WEP Warwick Evans Publishing

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

Research on genetic disease identification and classification model based on combined sampling and random forest feature selection

Jianzhang Li*

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China, 215123

* Corresponding Author Email: Jianzhang.Li23@student.xjtlu.edu.cn

Abstract. Machine learning classification models have been widely used in gene identification tasks. However, high-dimensional gene loci and label imbalance are still urgent problems to be solved. To this end, this paper proposes a classification model that combines combinatorial sampling and random forest feature selection. This method applies random forest to randomly selected subsamples for feature selection, aiming to retain effective prediction information and alleviate the impact of the curse of dimensionality. In addition, the random subsamples after feature selection can be used in combination with any classification model. Experiments show that the proposed method is superior to a variety of classic classification models in terms of accuracy and efficiency.

Keywords: Gene Identification; Genetic Disease Testing; Class Imbalance; Curse of Dimensionality.

1. Introduction

Gene loci refer to deoxynucleotides located at specific positions. Variations at these positions are the direct source of DNA diversity, so we call these special deoxynucleotides gene loci. The differences in gene loci lead to biological diversity, and the differences at a single locus are called single nucleotide polymorphisms (SNPs) [1]. In modern medicine, genome-wide association studies (GWAS) identify high-density molecular markers and polymorphic markers for millions or even tens of millions of SNPs, and then examine the relationship between SNPs and specific diseases to screen out pathogenic genes [2]. Obviously, the processing of gene loci data is crucial for genetic research and the prevention and treatment of genetic diseases. In order to efficiently obtain the characteristic information contained in the locus data, it is urgent to quickly identify, detect and process high-dimensional locus data. We can abstract this medical problem into a mathematical classification problem. In previous medical research, Li et al. proved the superiority of fuzzy decision tree models in summarizing medical information system data [3]. Hao et al. used regression models to predict the carcinogenicity of different drugs to humans [4]. These studies show that different classification models have their own advantages in biomedical research. However, in the task of detecting genetic features of genetic diseases, we must face two challenges: imbalanced sample categories and the curse of dimensionality.

For the problem of sample class imbalance, previous studies have adopted a variety of models and algorithms to achieve different purposes. For example, when dealing with the problem of unbalanced time classification of air handling unit data, Huotari et al. first divided the data into a training set and a test set, initialized the training data through a specific number of label arrays, and then obtained the target label from the training set through the GetTarget function and initialized other label arrays. Then, the entire training set was labeled in a loop, the label status of the selected samples was checked, and if the conditions were met, the labels were selected and updated, and finally undersampling was implemented to solve the problem of sample class imbalance [5]. Peng et al. Considered using meta-learning methods to solve the problem of discarding classifier information during data sampling based on undersampling data, thereby improving the stability of data processing [6]. In addition to undersampling, Roweida et al. reviewed the problems caused by unbalanced data sets using a variety of experiments and comprehensively described the advantages and disadvantages of oversampling compared to undersampling in different situations [7]. When faced with the curse of dimensionality, Leygonie et al. compared the conventional dimensionality reduction methods such as PCA in the

multivariate time series classification task, converted the multidimensional data into images, and combined the convolutional neural network (CNN) and long short-term memory network (LSTM) model to extract features, thereby avoiding the information loss and complexity problems caused by the large data dimension [8]. In addition, Aremu et al. analyzed in detail the shortcomings of traditional dimensionality reduction methods in dealing with discontinuous high-dimensional data and proposed a dimensionality reduction framework based on machine learning (ML) to solve the problem of the curse of dimensionality [9]. At the same time, Chandra et al. demonstrated the practicality of clustering models such as Bayesian mixture models in dealing with the curse of dimensionality by studying random partitioning posteriors in non-standard settings with fixed sample size and increased data dimension [10].

