SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso

Yanhao Huo , Lihui Xin , Chuanze Kang , Minghui Wang , Qin Ma , Bin Yu

PII: S0022-5193(19)30467-9
DOI: https://doi.org/10.1016/j.jtbi.2019.110098
Reference: YJTBI 110098

To appear in: *Journal of Theoretical Biology*

Received date: 15 May 2019
Revised date: 27 November 2019
Accepted date: 28 November 2019

Please cite this article as: Yanhao Huo , Lihui Xin , Chuanze Kang , Minghui Wang , Qin Ma , Bin Yu , SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso, *Journal of Theoretical Biology* (2019), doi: https://doi.org/10.1016/j.jtbi.2019.110098

Highlights

- A novel method (SGL-SVM) for tumor classification on two-class and multi-class tumor datasets.
- Gene primary selection reduces the computational complexity of the selection largely.
- Sparse group Lasso considers the group effect variables and select more effective feature genes.
- We investigate different feature gene selection methods and classifiers on the results.
- The proposed method has better performance than other methods.

SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso

Yanhao Huo[a,b,1], Lihui Xin[c,1], Chuanze Kang[a,b], Minghui Wang[a,b], Qin Ma[d], Bin Yu[a,b,e,*]

[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 Science, Dalian University of Technology, Panjin 124221, China*

[d] *Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, Ohio 43210, USA*

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

Abstract: At present, with the in-depth study of gene expression data, the significant role of tumor classification in clinical medicine has become more apparent. In particular, the sparse characteristics of gene expression data within and between groups. Therefore, this paper focuses on the study of tumor classification based on the sparsity characteristics of genes. On this basis, we propose a new method of tumor classification—Sparse Group Lasso (least absolute shrinkage and selection operator) and Support Vector Machine (SGL-SVM). Firstly, the primary election of feature genes is performed on the normalized tumor datasets using the Kruskal-Wallis rank sum test. Secondly, using a sparse group Lasso for further selection, and finally, the support vector machine serves as a classifier for classification. We validate proposed method on microarray and NGS datasets respectively. Formerly, on three two-class and five multi-class microarray datasets it is tested by 10-fold cross-validation and compared with other three classifiers. SGL-SVM is then applied on BRCA and GBM datasets and tested by 5-fold cross validation. Satisfactory accuracy is obtained by above experiments and compared with other proposed methods. The experimental results show that the proposed method achieves a higher classification accuracy and selects fewer feature genes, which can be widely applied in classification for high-dimensional and small-sample tumor datasets. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/SGL-SVM/.

* Corresponding author.
E-mail address: yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

Keywords: Tumor classification; Gene expression data; Feature selection; Sparse group Lasso; Support vector machine.

1. Introduction

At present, the DNA microarray technology of gene expression data from thousands of genes is becoming more and more mature, which is widely used in the field of biomedicine, especially in tumor research (BolonCanedo et al., 2014; Margalit et al., 2005; Lv et al., 2016; Salem et al., 2017). The feature genes extracted from the DNA microarray data can not only be used as tumor classification markers but can be used for more accurate diagnosis, prevention, and treatment of tumors (Shipp et al., 2002; Latkowski et al., 2015; Northcott et al., 2017; Han et al., 2017; Wang et al., 2018). Microarray data is characterized by small samples, high dimensionality, and unbalanced distribution (Zhao et al., 2016; Kang et al., 2017; Jain et al., 2018). There are redundancy and noise in microarray data. To avoid dimensionality catastrophe, the selection of effective feature genes from high-dimensional and massive tumor genes data is the key to tumor classification based on DNA microarray technology (Chandra et al., 2011; Piao et al., 2012; Yu et al., 2013; Bharat et al., 2014). Therefore, building an effective tumor classification model based on the analysis of tumor gene expression profile has become a hot topic in bioinformatics in recent years (Chan et al., 2015; Lu et al., 2017).

In recent years, NGS technology has been widely used in the field of gene expression research. This technology has completely changed the genome research and is replacing microarray technology. NGS technology has many advantages, for example: as the length of NGS read genes increases, the ability to detect repeated regions of the genome will also increase. In terms of experiments, the use of nanogram materials for NGS can reduce or eliminate the dependence on PCR amplification of materials. Because the data is collected across the genome, researchers can simultaneously monitor RNA from all known and undefined genomic characteristics (ie, promoters, exons, non-coding RNAs, and enhancers) or the factors that bind to it, and etc. Microarray technology requires prior knowledge of the genome or genomic characteristics. If the genome annotation is incomplete, incorrect, or outdated, it will directly affect the validity of the array. The main obstacles to microarray analysis come from similar

sequences, which limits microarray analysis to non-repeating parts of the genome and complicates the analysis of related genes (or features). However, after more than 30 years of development, microarray technology has become an effective basic and clinical medical research method, and has become more mature in clinical medical research and application. Therefore, the use of microarray technology is still of great significance for the study of tumor classification. This paper considers the performance of the model on the gene expression data of the two technologies, and proposes a comprehensive model for tumor classification research (Elingaramil et al., 2013; Castillo et al., 2017).

The hypothesis that the microarray data do not satisfy the Gaussian distribution is due to noise interference and data measurement errors in the gene expression data. Kruskal et al. (1952) proposed the Kruskal-Wallis rank sum test as a non-parametric test, which is suitable for small sample datasets. Therefore, to avoid the hypothesis that the data should follow the Gaussian distribution, genetic screening using the Kruskal-Wallis rank sum test may be a very desirable and feasible method.

In recent years, more and more scholars are keen to use sparse expression method as a feature gene selection method for tumor classification (Borgi et al., 2015; Gao et al., 2010; Algamal et al., 2015; Kolali et al., 2016), because it can not only effectively solve dimensional disaster problems, but eliminate redundant genes and significantly improve tumor classification. In the process of sparse theory research, Frank et al. (1993) proposed a biased regression method—Ridge Regression. This method is essentially an improved least squares estimation method that can effectively solve non-full the problem that the rank matrix does not exist in the inversion, which will provide new ideas and new methods for the study of high-dimensional small-sample data. Because this method uses l_2 penalty terms, it does not have the sparsity of variable selection. Based on the "ridge regression" and the "Nonnegative Garrote" proposed by Breiman (1995), the statistician Tibshirani et al. (1996) proposed a new regularization method called Lasso, which is based on the linear regression model for dimensionality reduction. The method has been successfully applied to the selection of variables in the COX model. To achieve the purpose of variable selection, this method compresses the regression coefficients of variables and makes the regression coefficients of some variables to zero, which overcomes the shortcomings of traditional methods in variable selection. However, this method is rather complicated. Several important

algorithms have been proposed to solve this problem. Fu et al. (1998) proposed the Shooting algorithm. Osborne et al. (2000) found that the path to the solution of the Lasso method is piecewise linear, and thus proposed a corresponding homotopy algorithm. The LARS (Least Angle Regression) algorithm proposed by Efron et al. (2004) found the entire regularized path at the cost of a single matrix transformation. The above algorithm effectively solves the Lasso calculation problem.

