Hindawi
Complexity
Volume 2023, Article ID 9825382, 1 page
https://doi.org/10.1155/2023/9825382

Retraction

Retracted: Research on Complex Classification Algorithm of Breast Cancer Chip Based on SVM-RFE Gene Feature Screening

Complexity

Received 22 August 2023; Accepted 22 August 2023; Published 23 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] G. Chen, X. Xie, and S. Li, "Research on Complex Classification Algorithm of Breast Cancer Chip Based on SVM-RFE Gene Feature Screening," *Complexity*, vol. 2020, Article ID 1342874, 12 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 1342874, 12 pages
https://doi.org/10.1155/2020/1342874

Research Article

Research on Complex Classification Algorithm of Breast Cancer Chip Based on SVM-RFE Gene Feature Screening

Guobin Chen,[1] Xianzhong Xie,[1] and Shijin Li[2]

[1]*College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China*
[2]*Academic Affairs Office, Yunnan University of Finance and Economics, Kunming 650221, Yunnan, China*

Correspondence should be addressed to Shijin Li; shijin_lee@ynufe.edu.cn

Received 1 April 2020; Revised 10 May 2020; Accepted 23 May 2020; Published 13 June 2020

Guest Editor: Zhihan Lv

Screening and classification of characteristic genes is a complex classification problem, and the characteristic sequences of gene expression show high-dimensional characteristics. How to select an effective gene screening algorithm is the main problem to be solved by analyzing gene chips. The combination of KNN, SVM, and SVM-RFE is selected to screen complex classification problems, and a new method to solve complex classification problems is provided. In the process of gene chip pretreatment, LogFC and P value equivalents in the gene expression matrix are screened, and different gene features are screened, and then SVM-RFE algorithm is used to sort and screen genes. Firstly, the characteristics of gene chips are analyzed and the number between probes and genes is counted. Clustering analysis among each sample and PCA classification analysis of different samples are carried out. Secondly, the basic algorithms of SVM and KNN are tested, and the important indexes such as error rate and accuracy rate of the algorithms are tested to obtain the optimal parameters. Finally, the performance indexes of accuracy, precision, recall, and F1 of several complex classification algorithms are compared through the complex classification of SVM, KNN, KNN-PCA, SVM-PCA, SVM-RFE-SVM, and SVM-RFE-KNN at $P = 0.01, 0.05, 0.001$. SVM-RFE-SVM has the best classification effect and can be used as a gene chip classification algorithm to analyze the characteristics of genes.

1. Introduction

Since the birth of gene chip technology, a large number of feature selection methods for gene expression microarray data have emerged in academia. Most of these methods focus on the quality of the selected genes, while few people pay attention to the efficiency of the algorithm itself. Gene expression microarray data has a large number of characteristic genes. If it is not an efficient characteristic selection method, the whole process of key gene selection will become very long. Many existing classical feature selection methods have low efficiency, and some even reach unacceptable levels. Among them, representatives include CFS, mRMR, and SVM-RFE. Especially in SVM-RFE, the whole selection process is very time-consuming. Some researchers have also improved SVM-RFE, but the time-consuming problem has not been fundamentally improved. This chapter takes SVM-RFE as the research object, and SVM and RFE are improved, respectively. By introducing a more efficient implementation of the classical linear support vector machine to reduce the time consumption of the basic feature selection process and proposing a recursive feature elimination strategy with variable step size to reduce the iteration times of the basic feature selection process, the combination of the two finally attempts to fundamentally solve the inefficiency problem of SVM-RFE.

A support vector machine based on recursive feature eligibility (SVM-RFE) [1] was proposed by Guyon et al. In 2002. This method makes full use of the characteristics of SVM; that is, it can rank and score all genes according to their importance while training the SVM classification method and combine the recursive feature elimination strategy to make feature selection. Duan et al. [2] improved SVM-RFE to deal with only two classification problems and proposed a one-to-one and one-to-many multiclassification SVM-RFE method, which enables SVM-RFE to deal with multiclassification problems.

Aiming at the low efficiency of SVM-RFE feature selection process, Ding and Wilkins [3] improved the iterative process of RFE, from deleting one feature at a time to deleting several, which improved the efficiency of the algorithm without losing the classification accuracy. Yoon and Kim [4] proposed a SVM-RFE method based on mutual information, which solves the problem that the SVM-RFE method does not consider feature correlation in the process of feature selection to a certain extent. Tang et al. [5] divided SVM-RFE into two stages. In the first stage, rough selection is carried out on the features to be selected, filtering out irrelevant features, redundant features, and noise features. In the second stage, finer feature selection is carried out on the basis of the first stage. The next two sections will introduce SVM-RFE in detail and analyze the reasons for its inefficiency in depth. Tang et al. [6] feature clustering SVM-RFE (FCSVM-RFE) feature clustering to enhance SVM-RFE gene selection. The proposed method first roughly selects genes and then ranks the selected genes. Clustering algorithm is used to cluster genes into genomes, in which each gene has a similar expression profile. SVM-RFE was used to rank these representative genes. FCSVM-RFE reduces computational complexity and redundancy. Although SVM-RFE can effectively delete irrelevant functions, it cannot handle most redundant functions [7]. In order to overcome this shortcoming, this paper develops a new feature selection method, the core of which is to delete redundant features according to the correlation between features before using SVM-RFE. The proposed method was tested on a pancreatic cancer microarray dataset. The method is much better than baseline SVM-RFE in classification accuracy. In order to improve the accuracy of classification, radial basis function (RBF) kernel is also introduced [8]. Chen and Zhu [9] proposed a feature selection method based on support vector machine recursive feature elimination (SVM-RFE) and binary particle swarm optimization (BPSO) algorithm. SVM-RFE removes some irrelevant features to reduce the data dimension and then continues to search for the best subset and uses some better SVM-RFE subsets as part of the initial PSO population and has a good starting point. SVM-RFE not only reduces the search space of particles but also provides prior experience, thus improving the search efficiency and accuracy of the algorithm. Anaissi et al. [10] used ESVM recursive feature elimination (ESVM-RFE) for gene selection. It follows the concepts of integration and bagging used in random forest but adopts backward elimination strategy, which is the basic principle of RFE algorithm. The principle behind this is that using randomly drawn boot program samples from the training set to build an integrated SVM model will generate different feature levels, which will then be summarized into one feature level. As a result, the decision to delete features is based on the ranking of multiple SVM models, rather than selecting a specific model. However, in the classification of unbalanced datasets, imbalance is a common problem in gene expression microarray data [11]. Generally speaking, people are only interested in a few categories because the few categories are usually patients, while normal people often account for the majority. For the classification method, too few samples in a certain category means that the category contains less information, so the classification model finally learned by the classification algorithm can easily predict patients among normal people when making classification prediction [12]. Especially for small sample data such as gene expression microarray data, it becomes more important to solve the problem of category imbalance. The most basic methods to solve the problem of category imbalance are upsampling and downsampling. Zhou and Wang [13] proposed a feature selection method combining relief-F and SVM-RFE algorithm. The algorithm integrates the weight vector from relief-F into the SVM-RFE method. In this method, relief-F filters out many noisy functions in the first stage. Then, a new sorting criterion based on the SVM-RFE method is applied to the final feature subset. A SVM classifier is used to evaluate the final image classification accuracy. A new method for multiclass gene selection and classification based on multiple supports vector machines recursive feature elimination (SVM-RFE) is proposed [14]. For the multiclass DNA microarray problem, we solve it as a multibinary classification problem. The "one-to-all" method is used to decompose multiple types of tasks into multiple binary problems, and SVM-RFE selects genes for each binary problem. The SVM classifier is used to train selected gene data for binary problems. Firstly, the basic method of SVM is introduced, and the application of RFE algorithm is explained in detail. Secondly, the chip GSE76275 screens and classifies different P values under SVM-RFE algorithm. Finally, the classification effect of SVM-RFE algorithm after filtering with different P values is illustrated by comparative research under different SVM-RFE-KNN, SVM-RFE-SVM, and other four algorithms.

