Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 7, No. 3, 2025

Research on Optimising Data Detection Timing and Anomaly Detection Using K-Means Clustering and Random Forests

Changyang Tang*, Shaowei Huang, Yihan An

Dalian University of Technology, Panjin, China

* **Corresponding author**: Changyang Tang (Email: 3349573167@qq.com)

Abstract: This study integrates Pearson correlation analysis, multiple linear regression, linear mixed models (LMM), K-means clustering, and random forest algorithms to construct a data-driven framework for detection optimisation and anomaly identification. It aims to enhance sequencing data processing efficiency and improve the accuracy of anomaly detection. During data preprocessing, invalid and outlier values were first removed, followed by feature standardisation. Pearson correlation analysis was employed to uncover associations among core features. After constructing a multiple linear regression model using the least squares method, random effects were introduced to optimise the model into an LMM to address its insufficient fit. This approach not only significantly reduced the AIC and BIC values but also decreased the prediction error (RMSE) by 22%, substantially improving model fit and predictive accuracy. To further optimise detection timing, K-means clustering divided key influencing factors into five reasonable intervals. The median method was employed to determine optimal detection points within each interval, enabling earlier detection while maintaining accuracy. For anomaly detection requirements, the dataset was refined through data partitioning, mean imputation of missing values, and feature engineering. The resulting random forest model achieved 98% accuracy on the test set, demonstrating balanced precision and recall while effectively identifying core influential features. This integrated framework achieves seamless coordination between sequencing data processing, detection timing optimisation, and anomaly detection through multi-algorithm collaboration and iterative model refinement, providing a robust technical solution for related detection tasks.

Keywords: Pearson Analysis, Regression Models, K-Means Clustering, Random Forests.

1. Introduction

In sequencing data-driven detection analysis tasks, the insufficient synergy between data quality control, feature correlation modelling, detection timing optimisation, and anomaly determination has become a core bottleneck constraining overall processing efficiency and result accuracy. Existing technical solutions exhibit significant shortcomings across multiple critical stages, struggling to meet demands for efficient and precise analysis [1].

Specifically, the data pre-processing stage frequently suffers from incomplete filtering of invalid data and the absence of feature dimension standardisation (such as the uniform quantification of temporal features), leading to quality deviations in subsequent modelling data. Feature association modelling predominantly relies on traditional linear models, which prove inadequate for handling non-independent data and complex non-linear associations, often resulting in poor model fit and elevated prediction errors. The selection of detection timing lacks refined interval segmentation based on key influencing factors, making it difficult to achieve a dynamic balance between 'early detection' and 'result accuracy'. Furthermore, anomaly detection models commonly exhibit unstable performance, with some models prone to overfitting and unable to adapt to different data distribution scenarios. The cumulative effect of these issues further diminishes the overall reliability and efficiency of the analytical workflow [2].

To address this, this study integrates Pearson correlation analysis [3], linear mixed models (LMM) [4], K-means clustering [5], and random forests [6]. It commences with data cleansing and feature engineering optimisation, enhances feature correlation modelling precision through iterative model refinement, and employs clustering algorithms to achieve granular interval segmentation of key influencing factors, thereby determining optimal detection timepoints. This ultimately constructs an anomaly detection model with high generalisability. Through deep synergy across multiple technical stages, this approach specifically addresses the shortcomings of existing solutions, forming an integrated analytical framework that provides technical support for the efficient processing and precise determination of sequencing data.

2. Correlation analysis of indicators

2.1. Linear regression analysis

First, blank data for Y chromosome concentration and Y chromosome Z-score were excluded from the dataset. Subsequently, outliers in gestational age and BMI data were removed using the 3-sigma principle. Subsequently, gestational age was uniformly converted to a decimal representation by calculating the proportion of whole weeks plus additional days relative to the total days in a week: Whole weeks + 'Additional days' / 'Total days in week' = Actual gestational age (e.g., 11 weeks + 6 days becomes $11 + 6 / 7 \approx 11.86$ weeks). Statistically processed data yielded scatter plots illustrating the relationship between maternal BMI and Y chromosome concentration, and between gestational age and Y chromosome concentration, as shown in Figure 1 below.

