Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine

Chuanze Kang , Yanhao Huo , Lihui Xin , Baoguang Tian , Bin Yu

PII: S0022-5193(18)30600-3
DOI: https://doi.org/10.1016/j.jtbi.2018.12.010
Reference: YJTBI 9749

To appear in: *Journal of Theoretical Biology*

Received date: 2 June 2018
Revised date: 3 November 2018
Accepted date: 6 December 2018

Please cite this article as: Chuanze Kang , Yanhao Huo , Lihui Xin , Baoguang Tian , Bin Yu , Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine, *Journal of Theoretical Biology* (2018), doi: https://doi.org/10.1016/j.jtbi.2018.12.010

Highlights

- A new method (rL-SVM) for tumor classification on two-class and multi-class tumor datasets.
- Relaxed Lasso reduces the biased estimate of Lasso.
- We select the optimal parameter of GenSVM by 10-fold cross-validation grid search.
- GenSVM uses regularization term to avoid overfitting and achieves better accuracy.
- We investigate different feature gene selection methods and classifiers on the results.
- The proposed method has better performance over several methods.

Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine

Chuanze Kang [a,b], Yanhao Huo [a,b], Lihui Xin [a,b], Baoguang Tian [a,b], Bin Yu [a,b,c,*]

[a] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

[b] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China

[c] School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

Abstract: At present, the study of gene expression data provides a reference for tumor diagnosis at the molecular level. It is a challenging task to select the feature genes related to the classification from the high-dimensional and small-sample gene expression data and successfully separate the different subtypes of tumor or the normal from the patient. In this paper, we present a new method for tumor classification—relaxed Lasso (least absolute shrinkage and selection operator) and generalized multi-class support vector machine (rL-GenSVM). The tumor dataset is first z-score normalized. Secondly, using relaxed Lasso to select feature gene sets on training set, and finally, generalized multi-class support vector machine (GenSVM) serves as a classifier. We select four two-class datasets and four multi-class datasets for experiments. And four classifiers are used to predict and compare the classification accuracy on test set. To compare with other proposed methods, we obtain satisfactory classification accuracy by 10-fold cross-validation on all samples of each dataset. The experimental results show that the method proposed in this paper selects fewer feature genes and achieve higher classification accuracy. rL-GenSVM uses regularization parameters to avoid overfitting and can be widely applied to high-dimensional and small-sample tumor data classification. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/rL-GenSVM/.

Keywords: Tumor classification; Gene expression data; Feature genes; Relaxed Lasso; Generalized multi-class support vector machine

1. Introduction

DNA microarray technology provides a series of gene expression profile data at the

* Corresponding author.
E-mail address: yubin@qust.edu.cn (B. Yu).

molecular level and has been widely applied in research of cancer classification prediction. The rapid development of microarray technology has made it easier to monitor the expression data of millions of genes. Researchers have used gene expression data to analyze and identify normal people and patients or different types of cancer (Lu et al., 2017; Lv et al., 2016; Salem et al., 2017). Gene expression usually contains tens of thousands of genes and a small number of samples, having dimensional disasters. If all genes are applied to tumor classification, the efficiency will decline under the influence of redundant genes. Therefore, removing redundancy and selecting related genes from high-dimensional and small-sample microarray data can improve the learning efficiency of the classification model and classification accuracy, which effectively helps us diagnose, predict and treat clinical cancer (Wong and Liu, 2010).

In the view of the high-dimensional characteristic of gene expression data, there existed various gene selection techniques to reduce the dimensionality and select related genes to improve the accuracy. Gene selection techniques are usually divided into feature extraction and feature selection. The feature extraction method converts original variables into lower dimensional space by using a combination of them. According to mid-subset evaluation criteria and the combination of the follow-up learning algorithm, the feature selection methods are divided into three categories: filter, wrapper and embedded. Filter method selects features based on the statistical information of the data, such as the mRMR based on mutual information (Peng et al., 2005), and Kruskal-Wallis rank sum test (Kruskal and Wallis 1952) that tests whether there are significant differences in the distribution of multiple populations. There are also ReliefF (Kononenko 1994) that can assign different weights to features based on their relevance to each feature and category. The wrapper method is a repetitive search process in which the results of the learning algorithm in each repetition are used to guide the search process, such as SVM-RFE (Guyon et al., 2002).

Embedded methods contain feature selection during training. For example, the regularization method proposed by Tibshirani (1996), the least absolute shrinkage selection operator (Lasso), uses the penalty term to shrink the coefficients of certain variables to zero to achieve feature selection, which provides reference to the development of related methods. However, when processing high-dimensional and small-sample data with a lot of noisy information, Lasso performs the same degree of shrinkage on all variables. The redundancy variables are shrinking

towards 0, meanwhile, the related variables are affected on the same extent, generating a biased estimate. Lasso also has a slow convergence rate and tends to choose more variables. For Lasso biased estimates, the researchers proposed a series of sparse models with approximate unbiased estimates. SCAD that constructs confidence intervals for estimated parameters can select variables and estimate coefficients simultaneously. (Fan and Li, 2001). Relaxed Lasso uses two separate parameters to shrink the coefficients of the variables twice, providing a solution for continuous solutions including soft-threshold and hard-threshold estimators (Meinshausen, 2006). It uses the subset of variables chosen by Lasso for parameter estimation and variable selection for the second time. Zou (2006) proposed adaptive Lasso, where adaptive weights are used for penalizing different coefficients in the L_1 penalty.

Classification is an important part of microarray data processing and analysis, mainly based on the relevance of gene expression data for classification. Many machine learning methods have been widely used in the study of tumor classification problems. Classification methods are mainly divided into the unsupervised and supervised. Unsupervised method finds the structural features of the sample, based on sample similarity measure clustering, such as the K-means method. Ignoring sample class information is the shortcoming of it. Supervised classification methods include K-nearest neighbors, support vector machines, decision trees, random forests, and neural networks. The advantage is that it always uses the sample information for learning to minimize the loss function. For tumor classification, SVM is one of the most common classifiers. It has shorter calculation time and better performance for classification. SVM has been successfully applied to tumor classification (Vapnik, 1999; Wang et al., 2017). The SVM maps the sample space into a high-dimensional feature space (Hilbert space), which transforms the problem of nonlinear separability in the original sample space into a linearly separable problem in the feature space. An optimal hyperplane is constructed in the feature space to maximize the boundaries between the support vectors (critical boundary samples for each class) and different classes (Shahbeig et al., 2017).

Due to its outstanding characteristics, SVM is extended to multi-classification problems. In general, these extended methods have obvious differences, mainly distinguishing three types of multi-class SVM (MSVM) methods. First, using the binary SVM as the basic classifier, the K

-class problem is decomposed into multiple binary problems by the heuristic method. The most commonly used heuristic method is the one-to-one (OVO) method. If there are many classes, the one-to-one method calculates the $K(K-1)/2$ binary SVM problem, which creates high computational time, and the criteria may be problematic. Another heuristic approach is the one-to-all (OVA) approach, which isolates a class of samples at a time and constructs class boundaries relative to other class samples, constructing a total of K classification boundaries (Vapnik, 1995). However, both OVO and OVA methods have unpredictable space. The second type use error correction codes. In these methods, the problem is decomposed into multiple binary classification problems based on a constructed coding matrix, which determines the grouping of classes in a particular binary subproblem (Allwein et al., 2000; Crammer et al., 2002). The third type optimizes a single loss function to simultaneously estimate the boundaries of all classes (Rifkin and Klautau 2004). Many related methods have been proposed by Crammer and Singer (2001), Guermeur and Monfrini (2011) et al. The idea of converting a multi-class SVM problem to the $K-1$ dimension is novel because it reduces the dimensionality of the problem and also appears in other multi-class classification methods, such as polynomial regression and linear discriminant analysis (Van and Groenen 2016). In addition, regularization terms have also been extensively studied and applied in SVM. Hsieh (2008) proposed a new dual coordinate descent method for L_2 - regularized L_1 - and L_2 - loss SVM. Yuan (2010) discussed the advantages and disadvantages of different solutions for L_1 - regularized L_2 - loss SVM.

Van and Groenen (2016) proposed a generalized multi-class SVM, called GenSVM. This method uses the simplex encoding to formulate the multiclass SVM problem as a single optimization problem. The multiclass SVM problem is reduced to the binary SVM when $K=2$. By using a flexible hinge function and the l_p norm of error, the GenSVM loss function incorporates three existing multi-class SVMs that use the sum of hinge errors and extends these methods. The use of kernel functions similar to the binary SVM. Regularization has been required to avoid over-fitting in classification problems, especially when there is only a small number of training examples, or when there are a large number of parameters to be learned (Lee et al., 2006). In particular, L_1 -regularized regression is often used for classification problem. The GenSVM loss function is general and flexible, so it can be adjusted for specific datasets. In its loss function,

flexibility is increased by the use of Huber hinge function rather than an absolute hinge function, and the l_p norm of the hinge error is calculated.

Today, researchers have proposed many classification models using statistical and machine learning based on microarray data. Lv et al. (2016) proposed an improved univariate marginal distribution algorithm for gene expression data classification. Salem et al. (2017) combined information gain and standard genetic algorithms for feature selection, and finally used Genetic Programming for cancer classification. Lu et al. (2017) proposed a hybrid feature selection method that combines mutual information maximization and adaptive genetic algorithm. Aziz et al. (2017) proposed a new hybrid search technique for feature selection using Independent component analysis (ICA) and Artificial Bee Colony (ABC) called ICA + ABC, to select informative genes based on a Naïve Bayes (NB) algorithm. For the two-class tumor classification problem, Nanni and Lumini (2010) proposed a classification method based on Orthogonal Linear Discriminant Analysis (ODA), Sequential Forward Floating Selection (SFFS) and Support Vector Machine (SVM). In order to increase the dimensionality of the ODA subspace and increase the accuracy, it combines "original" features to obtain new features. The features are combined into groups of K each new feature being obtained by mapping a K-dimensional feature space into projections of a single dimension. In a newly generated space with only a few hundred features, a set of related features is selected using a wrapper feature selection method. Finally, the radial basis function SVM is trained by these features. For a multi-class tumor classification problem, Liu et al. (2017) proposed a hybrid method based on WELM to deal with multi-class imbalance problems of cancer microarray data at both feature and algorithm level. At the feature level, a feature-oriented search for each class is performed using a class-oriented feature selection method to explicitly select the features related with minority class. At the algorithm level, multiple modified WELM models are trained on datasets characterized by different subsets of features. In order to promote diversity of sets, models with lower similarity are removed and the retained models are combined into an integrated model. Moreover, based on the regularization method, many researchers proposed a series of tumor classification models. Guo et al. (2016) used a kernel-based approach to estimate the class centroid to define both the between-class separability and the within-class compactness for the criterion, and formulate the selection problem as a L_1- regularized optimization problem.

Chen et al. (2018) adopted four kinds of weighting methods based on feature and label correlation to enhance the discriminative performance of $l_{1,2}$ norm joint sparsity, and in the method increased the F -norm regularization extended from multi-class elastic networks. Zhang et al. (2017) proposed a gene interaction regularization elastic network (GIREN) model that integrates multiple data types to predict clinical outcomes. Becker et al. (2011) proposed an Elastic SCAD support vector machine classification model, where the Elastic SCAD is a combination of SCAD and ridge penalties, overcoming the limitations of the separate penalty. Bakir et al. (2016) proposed a range search technique for effectively estimation of optimal regularization parameters to improve the performance of regularized linear discriminant analysis and apply on tumor classification.