2. Relevant Theoretical Works

2.1. Support Vector Machine. A support vector machine (SVM) is recognized as one of the most classical machine learning algorithms. Its essence is the maximum interval classification method. At this time, a support vector machine can only deal with linearly separable data classification problems and is called hard interval support vector machine. Soft interval support vector machine was proposed in 1995. At this time, a support vector machine can deal with data classification problems that are approximately linearly separable. Subsequently, support vector machines have been further developed. Support vector machines, support vector regression machines, and multiclassification support vector machines based on kernel techniques have been proposed one after another. At this time, the support vector machine has formed a very complex and complete theoretical system, which can not only deal with linear separable problems but also classify nonlinear separable data, becoming very powerful. The SVM-RFE algorithm uses a support vector machine based on linear kernel. The support vector machine model is shown in Figure 1.

The algorithm idea of the support vector machine is actually very simple. For hard interval support vector machines, the whole process is divided into three steps: first, the dataset is linearly separable; second, finding two hyperplanes requires that no data points fall between the two planes.

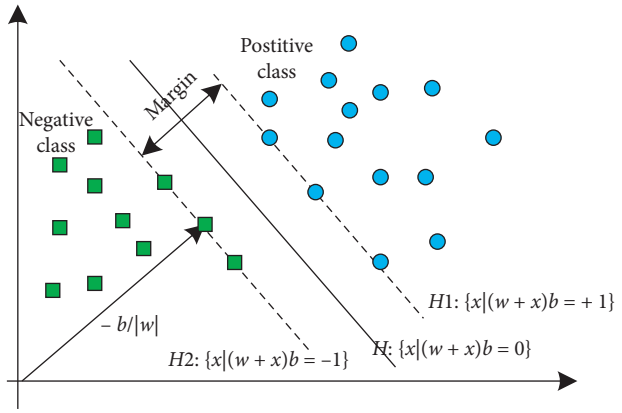

Figure 1: Support vector machine model.

Third, maximize the distance between the two planes. The objective function at this time is

$$\max \frac{2\eta}{\|w\|} \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq \eta, \quad i = 1, \ldots, n, \tag{1}$$

where $(2\eta/\|w\|)$ is the distance between the two hyperplanes and is the target optimization value, x_i and y_i represent the i-th sample and the corresponding label, respectively, and $y_i(w \cdot x_i + b)$ represents the distance from the point x_i to the nearest hyperplane. $y_i(w \cdot x_i + b) \geq \eta$ means that point x_i cannot fall between two hyperplanes. In order to facilitate the solution, formula (1) is usually transformed into a quadratic programming problem:

$$\min \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, \ldots, n. \tag{2}$$

In all classifications, the classification interval of the optimal plane is the largest; at this time, $\|w\|^2$ is the smallest, H is called the optimal classification line, and the training samples on H1 and H2 are called support vectors. The Lagrange optimization method is used to obtain it. Assuming that $b = (b_1, b_2, \ldots, b_n)$ and equation (2) constitute Lagrange multiple terms, the maximum value is taken.

$$W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{j=1}^{n} a_i a_j y_i y_j x_i x_j, \tag{3}$$

where $a_i \geq 1, \sum_{i=1}^{n} y_i a_i = 0$ is quadratic programming that can optimize equation (3). Assuming that there is a maximum vector $a^0 = (a_1^0, a_2^0, \ldots, a_n^0)$ of equation (3) and the optimal hyperplane is described by (w_0, b_0), then w_0 is shown in the following equation:

$$w_0 = \sum_{i=1}^{n} a_i^0 y_i x_i. \tag{4}$$

If the restriction condition is proposed in equation (4), the decision function of the optimal classification is shown in in the following equation:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} a_i^0 y_i x_i + b_0\right) = 0. \tag{5}$$

Equation (5) introduces Lagrange's equation:

$$L(w, b, a) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{n} a_i y_i (w x_i + b) - 1, \tag{6}$$

where a is the Lagrange coefficient. By differentiating w and b, we obtain the quadratic programming problem:

$$\begin{cases} \min & \frac{1}{2}\sum_{j=1}^{n} a_i a_j y_i y_j \left[\phi(x_i) \cdot \phi(x_j)\right] - \sum_{i=1}^{n} a_i \\ \text{s.t.} & \sum_{i=1}^{n} y_i a_i = 0 \\ & a_i \geq 0, \quad (i = 1, 2, \ldots, n). \end{cases} \tag{7}$$

2.2. Recursive Feature Elimination (RFE). The main idea of recursive feature elimination is to repeatedly build models (such as SVM or regression models). The importance of each feature is obtained through the attribute value returned by the learner or the importance score of the feature. Then, the least important feature variable is removed from the current feature set. Then, the model is constructed on the remaining characteristic variables. Repeat the abovementioned process until there is only one feature variable left. This process constructs a model of feature number minus one time. The order in which features are eliminated is the importance ranking of features. This is a greedy algorithm to find the optimal feature subset, which requires a lot of computation and requires high hardware requirements of computers. The stability of RFE depends to a large extent on which a model is selected at the bottom during iteration. For example, if the ordinary linear regression adopted by RFE is unstable without regularization, then RFE is unstable. If a ridge is used and the regression regularized by the ridge is stable, then RFE is stable. For example, a linear kernel support vector machine SVM-RFE, as an effective feature selection method, has been successfully applied to fault diagnosis. However, some problems may be nonlinear.

SVM-RFE is a supervised sequential backward selection algorithm. In the linear classifier, it takes the discriminant information of each feature to the objective function as the sorting coefficient. That is, the contribution of the weight vector to the classification surface $y_i(w \cdot x_i + b)$ is used to construct the feature ranking table. If the weight corresponding to the feature is larger, the decision function will be affected more, and the feature with larger weight has more discrimination information. Each iterate removes a feature with the smallest weight and then retrains the classifier until the feature ranking table is completed. The sorting principle can also be analyzed from the objective function of the following formula:

Input: the training samples: $X_0 = [x_1, x_2, ..., x_i]^T$, x_i is the samples of d-dimensional space
Category label: $y = [y_1, y_2, ..., y_i]^T$
Initialization: feature sort $r = []$, current feature index sequence $s = [1, 2, \ldots, d]$
Feature sorting: iterates in a loop until $s = []$;
Step 1: obtain a new data sample according to the current feature: $X = X_0(:, S)$
Step 2: train SVM with a new sample set to obtain support vector related parameters: $a = \text{SVMtrain}(X, y), X = X_0(:, s)$
Step 3: calculate the sorting factor
Step 4: find out the feature $f = \arg\min(\text{Rank}(i))$ with the smallest sorting criterion and add it to the feature sorting table: $r = [s(f), r]$
Step 5: remove the feature with the smallest sorting coefficient from the current remaining dataset: $s = s(1: f-1, f+1: \text{length}(s))$
Output: sorted list of features.