Based on this, this paper pioneered the use of a combined sampling method to solve the problem of imbalanced sample data categories, combining randomized samples with features, and using a random forest classification model for feature selection to deal with the dimensionality curse. In addition, the random subsamples after feature selection can be used in combination with any classification model. By adjusting the number of samples, feature ratios, and imbalanced sample ratios, this paper conducted a large number of simulation experiments, and used the gene loci data of osteoarthritis genetic disease as an empirical object to test the effectiveness of the proposed method in practical applications. The experimental results show that the proposed method has higher prediction accuracy than some classic classification models, and can effectively deal with the two problems of sample imbalance and dimensionality curse. In addition, when the submodel is interpretable, we can further analyze the correlation and importance of specific gene loci with osteoarthritis, which provides important support for in-depth research on this genetic disease in medicine and biology.

2. Theory and Methods

In binary classification problems, we usually face problems such as sample category imbalance and high-dimensional features. Specifically, assume that the prediction vector is $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{p})^\mathrm{T}$ and the response variable is $Y = y \in \{0, 1\}$. First, we perform combined sampling on the data to eliminate the sample imbalance problem. When the sample label ratio is extremely unbalanced, we use the oversampling method. For example, the SMOTE algorithm [11] generates new synthetic samples between minority class samples to balance the category distribution. The specific generation formula is:

$$x_{\mathrm{new}} = x_i + \lambda \cdot (x_{NN} - x_i), \quad \lambda \sim U(0,1) \tag{1}$$

Where x_i is the minority class sample and x_{NN} is its nearest neighbor sample. λ is a random number that follows a uniform distribution and controls the position of the new sample. Then, sampling is performed.

$$(X_\mathrm{r}, y_r) = \mathrm{SMOTE}(X^*, y^*) \tag{2}$$

Repeating the above sampling process several times will form several sub-sample data sets, which are denoted as $\{(X_\mathrm{r}, y_r)\}_{\mathrm{r}=1}^{\mathrm{R}}$. When the sample label ratio is generally unbalanced, we use the downsampling method. Downsampling based on clustering algorithm is a method that groups data through clustering algorithm and then selects representative samples from each cluster to reduce the data size. It is mainly used to reduce data redundancy and sample size while retaining the representativeness and diversity of the data set. We first cluster the majority class samples, and the commonly used algorithm is K-means clustering [12]:

2.1. Clustering process

Given a majority class sample set D_M, the goal of the clustering algorithm is to divide D_M into K clusters $C_1, C_2, \ldots, C_K$, where each cluster C_k consists of a set of samples $x_i \in D_M$. K-means clustering completes clustering by iteratively optimizing the following objective function:

$$\min_{C_1,C_2,\ldots,C_K} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{3}$$

Among them, μ_k is the centroid of the kth cluster, that is, the mean of all samples in the cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{4}$$

By minimizing the distance from each sample to the centroid of its cluster, clustering algorithms are able to group similar samples together.

2.2. Select representative samples

After clustering is completed, the number of samples in each cluster C_k is $|C_k|$. For downsampling, we select one or several representative samples from each cluster. Common strategies include:

1. Select centroid sample: You can directly select the centroid μ_k of each cluster as the representative of the cluster.

2. Select the sample closest to the centroid: Select the sample closest to the centroid from each cluster, that is:

$$x_{\text{rep}} = \arg \min_{x_i \in C_k} \|x_i - \mu_k\| \tag{5}$$

3. Random selection: Randomly select a certain number of samples from each cluster to ensure data diversity. By selecting representative samples from each cluster, the size of the final downsampled dataset $D' \subseteq D_M$ is significantly reduced while maintaining the diversity and representativeness of the dataset.

Next, we use the random forest algorithm [13] to select features for each random subsample. Random forest is an ensemble learning method that is mainly used for classification and regression problems. It improves accuracy and robustness by combining multiple decision tree models. Each decision tree recursively divides the data into features until the termination condition is met. The core idea of the decision tree is to maximize information gain, or use Gini impurity to select the optimal partitioning features. For a given sample set D, Gini impurity Gini(D) is defined as:

$$\text{Gini}(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{6}$$

Where p_k is the proportion of samples of the kth class in the sample set, and K is the number of classes. The core idea of random forest is to build multiple decision trees through the bagging method: multiple sample subsets are randomly extracted from the original training data set with replacement, and a decision tree is trained for each subset. Each sample may appear repeatedly in different subsets. Each decision tree grows independently without pruning. Specifically, from the total p features, p_{try} features are randomly selected, the splitting criteria (such as Gini impurity) of each candidate feature is calculated, and the optimal feature is selected:

$$\text{Gini}_{\text{split}}(D) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \tag{7}$$

