanism for model estimation, effectively avoiding model selection bias, and robustly and effectively improving the accuracy of model estimation.

Based on this, this paper combines the Bagging algorithm with the Gaussian Process Regression model to form an ensemble learning framework with uncertainty quantification. Specifically, by independently training Gaussian Process models based on various kernel functions on independently resampled sub-datasets, this method not only significantly enhances the adaptability and robustness of the model's predictions but also breaks through the limitations of single model selection, reflecting the advantages of model averaging. To further optimize the predictive performance, we introduce an adaptive weight distribution mechanism based on cross-validation. This mechanism dynamically adjusts the weight of the sub-models according to their predictive performance on the validation set data, ensuring that models with excellent performance play a leading role in the prediction task, while the contribution of poorly performing models is correspondingly reduced [6]. Simulation data and actual data analysis have proven that the proposed method is highly competitive compared to some comparative methods.

2. Theory and Methodology

2.1. Gaussian Process

Gaussian Process Regression [7] is a non-parametric Bayesian method where a finite number of random variables are all consistent with the Gaussian distribution. Gaussian Process Regression uses a Gaussian process prior for regression analysis of the data. Due to the effect of noise, the output Y can be represented by the input values X and the noise ϵ as follows: Y=f(x)+ε, Assuming the noise follows a normal distribution, that is: $\varepsilon \sim N(0, \sigma_n^2)$. Given the output Y$Y$, if the Gaussian prior distribution is assumed, and given the output YY and the predicted values y, the joint Gaussian prior distribution would be described as follows:

$$Y \sim N\left[0, K(X, X) + \sigma_n^2 I_n\right] \tag{1}$$

$$\begin{bmatrix} Y \\ y \end{bmatrix} \sim N\left(0, \begin{bmatrix} \boldsymbol{K}(X, X) + \boldsymbol{\sigma}_n^2 I_n & \boldsymbol{K}(X, x_*) \\ \boldsymbol{K}(x_*, X) & \boldsymbol{K}(x_*, x_*) \end{bmatrix}\right) = N\left(0, \begin{bmatrix} K & K_*^{\mathrm{T}} \\ K_* & K_{遍} \end{bmatrix}\right) \tag{2}$$

In the formula: K(x, x')-a symmetric positive definite covariance matrix, which measures the correlation between x and x' through the kernel function; K(X, X)-the covariance matrix between the training set; I_n — the n -dimensional identity matrix; K(X_* ,X)=K(X, X_*)-The covariance matrix between the test set X_* and the training set X ;K(X_* , X_*)-The covariance matrix among the training set. Among them, $K(x, x') = p_1 \cdot \exp\left[-\frac{(x-x')^2}{2p_2}\right]$, In the formula: p1, p2—adjustable parameters. The posterior distribution of the predicted value y [8] is as follows:

$$\begin{cases} y|Y \sim N(y, \sigma_y^2) \\ y = K_* K^{-1} Y \\ \sigma_y^2 = K_{\star\star} - K_* K^{-1} K_*^{\mathrm{T}} \end{cases} \tag{3}$$

Therefore, the probability density function for the i-th predicted value is as follows:

$$p(y_i) = \frac{1}{\sqrt{2\pi}\sigma_{yi}} \exp\left(-\frac{y_i - \bar{y}_i}{2\sigma_{yi}^2}\right) \tag{4}$$

The kernel function [9] indeed plays an essential role in Gaussian Process Regression (GPR). Its key function is to reveal the inherent complexity of the data, such as smoothness, periodicity, or dynamic variability, based on our prior assumptions about the data. Specifically, GPR uses the kernel function to express the properties of the dataset in a prior mathematical form. This process can map nonlinear relationships into linear relationships in a high-dimensional feature space, thereby simplifying complex nonlinear prediction problems into relatively simpler linear problems. Common kernel functions in GPR include the Radial Basis Function (RBF) kernel, the Matern kernel, the linear kernel, and the periodic kernel.

(1)Radial Basis Function (RBF) Kernel:

$$k_{\mathrm{RBF}}(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x - x')^2\right) \tag{5}$$

RBF, also known as the Gaussian kernel or squared exponential kernel, produces very smooth functions. Here, σ_f^2 is the signal variance, and l is the length scale parameter, which determines the rate at which the function changes.

(2) Matern Kernel Function

$$k_{\mathrm{Matern}}(x, x') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,|x - x'|}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\,|x - x'|}{l}\right) \tag{6}$$

$\Gamma(\nu)$ represents the gamma function, K_ν denotes the modified Bessel function of the second kind, and l stands for the length scale parameter.