Algorithm 1: SVM-RFE algorithm flow.

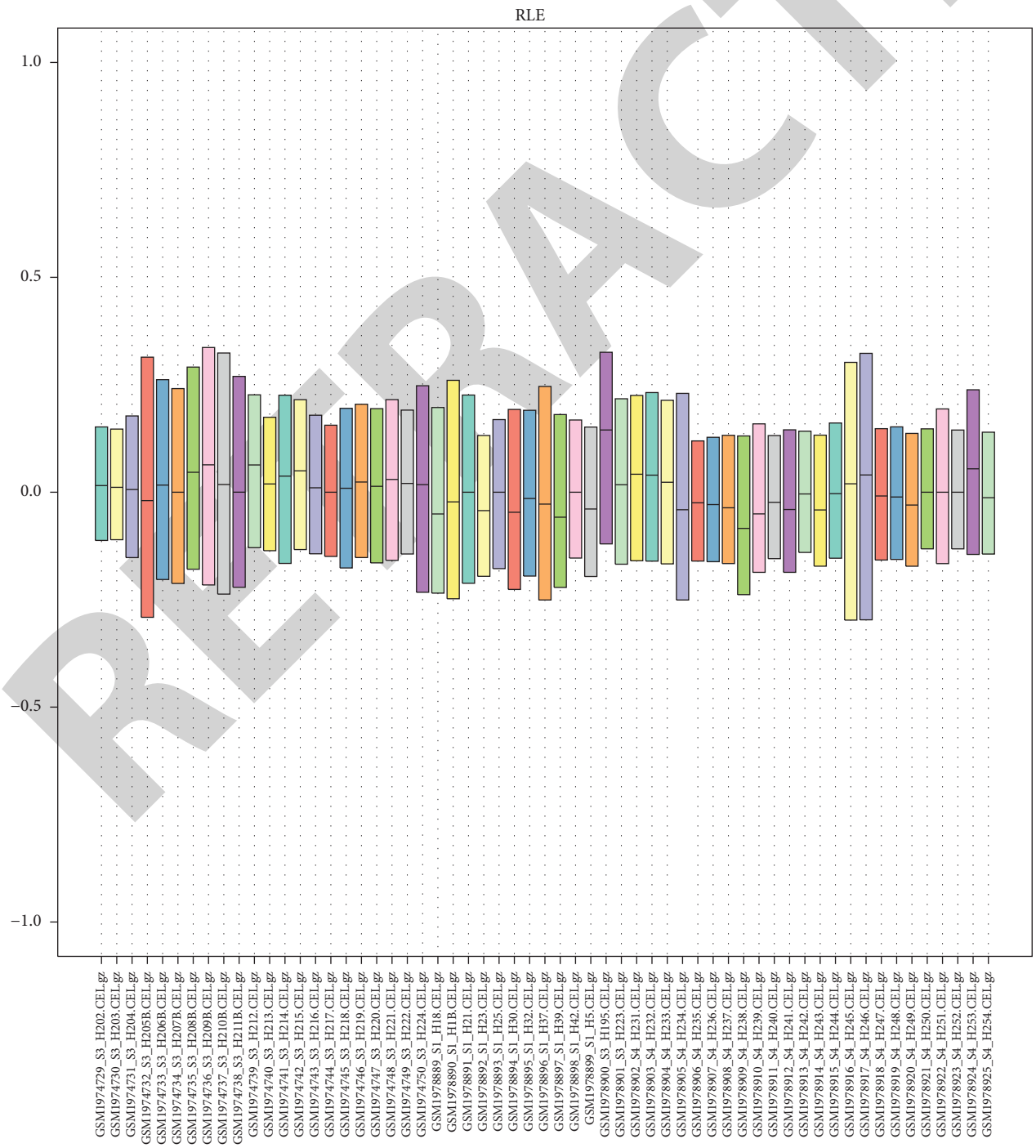

Figure 2: RLE box chart.

Table 1: Probes number and genes number under different P values.

	$P \le 0.1$	$P \le 0.05$	$P \le 0.01$	$P \le 0.001$
Probes number	24017	6251	6022	5370
Genes number	16383	5561	5361	4792

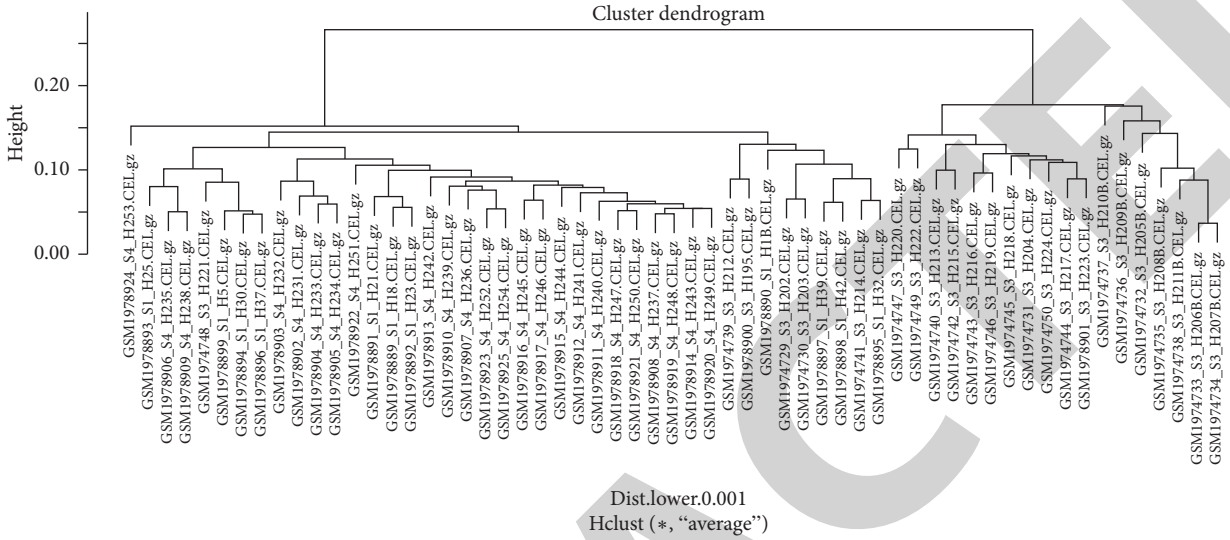

Figure 3: $P < 0.0001$ chip cluster diagram.

$$J = \frac{1}{2}\|w\|^2. \tag{8}$$

Calculating that the i-th feature removal is the change of J,

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2, \tag{9}$$

where w_i also means that the i-th feature is deleted; so, as a sorting criterion, the later the feature means that the less information it contains, and the more it will be deleted first. The algorithm is a circular process.