Ultimately, through these steps, random forests can build multiple different trees, thereby improving the robustness and prediction performance of the model. Finally, each random subsample set after feature selection is brought into the classification model for prediction, and the predicted probability is combined and weighted with the normalized AUC score obtained on the validation set to obtain the final predicted probability for classification. The specific algorithm process of this model is as follows:

Step 1: Randomize the samples using the combined sampling method to form L sub-sample sets $\{X_i, Y_i\}$.

Step 2: Use the random forest model for each sub-sample set $\{X_i, Y_i\}$, generate K random forest models and define each sub-model as M_i, and define each important feature set selected by the model as X_{M_i}.

Step 3: Use the classification model to predict the random sub-sample $\{X_i, Y_i\}$ after feature selection, and use the AUC score on the validation set to perform weighted average of the prediction probabilities of several sub-models to obtain the final prediction probability, and use the corresponding threshold for classification judgment.

3. Data analysis

3.1. Simulation experiment

The experimental data in the simulation experiment of this article are all generated by the function in the sklearn.model_selection module, and multiple adjustable feature value parameters (such as the number of samples, the number of features, the ratio of effective features and the ratio of unbalanced samples, etc.) are set based on this function to facilitate the control of variables for multiple experiments. In addition, for the judgment of the quality of the classification results, this experiment uses the AUC indicator for judgment. As for the AUC indicator, its working principle can be briefly summarized as assuming that there are m positive samples and n negative samples. For a two-classification model, its prediction result can be expressed by a score. For any pair of positive and negative samples, the prediction score of the positive sample isx_i, and the prediction score of the negative sample is y_j. If the model prediction effect is good, then the score of the positive sample should tend to be higher than the score of the negative sample. The formula is expressed as:

$$\mathrm{AUC} = \frac{\sum_{i=1}^{m}\sum_{j}^{n} I(x_i > y_j) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I(x_i = y_j)}{m \times n} \tag{8}$$

Where $I(\cdot)$ is an indicator function, which is 1 when the condition in the brackets holds and 0 when it does not. At the same time, this experiment was repeated 20 times, and the AUC mean and standard deviation were calculated after 20 repeated experiments as the final evaluation results. As for the comparison method of subsequent experiments, we used the random forest model [11] for subsequent experiments. In this experiment, we set the basic parameters as 1000 samples, 100 features, 2 valid features, 15 redundant features, 2 categories, and weights (0.95:0.05). On this basis, we adjusted the number of samples, the number of features, and the parameters related to the weight of the imbalanced sample ratio, conducted multiple experiments, and drew relevant data tables and comparison charts. The experimental results are shown in Table 1-3 and Figure 1-4:

Table 1. Experimental data table with sample size as the control variable

Method	N=600	N=1000	N=1400
ENLM	0.931763 (0.056526)	0.930613 (0.038820)	0.911730 (0.031538)
Logistic model	0.866454 (0.074802)	0.875602 (0.061706)	0.848407 (0.053854)
SVM model	0.830451 (0.087663)	0.851002 (0.069588)	0.836860 (0.057130)
KNN model	0.875535 (0.053982)	0.898235 (0.053792)	0.874909 (0.047134)
RF model	0.872818 (0.086044)	0.883052 (0.057621)	0.858135 (0.063863)
Bayes model	0.917390 (0.064699)	0.924305 (0.038556)	0.910517 (0.029242)

Table 2. Experimental data table with control variables as the number of effective features

Method	p=50	p=100	p=200
ENLM	0.906770 (0.044664)	0.930613 (0.038820)	0.911190 (0.043096)
Logistic model	0.872372 (0.047610)	0.875602 (0.061706)	0.846010 (0.064406)
SVM model	0.859508 (0.067946)	0.851002 (0.069588)	0.829526 (0.066593)
KNN model	0.873636 (0.059311)	0.898235 (0.053792)	0.821997 (0.075261)
RF model	0.855987 (0.060051)	0.883052 (0.057621)	0.851669 (0.089262)
Bayes model	0.896686 (0.050107)	0.924305 (0.038556)	0.905445 (0.041864)