The algorithm of Lasso does not effectively solve the difference between the two regression coefficients, which leads to the great volatility of the estimation. Tibshirani et al. (2005) added new constraints to improve Lasso's method to better control the volatility of regression coefficients and make it possible to analyze and study high-dimensional and small-sample data more conveniently. Adaptive Lasso where adaptive weights are used to penalizing different coefficients in the l_1 penalty, resolved the problem that the Lasso is inconsistent for variable selection in specific scenarios (Zou 2006). Wang et al. (2014) studied theoretical results of the adaptive Lasso for generalized linear models in terms of diverging number of parameters and ultrahigh dimensionality. For variable selection, Lasso that tends to select more variables is not an ideal method. Smoothly clipped absolute deviation (SCAD) selects variables and estimate coefficients simultaneously. It overcomes the shortcomings of Lasso's biased estimation and improves consistency of parameter estimation and variable selection (Fan et al., 2001). Wang et al. (2016) studied the variable selection of quadratic approximation via the SCAD penalty with a diverging number of parameters. Lasso is only suitable for selecting a single input variable and is not applicative when there exist group variables with high correlations. Different researchers put forward the sparse representation of various deformation. Wang et al. (2017) proposed an adaptive group bridge method. Zou et al. (2005) proposed the Elastic Net. This method is like Lasso. The elastic network can simultaneously perform automatic variable selection and continuous contraction and select multiple sets of related variables. Combining the nonconvex penalty and ridge regression, Wang et al. (2016) proposed the weighted elastic-net to deal with the variable selection of generalized linear models on high dimension. Yuan et al. (2006) proposed that the group Lasso method provides a new idea for analyzing mass data. Group Lasso lacks some theoretical support, and it is not sparse in the group. That is, if a group parameter is non-zero, all parameters in the group are non-zero. For this problem, Liu et al. (2010) proposed the overlap

group Lasso method. Wang et al. (2008) proposed an adaptive group Lasso method. The above methods for the group Lasso methods have made some improvements. There are also some scholars who use group Lasso in combination with other methods and have achieved good results. Ma et al. (2007) proposed a method of supervising group Lasso combining Lasso and group Lasso, which considers the structure of gene expression data clusters. For gene selection and prediction model establishment. Xu et al. (2015) proposed a Bayesian group Lasso variable selection method. Kang et al. (2019) proposed a new method relaxed Lasso and generalized multi-class support vector machine for tumor classification. Foygel et al. (2010) proposed a new linear regression method-single line search method. This method uses a single-variable line search to calculate the exact optimal value of each group and develops the SSLS (Signed Single Line Search) algorithm based on the SLS (Single Line Search) algorithm. The theoretical results provide strong support for sparse group Lasso algorithm. Negahban et al. (2009) also provided a unified framework for sparse group Lasso methods to establish the consistency and convergence rate of regular estimates in high-dimensional situations. The framework provides the possibility for the theoretical analysis of sparse group Lasso and provides a better method for the analysis of high-dimensional and massive data and the selection of the important factors from it.

In the study of tumor classification using gene expression profile data, many new data mining algorithms have been used for tumor classification, which are similar to pattern classification methods. Tumor classification often uses unsupervised and supervises two classification methods. Unsupervised classification methods mainly include Cluster Analysis (CA), Principal Component Analysis (PCA), Independent Component Analysis (ICA) and so on. The advantage of unsupervised classification is that it is possible to find new tumor subtypes and provides a biomedical research direction. The disadvantage is that the existing simple type information is not utilized. Supervised classification methods: k-nearest neighbor (KNN), support vector machine (SVM), Naïve Bayesian (NB), artificial neural networks (ANN), C4.5 decision tree, random forest classification method, have the advantage of being able to learn based on known sample type information, extracting sample classification knowledge. These methods have been successfully applied to the field of tumor classification. For example, Yang et al. (2012) proposed a tumor classification method based on clustering algorithm and Euclidean distance. Luo et al. (2015) propose a Sparse Semi-Supervised Extreme Learning Machine (S3ELM) via joint

sparse regularization for classification. Kar et al. (2015) proposed particle swarm optimization (PSO) methods and adaptive k-nearest neighbor gene selection techniques to achieve tumor classification. Lv et al. (2016) proposed an improved univariate marginal distribution algorithm for gene expression data classification. Guo et al. (2018) proposed a deep learning model, so-called BCDForest, to address cancer subtype classification on small-scale biology datasets. Behrmann et al. (2017) proposed an adapted architecture based on deep convolutional networks to mainly handle the characteristics of mass spectrometry data. For the classification of tumor in two categories, Salem et al. (2017) proposed a methodology that combines both Information Gain (IG) and Standard Genetic Algorithm (SGA) for the tumor classification. Firstly, this method uses information gain for feature selection; secondly it uses a genetic algorithm to reduce feature genes, and finally, it uses genetic programming to perform tumor classification. Genetic algorithms generally improve the classification performance of other classifiers. For multiple categories of cancer classification problems, Liu et al. (2017) proposed a hybrid method based on WLEM to deal with multi-class imbalance problems of cancer microarray data at the feature and algorithm level. At the feature level, a feature-oriented search of each class is performed using the class-oriented feature selection method, so that the features associated with the minority classes are explicitly selected. At the algorithmic level, a further improved WELM strengthens input nodes with high discrimination and integrates the model for generalization. Although the above algorithms have achieved good results in tumor classification, there are some shortcomings in those methods. For example, the PSO algorithm provides a global search method, but this does not guarantee that the result converges to the global best. And it is a heuristic bionic optimization algorithm, and there is no strict theoretical basis. In addition, the efficiency of genetic algorithms is lower than other traditional optimization methods. There is no effective quantitative analysis method for a genetic algorithm in terms of accuracy, feasibility, and computational complexity.

In this study, based on high-dimensional and small-sample tumor datasets and the sparsity of genes between groups and within a group, we present a new tumor classification method—sparse group Lasso-SVM (SGL-SVM). First, we perform a feature gene primary selection on the normalized datasets using the Kruskal-Wallis rank sum test. Second, using sparse group Lasso for further feature gene selection and support vector machines serves as the classifier. To validate the effectiveness of the proposed classification method, we use SVM, KNN, Naïve Bayes, and

Random Forest as the classifier to compare three two-class and five multi-class microarray datasets by 10-fold cross-validation. To validate the effectiveness of our method for NGS data, this paper uses the SGL-SVM method to test two tumor datasets from TCGA, Breast Invasive Carcinoma (BRCA) and Glioblastoma Multiforme (GBM), through 5-fold cross-validation. Results indicate that SGL-SVM method has better performance and selects fewer feature genes, which can significantly distinguish patients from non-patients and different tumor subtypes.

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

2. Materials and methods

For convenience, the new tumor classification method proposed in this paper is called SGL-SVM. The general framework is shown in Fig.1. We have implemented it in R version 3.3.3 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.

The steps of the SGL-SVM tumor classification method are described as follows:

(1) Normalize each initial dataset;

(2) The normalized datasets are pre-selected using the Kruskal-Wallis rank sum test. The p -value is calculated for each gene using Eq. (2), and arranged in ascending order selecting corresponding Top k genes;

(3) Reducing the dimensions of the original datasets to a matrix of $X_{N\times k}$ with N samples and k variables;

(4) SGL is used for the $X_{N\times k}$ data matrix, and the genes are grouped using Eq. (3) to obtain a sparse coefficient vector with a length of k . The feature genes are selected by finding non-zero components of the coefficient vector;

(5) Using support vector machine as a classifier, the training of feature genes is carried out by using Eq. (11). ACC, AUC, and Kappa are calculated by 10-fold cross-validation to evaluate classification methods.