(3) Linear Kernel Function:

$$k_{\mathrm{linear}}(x, x') = x^T x' \tag{7}$$

(4) Periodic Kernel Function:

$$k_{\mathrm{periodic}}(x, x') = \sigma_f^2 \exp\left(-\frac{2\sin^2(\pi\,|x - x'|/p)}{l^2}\right) \tag{8}$$

The periodic kernel function is specifically designed to handle data with periodic patterns, where p is the period length and l is the length scale parameter.

2.2. Gaussian Process Regression Model Based on Cross-Validation Weighted Bagging Algorithm

Based on this, the paper introduces a novel algorithm—the Cross-Validation Weighted Ensemble Gaussian Process Regression model. This algorithm consolidates multiple Gaussian Process Regression (GPR) models, employs least squares to calculate the weights[10], and dynamically assigns these weights, resolving the issue of kernel function selection and improving the accuracy and generalization of predictions. The detailed process is depicted in Figure 1.

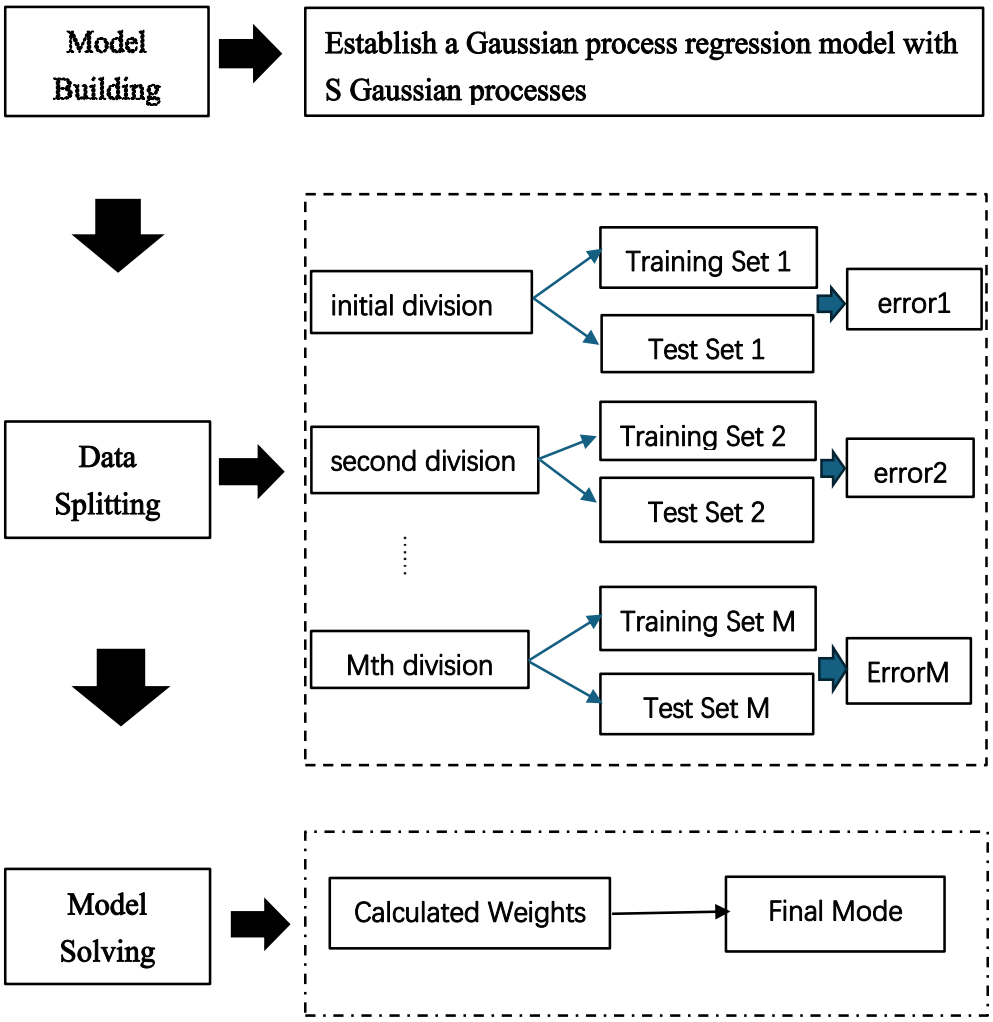

Figure 1. Weighted Bagging Algorithm Flowchart

3. Data Analysis

3.1. Data Sources

The spectral dataset of meat samples can be downloaded from http://lib.stat.cmu.edu/datasets/teccator.This dataset contains n=240 samples, each with the moisture, fat, and protein content of the meat, as well as the absorption spectrum measured by a near-infrared spectrometer in the 850-1050 nanometer (nm) range. With intervals of two wavelengths, a total of 100 absorption spectrum data points were recorded.