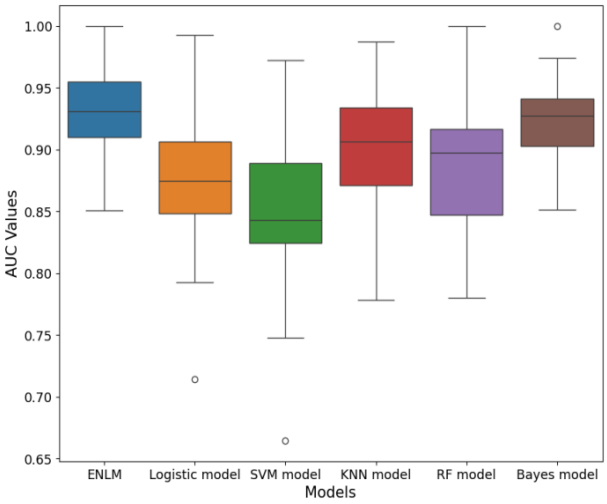

Figure 1. Comparison of AUC mean and standard deviation of each method when N=1000, n=100, weights: 0.95: 0.05

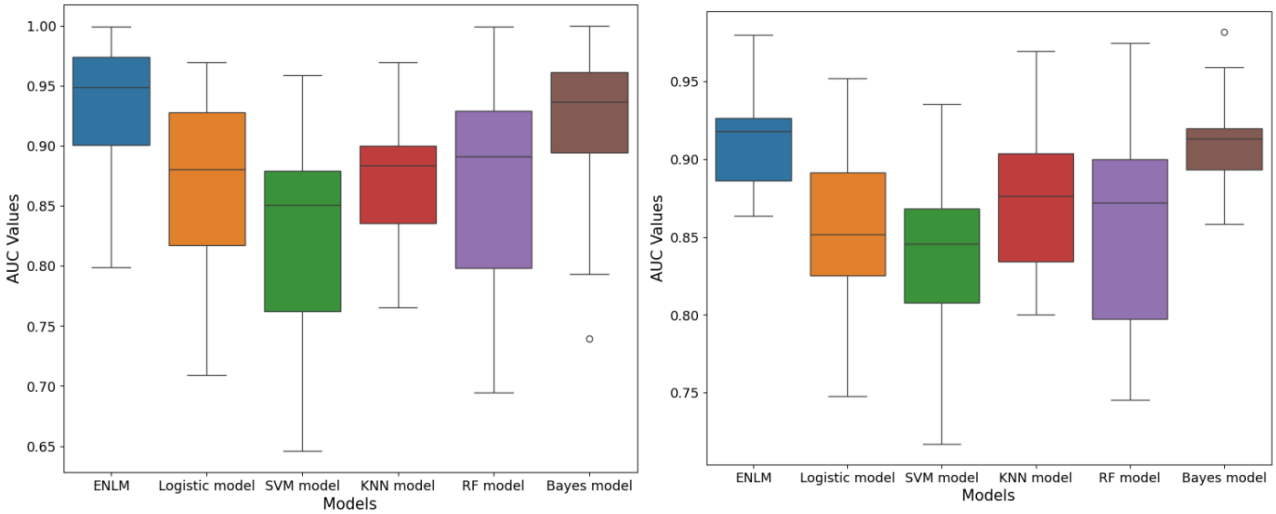

Figure 2. Comparison of AUC mean and standard deviation of each method when N=600 and N=1400

Table 3. Experimental data table with control variables as unbalanced sample proportion weights