For high-dimensional and small-sample tumor datasets, we propose a new tumor classification method—relaxed Lasso-GenSVM (rL-GenSVM). The tumor dataset is first z-score normalized. Secondly, using relaxed Lasso to select feature gene sets on training set, and finally, generalized multi-class support vector machine (GenSVM) serves as a classifier. We select the optimal parameter λ , γ and κ of GenSVM to determine the final classification model on training set. To test the validation of the proposed classification method, we apply it to four two-class datasets and four multi-class datasets. GenSVM was compared with KNN, L_1 -regularized regression, L_2 -regularized regression according to test accuracy. To make comparison with other proposed methods, we then obtained classification accuracy by 10-fold cross-validation on all samples of each dataset. Results indicate that the proposed method in this paper uses regularization parameters to avoid overfitting and can use fewer feature genes to reach higher classification accuracy, which can effectively deal with the tumor classification of high-dimensional and small-sample data.

This paper is divided into four sections and the rest is organized as follows: Section 2 includes descriptions of tumor datasets, feature selection methods, classification algorithms and concepts related to evaluation indicators. Section 3 discusses the experimental results of the proposed method and the process of parameter selection. We present the conclusions and discussion of the future work in Section 4.

2. Materials and methods

2.1. Datasets

To estimate the validation and predictive performance of model, we applied the classification method to eight tumor datasets. Brain datasets is from website: http://www.gems-system.org, MLL and Ovarian datasets are from website: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html, TOX_171 is from website: http://featureselection.asu.edu/datasets.php, the rest are from website: http://www.biolab.si/supp/bicancer/projections/index.html. Table 1 shows the description of tumor datasets. To evaluate proposed method, all data sets are divided into training sets and test sets according to 8:2.

Table 1

Information about the test microarray datasets.

Dataset	No. of Sample	No. of Genes	No. of Class	Train:Test	Reference
DLBCL	77	7129	2	62:15	Shipp et al. (2002)
CNS	60	7129	2	48:12	Pomeroy et al. (2002)
Lung	86	7129	2	69:17	Beer et al. (2002)
Ovarian	253	15154	2	202:51	Petricoin et al. (2002)
Brain	50	10367	4	40:10	Nutt et al. (2003)
Lymphoma	62	4026	3	50:12	Alizadeh et al. (2000)
MLL	72	12582	3	58:14	Armstrong et al. (2002)
TOX_171	171	5748	4	137:34	Stienstra et al. (2010)

2.2. Methods

2.2.1. Relaxed Lasso

Meinshausen (2007) proposed relaxed Lasso as a generalization of soft-thresholding and hard-thresholding. The variable selection and shrinkage of coefficients is controlled by two separate parameters λ and ϕ.

$$\hat{\beta}^{\lambda,\phi} = \arg\min_{\beta} n^{-1}\sum_{i=1}^{n}\left(Y_i - X_i^T\{\beta \cdot 1\omega_\lambda\}\right)^2 + \phi\lambda\|\beta\|_1 \tag{1}$$

where $1\omega_\lambda$ is the indicator function on the set of variables $\omega_\lambda \subseteq \{1,\ldots,p\}$. For all $k \in \{1,\ldots,p\}$

$$\{\beta \cdot 1\omega_\lambda\}_k = \begin{cases} 0 & k \notin \omega_\lambda \\ \beta_k & k \in \omega_\lambda \end{cases} \tag{2}$$

where the variable in ω_λ is selected by ordinary Lasso. Relaxed Lasso only uses element of ω_λ for estimation. The parameter λ controls variable selection as in ordinary Lasso. The new parameter ϕ controls the shrinkage of coefficients. If $\phi = 1$, relaxed Lasso is identical to Lasso. The shrinkage of relaxed Lasso with $\phi < 1$ compared to ordinary Lasso is reduced. We define the relaxed Lasso for $\phi = 0$ as the limitation of the above definition for $\phi \to 0$. The relaxed Lasso provides continuous solutions that include both soft- and hard- thresholding of

estimators. It's similar to the Bridge estimators $\|\beta\|_\gamma$ when varying γ in the range $[0,1]$.

The algorithm of relaxed Lasso developed by Meinshausen (2007) is as following:

Step 1) Compute all ordinary Lasso solutions with the Lars algorithm in Efron et al. (2004). Let $\omega_1,\ldots,\omega_s$ be the variable sets of results. Let $\lambda_1 > \ldots > \lambda_s = 0$ be a sequence of penalty term. When $\omega_\lambda = \omega_k$, if and only if $\lambda \in (\lambda_k, \lambda_{k-1}]$.

Step 2) For each $k = 1,\ldots,s$, let $f(k) = \left(\hat{\beta}^{\lambda_k} - \hat{\beta}^{\lambda_{k-1}}\right) \Big/ \left(\lambda_{k-1} - \lambda_k\right)$. This is the direction of ordinary Lasso solutions. Let $\tilde{\beta} = \hat{\beta}^{\lambda_k} + \lambda_k f(k)$. If there is at least one component l so that $sign\left(\tilde{\beta}_l\right) \neq sign\left(\hat{\beta}_l^{\lambda_k}\right)$, for $\lambda \in \Lambda_k$ compute all Lasso solutions on the set ω_k, varying the penalty parameter between 0 and λ_k. Otherwise, relaxed Lasso solutions for $\lambda \in \Lambda_k$ and $\phi \in [0,1]$ are linear interpolation between $\hat{\beta}^{\lambda_{k-1}}$ and $\tilde{\beta}$.

The response variable is a linear combination of the predictor variables,

$$Y = X^T \beta + \varepsilon \tag{3}$$

where $\varepsilon \sim N\left(0, \sigma^2\right)$, the loss function of relaxed Lasso under parameter λ and ϕ is defined as following:

$$L(\lambda, \phi) = E\left(Y - X^T \hat{\beta}^{\lambda,\phi}\right)^2 - \sigma^2 \tag{4}$$

Meinshausen (2007) have shown that the convergence rate of relaxed Lasso is not affected by the number of variables p. As the gaining rate of p increases, the convergence rate of Lasso drops sharply. And if the redundancy variable increases too much, the ordinary Lasso's convergence rate will also decrease. On the contrary, relaxed Lasso is more suitable for sparse high-dimensional data.

2.2.2. Kruskal-Wallis rank sum test

The Kruskal-Wallis rank sum test is a non-parametric test (Kruskal and Wallis 1952). It can test whether there is a significant difference in the distribution of multiple populations.

There are k independent populations $X_i \sim F(x - \beta_i)$. Their distributions have the same form except a positional parameter, where the $F(x)$ is required to be a continuous function, and the hypothesis test is established without considering the specific form. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$,

$H_1 : \beta_1, \beta_2, ..., \beta_k$ are not all equal.

Let $X_{i1}, X_{i2}, ..., X_{in_i}$ be a sample of type X_i and $N = \sum_{i=1}^{k} n_i$ $(i = 1, 2, ..., k)$. The rank in X_{ij} is r_{ij} $(j = 1, 2, ..., n_i)$, the rank sum of X_i samples is $r_i = \sum_{j=1}^{n_i} r_{ij}$, and the average rank is $\bar{r}_i = r_i / n_i$. The overall average rank is

$$\bar{r} = \sum_{i=1}^{k} r_i / N = (N+1)/2 \tag{5}$$

The sufficient condition for the null hypothesis is that if these $\bar{r}_i$ differ largely. Statistics are now constructed

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i (\bar{r}_i - \bar{r})^2 \tag{6}$$

Calculate $P = \Pr(\chi^2_{n-1} > H)$ and select the population with the most significant difference in expression based on the P value.

2.2.3. Information gain

Information gain is a measure of entropy that is the uncertainty of information. The IG value of each variable determines whether this feature is selected. Therefore, setting the threshold, the variable whose IG value is larger than the threshold will be selected, otherwise it will not (Dagliyan et al., 2011).

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 \left(p(y) \right) \tag{7}$$

where Y and X are the features, $p(y)$ is the marginal probability density function. When expression values of Y are affected by X, and the entropy of Y prior to partitioning is higher than the entropy of Y with respect to the partitions induced by X, there is a relationship between feature Y and X. The entropy of Y after observing X is given in Eq. 8.

$$H(Y|X) = -\sum_{y \in Y} p(x) \sum_{y \in Y} p(y|x) \log_2 \left(p(y|x) \right) \tag{8}$$

$$InfGain = H(Y) - H(Y|X) \tag{9}$$

where $p(y|x)$ is the conditional probability. Information Gain is a measure of the amount that the entropy of Y decreases given X.

2.2.4. ReliefF

The ReliefF (Kononenko 1994) randomly fetches a sample T from training samples each time. It finds k nearest neighbor samples from samples of the same class as T and samples of different classes from T, respectively, to update the weight of each feature:

$$W_R = W_R - \sum_{j=1}^{k} diff(R,T,H_j)/mk + \sum_{A \in class(T)} [\frac{p(A)}{1-p(Class(A))} \sum_{j=1}^{k} diff(R,T,M_j(A))]/mk \quad (10)$$

Where $M_j(A)$ denotes the jth nearest neighbor sample in class A and $diff(f,s_1,s_2)$ denotes the difference between sample s_1 and sample s_2 in feature f. If feature f is continuous, then $diff(f,s_1,s_2) = \frac{|s_1[f]-s_2[f]|}{\max(f)-\min(f)}$. If the feature f is discrete, the formula for $diff(f,s_1,s_2)$ is as follows:

$$diff(f,s_1,s_2) = \begin{cases} 0 & s_1[f] = s_2[f] \\ 1 & s_1[f] \neq s_2[f] \end{cases} \quad (11)$$

The series of Relief algorithms is highly efficient and has no restrictions on data types. It is a feature weight algorithm. The algorithm will give higher weight to features which have high correlation with the class, so it cannot effectively remove the redundant features.

2.2.5. Generalized multi-class support vector machine

Van and Groenen (2016) proposed a generalized multi-class support vector machine called GenSVM. The classification boundary of the problem in the method is established using simplex coding in dimensional space. This dimensional space is called a simplex space. The mapping from the original space to the simple space is optimized by minimizing the errors of the classification. The classification error is calculated by measuring the distance between the object and the decision boundary in the simplex space. The use of simplex coding ensures that a class is predicted for each point in the predictor space, so there is no ambiguity area in the GenSVM solution.