The spectral dataset of corn samples can be downloaded from http://www.eigenvector.com/data/Corn/index.html.Originally from Cargill, USA, this dataset consists of n=80 corn samples, with a wavelength range of 1100–2498 nm, measured at intervals of 2 nm, totaling 700 absorption spectrum data points. The oil, moisture, and starch content of each corn sample were also recorded.

In this experiment, we used a systematic approach to ensure the consistency of data division and the reliability of model evaluation. Specifically, the original data was divided into a training set of 80% and a test set of 20%. This ratio ensured that most of the data was used to construct the model, while retaining a portion of independent data to verify the model's generalization ability. To avoid the impact of randomness in data division on model performance evaluation, we adopted a design of 10 repeated experiments, with the data set randomly divided in each experiment, which helps to reduce bias and improve the robustness of the evaluation. The predictive performance of the model is measured by calculating the root mean square error (RMSE) on the test set. RMSE, as an important indicator in regression analysis, can quantify the average difference between predicted values and actual observations. We focus on the average and standard deviation of RMSE in 10 experiments, where the average of RMSE reveals the overall accuracy of the model's predictions, and the standard deviation reflects the stability of the model's performance across different experiments and data divisions. In terms of simulated data generation, we generated a dataset containing 7 features (x1 to x7), each with 150 samples, and each feature follows a truncated normal distribution with a mean of 0.5 and a variance of 0.1, limited between 0 and 1, y is a nonlinear combination of x1 - x4, that is:

$$y = \sin(\pi \cdot x_1) + e^{x_2} - \ln(x_3 + 1) + 3 \cdot x_4^3 + \varepsilon \tag{9}$$

We introduced random noise ε into the dataset, which follows a normal distribution, denoted as $\varepsilon \sim N(0, \sigma_n^2)$, And we set three irrelevant variables x5,x6,x7$x5,x6,x7$ to verify that our Gaussian-weighted model can mitigate the impact of irrelevant variables on the accuracy of y. To make specific comparisons, we can use the following formula:

$$\left|\frac{\mathrm{RMSE}_{\mathrm{other}} - R\mathrm{MSE}_{\mathrm{new}}}{\mathrm{RMSE}_{\mathrm{other}}}\right| \times 100\% \tag{10}$$

In the formula, RMSEother the mean of the Root Mean Square Error (RMSE) from other models, and RMSEnew is the mean of the RMSE from the model used in this experiment.

3.2. Comparison Results

The following tables list the Root Mean Square Error (RMSE) mean and standard deviation for simulated data, meat data, and corn data on various models such as K-Nearest Neighbors Regression, LASSO Regression, BP Neural Network, and CVGP.

Table 1. Simulated Data

Model Name	RMSE MEAN	Standard Deviation
Random Forest	1.2861e-1	3.0249e-2
K-Nearest Neighbors Regression	2.0075e-1	3.5652e-2
LASSO REGRESSION	3.4678e-1	3.9068e-2
BP Neural Network	3.1109e-1	4.1436e-2
CVGP	9.8015e-2	5.3371e-2

Table 1 offers a comparative analysis of various predictive models, with the CVGP model outperforming others in predictive accuracy, boasting the lowest average Root Mean Square Error (RMSE) of 0.0980. Although its standard deviation of 0.05337 is not the lowest, the model demonstrates commendable stability, ensuring reliable predictions. In contrast, the K-Nearest Neighbors Regression and the Backpropagation Neural Network exhibit higher average RMSEs of 0.2008 and 0.3111, respectively. The LASSO regression lags significantly with an average RMSE of 0.3468. The findings highlight the CVGP model's superiority in both accuracy and stability in the simulated dataset prediction task, surpassing other models.

Table 2. Meat Data

	Model Name	RMSE MEAN	Standard Deviation
Fat	Random Forest	7.1725	0.6891
	K-Nearest Neighbors Regression	8.3312	1.2847
	LASSO REGRESSION	4.3374	0.3299
	BP Neural Network	11.8010	0.5308
	CVGP	2.8522	1.2489
Water	Random Forest	6.2869	0.5246
	K-Nearest Neighbors Regression	7.1378	0.4541
	LASSO REGRESSION	4.0982	0.2766
	BP Neural Network	17.6205	2.3317
	Random Forest	2.2026	0.1935
Protein	K-Nearest Neighbors Regression	2.4981	0.2259
	LASSO REGRESSION	2.8243	0.1839
	BP Neural Network	5.4206	0.5862
	CVGP	1.1695	0.3614