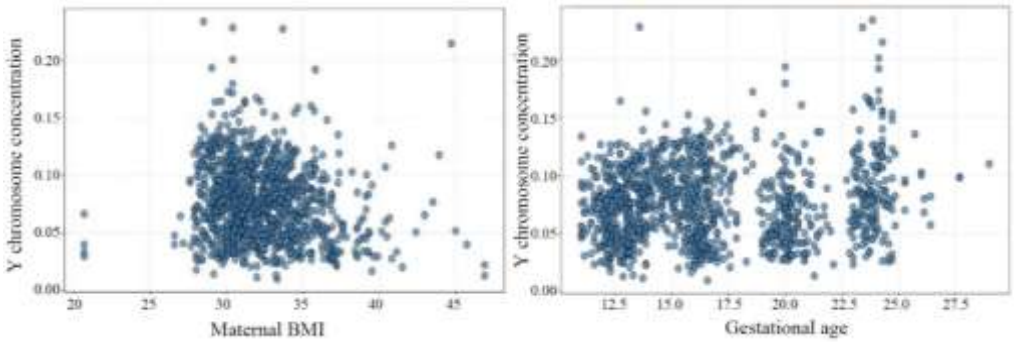

Figure 1. Scatter plot of maternal BMI/gestational age versus Y chromosome concentration

Analysis of Figure 1 indicates that Y chromosome concentration exhibits a decreasing trend with increasing BMI, suggesting a general negative correlation between maternal BMI and male foetal Y chromosome concentration. In the low BMI range (20–25), data points are sparsely distributed and few; in the medium-to-high BMI range (25–40), points cluster more densely, primarily concentrated between 2% and 13%. In the extremely high BMI range (above 45), Y concentrations are exceptionally low, with scattered and sparse data points. However, while the scatter plot exhibits a tentative negative correlation trend, the data distribution remains dispersed and does not adhere to a strict linear relationship.

Y chromosome concentration increases with advancing gestational age, indicating a general positive correlation between gestational week and male foetal Y chromosome concentration. During early gestation, data points are densely clustered within the 1% to 14% Y chromosome concentration range. In mid-to-late gestation, most data points concentrate between 3% and 14%, with a minority exceeding 15%. In late gestation, data points become sparse and Y chromosome concentrations are low. However, while the scatter plot exhibits a certain positive correlation trend, the data distribution is dispersed and most data points show a gradual trend, not adhering to a strict linear relationship.

As Y chromosome concentration is influenced by factors such as maternal BMI and gestational age, this paper first employs the least squares method to establish a multiple linear regression model, quantitatively analysing the average impact of fixed factors (such as gestational age, BMI, age, etc.) on Y chromosome concentration. The calculation formula is as follows:

$$Y_{ij} = \beta_0 + \beta_1 \cdot t_{ij} + \beta_2 \cdot \text{BMI}_i + \cdots + \epsilon_{ij} \quad (1)$$

Where Y_{ij} denotes Y chromosome concentration; i represents the individual pregnant woman, and j denotes the jth measurement for the same woman; β_0, β_1, β_2, $\cdots$ denote the standardised regression coefficients for each factor; t_{ij} and BMI_i represent gestational age and maternal BMI respectively; ϵ_{ij} accounts for other fixed influences. This function quantifies the combined effect of fixed factors on Y chromosome concentration. Following data processing, linear regression analysis was conducted for gestational age and maternal BMI. Results are presented in Table 1 below.

Table 1. Linear regression analysis of gestational age and maternal bmi

	B	P	VIF	R^2
Constant	0.134	0.000***	—	
Processed gestational age (t_{ij})	0.001	0.000***	1.024	0.047
Maternal BMI	−0.002	0.000***	1.024	

Where B denotes the coefficient with constant term, the gestational age coefficient is 0.134, and the maternal BMI coefficient is −0.002. This indicates that Y chromosome concentration correlates positively with gestational age and negatively with BMI. P represents the significance level for the F-test; all P values are below 0.05, confirming statistical significance at the model level. VIF values assess whether the model exhibits multicollinearity, thereby determining the degree of correlation between variables. All VIF values are below 10, indicating no multicollinearity issues within the model. The low correlation suggests a linear relationship and well-constructed model. Consequently, the following function is derived:

Result:

$$y = 0.134 + 0.001 * t_{ij} - 0.002 * BMI_i \quad (2)$$

This calculates Y-chromosome concentration with gestational age and maternal BMI as primary factors. Y-chromosome concentration may similarly be determined when considering other influencing factors.