The classic is a linear kernel function with sorting coefficients of

$$\begin{cases} \text{Rank}(i) = (w_i)^2, \\ w = \sum_{i=1}^{l} a_i y_i x_i. \end{cases} \tag{10}$$

In the case of nonlinearity, it is assumed that in the training sample matrix, when a certain feature is removed, the median of quadratic programming remains unchanged; that is, the obtained classifier does not change. On the premise of this assumption, the contribution value of each feature to the objective function, i.e., the ranking coefficient, is

$$\begin{cases} \text{Rank}(i) = \frac{1}{2}a^T Q a - \frac{1}{2}a^T Q(-i)a, \\ Q_{ij} = K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j). \end{cases} \tag{11}$$

This assumption is also reasonable and feasible in practical application, where $a = [a_1, a_2, ..., a_l]$, $Q(-i)$, means the Q matrix value calculated when the i-th feature is assumed to be removed. In practical applications, nonlinear kernels and linear kernels often produce similar results. The SVM-RFE method executes this process iteratively and finally obtains a feature sorting table. Using the sorting list, several nested feature subsets are defined to train SVM, and the advantages and disadvantages of these subsets are evaluated according to the prediction accuracy of SVM, thus obtaining the optimal feature subset. It should be noted that the single feature in the front row does not necessarily make the SVM classifier to obtain the best classification performance, but the combination of multiple features makes the classifier to obtain the best classification performance. Therefore, SVM-RFE algorithm can select complementary feature combinations. The objects targeted by the two formulas are different, corresponding to linear and nonlinear kernels, respectively, but in fact the difference in the final selection of eigenvalues is not obvious.

SVM-RFE algorithm can define a set of nested feature subsets $F_1 \subset F_2 \subset F_3 \ldots F_n$ according to its feature sorting table. The prediction accuracy of SVM is used to evaluate the advantages and disadvantages of these subsets, so as to obtain the optimal feature subset. $F_i (i = 1, 2, ..., n)$ means that i-th features with the highest ranking are selected from the feature set as subsets to ensure that each subset contains features with relatively important information, and then the classifier is designed with the selected optimal subset. The algorithm is as follows in Algorithm 1.

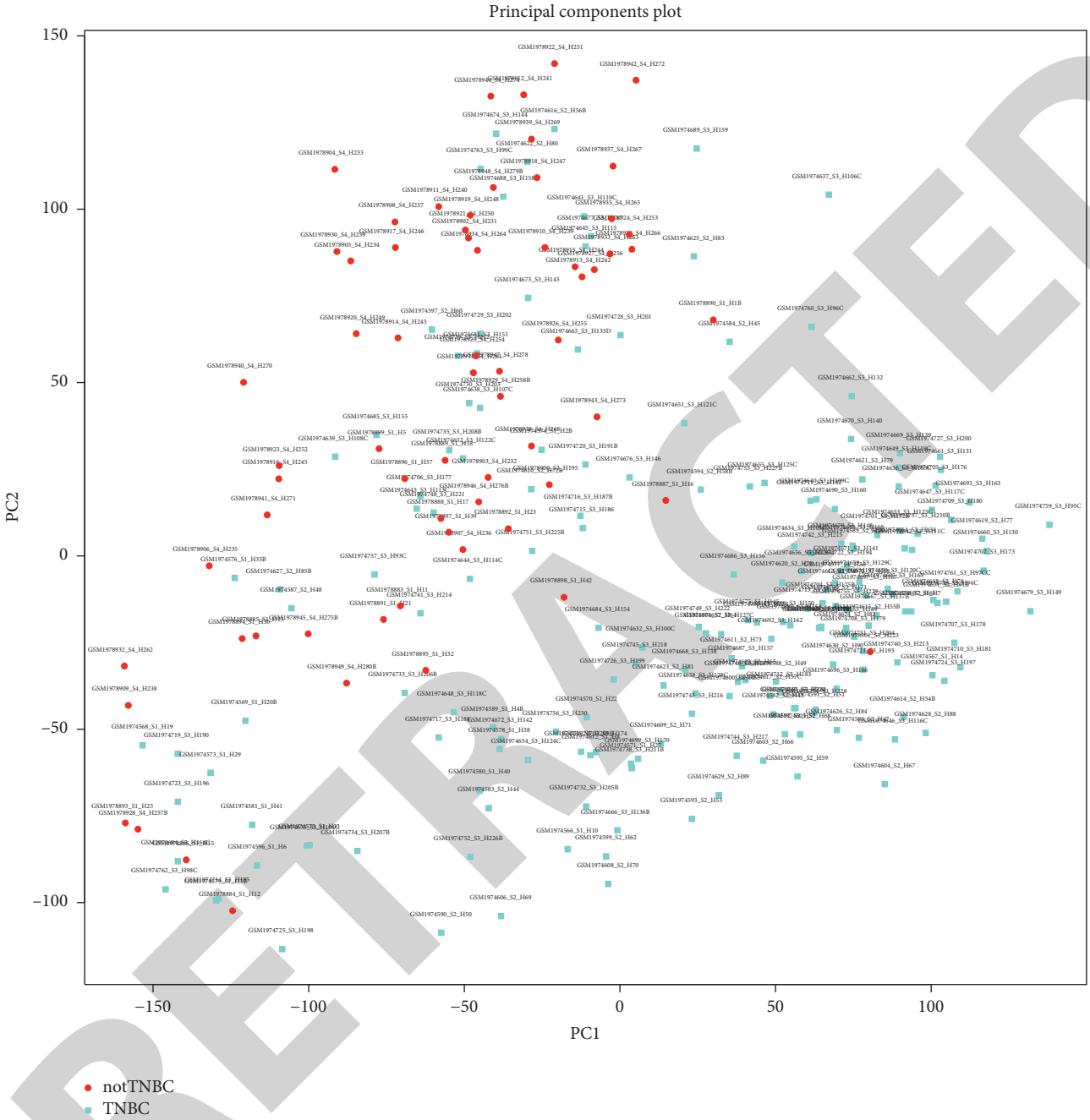

Figure 4: PCA differential expression diagram.

3. Result Analysis

In this paper, the gene chip GSE76275 is used as the research basis and the relevant parameters of the chip are described. The GSE76275 dataset contains 265 samples, including 198 TNBC and 67 non_TNBC, with a total of 54613 gene expression values. In the experiment, the relevant basic data are analyzed uniformly, and the expression level of most genes can be kept consistent. The relative logarithmic expression (RLE) box chart can reflect the abovementioned trend. It is defined as the logarithm of the expression value of a probe group in a certain sample divided by the median of the expression value of the probe group in all samples. The distribution of RLE of all probe groups in a sample can be represented by a box chart commonly used in statistics, and the center of each sample should be very close to the position of ordinate 0. The RLE box chart of this experiment meets this requirement, as shown in Figure 2.