Let $x_i \in R^p$ be a vector with p features and $y_i \in \{1,\ldots,K\}$ denote the class of $i \in \{1,\ldots,n\}$ samples. In addition, let $\mathrm{W} \in \mathrm{R}^{p\times(K-1)}$ be the weight matrix and define the translation vector $t \in \mathrm{R}^{K-1}$ as a bias term. The sample i is projected to the $(K-1)$ dimensional space $z_i' = x_i'\mathrm{W} + t'$ by a linear function. For kernel changes in the original space in GenSVM, preprocessing is required on the kernel matrix. Let $k: \mathrm{R}^p \times \mathrm{R}^p \to \mathrm{R}^+$ denote a positive definite nucleus that satisfies Mercer's theorem, and let H_k denote the corresponding reproducing core Hilbert space. In

addition, the definition maps to $\psi(x)=k(x,\cdot)$, $k(x_i,x_j)=\langle\psi(x_i),\psi(x_j)\rangle_{\mathrm{H}_k}$ under k action. Therefore, the kernel matrix K is defined as the $n\times n$ matrix with $k(x_i,x_i)$, and ψ is defined as the $n\times l$ matrix with row $\psi(x_i)$, then $K=\psi'\psi$. Correspondingly, the simplex space maps to

$$Z=\psi W+1t' \tag{12}$$

The error of sample i is calculated by determining the distance to each classification boundary. The distance from sample i to class k and j is

$$q_i^{(kj)}=(x_i'\mathrm{W}+t')(g_k-g_j) \tag{13}$$

Huber hinge loss is defined as following:

$$h(q)=\begin{cases}1-q-\dfrac{\kappa+1}{2} & \text{if } q\le-\kappa\\ \dfrac{1}{2(\kappa+1)}(1-q)^2 & \text{if } q\in(-\kappa,1]\\ 0 & \text{if } q>1\end{cases} \tag{14}$$

For each sample, the error is summed by the l_p norm to give the total error as

$$\Big(\sum_{\substack{j=1\\ j\neq y_i}}^{K}h^p\big(q_i^{(y_i,j)}\big)\Big)^{1/p} \tag{15}$$

Let $\omega_i=\dfrac{n}{n_kK}, i\in G_k$ denote optional sample weights, correcting different group sizes, or assigning extra weights to certain classification of errors. Where $G_k=\{i:y_i=k\}$ is the set of samples which belong to class k, and n_k is the number of G_k samples. The GenSVM total loss function is

$$L_{MSVM}(\mathrm{W},t)=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in G_k}\omega_i\Big(\sum_{j\neq k}h^p\left(q_i^{(kj)}\right)\Big)^{1/p}+\lambda\mathrm{trW}'\mathrm{W} \tag{16}$$

where $\lambda\mathrm{trW}'\mathrm{W}$ is a penalty term to avoid overfitting and λ is a regularization term. The effect of this penalty term is similar in Ridge Regression, forcing the l_2 norm of the row vector in W to approach zero. When $K=2$, the penalty term becomes $\lambda\mathrm{W}'\mathrm{W}$ and the loss function Eq. (16) is improved on the basis of the Huber hinge loss of the two-class SVM defined by Groenen et al. (2008).

For the unknown sample x_m, the optimal $z_m=x_m'\mathrm{W}+t'$ is mapped into the simplex space. Predict the class label of x_m:

$$\hat{y}_m = \arg\min_k \left\| z'_m - g'_k \right\|^2 \quad \text{for } k = 1, \ldots, K \tag{17}$$

2.2.6. L_1 and L_2 regularized logistic regression

L_1 -regularized logistic regression solves the following unconstrained optimization problem:

$$\min \quad \|\omega\|_1 + C\sum_{i=1}^{l} \log(1 + e^{-y_i \omega^T x_i}) \tag{18}$$

where $C > 0$ is the penalty coefficient, $\|\cdot\|_1$ donates the l_1 norm, $\|\omega\|_1$ regularization term avoids overfitting, and l_1 norm regularization produces the sparse solution of (18). Yuan et al. (2012) improved the GLMNET algorithm proposed by Friedman et al. (2010) and redefined the optimization problem as

$$f(\omega) = \|\omega\|_1 + C(\sum_{i=1}^{l} \log\left(1 + e^{-\omega^T x_i}\right) + \sum_{i:y_i=-1} \omega^T x_i) \tag{19}$$

which improves the rate of calculation and adjust the stopping condition for sub-problems.

L_2 -regularized logistic regression minimizes the following regularized negative log-likelihood (Yu et al., 2011) ：

$$\min \quad \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{l} \log(1 + e^{-y_i \omega^T x_i}) \tag{20}$$

The dual problem:

$$\min \quad \frac{1}{2}\alpha^T Q\alpha + \sum_{i:\alpha_i > 0} \alpha_i \log \alpha_i + \sum_{i:\alpha_i < C} (C - \alpha_i)\log(C - \alpha_i) - \sum_{i=1}^{l} C \log C \tag{21}$$

$$\text{subject to } \ 0 \le \alpha_i \le C, \ i = 1, \ldots, l \text{ , where } \ Q_{ij} = y_i y_j x_i^T x_j, \ \forall i, j \text{ .}$$

Yu et al. (2011) proposed the dual-coordinate descent method that can effectively solve the dual problem by applying many or few Newton iterations to solve the sub-problem, respectively.

2.2.7. K-nearest neighbor

The K-nearest neighbor is a practical supervised learning method whose mechanism is very simple. The observations have p variables. Based on the Euclidean distance, each sample is classified according to its K-nearest neighbor's class. If all K-nearest neighbors of the samples belong to the same class, the sample is classified as the class of neighbors. Otherwise, the sample is considered non-classifiable (Li et al., 2001). Let the sample class be C and the i th class contain the n_i sample. The test sample discriminant function is:

$$f(a) = \min \|a - a_i\| \quad (i = 1, \ldots, C) \tag{22}$$

2.3. Evaluation index and model building

In experiments, the performance of the classifier is usually tested using an independent test. It mainly includes: k-fold cross-validation, leave-one-out method, and retain method. This paper uses the 10-fold cross-validation to calculate accuracy rate. The following different performance indicators are:

Positive samples and negative samples, True positives (TP) denote the number of correctly predictive positive samples. True negatives (TN) denote the number of correctly predictive negative samples. False positives (FP) denote the number of wrongly predictive positive. False negatives (FN) denote the number of wrongly predictive negative. These performance indicators are used to calculate the classification accuracy (ACC):

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{23}$$

For the two-class datasets, the classifier performance is evaluated using AUC (Sing et al., 2005) which is not sensitive to sample class balance. For multi-class data sets, we evaluate classifier performance using Kappa (Cohen, 1960) values that reflect the consistency of actual and predicted values.

For convenience, the new tumor classification method proposed in this paper is called rL-GenSVM. The general framework is shown in Fig. 1 We have implemented it in R version 3.4.4 programming in windows 10 running on a PC with system configuration Intel (R) Core (TM) i7-4510U CPU @ 2.00 GHz 2.60 GHz with 8.00GB of RAM.

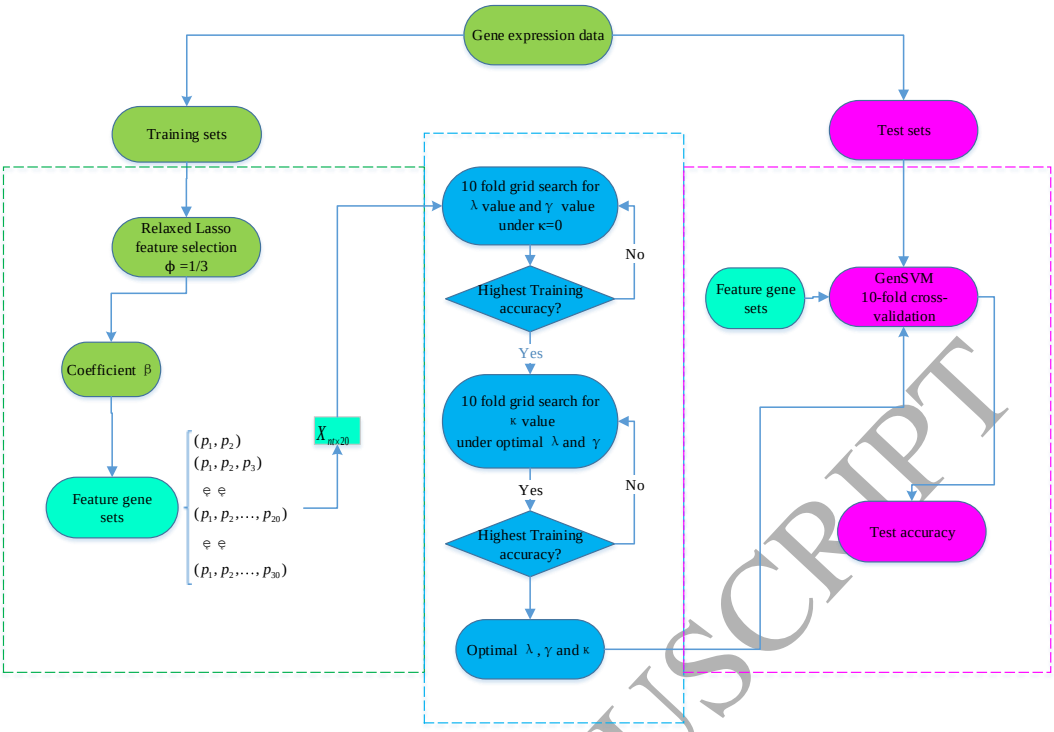

Fig. 1. Flowchart of feature gene selection and classification prediction.

The steps of the rL-GenSVM tumor classification method are described as follows:

(1) Z-score standardization of tumor data sets;

(2) Using the Lars (Efron et al., 2004) algorithm to solve the ordinary Lasso solution, obtain $\lambda_1 > \ldots > \lambda_{30} = 0$ penalty term sequence and $\omega_1, \ldots, \omega_{30}$ variable set on training set;

(3) Use the relaxed Lasso Eq. (1) to perform parameter estimation with variables in ω_k and obtain the feature gene set by shrinking the coefficient with penalty term λ_k, generating matrix $X_{nt\times 2}, \ldots, X_{nt\times 30}$ (nt indicates the number of training samples);

(4) GenSVM with the radial basis kernel function serves as the classifier, calculate Eq. (15) by l_1 norm and find optimal parameter λ, γ and κ through 10-fold cross-validation grid search on matrix $X_{nt\times 20}$;

(5) Fix optimal regularization parameter λ, radial basis kernel function parameter γ and Huber hinge loss parameter κ, using GenSVM as a classifier, training classification model based on Eq. (12) and Eq. (16) with feature gene set;

(6) Calculate the ACC, AUC, and Kappa by 10-fold cross-validation on training set and evaluate the classification effect on test set.

3. Results and discussion

3.1. Effect of feature gene selection on results

The idea of the classification method proposed in this paper is to select feature genes with the highest classification accuracy rate under the combined effect of feature selection methods and classifiers. An excellent feature gene selection method can select fewer feature genes and reach higher classification accuracy. First, we processed tumor datasets by z-score standardization. Second, the relaxed Lasso was used to select the feature genes through hard-thresholding and soft-thresholding of the coefficients on training set. $\phi\lambda$ is a penalty for relaxed Lasso, which is compared with Lasso by adding a hard-thresholding $\phi \in (0,1)$. If $\phi = 1$, relaxed Lasso estimator is equivalent to Lasso. When the value of λ increases, $\|\beta\|_1$ decreases with it, and the component corresponding to the irrelevant variable is gradually shrank towards 0, which achieves the target of relevant variable selection.