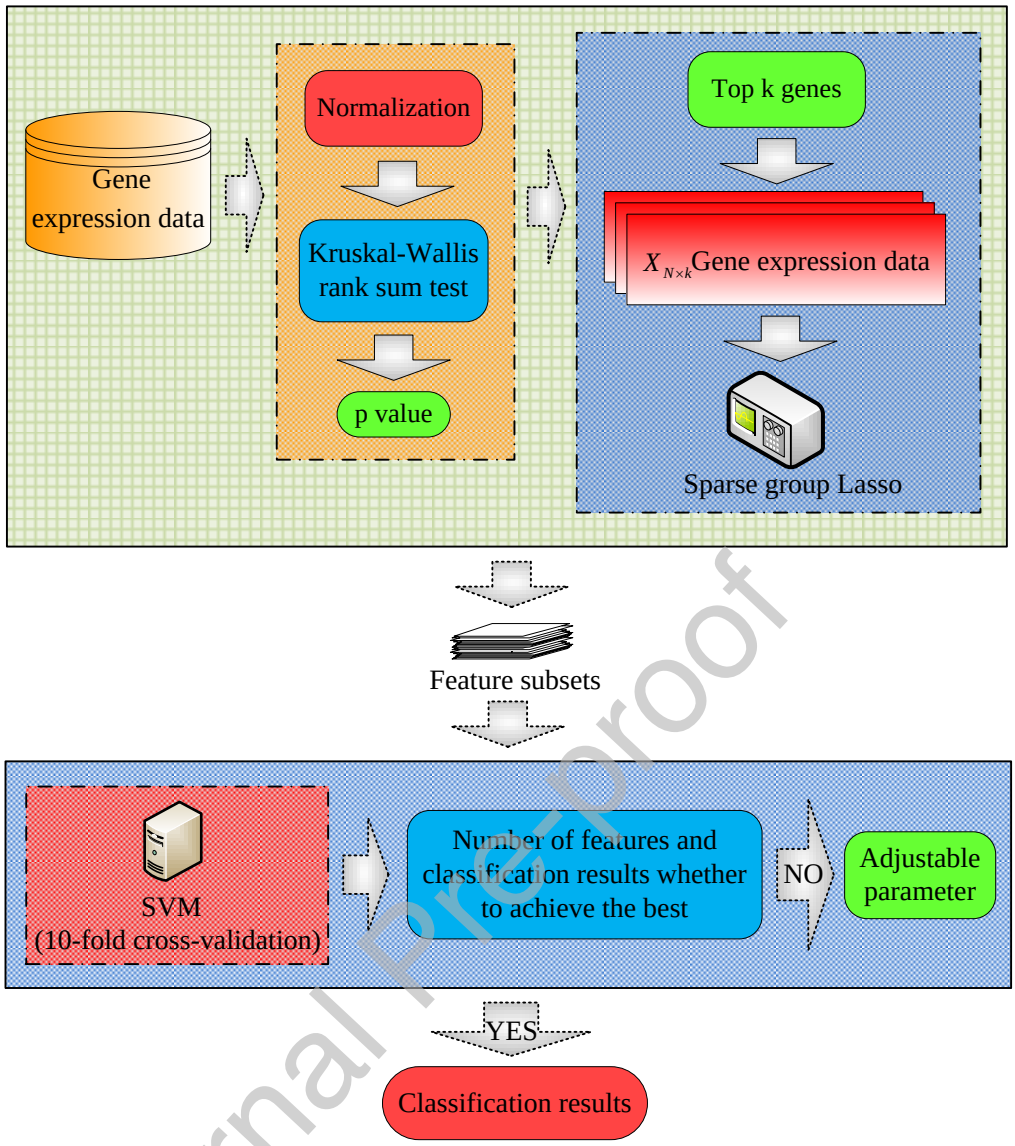

Fig. 1. Flowchart of features gene selection and classification prediction.

2.1. Datasets

To validate the effectiveness of the method, we test the classification method on eight tumor datasets. The tumor dataset has the following characteristics: (1) All experimental datasets are real tumor datasets of high dimensionality and small sample. Among them, the dimensionality of four tumor datasets exceeds 10000, and particularly one tumor dataset has seven categories. (2) The number of genes is greater than the number of samples. (3) These datasets contain a large number of redundant and unrelated genes.

Lung and MLL datasets are from the website: https://github.com/sqsun/kernelPLS-datasets and the rest datasets are from the website: http://www.biolab.si/supp/bicancer/projections/index.html. Table 1 shows the description of the

tumor datasets. The details of the tumor datasets see Supplementary materials Section 1.

Table 1

Information about the test microarray datasets.

Datasets	No. of Sample	No. of Genes	No. of Class	#instance per class	Reference
ALLAML	72	7129	2	47/25	Golub et al. 1999
DLBCL	77	7129	2	58/19	Shipp et al. 2002
Prostate	102	12600	2	52/50	Singh et al. 2002
Stjude	215	12558	7	9/18/42/14/28/52/52	Yeoh et al. 2002
SRBCT	83	2308	4	29/11/18/25	Khan et al. 2001
MLL	72	12533	3	24/28/20	Armstrong et al. 2002
Lymphoma	62	4026	3	42/9/11	Alizadeh et al. 2000
Lung	203	12600	5	139/21/20/6/17	Bhattacharjee et al. 2001

2.2. Methods

The description of ReliefF algorithm, Lasso, Elastic Net, Group lasso, K-Nearest Neighbor, Naïve Bayes, and Random Forest methods see Supplementary materials Section 2.

2.2.1. Kruskal-Wallis rank sum test

The Kruskal-Wallis rank sum test is a method to test whether there is a significant difference in the distribution of multiple overall distributions (Kruskal et al., 1952). Checking whether variables have the same distribution requires constructing statistics which can reflect the differences in the position parameters of these samples and have an exact or an approximate distribution.

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

where $N = \sum_{i=1}^{n} \lambda_i$ is the total number of samples. λ_i is the number of samples of the variable i. $\bar{r}_i = (\sum_{j=1}^{\lambda_i} r_{ij}) / \lambda_i$ is the average rank of each variable. r_{ij} is the rank of the variable i in the sample j. The observed value i of the sample j is arranged in order from small to large, and the foot mark r_{ij} of i is the rank of i. $\bar{r} = 1/2(N+1)$ is the average of the rank sum. The definition of the p value.

$$p = \Pr(\chi^2_{n-1} > H) \tag{2}$$

Finally, the Top k variable is selected directly according to the p value from small to large.

2.2.2. Sparse group Lasso

Group Lasso only has sparsity between groups, and there is no sparsity within the group. The so-called sparsity between groups refers to that at least one set of coefficients is non-zero. The sparsity within group refers to that the components of the coefficient vector is sparse. However, in many practical problems, the effects of variables within groups tend to differ which may limit its application. Simon et al. (2013) proposed sparse group Lasso criterion:

$$\min_{\beta} \frac{1}{2n}\left\| y-\sum_{l=1}^{m} X^{(l)}\beta^{(l)} \right\|_2^2 + (1-\alpha)\lambda\sum_{l=1}^{m}\sqrt{p_l}\left\|\beta^{(l)}\right\|_2 + \alpha\lambda\|\beta\|_1 \tag{3}$$

where $X_{N\times p}$ is divided into m submatrix $X^{(1)}, X^{(2)},\ldots, X^{(m)}$, $X^{(i)}$ is a submatrix of X . $\beta^{(i)}$ coefficient vector in group i with p_i length. $\alpha \in [0,1]$, the penalty of it is a convex combination of Lasso and group Lasso penalty. Based on Nesterov's method for generalized gradient descent, an algorithm is developed to calculate $\beta^{(i)}$. Available R algorithm is provided publically in the package SGL.