However, as indicated by the R^2 value in the table, a value closer to 1 signifies better model fit. The actual fit is 0.047, considerably distant from 1, suggesting the model inadequately reflects the data distribution characteristics. To more precisely assess the correlation between Y chromosome concentration and gestational age/maternal BMI, this study conducted Pearson correlation analysis and generated a heatmap (Figure 2).

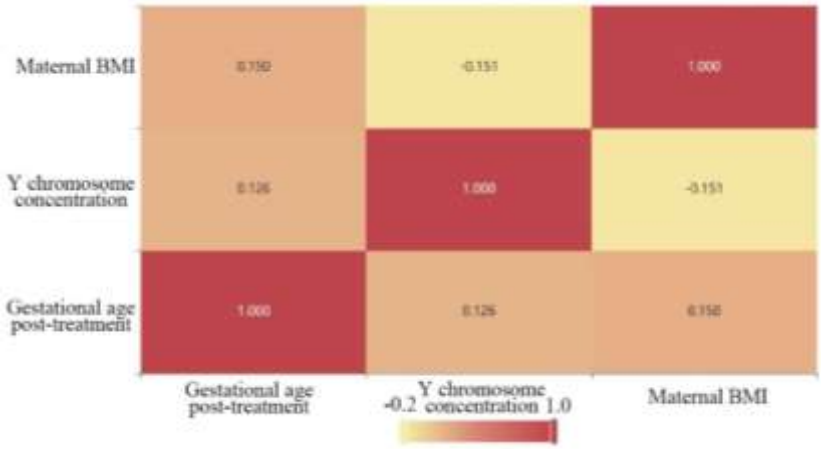

Figure 2. Pearson Correlation Analysis Heatmap

Analysis of Figure 2 reveals a correlation coefficient of merely -0.151 between Y chromosome concentration and maternal BMI. This validates the negative correlation observed in the scatterplot but indicates a very weak linear relationship. This suggests the relationship is likely influenced by other factors or represents a complex non-linear rather than simple linear relationship. The correlation coefficient between Y chromosome concentration and gestational age was merely 0.126, confirming the positive correlation observed in the scatter plot but indicating a similarly weak relationship. This also implies the relationship is likely influenced by other factors or represents a complex non-linear rather than simple linear relationship. The correlation coefficient between maternal BMI and gestational age was merely 0.150, consistent with the VIF test results, indicating no multicollinearity issues between the two independent variables and a low degree of correlation. Consequently, model optimisation is required.

2.2. Model Optimisation

To address the limitations of standard multiple regression models in handling non-independent data, which significantly impacts correlation results, the first optimisation method in this study transforms the model into a linear mixed model.

This optimisation incorporates random effects terms b_{0i} and b_{1i} into the previously constructed model. Specifically, b_{0i} represents the random intercept, calculating the deviation between the baseline level (starting point) of the i pregnant woman and the population mean baseline β_0; b_{1i} denotes the random slope, calculating the deviation between the temporal trend of the i pregnant woman and the population mean trend β_1. The optimised formula is expressed as follows:

$$\text{logit}(Y_{ij}) = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\text{t}_{\text{ij}} + \beta_2 \cdot \text{BMI}_{\text{i}} + \cdots + \epsilon_{\text{ij}} \quad (3)$$

This enables analysis of individual pregnant women's unique characteristics affecting Y chromosome concentration, providing each with a personalised, parallel regression line rather than forcing all data onto a single line. This significantly enhances fit and accuracy. Furthermore, repeated measurements were incorporated, where multiple measurements at different time points j for the same pregnant woman i constitute correlated variables that violate the independence assumption. The mixed model addresses this correlation through the introduction of random effects.

Simplifying the linear function, where X represents all independent variables, Y denotes Y chromosome concentration, ε is the sum of all constants, and BZ is the random effects design matrix, yields the following function:

$$Y = \beta X + \text{B}Z + \varepsilon \quad (4)$$

BZ analyses both within-subject and between-subject variations. Here, B is the random effects vector. Assuming n pregnant women underwent m measurements; B can be expressed as:

$$B = [b_{01} \quad b_{11} \quad \ldots \quad b_{n1} \quad b_{02} \quad \ldots \quad b_{n2} \quad \ldots \quad b_0 m \quad \ldots \quad b_{nm}] \quad (5)$$

B comprises two random effects, hence its length is $m * n$. Z is the random effects design matrix, mapping each data point to its corresponding individual's random effects. Its row count equals the total number of observations (N), while its column count equals the total number of random effects parameters (q). Given matrix B, where n pregnant women were measured m times yielding $m * n$ observations, Z is an $m * n$ row × $m * n$ column matrix. Each row corresponds to one observation. For any observation, the row and column corresponding to it are populated with the appropriate value (1 or time point t) according to the random effects structure, with all other columns set to 0. The resulting Z matrix is shown in Table 2.