Method	0.95: 0.05	0.90: 0.10	0.80: 0.20
ENLM	0.930613 (0.038820)	0.926242 (0.041910)	0.941669 (0.026289)
Logistic model	0.875602 (0.061706)	0.883524 (0.048757)	0.909061 (0.033888)
SVM model	0.851002 (0.069588)	0.868949 (0.049615)	0.901241 (0.033644)
KNN model	0.898235 (0.053792)	0.893154 (0.049834)	0.937218 (0.024584)
RF model	0.883052 (0.057621)	0.900848 (0.057453)	0.932462 (0.025727)
Bayes model	0.924305 (0.038556)	0.921713 (0.040137)	0.939836 (0.026260)

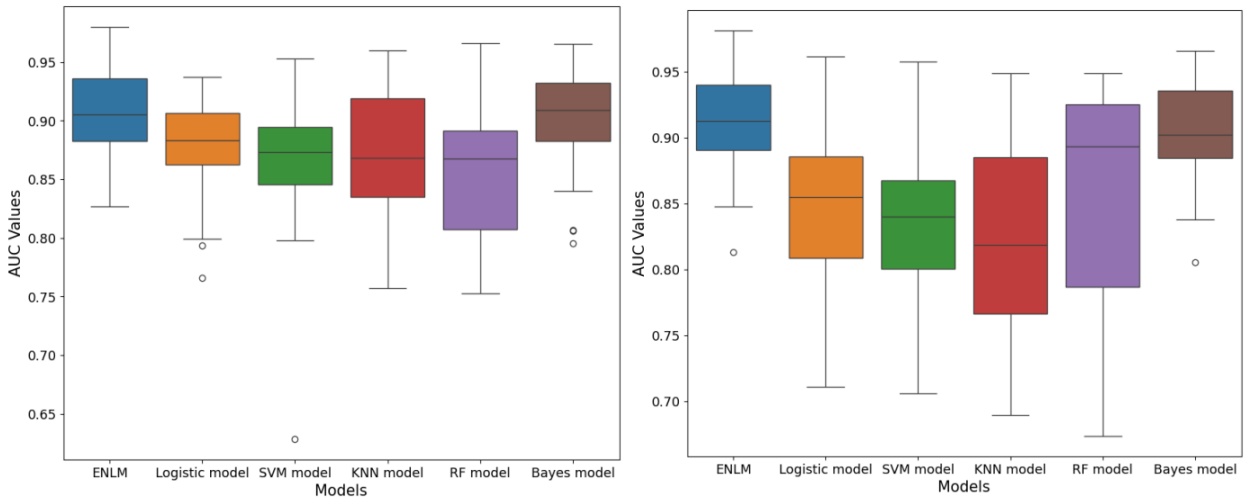

Figure 3. Comparison of mean and standard deviation AUC of various methods when p=50 and p=200

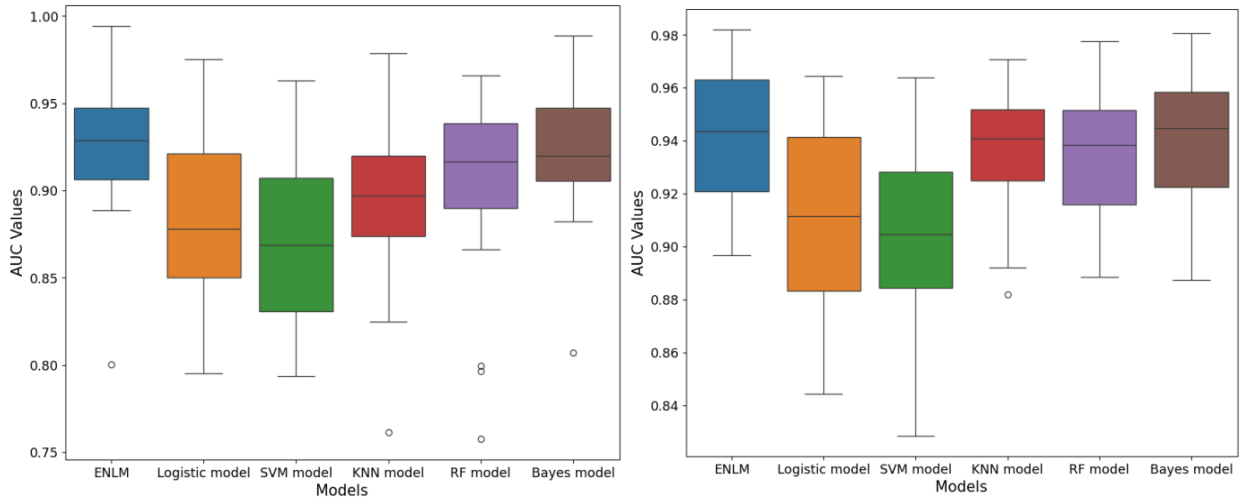

Figure 4. Comparison of AUC mean and standard deviation of each method when the weight is 0.90:0.10 and 0.80:0.20