Based on the Lars algorithm proposed by Efron et al. (2004), the shrinkage process of the coefficient vector corresponding to $\phi = 1/3$ and $\phi = 1$ with variation of its l_1 norm is shown in Fig. 2, where each line corresponds to the shrinkage process of the coefficient component. The abscissa is $\|\beta\|_1 / \max\|\beta\|_1$, normalizing the $\|\beta\|_1$ interval to the $[0,1]$ interval, and $\|\cdot\|_1$ is the l_1 norm. Take the DLBCL dataset as an example. For ease of observation, we set the number of iterations of relaxed Lasso to 10, the number of shrank variables to 9, and the coefficient vector to $\beta = (\beta_1, \beta_2, \ldots, \beta_9)$.

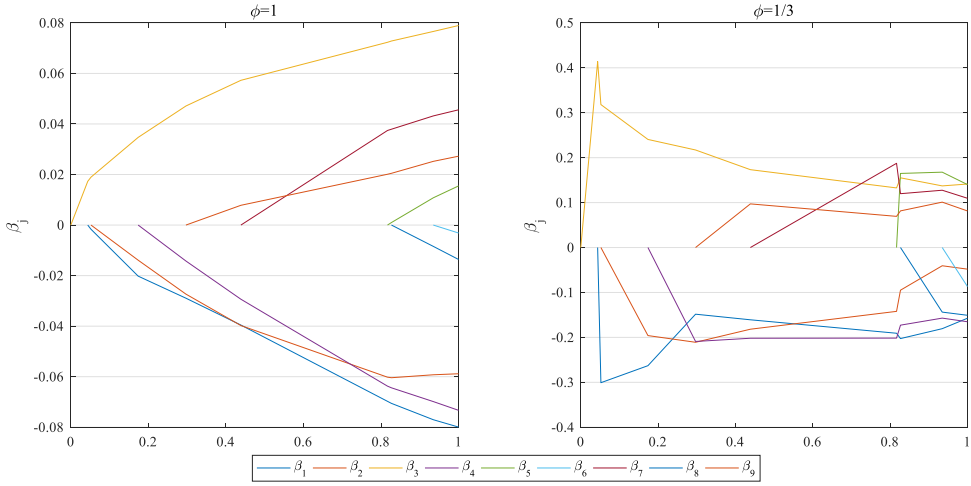

Fig. 2. The shrinkage process of different ϕ.

Fig. 2 shows that during the decrease of $\|\beta\|_1$, its component β_j is gradually shrank to 0.

Compared with Lasso's smoother shrinkage process, adding a hard-thresholding makes the shrinkage process appear angular and accelerates the convergence rate. The relaxed Lasso shrinks $\|\beta\|_1$ and sets the excluded variable towards 0 to implement variable selection.

In order to test the validation of the relaxed Lasso (rL) feature selection method, we let $\phi = 1/3$ and the number of iterations $\max.steps = 2nt$. Three feature gene selection methods, Information Gain (IG), ReliefF (RF) and Kruskal-Wallis rank sum test (KW), are selected for comparison. For all datasets, GenSVM with RBF kernel function is serves as a classifier. Parameters are not subject to artificial intervention and are determined by its own iterative algorithm. We obtain the training accuracy by 10-fold cross-validation. Results indicate that through several experiments, the classification accuracy rate has reached the highest when the number of feature genes is less than 30 for all datasets. Therefore, it is reasonable to compare the performance of methods within the range of 1-30. The relationship between the number of feature genes (NF) in the 1-30 range and the classification accuracy (ACC) is shown in Fig. 3.

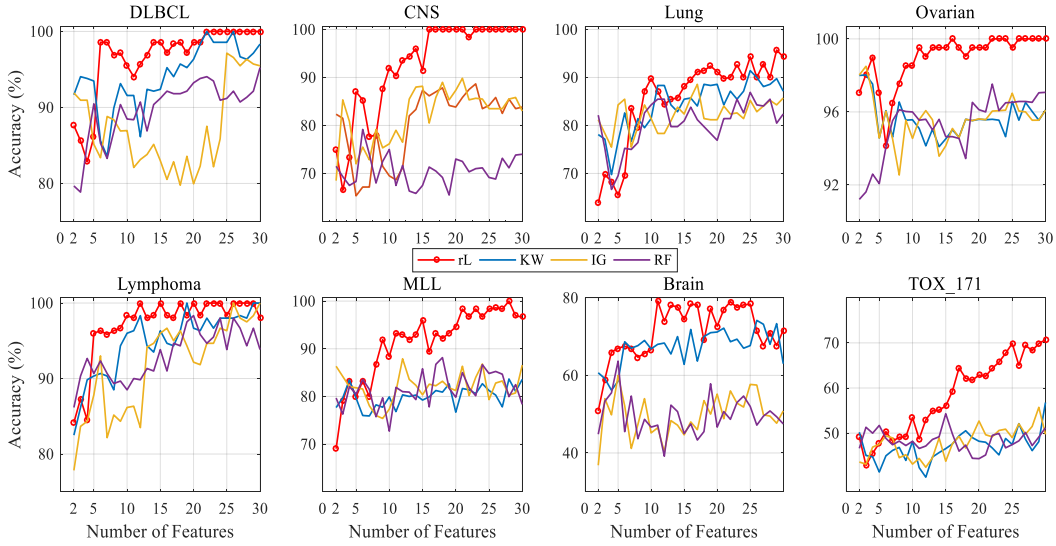

Fig. 3. Relationship between the number of features and accuracy rate of different feature selection methods on training set.

Fig. 3 shows that, rL has the highest ACC for two-class datasets. For the DLBCL dataset and the CNS dataset, the ACC reaches 100%. While other methods have more fluctuations in ACC, there has no breakthrough with increasing of NF. For Lung datasets, rL has the highest ACC. When NF is greater than 15, the genes selected by other three methods do not maintain the

stronger correlation with the classification, resulting in fluctuations of accuracy. For Ovarian datasets, rL only achieved 100% classification accuracy and other methods select more redundant genes. For Lymphoma and MLL datasets, the ACC of rL also reaches 100%. For Brain dataset, the ACC has remained relatively high, in the 10-25 range. For TOX_171 datasets, four methods show an overall upward trend, but the accuracy of rL is relatively high.

For all datasets, the ACC of rL is higher overall. The gap among IG, RF and rL is obvious. Although the ACC of KW is also competitive, with the increase of the NF, the continuous increase of ACC rarely occurs. However, KW's hypothesis of independent variables is often useful in feature selection. rL depends on the relevant variables selected by Lasso in ω_k to perform parameter estimation, and then shrinks all variables through penalty terms to remove redundant variables. This process reduces the negative impact on the relevant variables while shrinking the coefficients of variables, selecting more efficient feature genes, and achieving a higher ACC. With the increase of the NF, other methods probably select redundant genes, leading to reduction of ACC. Therefore, rL performs better and is more suitable for feature selection of high-dimensional and small-sample data.

3.2. Effect of classifier on results and biological significance of feature genes

The option of parameters plays an important role in the classifier. The classifier can be trained to be a better classification model by adjusting the parameters, which improves the classification accuracy. In order to select the optimal parameters, it is usually necessary to take ACC as the criterion. In this paper, we use the relaxed Lasso to select feature gene sets and train the GenSVM with them on training set. For GenSVM, we select the l_1-norm to calculate the total Huber hinge error. The kernel function is the radial basis kernel function, and the parameters λ, γ and κ are the main adjustment targets.

The above three parameters have different effects on GenSVM. The λ is a regularization parameter to avoid overfitting. For radial basis kernel function, the expression is $K(x_i, x_j) = \exp(\gamma \left\| x_i - x_j \right\|_2^2)$, where $\gamma = -\frac{1}{2\sigma^2}$. The default value of γ is $1/p$, which controls the radial range of the function. Increasing the value of κ makes the error close to the square of the Euclidean distance to the margin. Deviations can occur when using cross-validation (Cawley and Talbot, 2010), so Van et al. used nested cross validation (Stone, 1974) to eliminate errors. For all

datasets, 20 genes are uniformly selected as feature genes by the relaxed Lasso, generating an $X_{nt\times 20}$ data matrix (training set). We use 10-fold cross-validation grid search on $X_{nt\times 20}$ to evaluate different parameter combinations to predict classification accuracy and find the optimal parameters. First fix $\kappa = 0$, select $\lambda \in \{1e-1,\ldots,1e-10\}$, $\gamma \in \{1e-1,\ldots,1e-10\}$, there are a total of 100 kinds of parameter combinations, predicting the classification accuracy of the surface as shown in Fig. 4.

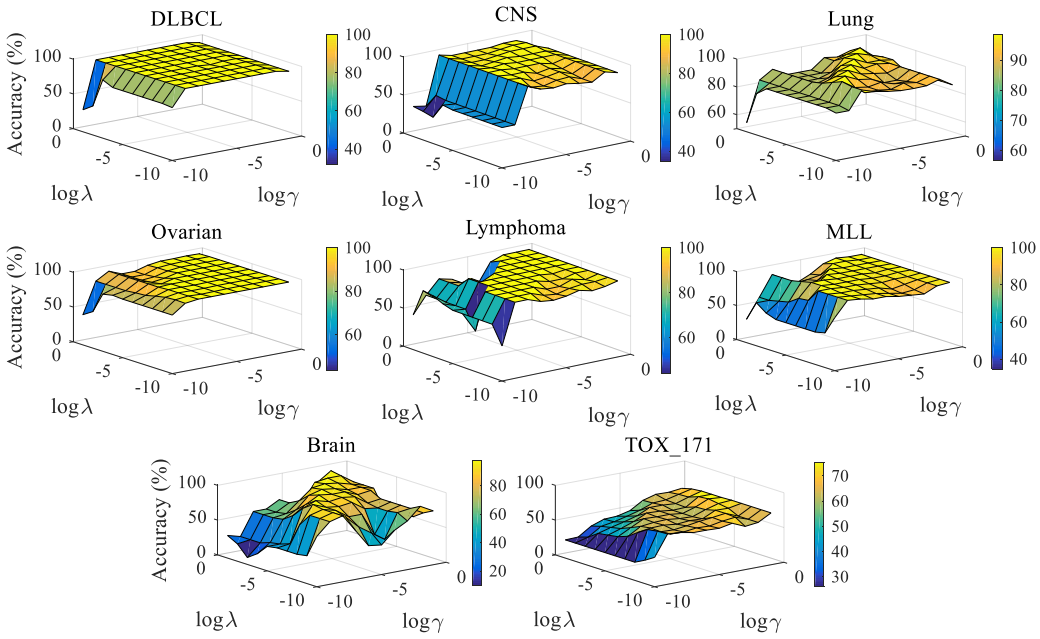

Fig. 4. The classification accuracy of GenSVM for different combinations (λ,γ).

Fig. 4 shows that the combination of the regularized parameter λ value and the radial basic kernel function parameter γ value have a great influence on the accuracy. For all datasets, the accuracy is low with larger γ and smaller λ. As the γ decreases, the accuracy begins to increase. For CNS, Lung, Brain and TOX_171 datasets, when we set smaller γ and larger λ, the accuracy is lower and unstable. As the λ decreases, the accuracy tends to be maximum. Therefore, when the γ and λ are close to each other, that is, on the diagonal line of surface, the accuracy rate can reach the maximum value (The specific accuracy rate see the Supplementary materials **Table S1-Tabel S8**). The experimental results show that when $\kappa = 0$, the combination of λ and γ has better accuracy. Through multiple tests, the optimal parameter combination and accuracy are shown in Table 2.