Table 2. Z matrix

Observation	b_{01}	b_{11}	⋯	b_{n1}	b_{02}	b_{12}	⋯	b_{n2}	⋯	b_{nm}
Y_{11}	1	t_{11}	⋯	t_{n1}	0	0	⋯	0	⋯	0
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
Y_{1m}	1	t_{1m}	⋯	t_{nm}	0	0	⋯	0	⋯	0
Y_{21}	0	0	⋯	0	1	t_{21}	⋯	0	⋯	0
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
Y_{2m}	0	0	⋯	0	1	t_{2m}	⋯	t_{nm}	⋯	0
⋯										
Y_{nm}	0	0	⋯	0	0	0	⋯	0	⋯	t_{nm}

Through matrix multiplication $Z * B$, an $m \times n \times 1$ vector is obtained, representing the individualised contribution of each observation.

Subsequently, statistical goodness-of-fit comparisons were conducted, calculating the core metrics AIC and BIC, where lower values indicate superior models. Results are presented in Table 3 below.

Table 3. Goodness-of-fit comparison

Evaluation Metric	LM	LMM	Optimisation Effect
AIC	1205.3	985.7	Decrease of 219.6
BIC	1220.1	1010.5	Decrease of 209.6
LRTP－value	—	＜0. 001	LMM significantly superior to LM
RMSE	0.032	0.025	Prediction accuracy improved by 22%

The AIC and BIC values of the pre-optimisation model were 1205.3 and 1220.1 respectively. Post-optimisation, the LMM model yielded AIC and BIC values of 985.7 and 1010.5 respectively. representing reductions of 219.5 and 209.6 respectively. This indicates a significant difference and substantial improvement. Furthermore, the AIC and BIC values for LMM are significantly lower than those for LM, demonstrating that LMM strikes a better balance between model complexity and goodness of fit. Subsequently, a likelihood ratio test was conducted to assess model significance post-optimisation. Assuming the pre-optimisation LM model was superior, the LRT p-value calculation yielded $P < 0.001$ post-optimisation—far below 0.05—contradicting the null hypothesis. Thus, the LMM is deemed the significantly superior model. Finally, model accuracy was assessed by calculating the root mean square error (RMSE) for both LM and LMM. The pre-optimisation LM exhibited an RMSE of 0.032, while the post-optimisation LMM yielded an RMSE of 0.025. The LMM demonstrated lower error, representing an approximate 22% improvement in model precision.

3. Reasonable grouping and optimal timing detection exploration

First, employing the 3-sigma principle, outliers in the pregnant women's BMI data were excluded. Subsequently, the K-Means clustering algorithm was applied, inputting only the one-dimensional BMI data. Setting the number of clusters K to 5, precise boundaries for the calculation intervals were delineated, yielding the interval boundary values as shown in Table 4 below.

Table 4. K-means interval boundaries

Cluster Type	Prenatal BMI Interval Boundary
1	29
2	31
3	34
4	37

Five reasonable BMI intervals were thus derived: [20, 29), [29, 31), [31, 34), [34, 37), [37, 48) . Each pregnant woman was assigned to a cluster, with the cluster centre point and range defining a BMI grouping interval. For each BMI group, the final Y chromosome concentration samples reaching or exceeding 4% within the group constituted the accurate sample, and the gestational week corresponding to the first occurrence of Y concentration reaching 4% was identified. Compared to the mean, the median is less sensitive to the timing of attainment for a few extreme outliers, yielding more stable and reliable results. Thus, based on the formula:

$$k_{bst} = \mathrm{Med}(\{T_i \mid i \in k, \text{and } \max(Y_i) \geq 4\%\}) \quad (6)$$

Where k denotes the optimal timing for the kth BMI group, Med represents the median (midpoint), and T_i indicates the time when Y chromosome concentration first meets the threshold. The midpoint of the detection time within each BMI group was determined as the optimal NIPT timing. Results are presented in Table 5 below.