For group k , $\hat{\beta}^{(k)}$ satisfies

$$\frac{1}{n}X^{(k)T}(y-\sum_{i=1}^{m}X^{(i)}\hat{\beta}^{(i)}) = (1-\alpha)\lambda g + \alpha\lambda h \tag{4}$$

where g and h , evaluated at $\hat{\beta}^{(k)}$, are secondary gradient of $\left\|\hat{\beta}^{(k)}\right\|_2$ and $\left\|\hat{\beta}^{(k)}\right\|_1$ respectively.

$$g = \begin{cases} \dfrac{\hat{\beta}^{(k)}}{\left\|\hat{\beta}^{(k)}\right\|_2} & \text{if } \hat{\beta}^{(k)} \neq 0 \\ \in \left\{ g : \|g\|_2 \le 1 \right\} & \text{if } \hat{\beta}^{(k)} = 0 \end{cases}$$

$$h_j = \begin{cases} \operatorname{sign}(\hat{\beta}_j^{(k)}) & \text{if } \hat{\beta}_j^{(k)} \neq 0 \\ \in \{h_j : |h_j| \le 1\} & \text{if} \hat{\beta}_j^{(k)} = 0 \end{cases}$$

For sparsity between groups, group k satisfies with $\hat{\beta}^{(k)} = 0$ if

$$\left\| f(X^{(k)T} r_{(-k)}/n, \alpha\lambda) \right\|_2 \le (1-\alpha)\lambda \tag{5}$$

where $r_{(-k)} = y - \sum_{i\neq k} X^{(i)}\hat{\beta}^{(i)}$ and $f(\cdot)$ is the coordinate-wise soft threshold operator:

$$(f(z,\alpha\lambda))_i = \operatorname{sign}(z_i)(|z_i| - \alpha\lambda)_+$$

If $\hat{\beta}^{(k)} \neq 0$, for sparsity within group, find $\hat{\beta}^{(k)}$ to minimize

$$\frac{1}{2n}\left\| r_{(-k)} - X^{(k)}\beta^{(k)} \right\|_2^2 + (1-\alpha)\lambda\left\|\beta^{(k)}\right\|_2 + \alpha\lambda\left\|\beta^{(k)}\right\|_1 \tag{6}$$

We define non-penalized loss function by $l(r_{(-k)},\beta)=\frac{1}{2n}\left\|r_{(-k)}-X^{(k)}\beta\right\|_2^2$.

The modern approach to gradient descent is to consider it as an optimal minimization scheme. We focus on the loss function, centered at a point β_0 by

$$l(r_{(-k)},\beta)\le l(r_{(-k)},\beta_0)+(\beta-\beta_0)^T\nabla l(r_{(-k)},\beta_0)+\frac{1}{2\omega}\left\|\beta-\beta_0\right\|_2^2 \tag{7}$$

where ω is sufficiently small so that the quadratic term dominates the Hessian of the loss function, $\nabla l(r_{(-k)},\beta_0)=-X^{(k)T}r_{(-k)}/n$. We add a penalty to Eq. (6), which majors the objective in Eq. (5).

We find $\hat{\beta}^{(k)}$ to minimize

$$\tilde{M}(\beta)=\frac{1}{2\omega}\left\|\beta-(\beta_0-\omega\nabla l(r_{(-k)},\beta_0))\right\|_2^2+(1-\alpha)\lambda\left\|\beta\right\|_2+\alpha\lambda\left\|\beta\right\|_1 \tag{8}$$

Combining the secondary gradient conditions, we get $\hat{\beta}^{(k)}\neq 0$ if

$$(1+\frac{\omega(1-\alpha)\lambda}{\left\|\hat{\beta}\right\|_2})\hat{\beta}=f(\beta_0-\omega\nabla l(r_{(-k)},\beta_0),\omega\alpha\lambda) \tag{9}$$

We assign new values to $\hat{\beta}$ by iterating and converge to the optimal solution for $\hat{\beta}^{(k)}$ given fixed values of the other coefficient vectors.

$$U(\beta_0,\omega)=(1-\frac{\omega(1-\alpha)\lambda}{\left\|f(\beta_0-\omega\nabla l(r_{(-k)},\beta_0),\omega\alpha\lambda\right\|_2})_+ f(\beta_0-\omega\nabla l(r_{(-k)},\beta_0),\omega\alpha\lambda) \tag{10}$$

The combination of different methods can select feature variables more effectively. Therefore, this paper combines the Kruskal-Wallis rank sum test (abbreviated as KW) with the regularization method SGL to obtain the feature selection method, which is called KW&SGL method. The algorithm of KW&SGL is described as follows:

Algorithm1 for the KW&SGL

Input: $\boldsymbol{X}_{N\times p}$ // Initial matrix

$\boldsymbol{y}$ // Response variables

Output: ***idx*** // The index of coefficient β vector of non-zero components

01: Suppose that each gene is a group.

Establish the hypothesis H_0 : all groups had the same distribution. H_1 : there were at least two genes had a different distribution. We build test statistic H , using Eq. (2) to obtain $\boldsymbol{p}$

02: Sort $\boldsymbol{p}$ in increasing order and obtain Top k features and matrix $X_{N\times k}$

03: Suppose that Top k features are divided into m groups with p_l variables in each group. We get submatrix $X_1, X_2, \ldots, X_m$

Out loop: Cyclically iterate through the groups

04: At each group k check if the group's coefficients are identically 0, by seeing if they obey Equation (5)

05: If not, within the group apply Inner loop

Inner loop: Until convergence iterate:

07: Update the center by $\theta \leftarrow \hat{\beta}^{(k)}$

08: Update $\hat{\beta}^{(k)}$ from Equation (10), by $\hat{\beta}^{(k)} \leftarrow U(\theta, t)$

09: **End**

10: **Return** ***idx***

2.2.3. Support vector machine

Support vector machine (SVM) is a supervised learning method proposed by Vapnik et al. (1995). It has a wide range of applications in statistical classification and regression analysis. It has many unique advantages in solving small samples, non-linear and high-dimensional pattern recognition, and can be widely applied to machine learning problems such as function fitting. The basic principle of this method is to find a fractal hyperplane for the training set V in the sample space, which will maximize the separation of categories. In addition, as a quadratic programming problem, SVM can find the globally unique optimal solution, so it can avoid the occurrence of the local minimum. The principle and solution process are as follows:

Given a sample dataset $V = \{(x_i, y_i) \mid x_i \in R^m, y_i \in \{-1, +1\}, i = 1, \ldots n\}$, then the support vector machine discriminant function is:

$$f(x) = sign(\sum_{i=1}^{n} A_i y_i K(x, x_i) + B) \tag{11}$$

where $K(x, x_i)$ is the kernel function, i is the number of support vector. The kernel function is very important in the training of support vector, which can effectively overcome the dimensionality disaster problem. The proper kernel function can improve the prediction accuracy

of the classification model. Commonly used kernel functions are linear kernel function, polynomial kernel function, radial basis function, Sigmoid kernel function and so on.

2.3. Performance evaluation

In statistical learning theory, jackknife, self-consistency, independent test, and k-fold cross-validation are commonly used to test the Validity of the model. This paper uses k-fold cross-validation to test the performance of the prediction model. And, we use the statistics of Accuracy, AUC (Sing et al., 2005) and Kappa (Cohen 1960) to evaluate the results of the prediction system. The above three evaluation indicators are defined in Section 3 of supplementary materials.

3. Results and discussion

3.1. Effect of further feature selection algorithm on results

In section 2, the KW&SGL feature selection method is proposed. Firstly, the Kruskal-Wallis rank sum test (the initial selection) is used for the normalized tumor datasets. Top k gene is defined in order that p-value and data matrix $X_{N\times k}$ is generated. The primary reason for the initial selection is that it can remove a large number of unrelated genes and significantly reduce the dimensionality of the dataset, which brings convenience to feature selection.

Choosing the appropriate k value as the dimension is the key to the initial selection. If the k value is too small, it will lead to the loss of important feature genes. If the value of k is too large, the redundant gene will be increased.

We take Top k genes as the feature genes, which vary from 1 to 100. The support vector machine (SVM) with radial basis function (RBF) as kernel function is used as the classifier. The Top k genes are tested by 10-fold cross-validation. As the value of k increases, the classification accuracy (ACC) changes as shown in Fig. 2.