According to the above charts, the AUC score means and standard deviation of the model proposed in this paper are better than those of other traditional classification models listed in the simulation experiments with different parameters. The chart also shows that no matter what indicator parameters are adjusted, the upper and lower edge numerical difference between the AUC score mean of the model proposed in this paper and the Bayesian model is the smallest, and the median is relatively high, which indicates that in the experimental classification results, the performance gap between the model proposed in this paper and the Bayesian model is the smallest, and it shows high stability and quality. On the other hand, we can observe that no matter how the parameters change, the median AUC scores of the Logistic Regression model and the Support Vector Machine are always low, reflecting that these two traditional classification models have obvious defects in dealing with dimensional disasters and sample category imbalance data. In addition, when processing the simulated experimental data, the classification effect and data accuracy of the model proposed in this paper are significantly better than those of the Logistic Regression model, Support Vector Machine, K Nearest Neighbor Algorithm and Single Random Forest Model, and slightly better than the Bayesian model. This shows that because the model in this paper adopts a combined classification strategy, its data processing accuracy is significantly stronger than other traditional classification models when dealing with sample category imbalance problems. However, due to the small parameter dimension set in the experiment, the advantages of the model in this paper in processing high-dimensional data are not fully reflected. Nevertheless, the Bayesian model's advantage in processing small-scale data makes its performance

in this simulation experiment not much different from that of the model proposed in this paper. In summary, this simulation experiment reflects that the model proposed in this paper can effectively handle and maintain high data accuracy and computational efficiency when facing these problems

3.2. Real data analysis

The genetic data of this experiment comes from the osteoarthritis (OA) microarray dataset GSE48556 in the GEO (Gene Expression Omnibus) database. The samples in the dataset are all human peripheral blood mononuclear cells. There are 139 samples in the GSE48556 dataset, including 106 peripheral blood mononuclear cell samples from OA patients and 33 peripheral blood mononuclear cell samples from normal people. Similar to the aforementioned simulation experiment, the experiment based on the real data used the same evaluation index and comparison method, and repeated the experiment twenty times to calculate the AUC mean and standard deviation. The experimental results are shown in Table 4 and Figures 5:

Table 4. Comparison of AUC mean and standard deviation data of two real experiments

Method	Standard parameter experiment	Adjusted parameter experiment
ENLM	0.946589 (0.028986)	0.946954 (0.030309)
Logistic model	0.838632 (0.051114)	0.834894 (0.051689)
SVM model	0.899323 (0.057750)	0.892625 (0.053245)
KNN model	0.823184 (0.074220)	0.825236 (0.067861)
RF model	0.681896 (0.095250)	0.709392 (0.073597)
Bayes model	0.757372 (0.092221)	0.773345 (0.065405)

In addition, in this experiment, based on the above steps, we used the Logistic regression model as a sub-model to further calculate the importance of each feature for the genetic disease OA. The experimental results are shown in Figures 6. It is found that under the influence of the three evaluation indicators, the gene feature with the greatest positive impact on OA is ILMN_1655595, while the feature with the greatest negative impact on OA is ILMN_1657515. This shows that the gene loci carrying these two features have a significant impact on the risk of OA. Based on this experimental result, we can further carry out causal research on OA. Through the results of simulation experiments and real data experiments, we can find that the classification model based on combined sampling and random forest feature selection proposed in this paper has the advantage of being more efficient and more stable in gene locus feature screening compared with other traditional classification models because it effectively solves the problems of sample class imbalance and dimensionality disaster.