Normalization processing: the purpose of normalization is to enable each group of measurements or measurements under experimental conditions to compare with each other and eliminate nonexperimental differences between measurements, which may come from sample preparation, hybridization process, or

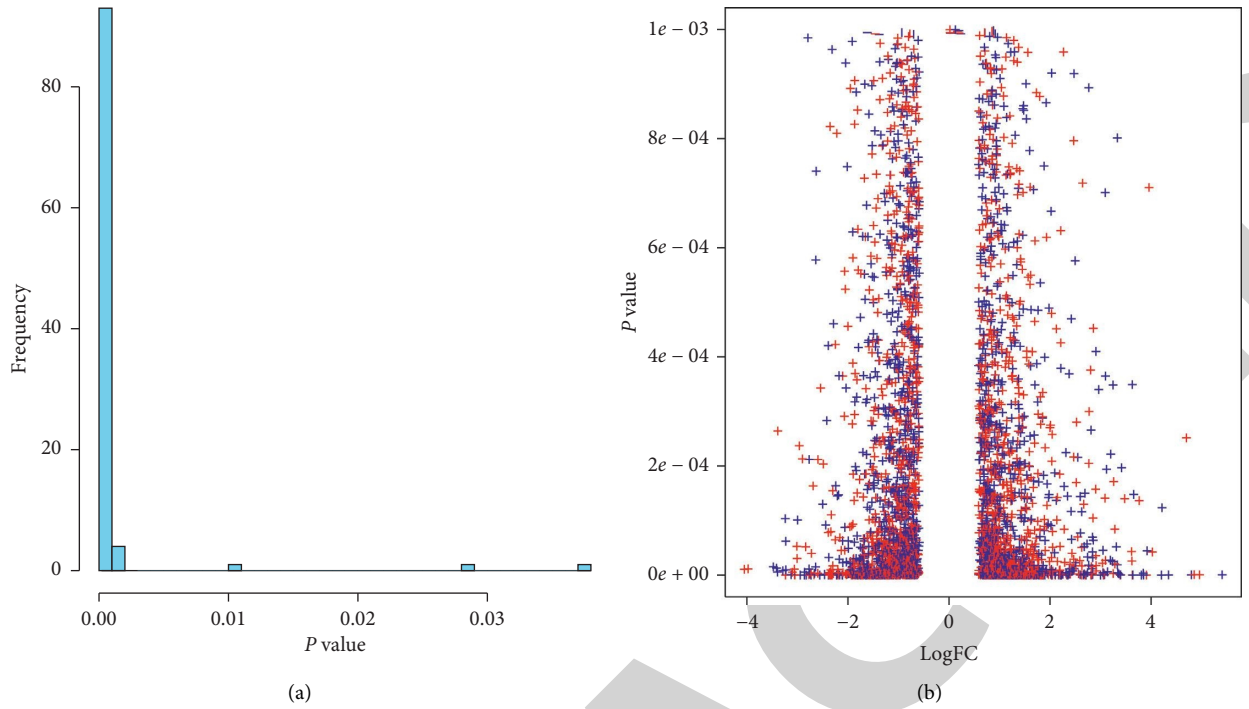

Figure 5: The genes number distribution of $P < 0.001$ and the correlation LogFC and P value.

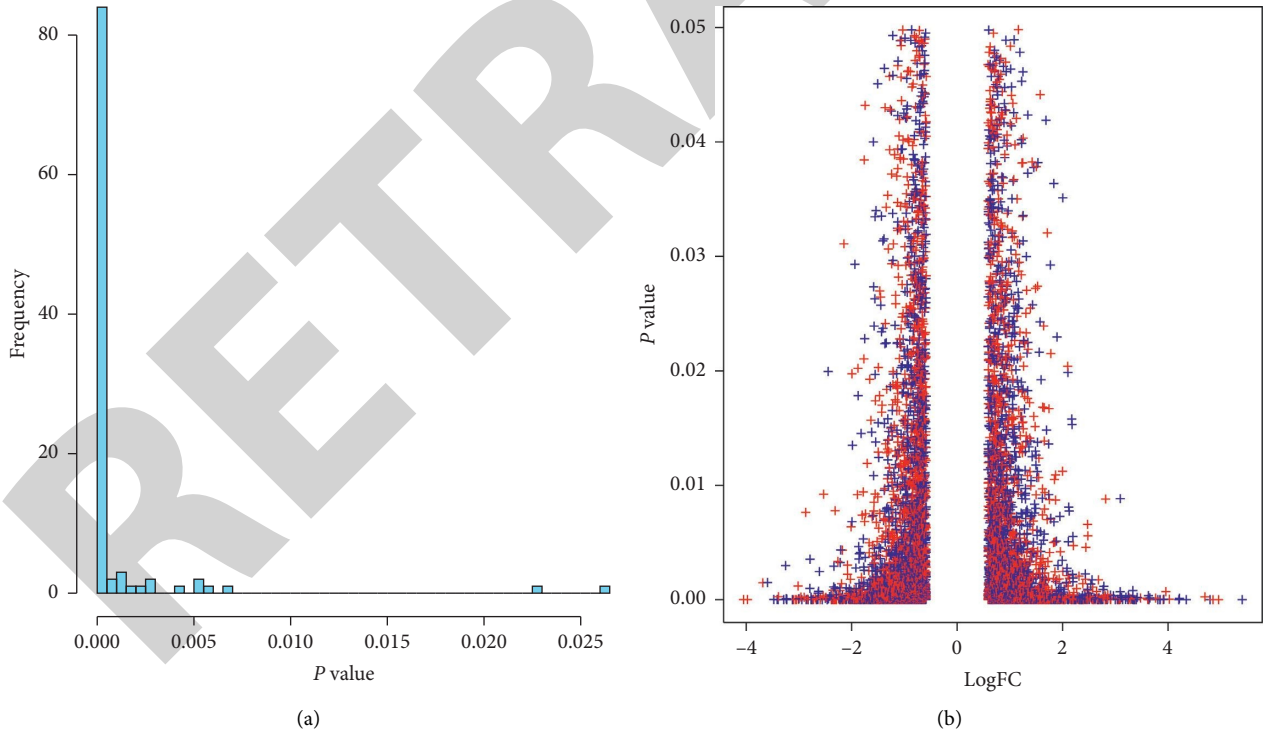

Figure 6: The genes number distribution of $P < 0.05$ and the correlation LogFC and P value.

hybridization signal processing. The process to normalize the abovementioned data can be implemented by the expresso function in the affy software package. In fact, the integrated algorithm using preset parameters is more reasonable and efficient.

3.1. Differentially Expressed Genes Selected. The first step in the significance analysis of gene expression differences is to select and express genes with significant differences. Generally speaking, the basic assumption of this kind of analysis is that the standardized chip data conform to normal