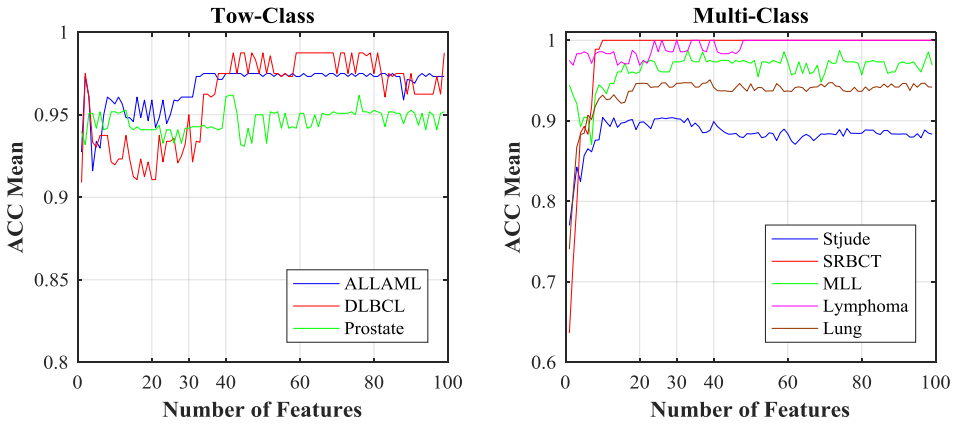

Fig. 2. Relationship between the number of genes selected by KW and classification accuracy on the two-class and multi-class datasets.

Fig. 2 shows that for the two-class datasets, when the k-value is greater than 40, the ACC reaches the maximum value. For multi-class datasets, the ACC reaches a maximum when the k-value is larger than 20. The experimental results show that for most datasets, the uniform k-value is 100, and selected genes have better ACC. However, the performances of MLL, Lung, and Stjude datasets are not desired. Through many experiments, the k-value of Lymphoma and Stjude datasets is placed at 150 and 500, respectively, so that the selected genes can achieve better ACC. To obtain higher ACC, we further use SGL to select feature genes.

The key of SGL is an appropriate penalty factor (λ value). The different λ-value corresponding to different coefficient vectors will have a corresponding loss of square error. To select the appropriate λ-value, take the ACC of the feature genes corresponding to the λ-value as the index. Therefore, we directly discuss the relationship between λ-value corresponding feature genes and ACC.

For all datasets, support vector machine with RBF kernel function is used as the classifier. Through a large number of experiments, 10-fold cross-validation training is conducted on feature genes to obtain the highest ACC. The classification accuracy has reached the maximum value for all datasets when the number of selected feature genes is lower than 30. The relationship between the number of feature genes (NF) and classification accuracy (ACC) is depicted in Fig. 3.

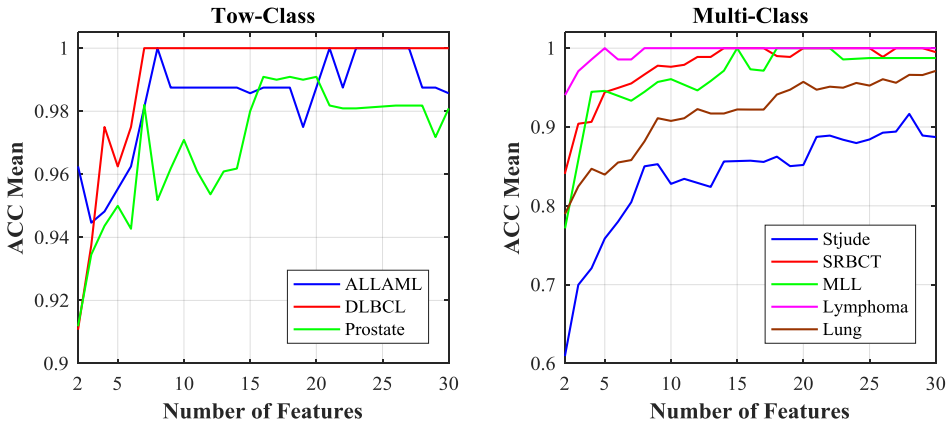

Fig. 3. Relationship between the number of genes selected by KW&SGL and classification accuracy on the two-class and multi-class datasets.

Fig. 3 shows that for ALLAML dataset, the classification accuracy is 100% using eight genes. For DLBCL dataset, the classification accuracy rate is 100% using seven genes. For SRBCT, MLL and Lymphoma datasets, the classification accuracy is 100% when NF were 15, 15 and 8, respectively. On the maximum ACC, taking the corresponding NF, KW and KW&SGL comparison results shown in Table 2, NF represents the number of feature genes.

Table 2
Comparison of ACC between KW and KW&SGL.

Datasets	KW		KW&SGL	
	NF	ACC (%)	NF	ACC (%)
ALLAML	100	97.32	8	**100**
DLBCL	100	97.5	7	**100**
Prostate	100	95.18	16	**99.09**
Stjude	500	88.85	28	**91.67**
SRBCT	100	**100**	15	**100**
MLL	100	96.90	15	**100**
Lymphoma	150	**100**	8	**100**
Lung	100	94.18	20	**95.73**

Table 2 shows that for most datasets, NF obtained by SGL is fewer, and the ACC is better. For ALLAML, DLBCL, and MLL datasets, the ACC is raised to 100%. For the SRBCT and Lymphoma datasets, fewer fifteen and eight feature genes are selected, maintaining 100% classification accuracy. To illustrate the sparsity of SGL selected features, we study the relative positions of feature genes in Top k genes. Rankings of the feature genes for the Top k genes are shown in Table 3.

Table 3

Rankings of the feature genes in Top k genes.

Datasets	Rankings of the feature genes
ALLAML	1,2,14,19,20,25,63,64
DLBCL	5,6,35,41,59,60,61
Prostate	1,2,4,5,6,7,9,13,21,25,39,41,56,59,62,75
Stjude	1,10,12,13,14,15,18,28,33,36,42,51,58,111,132,145,151, 170,173,202,212,249,260,380,389,426,430,476
SRBCT	1,2,3,4,7,8,13,17,19,20,23,24,39,45,87
MLL	5,6,7,8,11,13,14,19,25,26,27,28,40,51,83
Lymphoma	1,8,12,27,84,111,123,133
Lung	1,6,7,9,17,19,32,39,47,56,60,61,62,63,65,75,79,80,82,98

Table 3 shows that selected feature genes are more dispersed than expected in the Top k genes and not subject to the gene rankings. For the Stjude and Lymphoma datasets, the last of feature genes were respectively 476 and 133. The experimental results show that it is reasonable to extend the k-value to 500 and 150. KW considers only a single gene itself, simply arranged the genes by p-value. They are combined into a group of candidate genes, without considering the role of the genes on the classification of samples. While SGL pays attention to the characteristics of the group effect of genes. It takes all Top k genes into account and selects fewer and more effective feature genes. Therefore, the sparsity and feature selection are inextricably linked, which has an important impact on the classification results.

Feature genes have profound biological significance for tumor classification. We give descriptions of the selected feature genes of SGL and their biological significance (The specific biological significance of feature genes see Supplementary materials Section 4).

3.2. Effect of feature selection algorithm on results

This section discusses the impact of different feature selection algorithm on classification. The validity of the feature selection algorithm is determined by two factors, the number of feature genes and the classification accuracy. In this paper, four different feature gene selection methods, the Elastic Network (EN), the Group Lasso (GL), SGL, and ReliefF, are used to select Top k genes and obtained different feature genes. For all datasets, RBF is selected as the support vector machine for kernel function as the classification algorithm, and 10-fold cross-validation is utilized to train the feature genes and test the classification accuracy. After experimental verification, for all datasets, the number of selected feature genes is less than 30, the classification accuracy has

reached a maximum. We discuss the influence of different feature selection algorithm on classification in the range 1-30. The relationship between the number of feature genes (NF) and classification accuracy (ACC) of each dataset is shown in Fig. 4, Fig.5.