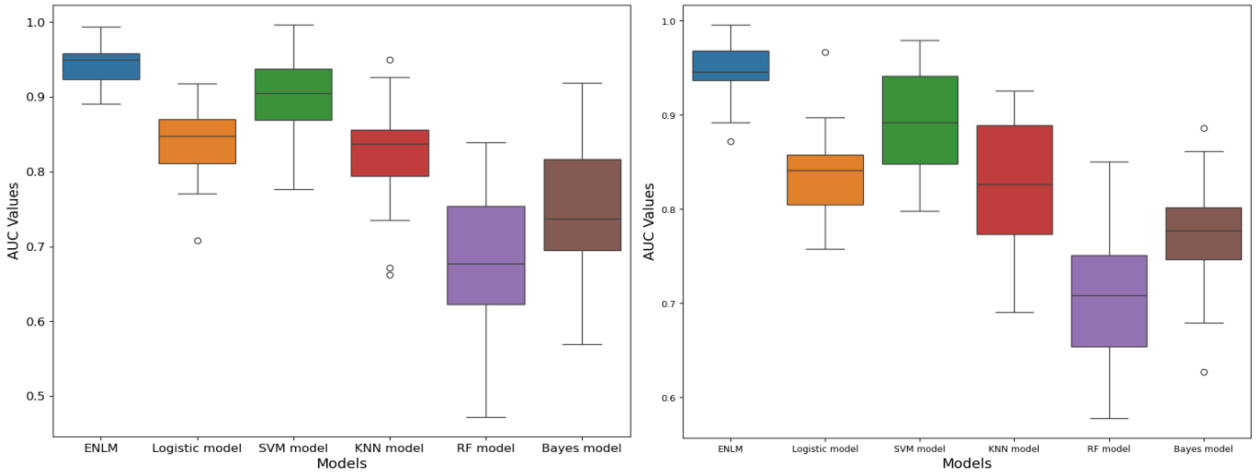

Figure 5. Box plot of standard parameter experimental data and box plot of experimental data after adjustment of parameters

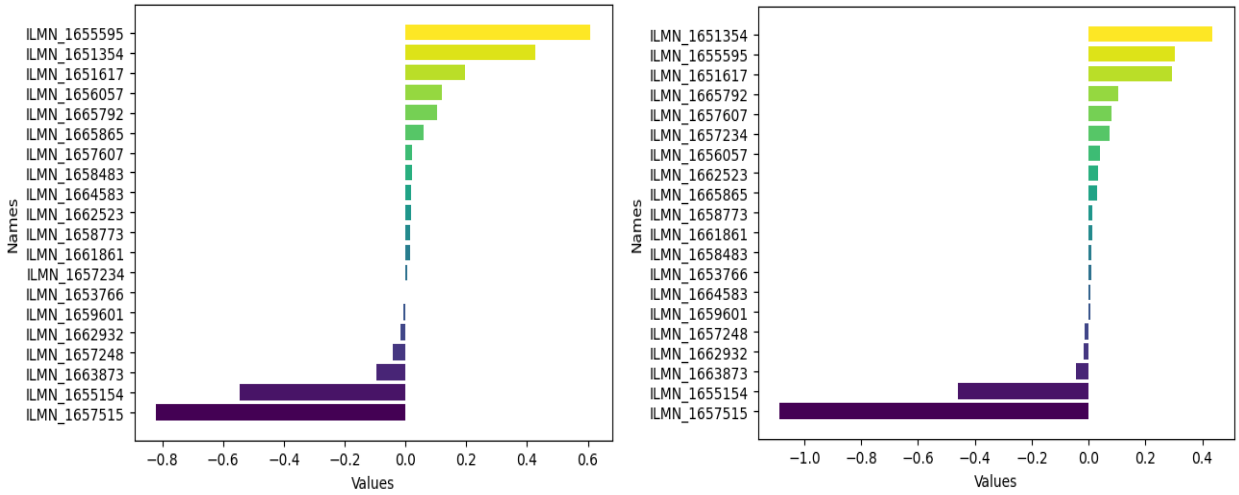

Figure 6. Importance of standard parameter experiment features and importance of experimental features after adjustment