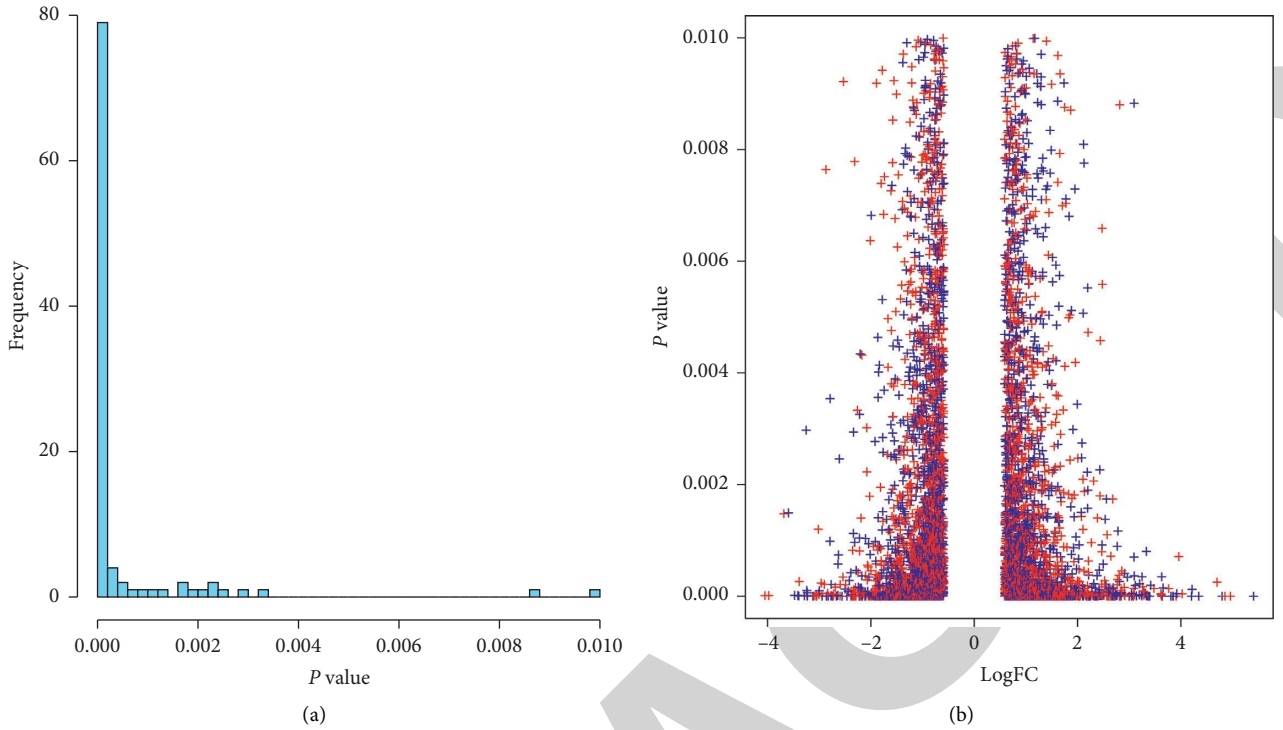

Figure 7: The genes number distribution of $P < 0.01$ and the correlation LogFC and P value.

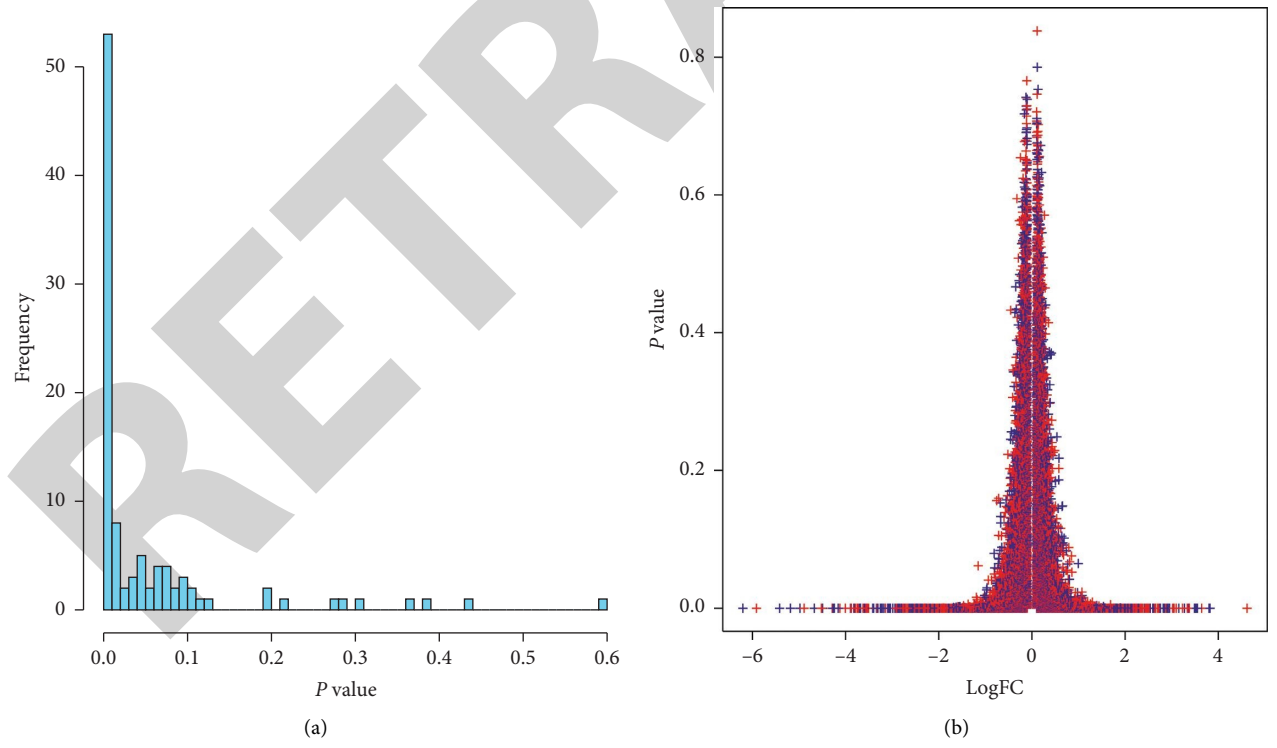

Figure 8: The genes number distribution of $P < 0.1$ and the correlation LogFC and P value.

distribution, so the statistical methods used are basically T test, F test, variance analysis, and the improved forms of these three statistical methods. In order to obtain standardized differential genes, the gene chip adopts the Bayesian method. Empirical Bayesian method is currently the most commonly used analysis method, which has been completely implemented by limma package of Bioconductor, as shown in Table 1.

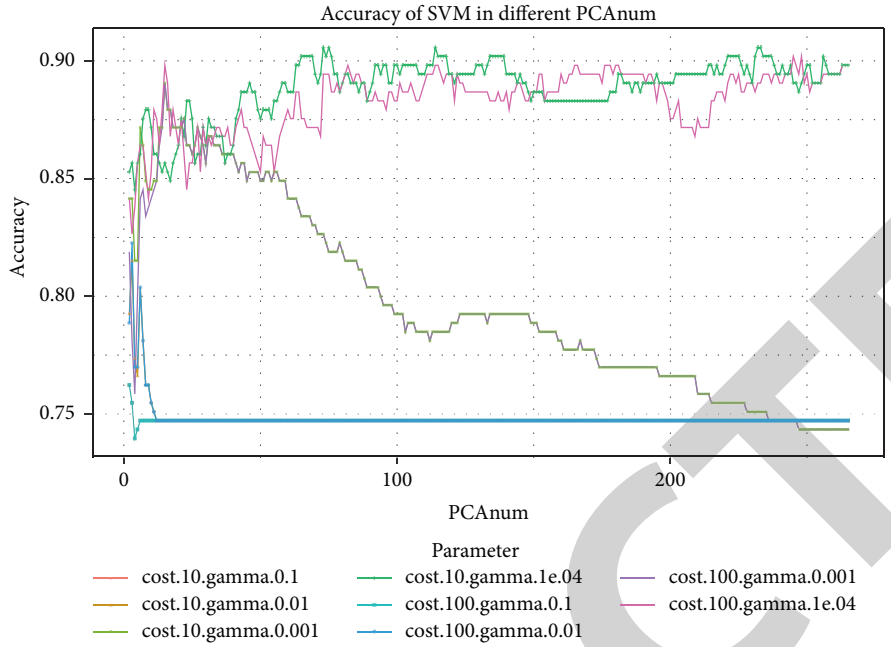

Figure 9: Comparison of correct rate of SVM optimization parameters.