Table 3. Corn Data

	Model Name	RMSE MEAN	Standard Deviation
Moisture	Random Forest	3.4861e-2	8.065e-3
	K-Nearest Neighbors Regression	1.5101e-1	3.247e-2
	LASSO REGRESSION	2.0833e-1	3.224e-2
	BP Neural Network	4.9869e-1	8.212e-1
	CVGP	3.600e-3	1.137e-2
oil	Random Forest	3.9095e-1	1.713e-1
	K-Nearest Neighbors Regression	8.7776e-1	1.459e-1
	LASSO REGRESSION	7.9614e-3	8.804e-4
	BP Neural Network	6.7093e+0	5.789e+0
	CVGP	1.6289e-10	5.510e-11
Protein	Random Forest	3.9266e-1	1.064e-1
	K-Nearest Neighbors Regression	9.1161e-1	9.384e-2
	LASSO REGRESSION	2.1399e-2	3.836e-3
	BP Neural Network	1.6327e+1	2.772e+0
	CVGP	1.2489e-10	3.572e-11
Starch	Random Forest	6.8924e-2	7.336e-3
	K-Nearest Neighbors Regression	9.7589e-1	1.900e-1
	LASSO REGRESSION	4.9564e-2	1.062e-2
	BP Neural Network	2.8551e+0	2.205e+0
	CVGP	1.8411e-7	5.810e-7

Analysis of Tables 2 and 3 reveals that the Conditional Variance Gaussian Process (CVGP) model has demonstrated exceptional performance in predicting nutritional indicators for both the meat and corn datasets. In the meat dataset, the CVGP model leads with a low RMSE of 2.8522 for fat content prediction and shows high accuracy with an RMSE of 1.1695 for protein content prediction. In the corn dataset, the CVGP model exhibits extremely high precision and stability in predicting moisture and oil content, with RMSE values of 0.0036 and 1.6289e-10, respectively. Additionally, the model maintains very low RMSE values and standard deviations for protein and starch content predictions, at 1.2489e-10 and 5.810e-7, respectively. Overall, the CVGP model outperforms other models in predicting various nutritional indicators, confirming its potential as a powerful predictive tool.

Use equation (10) to calculate the percentage improvement of the CVGP model in this experiment compared to other traditional models (Random Forest, K-Nearest Neighbors Regression, LASSO Regression, BP Neural Network).

Table 4. Percentage Improvement of CVGP on Simulated Data

Random Forest	K-Nearest Neighbors Regression	LASSO REGRESSION	BP Neural Network
23.78%	51.16%	71.72%	68.49%

Analysis of Table 4 reveals that the CVGP model significantly outperforms the Random Forest, K-Nearest Neighbors Regression, LASSO Regression, and Backpropagation Neural Network models in predictive performance on the simulated dataset. Specifically, the CVGP model shows a 23.78% improvement over the Random Forest, a 51.16% improvement over the K-Nearest Neighbors Regression, a 71.72% improvement over the LASSO Regression, and a 68.49% improvement over the Backpropagation Neural Network. These results indicate that the CVGP model has a significant advantage in enhancing the predictive accuracy of the simulated dataset.

Table 5. Percentage Improvement of CVGP on Meat Data

	Random Forest	K-Nearest Neighbors Regression	LASSO REGRESSION	BP Neural Network
Fat	60.16%	65.77%	38.55%	75.73%
Water	58.78%	63.79%	37.34%	85.33%
Protein	47.42%	53.11%	58.61%	78.41%

Table 6. Percentage Improvement of CVGP on Corn Data

	Random Forest	K-Nearest Neighbors Regression	LASSO REGRESSION	BP Neural Network
Moisture	90.24%	97.61%	98.27%	99.27%
Oil	99.99%	99.99%	99.81%	99.99%
Protein	99.68%	99.86%	99.41%	99.92%
Starch	99.97%	99.99%	99.63%	99.99%

Analysis of Tables 5 and 6 reveals that the Conditional Variance Gaussian Process (CVGP) model has shown significant performance improvements over traditional algorithms in the task of predicting nutritional components in both meat and corn datasets. In the meat dataset, the CVGP model achieved enhancements of 60.16% in fat, 58.78% in water, and 47.42% in protein content predictions, notably outperforming the backpropagation neural network with an 85.33% improvement in moisture prediction. In the corn dataset, the CVGP model realized near-perfect prediction improvements of 99.99% in oil, 99.92% in protein, 99.99% in starch, and 90.24% in moisture content, comprehensively leading other models. These results highlight the high precision and robustness of the CVGP model in food nutritional component analysis and underscore its potential for application in this field.