This method can not only quickly and intuitively analyze the degree of influence of different gene features on genetic diseases, but also provide a solid data foundation for studying the relationship between genetic diseases and gene loci. In summary, the method proposed in this paper can provide more accurate data and more detailed feature classification when processing high-dimensional and unbalanced genetic data compared with other commonly used classification models. The sub-model obtained by this method can intuitively reveal the degree of influence of each feature on the specific genetic disease OA. In addition, the sub-model obtained by this method has model freedom, which can facilitate the subsequent use of different models for more in-depth targeted research.

4. Conclusions

The classification model based on combined sampling and random forest feature selection proposed in this paper has successfully solved the two common problems of sample class imbalance and data dimensionality disaster in gene identification after multiple simulation experiments and experiments on real osteoarthritis gene loci data. In the experiment, the AUC score of the proposed model was compared with that of traditional classification models (such as Logistic regression model, support vector machine model, K nearest neighbor algorithm model, etc.), which intuitively demonstrated that the proposed model has faster data processing speed, more stable data dimensionality reduction effect and more accurate feature classification compared with the traditional single classification model in the same data processing task. In addition, this paper innovatively uses the Logistic regression model as a sub-model to further explore the importance of features in each sub-model to OA disease. This means that after processing gene locus data through the proposed model, not only can more stable and accurate experimental results be obtained, but also the degree of influence of features on genetic diseases can be directly identified, which effectively saves the time cost of data processing and provides convenience for subsequent research on genetic diseases.

In addition, this study shows that the freedom of the model provides flexibility for subsequent experiments when dealing with sub-models. Choosing which classification model can make the relationship between features and diseases more accurate has become a new topic worthy of in-depth discussion. At the same time, this article combines combined sampling with the random forest classification model. Whether other feature selection methods can be introduced in the future to obtain more accurate data results is also worthy of further study. On the other hand, the classification model proposed in this article not only has application value in gene identification research, but also shows an efficient classification model selection in the financial field, especially in credit issues, which deserves further consideration and application.

References

[1] Bichuan Liu. Exploration of pathogenic loci of genetic diseases and research on a type of high-dimensional small sample problem [D]. Nanjing: Nanjing Normal University, 2020.

[2] Zhengqiang Li. Screening and identification of gene loci considering interference information [D]. Guangxi: Guangxi Normal University, 2020.

[3] Meng Li. Research on data classification of hospital information system based on improved PSO and fuzzy decision tree [J]. Microcomputer Applications, 2024, 40(09): 194-196+201.

[4] Ning Hao. Machine learning classification and regression model prediction of human carcinogenicity and endocrine disrupting toxicity of typical organic chemicals [D]. Jilin: Jilin University, 2024.

[5] Huotari M, Främling K. Event Classification with Imbalanced and Missing Data for an Air-Handling Unit[C]//2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI). IEEE, 2022: 82-86.

[6] Peng M, Zhang Q et al. Trainable undersampling for class-imbalance learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 4707-4714.

[7] Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results[C]//2020 11th international conference on information and communication systems (ICICS). IEEE, 2020: 243-248.

[8] Leygonie R, Lobry S, Vimont G, et al. Transforming Multidimensional Data into Images to Overcome the Curse of Dimensionality[C]//2023 IEEE International Conference on Image Processing (ICIP). IEEE, 2023: 700-704.

[9] Aremu O O, Hyland-Wood D, McAree P R. A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data[J]. Reliability Engineering & System Safety, 2020, 195: 106706.

[10] Chandra N K, Canale A, Dunson D B. Esca** the curse of dimensionality in Bayesian model-based clustering[J]. Journal of machine learning research, 2023, 24(144): 1-42.

[11] Elreedy D, Atiya A F. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance[J]. Information Sciences, 2019, 505: 32-64.

[12] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.

[13] Parmar A, Katariya R, Patel V. A review on random forest: An ensemble classifier[C]//International conference on intelligent data communication technologies and internet of things (ICICI) 2018. Springer International Publishing, 2019: 758-763.