Table 2

Optimal parameter combination and classification accuracy of all datasets.

Datasets	DLBCL	CNS	Lung	Ovarian	Lymphoma	MLL	Brain	TOX_171
λ value	1e-7	1e-8	1e-8	1e-7	1e-7	1e-7	1e-7	1e-5
γ value	1e-8	1e-8	1e-8	1e-8	1e-7	1e-8	1e-8	1e-2
ACC	100%	100%	98.57%	100%	100%	100%	97.5%	75.39%

Table 2 shows that the use of 10-fold cross-validation grid search can get higher prediction accuracy. We selected $(1e-7,1e-8)$, $(1e-8,1e-8)$, $(1e-8,1e-8)$, $(1e-7,1e-8)$, $(1e-7,1e-7)$, $(1e-7,1e-8)$, $(1e-7,1e-8)$, $(1e-5,1e-2)$ from the combination of λ value and γ value through multiple experiments for the optimal combination of DLBCL, CNS, Lung, Ovarian, Lymphoma, MLL, Brain and TOX_171. After fixing the optimal λ and γ values, a 10-fold cross-validation grid search is used to evaluate the κ and test the accuracy on training set. We only consider the case of $\kappa \geq 0$, taking $\kappa \in [0, 2.5]$, and the interval is 0.1. The classification accuracy is shown in Fig. 5.

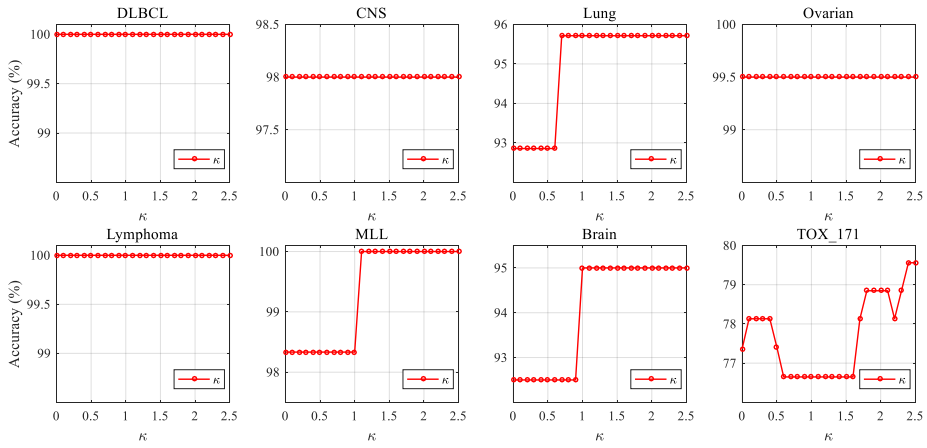

Fig. 5. The classification accuracy of GenSVM for different κ value.

Fig. 5 shows that, the change of κ value has less effect on the classification accuracy compared with λ and γ values. For DLBCL, CNS, Ovarian and Lymphoma dataset, the change in κ has no effect on accuracy. For Lung, MLL and Brain datasets, the classification accuracy reaches highest when $\kappa \geq 1.5$. For TOX_171 datasets, the classification accuracy excesses 79% when $\kappa \geq 2.4$. According to comparison in Fig. 5, the optimal κ and accuracy are shown in Table 3.

Table 3

Optimal κ value and accuracy of all datasets.

Datasets	DLBCL	CNS	Lung	Ovarian	Lymphoma	MLL	Brain	TOX_171
κ value	2	2	2	1	2.5	2	2.5	2.5
ACC	100%	98%	98.89	99.5%	100%	100%	95%	79.56%

Through multiple experiments we select 2, 2, 2, 1, 2.5, 2, 2.5 and 2.5 as optimal κ values for DLBCL, CNS, Lung, Ovarian, Lymphoma, MLL, Brain and TOX_171 respectively. The selection principle of optimal κ value reference experiment results. Before and after fixing the optimal parameters, the highest classification accuracy (ACC) of GenSVM both on training set and test set and the corresponding number of feature genes (NF) are shown in Table 4.

Table 4

The classification accuracy (%) before and after optimization under GenSVM.

Datasets	before optimization			after optimization		
	NF	Training ACC	Test ACC	NF	Training ACC	Test ACC
DLBCL	22	100%	100%	9	100%	93.33%
CNS	16	100%	58.33%	16	100%	75%
Lung	29	95.71%	82.35%	21	100%	76.47%
Ovarian	22	100%	100%	3	100%	100%
Lymphoma	12	100%	75%	8	100%	100%
MLL	28	100%	85.71%	14	100%	92.86%
Brain	11	79.17%	60%	28	96.17%	60%
TOX_171	30	70.68%	61.76%	29	86.03%	76.47%

Table 4 shows that for the trained GenSVM with fixed optimal parameters, the obtained model is tested by 10-fold cross-validation on training set. We find that for the dataset that has 100% pre-optimal training classification accuracy, the model with fixed optimal parameters achieves 100% ACC with fewer feature genes. For the DLBCL, Ovarian, Lymphoma, and MLL datasets NF reduces to 9, 3, 8, and 14 genes, respectively. Moreover, for CNS, Lymphoma, MLL and TOX_171 datasets, test ACC has also been improved. For CNS dataset, test ACC is increased by 16.667%, and for the Lymphoma dataset, test ACC is increased to 100% using 8 genes. For TOX_171 datasets, test ACC increases by 14.707% after fixing the optimal parameters. Most data sets have high-dimensional small sample characteristics. When they are divided into training set and test set, the number of training set samples is less than before. For datasets that are difficult to classify, such as CNS, Lung, Brain, and TOX_171, there are not enough samples to train method, which results in unpredictable results on the test set. For example, for the Ovarian dataset, the training set has 203 sample training models. On the test set, only 3 feature genes are used to achieve 100% classification accuracy. In general, it is very reliable to use 10-fold cross-validation

grid search to find the optimal parameters on training set. The GenSVM with optimal parameters has improved both in the NF and in test ACC aspects. Better performance classifiers use fewer feature genes to reach higher classification accuracy. We compare the validation of other three classifiers KNN, L_1, L_2 regularized logistic regression (L_1, L_2 logreg) to prove the effectiveness of GenSVM (Dhole et al., 2014). When KNN is used as a classifier, the best classification effect is obtained by continuous adjustment and we select the number of adjacent samples to be 5. We use relaxed Lasso to select feature genes and predict the ACC by 10-fold cross validation on training set. Test ACC of four classifiers is obtained on test set. The relationship between training ACC (%) and the number of feature genes (NF) is shown in Fig. 6.

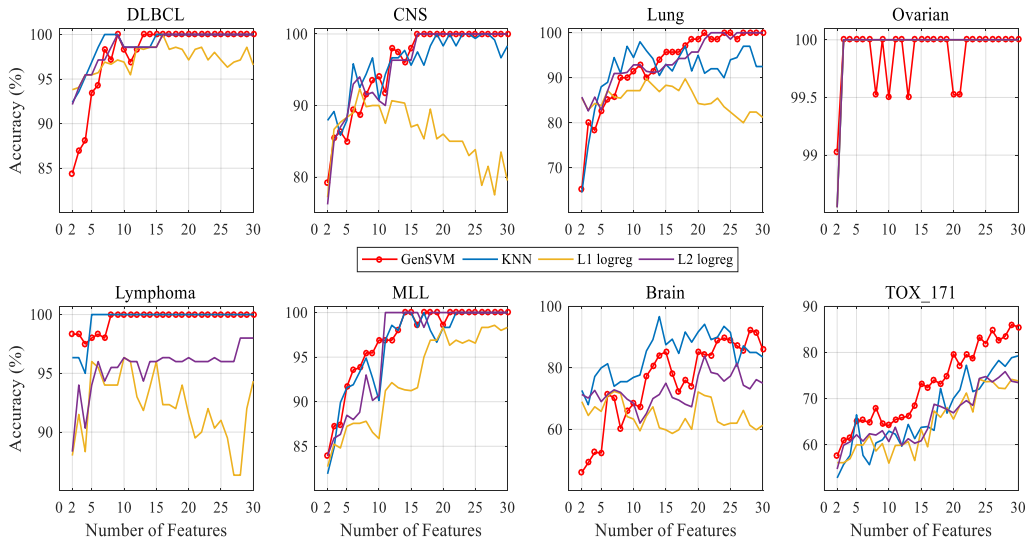

Fig. 6. Relationship between the number of features and training accuracy rate (%) of different classifiers.

Fig. 6 shows that ACC of the GenSVM and L_2 logreg reaches 100% for the two-class datasets. The ACC of L_1 logreg has large fluctuations. With the increase of the NF, the ACC cannot always keep at 100% in the DLBCL dataset. For CNS and Lung datasets, the ACC of L_1 logreg gradually decreases. KNN's ACC is lower for Lung dataset, but competitive for multi-class datasets. For Lymphoma and MLL datasets, the ACC of GenSVM, KNN reaches 100%. For Brain datasets, ACC of KNN is higher than other methods, when NF is less than 25. As the NF increases, L_1 and L_2 logreg fluctuate, but there is no breakthrough of increase. For TOX_171 datasets, ACC of GenSVM is overall higher than other methods.

According to highest training classification accuracy from Fig.6, we select feature genes of

each dataset. And we report the accuracy on test sets using feature genes. With the comparison of number of genes and best test accuracy, performance of GenSVM is more persuasive. For each dataset, the test accuracy (Test ACC (%)), number of feature genes (NF), AUC, and Kappa are shown in Table 5 and Table 6. Sun et al. (2014) indicated that in two-class datasets, AUC evaluated the generalization performance of two types of samples classification algorithm. In multi-class datasets, Kappa evaluated the generalization performance of multi-class samples classification algorithm.

Table 5

Comparison of GenSVM with other classifier methods for two-class tumor datasets.

Datasets	GenSVM		KNN		L_1 logreg		L_2 logreg	
	NF	Test ACC(AUC)	NF	Test ACC (AUC)	NF	Test ACC (AUC)	NF	Test ACC (AUC)
DLBCL	9	93.33%(0.958)	7	93.33%(0.958)	16	86.67%(0.917)	9	86.67%(0.917)
CNS	16	75.00%(0.75)	23	66.67%(0.547)	7	58.33%(0.5)	16	58.33%(0.563)
Lung	21	76.47%(0.673)	18	70.59%(0.827)	12	82.35%(0.625)	22	82.35%(0.712)
Ovarian	3	100%(1)	3	100%(1)	3	100%(1)	3	100%(1)

Table 5 shows that, for DLBCL dataset, GenSVM and KNN predict one sample in 14 test samples mistakenly. For CNS dataset, GenSVM predict three samples in 12 test samples mistakenly, using 16 genes. For Ovarian datasets, all methods use three genes achieving100% classification accuracy. For the two-class dataset, GenSVM and KNN performs better.

Table 6

Comparison of GenSVM with other classifier methods for multi-class tumor datasets.