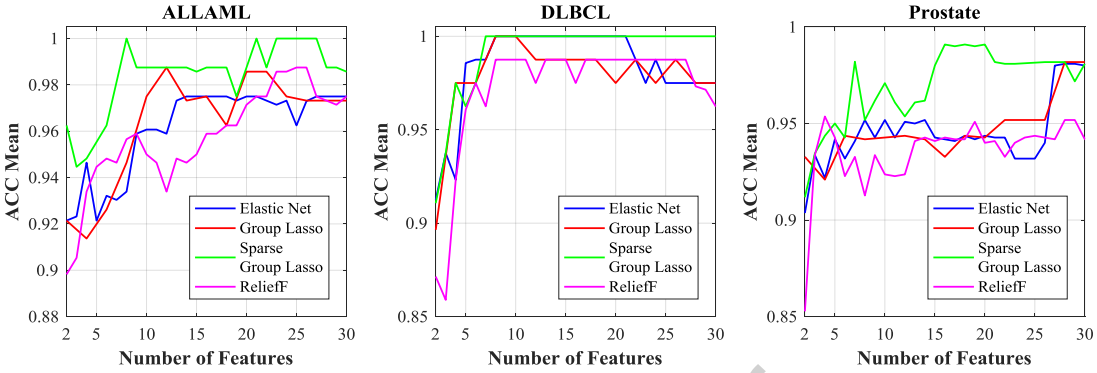

Fig. 4. Relationship between the number of features selected by different feature selection methods (EN, GL, SGL, ReliefF) and classification accuracy by 10-fold cross-validation on two-class datasets.

Fig. 4 shows that the classification accuracy of SGL is generally higher than that of EN, GL, and ReliefF for ALLAML dataset. For DLBCL dataset, the classification accuracy of SGL is stable at 100%, and the classification accuracy of EN is less than 90% when the number of feature genes is greater than or equal to 7. For the Prostate dataset, SGL has a classification accuracy of up to 99.09% and a maximum of 5% difference from ReliefF.

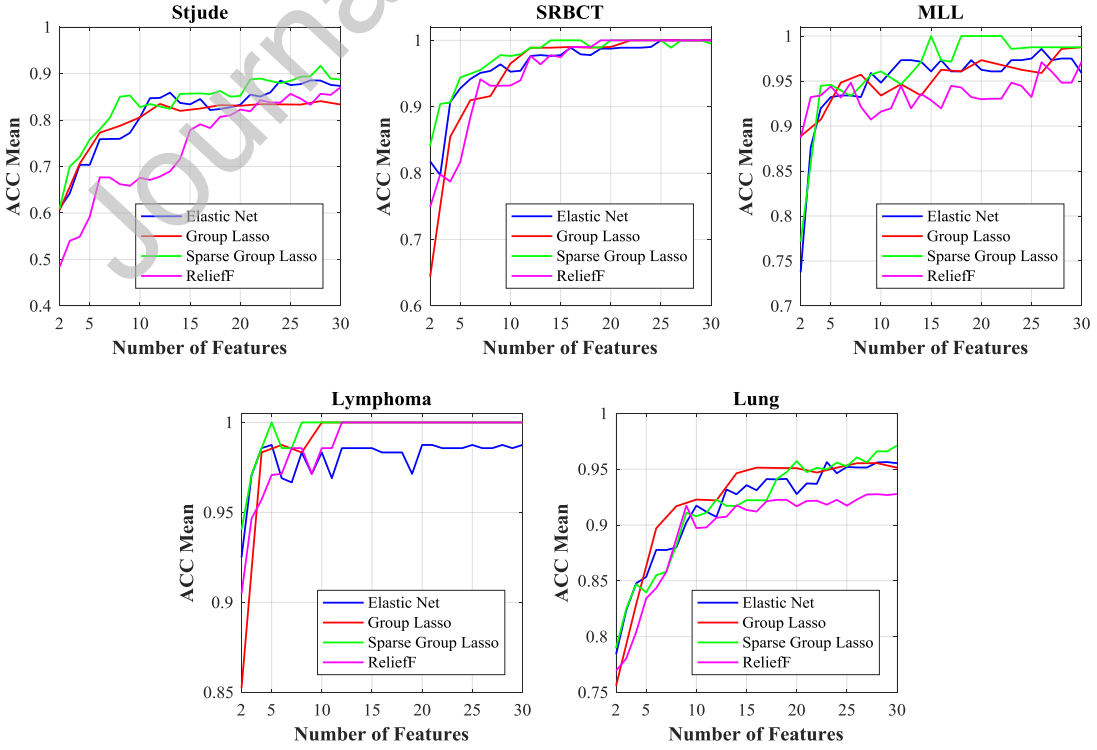

Fig. 5. Relationship between the number of features selected by different feature gene selection methods (EN, GL, SGL, ReliefF) and classification accuracy by 10-fold cross-validation on multi-class datasets.

Fig. 5 shows that for the Stjude dataset, the SGL classification accuracy is generally higher than EN, GL, and ReliefF. For the MLL dataset, the number of feature genes is 15. The classification accuracy of SGL is 100%, which is 5% higher than GL. For Lymphoma dataset, the classification accuracy is 100%, and the number of feature genes corresponding to SGL, GL, and ReliefF is 5, 10 and 12 respectively. For Lung dataset, the classification accuracy of EN and GL is higher than that of SGL when the number of feature genes is less than 20.

For two-class and multi-class datasets, the classification accuracy of EN, GL, SGL, and ReliefF fluctuate significantly with the increase of NF, and the SGL is more stable. Therefore, in general, the feature genes selected by SGL are more efficient and reliable, which achieves better classification accuracy.

For each dataset, the highest classification accuracy (%), the number of feature genes, AUC and Kappa for the four methods are given in Table 4 and Table 5. Sun et al. (2014) indicated that in two-class datasets, AUC evaluates the generalization performance of two types of the samples classification algorithm. In multi-class datasets, Kappa evaluates the generalization performance of multi types of the samples classification algorithm. To discuss the impact of different feature selection algorithms on the classification model. We evaluate EN, GL, SGL, and ReliefF from the perspective of classification accuracy and the number of feature genes.

Table 4
Comparison of SGL with EN, GL, and ReliefF for two-class tumor datasets.

Datasets	ReliefF		EN		GL		SGL	
	NF	ACC (AUC)	NF	ACC (AUC)	NF	ACC (AUC)	NF	ACC (AUC)
ALLAML	25	98.75(98.7)	22	97.5(100)	12	98.75(99.3)	**8**	**100(100)**
DLBCL	37	100(100)	8	**100(100)**	8	**100(100)**	**7**	**100(100)**
Prostate	28	95.18(96.8)	**8**	95.18(93.8)	28	98.18(98.7)	16	**99.09(98.8)**

Table 4 shows that for the ALLAML dataset, SGL only selects eight feature genes, and the classification accuracy rate reaches 100%. For the DLBCL dataset, SGL only selected seven feature genes, and the classification accuracy rate reached 100%. For the Prostate dataset, 16 feature genes are selected by SGL, the classification accuracy was 99.09%, 30 feature genes are

selected by GL, and the classification accuracy was only 98.18%. ReliefF selected 28 feature genes, and the classification accuracy was only 95.18%.

Table 5

Comparison of SGL with EN, GL, and ReliefF for the multi-class dataset.