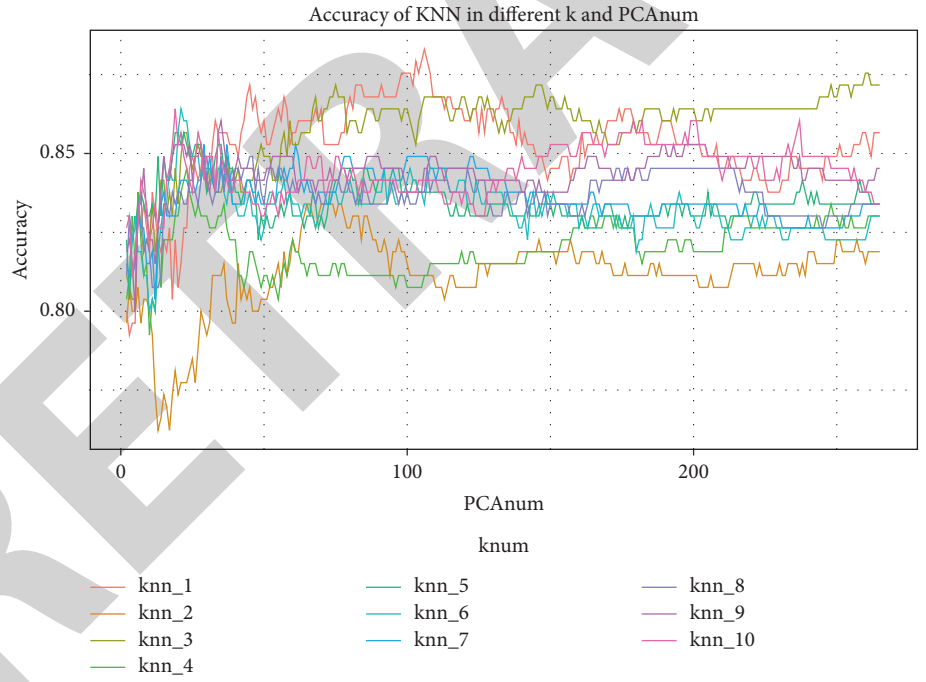

Figure 10: Comparison of correct rate of KNN optimization parameters.

The gene clustering of different genes with $P < 0.001$ was analyzed separately. Some samples are selected for clustering, and the same type of samples can be basically clustered together, as shown in Figure 3.

Select the differential gene expression data of sample $P < 0.001$ and make a PCA diagram. From the diagram, it can be seen that the classification of the two groups of samples is obvious; thus, it can be seen that the two types of samples have obvious differences, as shown in Figure 4.

When P is at different values, SVM-RFE shows the difference of screening genes. When P values are 0.001, 0.05, 0.01 and 0.1, the number and distribution of genes are shown in Figures 5–8:

As can be seen from the above Figures 5–8, when the maximum value of P becomes larger and larger, the number of genes distributed becomes more and more. However, there is a certain correlation between LogFC and P value. Most of the points are published between [−2, 2], which

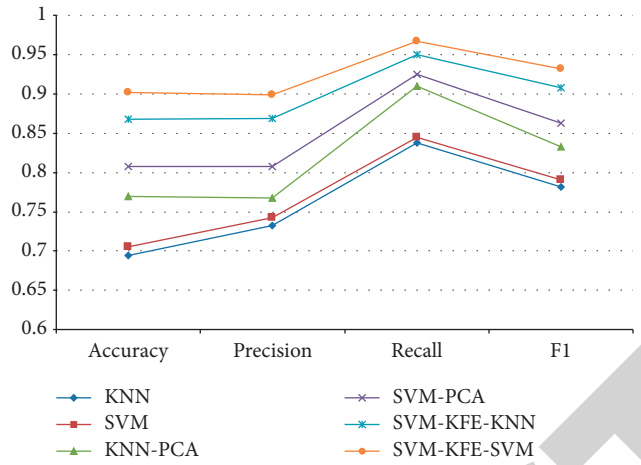

Figure 11: The comparison of algorithm performance when $P \leq 0.05$.

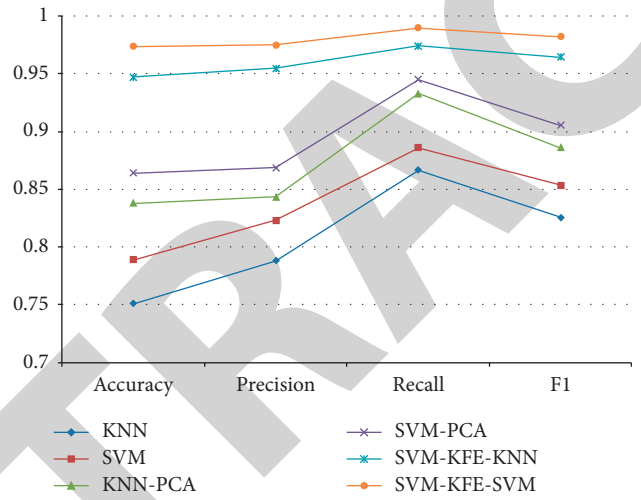

Figure 12: The comparison of algorithm performance when $P \leq 0.01$.

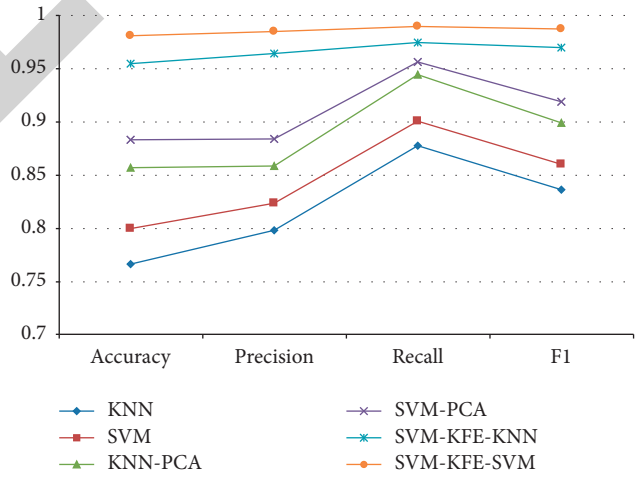

Figure 13: The comparison of algorithm performance when $P \leq 0.001$.

Table 2: Optimal values under SVM parameters.

Gamma	Cost	Error	Dispersion
10^{-6}	10	0.1510823	0.06924002
10^{-5}	10	0.1512987	0.07016349
10^{-4}	10	0.2686147	0.07255622
10^{-3}	10	0.2686147	0.07255622
10^{-2}	10	0.2686147	0.07255622
10^{-1}	10	0.2686147	0.07255622
10^{-6}	100	0.1274892	0.07413187
10^{-5}	100	0.1512987	0.07016349
10^{-4}	100	0.2686147	0.07255622
10^{-3}	100	0.2686147	0.07255622
10^{-2}	100	0.2686147	0.07255622
10^{-1}	100	0.2686147	0.07255622

accords with the characteristics of normal distribution. The larger the P value, the larger the number of genes screened out, which shows that selecting the appropriate value of P for screening can be effectively applied in SVM-RFE algorithm, In order to improve the effect and accuracy of algorithm classification, when |LogFC| approaches 0, the larger the range represented by P value, the more genes there are.

3.2. Complex Algorithm Parameter Selection. The parameter selection of the algorithm is an important part of the experiment, and better experimental results can be obtained by selecting better parameters. Therefore, some data in this experiment are selected for experimental parameter selection, and the final experimental comparison is carried out through the selected parameters.

In order to express the best parameter requirements, the cost of SVM is 10 and 100, the gamma is 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}, and the kernel function is radial. After comparing and optimizing the algorithms, the corresponding error values are obtained under different parameters, and the effects are shown in Table 2.