Datasets	GenSVM		KNN		L_1 logreg		L_2 logreg	
	NF	Test ACC (Kappa)	NF	Test ACC (Kappa)	NF	Test ACC (Kappa)	NF	Test ACC (Kappa)
Lymphoma	8	100%(1)	5	100%(1)	10	100%(1)	28	91.67%(0.846)
MLL	14	92.86%(0.891)	14	92.86%(0.891)	28	78.57%(0.659)	11	92.86%(0.887)
Brain	28	60%(0.487)	14	50%(0.324)	7	60%(0.428)	21	70%(0.595)
TOX_171	29	76.47%(0.684)	30	58.82%(0.449)	24	73.53%(0.644)	28	76.47%(0.684)

Table 6 shows that, for Lymphoma dataset, GenSVM uses 8 genes which are fewer than L_2 logreg and reaches 100% accuracy. For the Brain dataset, the test ACC of all classifiers is weak and there are more than 2 predicted error samples among the 10 test samples. For MLL and TOX_171 datasets, L_2 logreg performs better than GenSVM and KNN in NF and Test ACC.

In summary, for the multi-class datasets, the L_2 logreg classification works better and is similar to the GenSVM result. For two-class datasets, the average ACC of L_2 logreg is 4.36% lower than that of GenSVM. The reason may be that the nature of logistic regression is limited in multi-class problems. Compared with l_2 norm, the L_1 logreg classification effect is not good, and

the role of l_1 norm as a classification is far less than in feature selection. The KNN classification effect is more competitive in two-class datasets. The classification effect may be limited by the regularization feature selection method, and the feature selection method based on the distance principle is more suitable for KNN. For GenSVM, we use the 10-fold cross-validation grid search to find the optimal parameters. According to highest training accuracy from Fig.6, selecting corresponding feature genes, GenSVM achieves average accuracy 4% higher than other classifiers, depending on the advantage of regularization parameters and radial basis kernel functions.

Feature genes have profound biological significance for tumor classification. Taking the DLBCL and Lymphoma datasets as an example, we give descriptions of the selected genes of rL-GenSVM and their biological significance. (The specific biological significance of feature genes of CNS, Lung, Ovarian, Lymphoma, Brain and TOX_171 datasets see Supplementary materials **Table S9-Table S14.**)

The DLBCL dataset consists of diffuse large B-cell lymphoma and follicular lymphoma, having 58 DLBCL samples and 19 FL samples. Each sample contains 7129 gene expression data, and sample type labels are each labeled 0 and 1 (DLBCL: 58, FL: 19). Using rL to select 8 feature genes, GenSVM as a classifier, the ACC reaches 100%. Feature genes selected by rL-GenSVM for the DLBCL dataset and their biological significance are described in Table 7.

Table 7

List of feature genes for DLBCL dataset selected by rL-GenSVM.

Index no. of selected genes	Gene accession number	Gene description
699	D87119_at	tribbles homolog 2 (Drosophila), TRIB2
1367	L19314_at	hes family bHLH transcription factor 1, HES1
4028	X02152_at	lactate dehydrogenase A, LDHA
4194	X16983_at	"integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)", ITGA4
4459	X67683_at	keratin 4, KRT4
5077	Z21966_at	"POU domain, class 6, transcription factor 1", POU6F1
5198	Z70723_at	paraoxonase 1, PON1
6040	S57212_s_at	myocyte enhancer factor 2C, MEF2C
6179	M14328_s_at	"enolase 1, (alpha)", ENO1

The gene biological descriptions for DLBCL in Table 7 are from the bioinformatics website https://www.ncbi.nlm.nih.gov/gene. It is known from the website statistics, Mramor et al. (2007)

and Fan et al. (2006) mentioned D87119_at in the paper. Mramor et al. (2007) and Dagliyan et al. (2011) all mentioned X02152_at in the paper. X16983_at, M14328_s_at, and Z21966_at are all mentioned in the paper by Mramor et al. (2007). Among them, The D87119_at-encoding gene is one of the three members of the Tribbles family. The Tribbles member primarily induces apoptosis of hematopoietic origin as an oncogene inactivating the transcription factor C/EBPα (CCAAT/enhancer binding protein α) and causing acute myeloid leukemia. X16983_at encodes a member of the integrin alpha chain family of proteins. This integrin is a therapeutic target for the treatment of multiple sclerosis, Crohn's disease and inflammatory bowel disease. The encoded by S57212_s_at may play a role in maintaining the differentiated state of muscle cells. Mutations and deletions at this locus have been associated with severe cognitive disability, stereotypic movements, epilepsy, and cerebral malformation.

For the DLBCL dataset, we randomly select three feature genes {D87119_at, Z70723_at, M14328_s_at} from the feature genes {D87119_at, L19314_at, X02152_at, X16983_at, X67683_at, Z21966_at, Z70723_at, S57212_s_at, M14328_s_at}, as shown in Fig. 7.

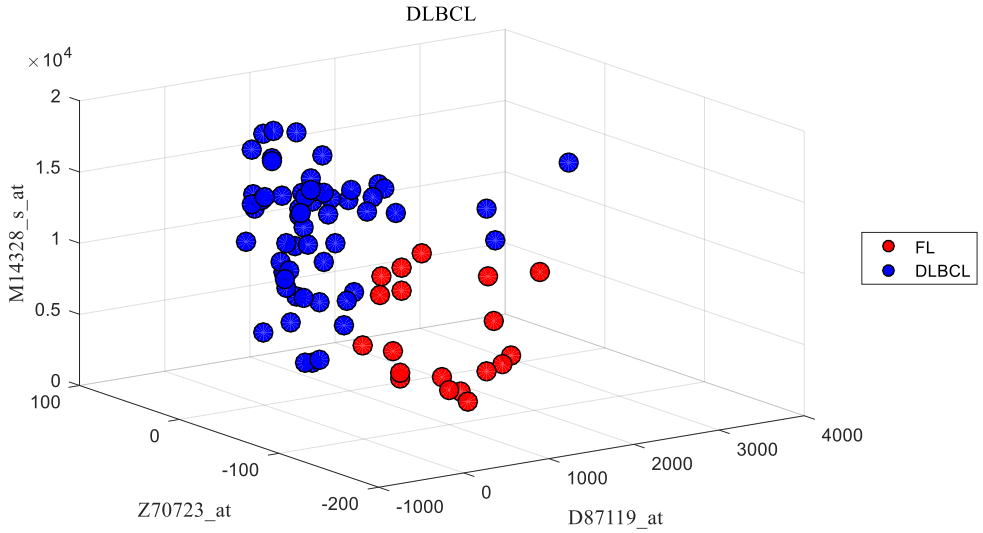

Fig. 7. The three-dimensional scatter plot of the three genes {D87119_at, Z70723_at, M14328_s_at} of the DLBCL dataset.

From Fig. 7 it can be seen that for the DLBCL dataset, the feature genes selected by rL-GenSVM can clearly distinguish DLBCL samples from FL samples. Therefore, the method rL-GenSVM proposed in this paper has good classification results for the two-class tumor dataset.

In the MLL dataset, 72 sample data are from patients. There are 24 acute lymphoblastic leukemia (ALL) samples, 20 myeloid lymphoid leukemia (MLL) samples and 28 acute myeloid leukemia (AML) samples. Each sample contains 12582 gene expression data. Using rL to select 13 feature genes, GenSVM as a classifier, the classification accuracy reaches 100%. The feature genes and their biological significance are described in Table 8.

Table 8

List of feature genes for MLL dataset selected by rL-GenSVM.

Index no. of selected genes	Gene accession number	Gene description
1119	34168_at	DNA nucleotidylexotransferase, DNTT
1316	35083_at	ferritin light chain, FTL
1696	32353_at	gap junction protein, beta 1, 32kDa (connexin 32, Charcot-Marie-Tooth neuropathy, X-linked), GJB1
3277	38242_at	B cell linker, BLNK
3634	39318_at	T-cell leukemia/lymphoma 1A, TCL1A
3768	39931_at	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 3, DYRK3
8428	36122_at	KIAA0391
8518	36571_at	DNA topoisomerase II beta, TOP2B
9084	38096_f_at	major histocompatibility complex, class II, DP beta 1, HLA-DPB1
9845	40533_at	baculoviral IAP repeat containing 5, BIRC5
9882	40570_at	forkhead box O1, FOXO1
9929	40617_at	THUMP domain containing 1, THUMPD1
10419	32541_at	protein phosphatase 3 catalytic subunit gamma,PPP3CC
12418	266_s_at	CD24 antigen (small cell lung carcinoma cluster 4 antigen), CD24

The biological description of the feature genes of the MLL in Table 8 is from the bioinformatics website https://www.ncbi.nlm.nih.gov/gene. According to the statistics of the website, Wang et al. (2012) mentioned 34168_at in the paper. Suárez-Fariñas et al. (2010) mentioned 35083_at in the paper. Wang et al. (2009) and Huang et al. (2009) mentioned 39318_at in their papers. Borczuk et al. (2005) mentioned 40617_at in the paper. Zennaro et al. (2012) mentioned 32541_at in the paper. Kar et al. (2015) and Guan et al. (2009) mentioned 266_s_at in the paper. 39931_at, 39318_at, 266_s_at are all mentioned by Mroror et al. (2007). Among them, 35083_at encodes the light subunit of the ferritin protein. A major function of ferritin is the storage of iron in a soluble and nontoxic state. 39318_at indicate that Overexpression of the TCL1 gene in

humans has been implicated in the development of mature T cell leukemia, in which chromosomal rearrangements bring the TCL1 gene in close proximity to the T-cell antigen receptor (TCR)-alpha (MIM 186880) or TCR-beta (MIM 186930) regulatory elements. 266_s_at encodes a sialoglycoprotein that is expressed on mature granulocytes and B cells and modulates growth and differentiation signals to these cells. This gene was missing from previous genome assemblies, but is properly located on chromosome 6.

For the MLL dataset, we randomly select three feature genes {40533_at, 35083_at, 266_s_at} from the feature genes {34168_at, 35083_at, 32353_at, 38242_at, 39318_at, 39931_at, 36122_at, 36571_at, 38096_f_at, 40533_at, 40570_at, 40617_at, 32541_at}, as shown in Fig. 8.