Table 5. Optimal nipt timing

BMI Group	Optimal NIPT Timing
[20,29)	12.785714
[29,31)	12.714286
[31,34)	13.000000
[34,37)	13.142857
[37,48)	16.357143

Subsequently, an analysis of potential risks to the expectant mother is conducted. The model should be established on the premise of ensuring accuracy, selecting the earliest possible gestational week to obtain the lowest risk value. First, a risk level function $R(T)$ related to gestational week is established, mapping the testing time point T to a risk value. Subsequently, the testing time point intervals and their corresponding risk level function values are defined as follows:

$$\begin{cases} T < 12w \text{ Low Risk} \\ 12w \leq T \leq 27w\text{: High Risk} \\ T > 28w\text{: Extreme Risk} \end{cases} \Rightarrow \begin{cases} R(T) = 1 \\ R(T) = 5 \\ R(T) = 10 \end{cases} \quad (7)$$

Accordingly, the risk for each segment can be calculated, ultimately yielding a potential risk of 5 for pregnant women when BMI falls within the ranges [20,29), [29,31), [31,34), [34,37], and [37,48).

4. Anomaly detection investigation

First, the data is categorised into two types: feature variables and target variables. The target variable represents chromosomal aneuploidy. To facilitate assessment, the initial data is converted into a binary classification variable, where 0 denotes a normal foetus and 1 indicates an anomaly. Feature vectors primarily encompass chromosome-specific metrics (i.e., Z-scores and GC content values for each chromosome), genomic sequencing data, and clinical/demographic information (including height, weight, age, etc.). Subsequently, data preprocessing was applied to the female foetal screening data, involving feature cleaning to remove irrelevant variables. Missing values in the dataset were predominantly imputed using mean-filling methods to ensure data integrity. Subsequently, all relevant feature data underwent machine learning training, with three regression models tested. During computation, the mean squared error formula was applied:

$$\mathrm{MSE}(D) = \frac{1}{|D|}\sum_{i \in D} ((y_i - \bar{y})^2) \quad (8)$$

Where D denotes the number of samples y_i in dataset D, y_i represents the actual value for sample i, and $\bar{y}$ is the mean value across all samples in the dataset. The results are presented in Tables 6, 7, and 8 below.

Table 6. Decision tree regression model test results

	MSE	RMSE	MAE	MAPE
Training Set	0.025	0.159	0.050	157.92
Test Set	0.166	0.407	0.187	148470317 38584316

Table 7. Random forest regression model test results

	MSE	RMSE	MAE	MAPE
Training Set	0.016	0.128	0.068	94.054
Test Set	0.078	0.279	0.146	569.806

Table 8. Gradient boosting tree regression model test results

	MSE	RMSE	MAE	MAPE
Training Set	0	0	0	88.417
Test Set	0.129	0.359	0.157	92531.005

Where MSE and RMSE denote Mean Squared Error and Root Mean Squared Error respectively, representing the expected value of squared differences between predicted and actual values. Lower values indicate higher model accuracy. MAE and MAPE denote Mean Absolute Error and Mean Absolute Percentage Error respectively, reflecting the actual magnitude of prediction errors. Lower values indicate higher model accuracy. Analysis indicates that compared to the decision tree regression model, the random forest regression model exhibits significantly lower MSE, RMSE, MAE, and MAPE values in test results. Consequently, the random forest regression model demonstrates superior performance. The gradient boosting tree regression model, whilst achieving zero error on the training set, exhibits substantial error on the test set, indicating severe overfitting. Therefore, the random forest decision model is determined to be the most robust choice.

Subsequently, all datasets were partitioned. To effectively evaluate model performance, 80% of the data was allocated as the training set, with the remaining 20% designated as the test set. Within the random forest model, 100 independent decision trees were constructed. These were trained on the training set using the $fit()$ function, learning to predict foetal abnormalities based on input data. The indicator function formula employed when constructing the 100 independent classification trees is as follows:

$$C(x) = \text{argmax} \sum_{k=1}^{K} I(C_K(x) = c) \quad (9)$$

Where $C(x)$ is the final predicted class of the random forest for sample x, $I()$ is the indicator function yielding 1 if the condition within parentheses is true, otherwise 0, and $C_K(x)$ denotes the predicted class of the kth decision tree for sample x. After multiple training iterations, the model achieved relatively stable results. The classification report is presented first, as shown in Table 9.