Datasets	ReliefF		EN		GL		SGL	
	NF	ACC (Kappa)	NF	ACC (Kappa)	NF	ACC (Kappa)	NF	ACC (Kappa)
Stjude	50	89.37(0.868)	**28**	88.5(0.857)	50	85.82(0.823)	**28**	**91.67(0.897)**
SRBCT	19	**100(1)**	25	**100(1)**	22	**100(1)**	**15**	**100(1)**
MLL	42	97.32(0.958)	26	98.57(0.979)	34	**100(1)**	**15**	**100(1)**
Lymphoma	12	**100(1)**	**5**	98.75(0.973)	10	**100(1)**	8	**100(1)**
Lung	45	94.21(0.886)	29	95.64(0.915)	28	95.55(0.912)	**20**	**95.73(0.919)**

Table 5 shows that for the Stjude dataset, 28 feature genes are chosen by SGL, and 50, 28, and 50 feature genes are selected by ReliefF, EN, and GL, respectively. SGL classification accuracy is 5.85% higher than GL and 3.17% higher than EN. For the SRBCT dataset, SGL selects 15 feature genes, and the ACC is 100%. For the MLL data set, the classification accuracy of SGL and GL reaches 100%, SGL only selects 15 feature genes, while GL selects 34 feature genes. For the Lymphoma dataset, SGL selects eight feature genes with ACC 100%. For the Lung dataset, SGL selects 20 feature genes, while ReliefF, EN, and GL select 45, 29, and 28 feature genes, respectively. SGL reached the highest ACC as 95.73%.

Regularized methods select subsets of candidate genes to obtain sparse coefficient vectors, and the feature genes are obtained by searching non-zero components of the vectors. The sparseness between groups and sparseness within groups make SGL more advantageous in the selection of feature genes. From the perspective of classification accuracy, the advantage of SGL for two-class tumor datasets is not clear. However, SGL performs better than ReliefF, EN, and GL on multi-class datasets. For the ALLAML, DLBCL, SRBCT, MLL and Lymphoma datasets, the classification accuracy reached 100%. From the perspective of the number of feature genes, the strength of SGL is more prominent. For the ALLAML, DLBCL, Stjude, MLL, and Lung data sets, a minimum of 8, 7, 28, 15, and 20 feature genes are chosen. Selecting SGL as a feature gene selection method for the classification model is both reasonable and effective.

In this study, DLBCL and Lymphoma datasets are taken as examples to illustrate the SGL classification effect intuitively with a three-dimensional scatter plot. The scatter plots of other tumor datasets see Supplementary materials Section 5.

For the DLBCL dataset, we randomly selected three feature genes {L19686_rna1_at, Z21966_at, X16983_at} from the SGL-selected feature genes {AB002409_at, D87119_at, L19686_rna1_at, X02152_at, X16983_at, Z21966_at, M14328_s_at}, as shown in Fig. 6.

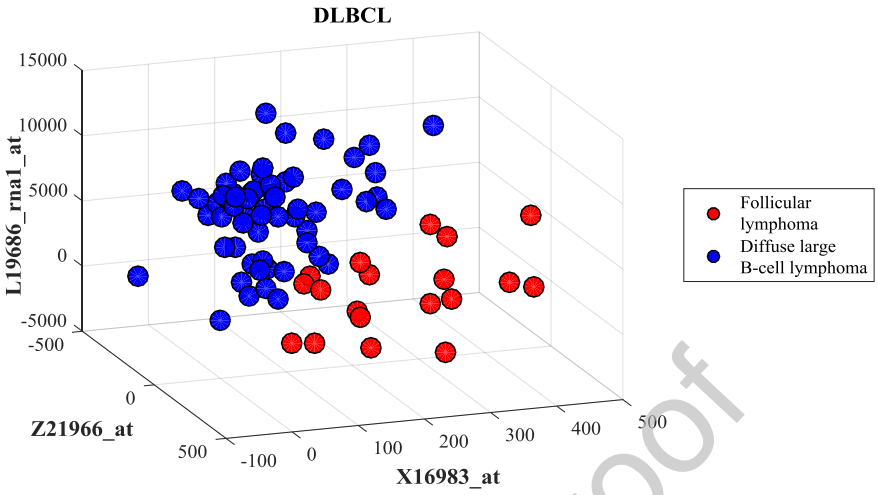

Fig. 6. The three-dimensional scatter plot of the three genes {L19686_rna1_at, Z21966_at, X16983_at} of the DLBCL dataset.

The DLBCL dataset consists of two subtypes of diffuse large B-cell lymphoma and follicular lymphoma, a dataset of 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 SGL to select seven feature genes, SVM as a classification algorithm, the ACC reaches 100%.

Fig. 6 shows that for the DLBCL dataset, a feature selected by SGL clearly distinguishes follicular lymphocyte samples from diffuse large B-cell samples. Therefore, SGL has an outstanding performance of feature selection, for two-class tumor datasets.

For the Lymphoma dataset, we randomly selected three feature genes {GENE2374X, GENE653X, GENE1553X} from the SGL-selected feature genes {GENE653X, GENE2374X, GENE833X, GENE639X, GENE1608X, GENE712X, GENE712X, GENE708X, GENE1553X}, as shown in Fig. 7.

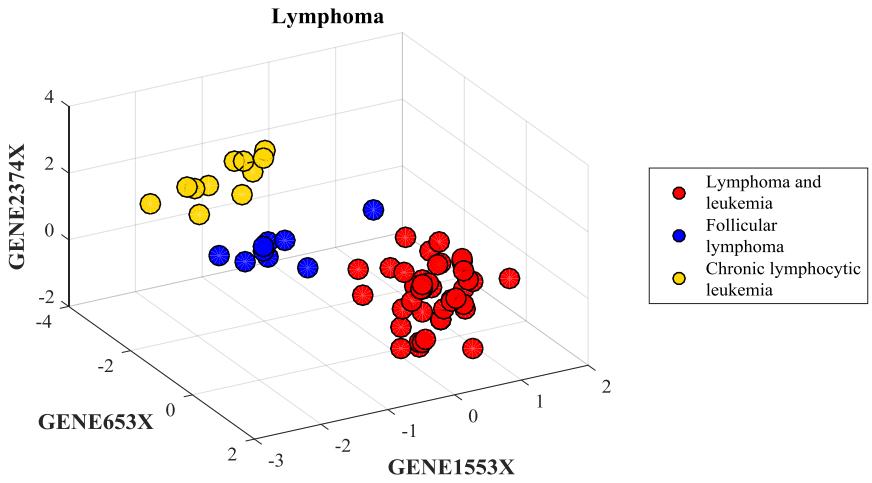

Fig. 7. The three-dimensional scatter plot of the three genes {GENE2374X, GENE653X, GENE1553X} of the Lymphoma dataset.

The Lymphoma dataset includes 62 samples that originate in the patient. The patients are divided into three classes: lymphoma and leukemia (DLCL, 42 samples), follicular lymphoma (FL, nine samples) and chronic lymphocytic leukemia (CLL, 11 samples). Each sample contains 4026 genes. Using SGL to select eight feature genes, SVM as a classification algorithm, the ACC reaches 100%.

Fig. 7 shows that the boundaries of the scatter plot are very clear. For the Lymphoma dataset, the selected feature genes can distinguish Lymphoma and leukemia samples, Follicular Lymphoma samples and chronic lymphocytic leukemia samples. Therefore, SGL has an outstanding performance of feature selection, for multi-class tumor datasets.

3.3. Effect of classification algorithm on results

The result of the tumor classification model is also limited by the classifier. Under the condition of the feature gene selection algorithm being determined, the classifier with stronger performance usually has a better classification effect. This paper investigates four classifiers: SVM, KNN, Naive Bayes and Random Forest.

In this study, we use 10-fold cross-validation on the tumor dataset to test ACC. In terms of parameter selection, support vector machine with RBF as a kernel function is used as the classifier. When KNN is used as the classifier, the best classification result achieved when five is selected by continuously adjusting the test. When Naive Bayes is used as the classifier, the Laplace parameter is chosen to be 1 owing to the incomplete class of the sample. When Random Forest is used as the

classifier, select $mtry = 1$ (the number of genes for each split point) and $B = 100$ (number of decision trees). When SGL is used as the feature gene selection method, the ACC (%), AUC and Kappa obtained by different classifiers are shown in Table 6 and Table 7.