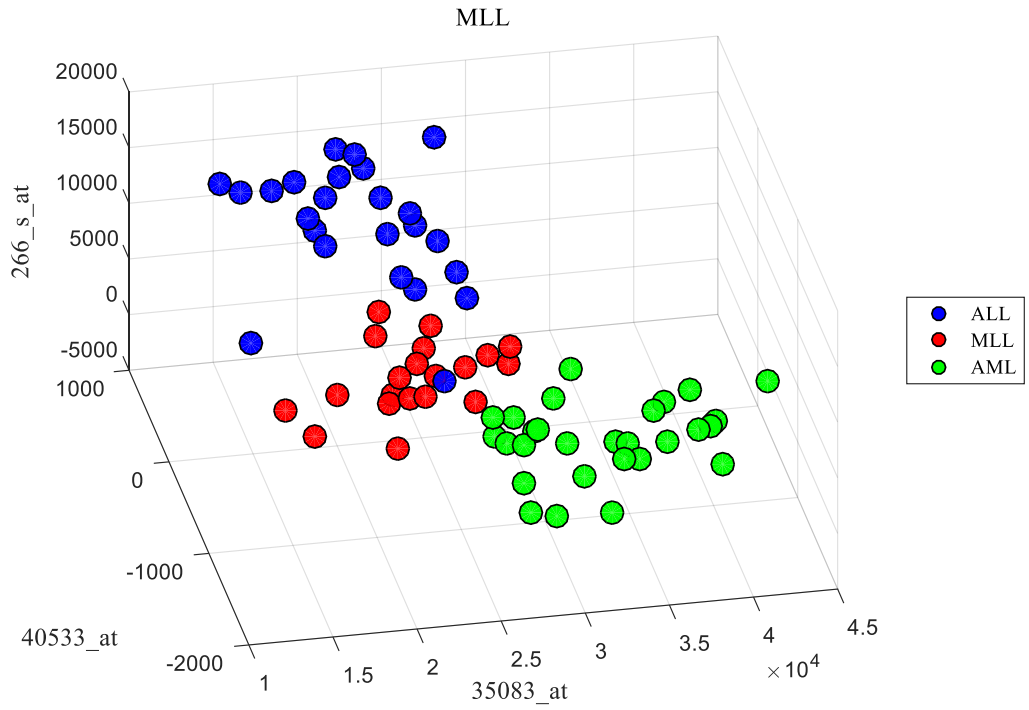

Fig. 8. The three-dimensional scatter plot of the three genes {40533_at, 35083_at, 266_s_at} of the MLL dataset.

From Fig. 8 it can be seen that for the MLL dataset, the feature genes selected by rL-GenSVM can clearly distinguish ALL samples, MLL samples and AML samples. Therefore, the method rL-GenSVM proposed in this paper has good classification results for the two-class tumor dataset. (CNS, Lung, Lymphoma and Brain datasets see Supplementary materials **Fig. S1-Fig. S6**).

3.3. Comparison with other methods

In order to demonstrate the validity of the proposed method, the accuracy and the number of feature genes were compared with other methods. Therefore, we select feature genes and optimal parameters on all samples of each dataset. Optimal parameter determined by 10-fold cross-validation grid search on all samples of each dataset, is shown in Table 9. We calculate classification accuracy by 10-fold cross-validation on all samples of each dataset, and select feature genes according to optimal accuracy. The biologic description of feature genes that was changed due to different processes of evaluation see Supplementary materials **Table S15-Table S20**.

Table 9

Optimal λ , γ and κ value of GenSVM by 10-fold cross-validation grid search on all samples of each datasets

Datasets	DLBCL	CNS	Lung	Ovarian	Lymphoma	MLL	Brain	TOX_171
λ	1e-7	1e-7	1e-8	1e-7	1e-7	1e-9	1e-8	1e-9
γ	1e-7	1e-7	1e-8	1e-7	1e-7	1e-8	1e-6	1e-7
κ	1.2	2	0.9	0	2.5	2	2.5	1.1

A total of 17 different methods were compared with rL-GenSVM. The comparison of rL-GenSVM and other methods in the ACC and the NF are shown in Table 10 and Table 11.

The other methods are described as following: RLR uses kernel-based expectations to estimate the centroid of a class and is used to define discriminant separability (Guo et al., 2016). IG/SGA uses information gain for feature selection, genetic algorithm for feature reduction, and finally uses Genetic Programming for cancer classification (Salem et al., 2017). MOEDA is a multi-objective optimization algorithm for gene expression data classification (Lv et al., 2016). CFS-iBPSO-NB combines correlated feature selection (CFS) with improved binary particle swarm optimization and naive Bayesian classification (Jain et al., 2018). kernelPLS is a kernel partial least squares-based filtering method (Sun et al., 2014). MTDT applies several univariate tests in each non-terminal node of the decision tree (Czajkowski et al., 2014). MRSRC is a maxdenominator reweighted sparse representation classification method to classify tumors (Li et al., 2017). DMFS is a discriminative multi-class feature selection method using weighted $l_{1,2}$ norms and extended elastic networks (Chen, S.B. et al., 2018). LDA-GA is an approach that is combination of a genetic algorithm (GA) with Fisher's linear discriminant analysis (LDA) (Huerta et al., 2010). Osareh et al. (2013) proposed an efficient ensemble learning method combining the

fast correlation-based feature selection method with ICA-based RotBoost ensemble. Genuer et al. (2010) proposed the increasingly used statistical method for classification and regression problems introduced by Leo Breiman in 2001 for random forests. KBCGS is a feature gene selection method based on learning the best gene weights in the clustering process (Chen, H. et al., 2016). MGNMF is a graph regularization subspace segmentation method in which the global optimal solution is solved by solving a Sylvester equation (Chen, X. et al., 2014). C-E-MWELM is a WELM-based hybrid method that can solve multi-class imbalance sample problems (Liu et al., 2017). Statnikov et al. (2005) proposed to use Kruskal-Wallis rank sum test to select feature genes and one-versus-rest (OVR) multi-classification support vector machine (MC-SVM) as a classifier to predict classification accuracy. DLSR-FS is a framework of discriminative least squares regression (LSR) for multiclass classification and feature selection (Xiang et al., 2012).

Table 10

Comparison rL-GenSVM with some other gene selection methods in two-class datasets

Datasets	Author	Method	ACC(10-fold)	NF
DLBCL	Guo et al. (2016)	RLR	93.04% (5-fold)	50
	Salem et al. (2017)	IG/SGA	94.80%	110
	Lv et al. (2016)	MOEDA	98.80%	4
	This paper	rL-GenSVM	100%	8
CNS	Lv et al. (2016)	MOEDA	72.00%	9
	Salem et al. (2017)	IG/SGA	86.67%	38
	Jain et al. (2018)	CFS-iBPSO-NB	96.67%	10
	This paper	rL-GenSVM	100%	17
Lung	Sun et al. (2014)	kernelPLS	77.4%	57
	Lv et al. (2016)	MOEDA	84.00%	10
	Czajkowski et al. (2014)	MTDT	100%	1000
	This paper	rL-GenSVM	100%	27
Ovarian	Huerta et al. (2010)	LDA-GA	100%	18
	Osareh et al. (2013)	FCBF-ICA_based RootBoost	99.40%	30
	Jain et al. (2018)	CFS-iBPSO-NB	100%	3
	This paper	rL-GenSVM	100%	3

Table 10 shows that, for DLBCL dataset, Salem et al. (2017) selected 110 genes using IG/SGA to obtain 94.80% ACC. Guo et al. (2016) used RLR to select 50 genes and obtained 93.04% ACC. In this paper, only 8 genes are selected and the ACC reaches 100%. IG/SGA does not effectively use genetic algorithms based on Information Gain. Compared with RLR, this paper

uses regularized feature selection method to select fewer feature genes and achieve higher ACC. For CNS dataset, Jain et al. (2018) used CFS-iBPSO-NB to select 10 genes to obtain 95.67% ACC. In this paper, the GenSVM with optimal parameters obtains 100% ACC using seventeen genes which is 28% higher than that of the MOEDA method. For Lung datasets, Lv et al. (2016) selected 10 genes using MOEDA and obtained 84% ACC by SVM. We select 27 genes and improve the classification accuracy to 100%. MTDT classification accuracy also reached 100%, but up to 1000 genes were selected. For Ovarian datasets, Huerta et al. (2010) and Osareh et al. (2013) selected 18 genes and 30 genes respectively. While we only select three genes and reach 100% classification accuracy.

Table 11

Comparison rL-GenSVM with some other gene selection methods in multi-class datasets

Datasets	Author	Method	ACC(10-fold)	NF
Lymphoma	Genuer et al. (2010)	Random forests	91%	12
	Chen, H. et al. (2016)	KBCGS	100%	11
	Jain et al. (2018)	CFS-iBPSO-NB	100%	24
	This paper	rL-GenSVM	100%	9
MLL	Li et al. (2017)	MRSRC	98.61%	625
	Chen, S. B. et al. (2018)	DMFS	99.8% (5-fold)	15
	Jain et al. (2018)	CFS-iBPSO-NB	100%	30
	This paper	rL-GenSVM	100%	13
Brain	Chen, X. et al. (2014)	MGNMF	54.20%	—
	Statnikov et al. (2005)	MC-SVM	85%	500
	Liu et al. (2017)	C-E-MWELM	91.67% (Leave-one-out)	200
	This paper	rL-GenSVM	96%	26
TOX_171	Xiang et al. (2012)	DLSR-FS	75.24%	50
	This paper	rL-GenSVM	81.38%	30

Note: The "-" literature does not mention the number of feature genes.

Table 11 shows that, for the Lymphoma dataset, Chen, H. et al. (2016) used KBCGS to select 11 genes and obtain 100% ACC by SVM. In this paper, the NF is reduced by the penalty term. We select 9 genes, and the classification accuracy reaches 100%. For the MLL dataset, Chen, S. B. et al. (2018) used weighted ENs extended in multi-class to select 15 genes and achieved 99.8% ACC. CFS-iBPSO-NB selected 30 genes and the ACC reached 100%. We use regularization parameters of GenSVM to avoid overfitting. Only 13 genes are selected and the ACC reaches 100%. For the Brain dataset, Statnikov et al. (2005) used Kruskal-Wallis rank sum test to select 500 genes and

obtained 85% ACC by MC-SVM. The GenSVM loss function incorporates three existing multi-class SVMs that use the sum of hinge errors, and extends these methods to improve classification performance of GenSVM. Therefore, only 26 genes are selected in this paper, and the ACC reaches 96%. Liu et al. (2017) used a WELM-based hybrid method to select 200 genes and achieved the ACC of 91.67%. For TOX_171 datasets, We only select 30 genes and ACC reaches 81.38%.

In summary, the above comparison results fully demonstrate that rL-GenSVM uses the l_1 regularization term to reduce the number of feature genes. The regularization parameters of the GenSVM and radial basis kernel functions improve the classification accuracy. The ACC reaches 100% on DLBCL, CNS, Lung, Ovarian, Lymphoma and MLL datasets, 96% on Brain dataset, and 81.38% on TOX_171 datasets. A satisfactory tumor classification result is given.

4. Conclusion

The big data of cancer can make us understand the pathology of tumor diseases more deeply and further improve the ability of early prediction and prevention of diseases. It is a hot research topic in tumor bioinformatics how to select feature genes and conduct in-depth research on disease-causing genes to reduce the mortality of cancer diseases and find better drug targets for precision medicine. For high-dimensional and small-sample tumor datasets, we propose a new tumor classification method—relaxed Lasso-GenSVM (rL-GenSVM). The tumor dataset is first divided into training and test sets and z-score normalized. Secondly, relaxed Lasso selects feature genes on training set. Finally, finding optimal parameters by the 10-fold cross-validation grid search on training set, GenSVM serves as the classifier. We obtain the training accuracy by 10-fold cross validation on training set and calculate test accuracy on test set. To test the validity of the rL-GenSVM method, this paper selects eight tumor datasets and compares test accuracy with other three classifiers. The experimental results show that the proposed method selects fewer feature genes and achieves higher classification accuracy. For feature gene selection method, relaxed Lasso uses the ordinary Lasso selected variable subset for parameter estimation, and achieves variable selection by the l_1 penalty term for the second shrinkage, which can reduce biased estimate of Lasso. As a feature selection method, it is more suitable for high-dimensional and small-samples dataset. For classifiers, GenSVM uses regularization parameters to avoid

overfitting, achieving higher classification accuracy with fewer feature genes. And using radial basis kernel functions on multi-class datasets, it maintains strong classification performance. Therefore, it is the future research direction how to combine the research results of this paper with some of the current advanced feature selection algorithms, such as semi-supervised manifold learning methods and sparse classification algorithms, and incorporate clinical biological information to further simplify the feature gene subset and apply the research results to the precision medical treatment of tumors.