The CVGP model has achieved significant performance improvements mainly due to the following aspects: First, it enhances adaptability and robustness in predictions by integrating multiple Gaussian processes and kernel functions. Second, the weight allocation mechanism of cross-validation ensures the dominant role of efficient models in predictions while reducing the negative impact of inefficient models. Additionally, the CVGP model can handle data noise and uncertainty, providing crucial predictive confidence intervals, which is essential for quantifying uncertainty. Through kernel functions, it transforms complex nonlinear problems into linear problems in high-dimensional space, simplifying the prediction process. Lastly, the CVGP model has demonstrated excellent generalization ability across various scenarios, and its application on both simulated and real datasets has verified its powerful predictive power and robustness. These advantages collectively form the core competitiveness of the CVGP model in handling complex data structures.

4. Conclusion and Extension

Faced with the subjectivity and limitations in the selection of the kernel function in Gaussian Process Regression, this paper innovatively approaches the issue from the perspective of model averaging, skillfully circumventing this core challenge. Specifically, the concept of ensemble learning is introduced, and a series of Gaussian process models are constructed by integrating various kernel functions, forming a model pool. This strategy not only dissolves the uncertainty of a single kernel function selection but also enhances the robustness and generalization capability of predictions through the integration of models. Furthermore, a cross-validation weighting mechanism is applied to finely calibrate each member of the model pool, ensuring the model's expressiveness on unknown data. This weighted averaging process is essentially an intelligent fusion strategy that dynamically allocates weights based on each sub-model's predictive effectiveness on an independent validation set, thereby selecting the most valuable model combination. This method not only significantly improves the overall accuracy of predictions but also effectively suppresses overfitting phenomena, ensuring the model's robust performance in complex data environments. The experimental results strongly prove the effectiveness of our method, as the CVGP model demonstrates excellent predictive performance on both idealized simulated datasets and complex real datasets.

This study not only injects theoretical innovation into the field of Gaussian Process Regression but also provides a powerful and flexible analytical tool for solving practical problems. The CVGP model shows a broad prospect for application, especially in cutting-edge fields such as financial market forecasting, bioinformatics analysis, and environmental monitoring systems, where highly accurate predictions and uncertainty assessments are required. The importance of the CVGP model is fully demonstrated in data science and machine learning research. For example, in financial market forecasting, the time series data that the model needs to handle often exhibit non-linear and non-stationary characteristics. The CVGP model, through its flexible kernel function selection, can effectively capture these complex patterns and provide more accurate predictive results. It is worth noting that although this paper mainly discusses data in Euclidean space, the flexibility of the CVGP model makes it equally applicable to datasets in non-Euclidean geometric spaces.

References

[1] Wang, Z. W., & He, G. Y. (2018). Research on kernel function selection methods. Journal of Hunan University (Natural Sciences), 45(10), 155-160. doi:10.16339/j.cnki.hdxbzkb.2018.10.021.

[2] Liang, L. M., Chen, M. L., Deng, G. H., et al. (2019). Selection of kernel functions for support vector machine under fractal theory. Science, Technology and Engineering, 19(13), 131-138.

[3] Zhong, Z. (2015). Kernel function selection for support vector machines under sparse representation (Master's thesis, Jiangxi University of Science and Technology).

[4] Gao, Y. T., & Jia, S. Q. (2024). A kernel function selection method based on Gram-Schmidt orthogonalization and HSIC. Computer Technology and Development, 34(06), 148-154. doi:10.20165/j.cnki.ISSN1673-629X.2024.0086.

[5] Chen, Q. R., & He, J. F. (2023). Research on model averaging assisted sampling estimation method. Statistics and Decision, 39(09), 35-41.

[6] Zhu, R., Zou, G. H., & Zhang, X. Y. (2018). Model averaging method for partially linear models. Systems Science and Mathematics, 38(07), 777-800.

[7] Li, C. R., Xiao, F., Fan, Y. X., et al. (2021). A lithium battery SOC estimation method based on Gaussian process regression. Journal of Naval University of Engineering, 33(01), 55-59.

[8] Deng, W. J., Xiao, H., Li, J. Z., et al. (2021). A photovoltaic power forecasting method based on improved long short-term memory network and Gaussian process regression. Electrical Equipment and Energy Saving Management Technology, (08), 51-57.

[9] Li, C. K., & Xu, M. C. (2015). The concept, properties, and construction of kernel functions. Computer Knowledge and Technology, 11(32), 171-173.

[10] Li, H. Y. (2020). Prediction method based on a new model weight and its application (Doctoral dissertation, Central China Normal University). doi:10.27159/d.cnki.ghzsu.2020.003027.