The distribution comparison of the algorithms shows that when gamma = 10^{-6} and cost = 100, the minimum error value is 0.1274892. There are only 49 samples to build the model, 41 of which are used as support vectors. The proportion of support vectors is too large (over 80%), which indicates that there are irrelevant and redundant features in the used features. It is suggested to use feature selection method RFE to eliminate redundancy and irrelevance and reduce dimension and then use SVM. It is also possible to consider reoptimizing the parameters, but since tune.svm () has been used to find parameters, the parameters found are not good. It is better to use fixed parameters and then use RFE for feature screening. Then, this model is used to classify the test set data and use the contingency table to count the accuracy rate, as shown in Figure 9.

When cost = 100 and gamma = 0.0001, cost = 100 and gamma = 0.1, and cost = 10 and gamma = 0.01, the average accuracy is only 75%. When cost = 10 and gamma = 0.001 and cost = 100 and gamma = 0.001, the accuracy of the algorithm is about 88%. With the increase of sample size, the accuracy also decreases. When the sample size is more than 200, the accuracy is less than 75%. When cost = 10 and gamma = 0.0001, the accuracy of the algorithm is relatively high, about 92%, and relatively stable, as shown in Figure 10.

In the initial stage of the algorithm, the overall accuracy rate is relatively low, only 80%, and the lowest is only 75%. Due to the small sample size, the classification effect is not very ideal. When KNN takes 3, the correct rate is about 90%. When KNN takes 2, the accuracy rate is relatively low, only about 82%. The average accuracy rate of the whole sequencing set is about 84%, and when the sample size increases, the average accuracy rate is relatively stable.

3.3. Comparison of Algorithms. When P takes different values, the differential gene expression data are screened and the selected results are classified. In this paper, several algorithms are selected to screen and analyze genes. The algorithms SVM, KNN, SVM-PCA, KNN-PCA, SVM-KFE-SVM, and SVM-KFE-KNN are used to compare and analyze the performance of accuracy, precision, recall, and F1.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{12}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{13}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{14}$$

In order to evaluate the advantages and disadvantages of different algorithms, the concept of F1 value is proposed on the basis of precision and recall to evaluate precision and recall as a whole. F1 is defined as follows:

$$F_1 = \frac{\text{precision} * \text{recall} * 2}{\text{precision} + \text{recall}}. \tag{15}$$

Through the comparative study of the performance indexes of the abovementioned algorithms, the effects are shown in Figures 11–13.

As can be seen from Figures 11–13, the overall effect of the 6 algorithms is relatively consistent at different P values. When the P value selected is smaller, the performance of the six algorithms is improved. In particular, the performance of SVM-RFE-SVM algorithm is obviously improved; accuracy, precision, recall, and F1 are close to 0.99. Among them,

KNN and SVM algorithms have the worst performance because they have no advantages in gene screening effect. SVM-RFE-SVM and SVM-RFE-KNN algorithms have the best results after gene screening and have obvious advantages in gene screening.

4. Conclusion

In this paper, SVM and KNN algorithms are tested, and important indexes such as error rate and accuracy rate of the algorithms are evaluated to obtain the optimal parameters. SVM-RFE-SVM was proved to be effective by screening and comparing SVM, KNN, KNN-PCA, SVM-PCA, SVM-RFE-SVM, and SVM-RFE-KNN binding genes. In the later research work, the effectiveness of the algorithm proposed in this paper is tested in different datasets, and normalization is carried out in unbalanced datasets for classification research. The effectiveness of the classification algorithm is analyzed by combining the number of exons and mutations of gene sequencing data in RNA-SEQ. Correlation analysis between different types of sequencing data is the ultimate goal of the research work.

Data Availability

The data used in this research are available in the website https://pan.baidu.com/s/1e3du8VbzjnvxHRhunF1o0Q (download code: yzjz).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Nature Science Foundation of China under contract numbers 61271259 and 61601070); the Chongqing Nature Science Foundation under contract numbers CTSC2011jjA40006, CSTC2010BB2415, and CSTC2016jcyjA0455; the Research Project of Chongqing Education Commission under contract numbers KJ120501, KJ12050, KJ1600411, and KJ110530; the Key Project of Science and Technology Research of Chongqing Education Commission (KJZD-K201800603 and KJZD-M201900602); the Chongqing Graduate Scientific Research Innovation Project under Grant no. CYB17131; and the Doctoral Candidate Innovative Talent Project of Chongqing University of Posts and Telecommunications under Grant no. BYJS2016003.

References

[1] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[2] K. B. Duan, J. C. Rajapakse, and M. N. Nguyen, "One-versus-one and one-versus-all multiclass SVM-RFE for gene selection in cancer classification," in *Proceedings of the Evolutionary Computation Machine Learning and Data Mining in Bioinformatics, European Conference, Evobio 2007*, pp. 47–56, Valencia, Spain, April 2007.

[3] Y. Ding and D. Wilkins, "Improving the performance of SVM-RFE to select genes in microarray data," *BMC Bioinformatics*, vol. 7, no. S2, 2006.

[4] S. Yoon and S. Kim, "Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms," *Pattern Recognition Letters*, vol. 30, no. 16, pp. 1489–1495, 2009.

[5] Y. Tang, Y.-Q. Zhang, and Z. Huang, "Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 365–381, 2007.

[6] X. Huang, B. L. Zhang, and Z. ZhangLi, "Feature clustering based support vector machine recursive feature elimination for gene selection," *Applied Intelligence*, vol. 48, no. 3, pp. 594–607, 2018.

[7] J. Yin, J. Hou, and Z. She, "Improving the performance of SVM-RFE on classification of pancreatic cancer data," in *Proceedings of the IEEE International Conference on Industrial Technology (ICIT)*, pp. 956–961, IEEE, Taipei, Taiwan, May 2016.

[8] R. Wang, R. Li, and Y. Lei, "Tuning to optimize SVM approach for assisting ovarian cancer diagnosis with photoacoustic imaging," *Bio-Medical Materials and Engineering*, vol. 26, no. s1, pp. S975–S981, 2015.

[9] C. Chen and H. D. Zhu, "Feature selection method based on parallel binary immune quantum-behaved particle swarm optimization," *Advanced Materials Research*, vol. 546–547, pp. 1538–1543, 2012.

[10] A. Anaissi, M. Goyal, D. R. Catchpoole, A. Kennedy, and P. J. Braytee, "Ensemble feature learning of genomic data using support vector machine," *PLoS One*, vol. 11, no. 6, Article ID e0157330, 2016.

[11] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Information Sciences an International Journal*, vol. 282, no. 5, pp. 111–135, 2014.

[12] T.-C. Chen, Y.-C. Hsieh, P.-S. You, and Y.-C. Lee, "Feature selection and classification by using grid computing based evolutionary approach for the microarray data," in *Proceedings of the 3rd International Conference on Computer Science and Information Technology (ICCSIT)*, pp. 85–89, IEEE, Chengdu, China, July 2010.

[13] X. Zhou and J. Wang, "Feature selection for image classification based on a new ranking criterion," *Journal of Computer and Communications*, vol. 3, no. 3, pp. 74–79, 2015.

[14] L. Zhang and X. Huang, "Multiple SVM-RFE for multi-class gene selection on DNA microarray data." in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, IEEE, Killarney, Ireland, July 2015.