Acknowledgements

The authors sincerely thank the anonymous reviewers for their many valuable comments. This work was supported by the National Natural Science Foundation of China (Nos. 61863010 and 11771188), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159), the College Students' Innovative Practice Training Program of Chinese Academy of Sciences, and the College Students' Innovative Entrepreneurial Training Program (No. 201710426046).

References

Alizadeh, A. A., Eisen, M.B., Davis, R. E., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Staudt, L. M., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G. E.,Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R. A., Levy, R., Wilson, W. H., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403, 503-511.

Allwein, E.L., Schapire, R.E., and Singer, Y., 2000. Reducing multiclass to binary: a unifying approach for margin classifiers. J. Mach. Learn. Res. 1, 113-141.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J., 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30, 41-47.

Aziz, R., Verma, C.K., Srivastava, N., 2017. A novel approach for dimension reduction of

microarray. Comput. Biol. Chem. 71, 161-169.

Bakir, D., James, A.P., Zollanvari, A., 2016. An efficient method to estimate the optimum regularization parameter in RLDA. Bioinformatics 32, 3461-3468.

Becker, N., Toedt, G., Lichter, P., Benner, A., 2011. Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. BMC Bioinform, 12, 138-138.

Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lizyness, M.L. Kuick, R., Hayasaka, S., Taylor, J.M., Iannettoni, M.D., Orringer, M.B., Hanash, S., 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. 8, 816-824.

Borczuk, A.C., Kim, H.K., Yegen, H.A., Friedman, R.A., Powell, C.A., 2005. Lung adenocarcinoma global profiling identifies type ii transforming growth factor-β receptor as a repressor of invasiveness. Am. J. Resp. Crit. Care. 172, 729-737.

Cawley, G.C., Talbot, N L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11, 2079-2107.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. Acm. T. Intel. Syst. Tec. 2, 1-27.

Chen, H., Zhang, Y., Gutman, I., 2016. A kernel-based clustering method for gene selection with gene expression data. J. Biomed. Inform. 62, 12-20.

Chen, S.B., Zhang, Y., Ding, C.H., Zhou, Z.L., Luo, B., 2018. A discriminative multi-class feature selection method via weighted l2, 1-norm and Extended Elastic Net. Neurocomputing 275, 1140-1149.

Chen, X., Jian, C., 2014. Gene expression data clustering based on graph regularized subspace segmentation. Neurocomputing 143, 44-50.

Chen, Y., Wang, L., Li, L., Zhang, H., Yuan, Z., 2016. Informative gene selection and the direct classification of tumors based on relative simplicity. BMC Bioinform. 17, 1-16.

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. 20, 37-46.

Crammer, K., Singer, Y., 2002. On the learnability and design of output codes for multiclass problems. Mach. Learn. 47, 201-233.

Crammer, K., Singer, Y., 2001. On the algorithmic implementation of multiclass kernel-based

vector machines. J. Mach. Learn. Res. 2, 265-292.

Czajkowski, M., Grześ, M., Kretowski, M., 2014. Multi-test decision tree and its application to microarray data classification. Artif. Intell. Med. 61, 35-44.

Dagliyan, O., Uney-Yuksektepe, F., Kavakli, I.H., Turkay, M., 2011. Optimization based tumor classification from microarray gene expression data. PloS ONE 6, e14579.

Dhole, K., Singh, G., Pai, P.P., Mondal, S., 2014. Sequence-based prediction of protein–protein interaction sites with L1-logreg classifier. J. Theor. Biol. 348, 47-54.

Efron, B., Hastie, T., Johnstone, I.M., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407-499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348-1360.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1-22.

Genuer, R., Poggi, J. M., and Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31, 2225-2236.

Guan, P., Huang, D., He, M., Zhou, B., 2009. Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method. J. Exp. Clin. Canc. Res. 28, 103-103.

Guermeur, Y., Monfrini, E., 2011. A quadratic loss multi-class SVM for which a radius–margin bound applies. Informatica 22, 73-96.

Guo, S., Guo, D., Chen, L., Jiang, Q., 2016. A centroid-based gene selection method for microarray data classification. J. Theor. Biol. 400, 32-41.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach Learn. 46, 389-422.

Hsieh, C., Chang, K., Lin, C., Keerthi, S.S., Sundararajan, S., 2008. A dual coordinate descent method for large-scale linear SVM. Proc. 25nd International Conference on Machine Learning 408-415.

Huang, L.T., 2009. An integrated method for cancer classification and rule extraction from microarray data. J. Biomed. Sci. 16, 1-10.

Huerta, E. B., Duval, B., and Hao, J. K., 2010. A hybrid LDA and genetic algorithm for gene selection and classification of microarray data. Neurocomputing 73, 2375-2383.

Jain, I., Jain, V.K., Jain, R., 2018. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. Appl. Soft. Comput. 62, 203-215.

Kar, S., Sharma, K.D., Maitra, M., 2015. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive k-nearest neighborhood technique. Expert. Syst. Appl. 42, 612-627.

Kononenko, I., 1994. Estimating attributes: analysis and extensions of RELIEF. Proc. ECML'94 171-182.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583-621.

Lee, S.I., Lee, H., Abbeel, P., Ng, A.Y., 2006, July. Efficient L1 regularized logistic regression. In AAAI-06 401-408.

Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G., 2001. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17, 1131-1142.

Li, W., Liao, B., Zhu, W., Chen, M., Peng, L., Wei, X., Gu, C., Li, K., 2017. Maxdenominator reweighted sparse representation for tumor classification. Sci. Rep. 7, 46030.

Liu, Z., Tang, D., Cai, Y., Wang, R., Chen, F., 2017. A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. Neurocomputing 266, 641-650.

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., Gao, Z., 2017. A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 256, 56-62.

Lv, J., Peng, Q., Chen, X., Sun, Z., 2016. A multi-objective heuristic algorithm for gene expression microarray data classification. Expert. Syst. Appl. 59, 13-19.

Meinshausen, N., 2007. Relaxed lasso. Comput. Stat. Data. An. 52, 374-393.

Mramor, M., Leban, G., Demsar, J., Zupan, B., 2007. Visualization-based cancer microarray data classification analysis. Bioinformatics 23, 2147-2154.

Nanni, L., Lumini, A., 2010. Orthogonal linear discriminant analysis and feature selection for

micro-array data classification. Expert. Syst. Appl. 37, 7132-7137.

Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., Black, P. M., von Deimling, A., Pomeroy, S.L., Golub, T.R., Louis, D.N., 2003. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 63, 1602-1607.

Osareh, A., and Shadgar, B., 2013. An efficient ensemble learning method for gene microarray classification. Biomed Res. Int. 2013, 478410.

Peng, H., Long, F., Ding, C., 2008. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.

Petricoin, E., Ardekani, A., Hitt, B., Levine, P., Fusaro, V., Steinberg, S., Mills, G., Simone, C., Fishman, D., Kohn, E., Liotta, L. A., 2002. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359, 572–577.

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovizky, G., Louis, D.N., Mesirov J.P., Lander E.S., Golub T.R., 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415, 436.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101-141.

Salem, H., Attiya, G., El-Fishawy, N., 2017. Classification of human cancer diseases by gene expression profiles. Appl. Soft. Comput. 50, 124-134.

Shahbeig, S., Helfroush, M.S., Rahideh, A., 2017. A fuzzy multi-objective hybrid TLBO–PSO approach to select the associated genes with breast cancer. Signal Process 131, 58-65.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., Gaasenbeek, M., Angelo, M., Reich, M.R., Pinkus, G.S., Ray, T.S., Koval, M., Norton, A.J., Lister, T.A., Mesirov, J.P., Neuberg, D., Lander, E.S., Aster, J.C., Golub, T.R., 2002. Diffuse large B-cell

lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 8, 68-74.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940-3941.

Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D.P., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21, 631-643.

Stienstra, R., Saudale, F., Duval, C., Keshtkar, S., Groener, J. N., Rooijen, van, Staels, B., Kersten, S., Mnller, M., 2010. Kupffer cells promote hepatic steatosis via interleukin-1beta-dependent suppression of peroxisome proliferator-activated receptor alpha activity, Hepatology 51, 511–522.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. B. 36,111-147.

Suárez-Fariñas, M., Shah, K.R., Haider, A.S., Krueger, J.G., Lowes, M.A., 2010. Personalized medicine in psoriasis: developing a genomic classifier to predict histological response to Alefacept. BMC Dermatol. 10, 1-8.

Sun, S., Peng, Q., Shakoor, A., 2014. A kernel-based multivariate feature selection method for microarray data classification. PloS ONE 9, e102541.

Tibshirani, R.J., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series. B. Stat. Methodo. 58, 267-288.

Van Den Burg, G.J., Groenen, P.J., 2016. GenSVM: a generalized multiclass support vector machine. J. Mach. Learn. Res. 17, 7964-8005.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural. Netw. 10, 988-999.

Wang, H., Zheng, B., Yoon, S.W., Ko, H.S., 2017. A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur. J. Oper. Res. 267.

Wang, S.L., Li, X.L., Fang, J., 2012. Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. BMC Bioinform. 13, 178.

Wang, X., Gotoh, O., 2009. Accurate molecular classification of cancer using simple rules. BMC Med. Genomics 2, 64.

Wong, T.T., Liu, K.L., 2010. A probabilistic mechanism based on clustering analysis and distance measure for subset gene selection. Expert Syst. Appl. 37, 2144-2149.

Xiang, S., Nie, F., Meng, G., Pan, C., and Zhang, C., 2012. Discriminative least squares regression for multiclass classification and feature selection. IEEE Trans. Neural Netw. Learn Syst. 23, 1738.

Yu, H. F., Huang, F.L., Lin, C.J., 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. Mach. Learn. 85, 41-75.

Yuan, G., Chang, K., Hsieh, C., Lin, C., 2010. A Comparison of optimization methods and software for large-scale L1-regularized linear classification. J. Mach. Learn. Res. 3183-3234.

Yuan, G.X., Ho, C.H., Lin, C.J., 2012. An improved glmnet for L1-regularized logistic regression. J. Mach. Learn. Res. 13, 1999-2030.

Zennaro, D., Scala, E., Pomponi, D., Caprini, E., Arcelli, D., Gambineri, E., Gambineri, E., Russo. G., Mari. A., 2012. Proteomics plus genomics approaches in primary immunodeficiency: the case of immune dysregulation, polyendocrinopathy, enteropathy, X-linked (IPEX) syndrome. Clin. Exp. Immunol. 167, 120–128.

Zhang, L., Liu, H., Huang, Y., Wang, X., Chen, Y., Meng, J., 2017. Cancer progression prediction using gene interaction regularized elastic net. IEEE/ACM Trans. Comput. Biol. Bioinform. 14, 145-154.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101, 1418-1429.