Table 9. Classification report

	Precision	Recall	$f_1 -$ score	support
0	0.97	1.00	0.99	105
1	1.00	0.81	0.90	16
Accuracy	-	-	0.98	121
Macro average	0.99	0.91	0.94	121
Weighted average	0.98	0.98	0.97	121

Here, accuracy denotes the proportion of correctly predicted instances, serving as the most intuitive performance metric. Precision denotes the proportion of samples correctly classified as abnormal when the model predicts abnormality. Recall represents the proportion of all genuinely abnormal samples correctly identified as such. The F1-score is the harmonic mean of precision and recall, providing a comprehensive measure of predictive performance. Support indicates the actual number of samples per category in the test set, offering insight into data distribution and enabling assessment of metric reliability.

The confusion matrix further reveals that among the 121 test set samples, 105 genuinely normal samples were correctly predicted as normal, while 13 genuinely abnormal samples were correctly identified as abnormal. However, three genuinely abnormal samples were incorrectly predicted as normal.

Finally, analysis of feature importance indicates that key characteristics influencing detection outcomes include X chromosome concentration; GC content of chromosomes 13, 18, and 21; and age.

5. Conclusions

This study, centred on the objectives of efficient processing and precise analysis of sequencing data, has established a comprehensive technical framework through multi-technique synergistic optimisation. This framework encompasses data pre-processing, feature association modelling, detection timing optimisation, and anomaly determination, effectively resolving key issues in traditional approaches such as insufficient fitting accuracy, coarse timing selection, and weak generalisability of determinations. At the feature association modelling level, incorporating random effects upgraded multiple linear regression to a linear mixed model (LMM). This not only reduced AIC and BIC values by 219.6 and 209.6 respectively but also decreased prediction error (RMSE) by 22%. This breakthrough overcame technical bottlenecks in handling non-independent data and fitting complex associations, significantly enhancing the stability and accuracy of core parameter predictions. Regarding detection timing optimisation, K-means clustering enabled a five-interval refinement of key influencing factors. Combined with the median method to determine optimal timing for each interval, this approach achieved earlier detection adaptation while maintaining data accuracy, providing technical support for process efficiency enhancement. Within the anomaly detection module, the constructed random forest model achieved 98% accuracy on the test set, balancing precision and recall. It effectively identified core influencing features such as X chromosome concentration and GC content on chromosomes 13, 18, and 21. This approach avoids the overfitting issues inherent in gradient-boosted tree models while providing clear direction for subsequent feature engineering optimisation. Overall, the developed technical framework achieves end-to-end coordination from data quality control to result adjudication. Its model optimisation strategy and interval partitioning methodology offer a reusable technical paradigm for similar sequencing data processing tasks, demonstrating strong practical application value.

References

[1] Xiang Chong, Chen Can. Optimisation of Genotype Imputation and Performance Analysis of Regression Models for Low-Depth Sequencing Data [J]. Hubei Agricultural Sciences, 2025, 64(07): 203-206. DOI: 10.14088/j.cnki.issn0439-8114.2025.07.035.

[2] Heng Hongjun, Dai Dongwei. A Multi-time Series Anomaly Detection Method Integrating Sparse Graph Attention [J].

Computer Engineering and Design, 2025, 46(03): 841-849. DOI:10.16208/j. issn1000-7024.2025.03.027.

[3] Zhang Yuhang. Research on Time-Delay Pearson Correlation Analysis and Key Variable Prediction Methods for Air Separation Equipment [D]. Hangzhou Dianzi University, 2023. DOI: 10.27075/d.cnki.ghzdc.2023.000752.

[4] Gai Yujie, Xie Yujiao, Wang Xiaodi. Parameter Estimation for Linear Mixed-Effects Models Based on Online Updating [J]. Applied Probability and Statistics, 2024, 40(03): 420-432.

[5] Xue Lei, Wang Tianfang. K-means Algorithm Based on Adaptive Dynamic Feature Weighting [J]. Journal of Jilin University (Science Edition), 2025, 63(05): 1404-1410. DOI: 10.13413/j.cnki. jdxblxb.2025001.

[6] Song Shijun, Fan Min. Design of a Big Data Anomaly Detection Model Based on the Random Forest Algorithm [J]. Journal of Jilin University (Engineering Science), 2023, 53(09): 2659-2665. DOI: 10.13229/j.cnki.jdxbgxb.20220598.