Table 6

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

Datasets	Feature	SVM		KNN		Naïve Bayes		Random Forest	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
ALLAML	8	**100**	100	96.25	96.33	98.75	100	98.75	99.33
DLBCL	7	**100**	100	98.75	100	98.75	98.3	96.25	100
Prostate	16	**99.09**	98.8	97.18	99	97.18	99.3	98.18	98.17

Table 6 shows that for ALLAML and DLBCL datasets, only classification accuracy using SVM as the classification algorithm can reach 100%. For the Prostate dataset, using SVM as a classification algorithm, the classification accuracy is 99.09%, which is 1.91% higher than that of KNN and Naïve Bayes. Therefore, for two-class datasets, SVM has better classification results.

Table 7

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

Datasets	Features	SVM		KNN		Naïve Bayes		Random Forest	
		ACC	Kappa	ACC	Kappa	ACC	Kappa	ACC	Kappa
Stjude	28	**91.67**	**0.897**	88.41	0.858	90.27	0.881	86.55	0.833
SRBCT	15	**100**	**1**	95.67	0.942	**100**	**1**	**100**	**1**
MLL	15	**100**	**1**	96.25	0.942	97.5	0.963	93.39	0.899
Lymphoma	8	**100**	**1**	**100**	**1**	**100**	**1**	98.33	0.96
Lung	20	95.73	0.919	96.05	0.917	**96.55**	**0.932**	93.12	0.849

Table 7 shows that for the Stjude dataset, the classification accuracy using SVM as a classification algorithm is 91.67%, which is 5.12% higher than that of Random Forest. For the SRBCT dataset, the classification accuracy is 100% using SVM, Naive Bayes, and Random Forest as the classification algorithm. For MLL dataset, using SVM and Random Forest as a classification algorithm, the classification accuracy is 100%, 6.61% higher than Random Forest. For the Lymphoma dataset, using SVM, KNN and Naïve Bayes as a classification algorithm, the classification accuracy is 100%. For the Lung dataset, using Naïve Bayes as a classification algorithm, the classification accuracy is 0.82% higher than SVM.

In summary, the experimental results show that SVM has the best performance. SVM achieves the highest ACC on other datasets, except for the Lung dataset. The ACC on the ALLAML, DLBCL, SRBCT, MLL, Lymphoma datasets reaches 100%. Therefore, it is reasonable and effective to select the SVM as the classifier.

3.4. Comparison with other methods

To prove the effectiveness of the proposed method, we compare the classification accuracy and the number of feature genes in other papers. A total of 11 different methods are selected to compare with SGL-SVM. SGL-SVM is compared with other methods in classification accuracy and feature gene numbers, as indicated in Table 8 and Table 9.

The eleven methods in Table 8 and Table 9 are described below. The expanded VizRank is a classification method for visualizing tumor datasets (Mramor et al., 2007). MDR (Multi-Dimensional Ranker) is a method of classifying tumors using a multidimensional sorter (Hewett et al., 2008). MSFCM+WNN is a tumor classification method that selects the position of the wavelet neural network translation vector through an improved point-symmetry-based fuzzy mean-value algorithm (Zainuddin et al., 2011). RNBC is a robust naïve Bayes classifier method (Chandra et al., 2011). HBE (Hyper-Box Enclosure) is a tumor classification method with mixed-integer linear programming (Dagliyan et al., 2011). ECBGS is an ensemble correlation-based gene selection method based on Symmetric Uncertainty (SU) and SVM (Piao et al., 2012). kernelPLS is a kernel partial least squares-based filtering method (Sun et al., 2014). MOEDA is a method for classifying gene expression data using a multi-objective heuristic algorithm (Lv et al., 2016). MRSRC is one of the maxdenominator reweighted sparse representation classification methods (Li et al., 2017). BagBoosting is an algorithm for tumor classification (Dettling 2004).

Table 8
Comparison proposed method with some other gene selection methods in two-class datasets.

Dataset	Experiments	Methods	10-fold CV ACC (%)	NF
DLBCL	Mramor et al. 2007	VizRank	90.89	8
	Dagliyan et al. 2011	HBE	92.25±1.46	—
	Glaab et al. 2012	BioHEL	95.00	—
	Sun et al. 2014	kernelPLS	97.50	—
	Lv et al. 2016	MOEDA	99.00	4
	This paper	**SGL-SVM**	**100**	**7**
Prostate	Glaab et al. 2012	BioHEL	94	—
	Lv et al. 2016	MOEDA	96	12
	Sun et al. 2014	kernelPLS	97.3	—
	Hewett et al. 2008	MDR	97.87	6
	Piao et al. 2012	ECBGS	98.04	46

	This paper	**SGL-SVM**	**99.09**	**16**
ALLAML	Li et al. 2017	MRSRC	83.87	—
	Piao et al. 2012	ECBGS	90.28	27
	Chandra et al. 2011	RNBC	94.29	—
	Sun et al. 2014	kernelPLS	97.50	—
	Mramor et al. 2007	VizRank	98.57	8
	Zainuddin et al. 2011	MSFCM+WNN	98.61	10
	This paper	**SGL-SVM**	**100**	**8**

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

It can be seen from Table 8, for the DLBCL tumor dataset, Mramor et al. (2007) used the extended VizRank method and selected eight feature genes to obtain only 90.89% classification accuracy. Our method selected seven feature genes, the classification accuracy rate reached 100%, and the classification accuracy rate is significantly higher than the extended VizRank method by 9.11% because our method considers the group structure and interactions between genes. For the Prostate tumor dataset, Piao et al. (2012) used the ensemble Correlation-Based Gene Selection (ECBGS) method. By analyzing the correlation between the signature genes, 46 Prostate tumor genes were selected as the signature genes. Utilizing SVM as a classifier to test, the classification accuracy rate reached 98.04%. In this paper, we analyzed the sparseness of intra-group distribution among the feature genes. Only 16 feature genes were selected and tested using SVM classifier. The classification accuracy rate reached 99.09%, which is 1.05% higher than that of ECBGS method. For the ALLAML tumor dataset, the method of this article only selects eight feature genes, and the classification accuracy rate reached 98.75%, which is higher than the classification accuracy rate of other methods.

Table 9

Comparison proposed method with some other gene selection methods in multi-class datasets.

Dataset	Experiments	Methods	10-fold CV ACC (%)	NF
MLL	Mramor et al. 2007	VizRank	91.61	8
	This paper	**SGL-SVM**	**100**	**15**
Stjude	Sun et al. 2014	ReliefF	86.20(5-fold)	1000
		mRMR	88.90(5-fold)	1000
		kernelPLS	89.90(5-fold)	1000
	This paper	**SGL-SVM**	**91.67**	**28**
SRBCT	Lv et al. 2016	MOEDA	96.00	4
	Mramor et al. 2007	VizRank	97.18	8
	Dagliyan et al. 2011	HBE	97.50	—

	Li et al. 2017	MRSRC	98.80	—
	This paper	**SGL-SVM**	**100**	**15**
	Dettling et al. 2004	BagBoost	98.39	10
Lymphoma	Piao et al. 2012	ECBGS	100	9
	This paper	**SGL-SVM**	**100**	**8**
	Chandra et al.2011	RNBC	88.50	—
Lung	Lv et al. 2016	MOEDA	96.00	25
	This paper	**SGL-SVM**	**95.73**	**20**

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

As can be shown in Table 9, for the MLL tumor dataset, Mramor et al. (2007) used the extended VizRank method to visualize tumor datasets and used the 10-fold cross-validation method to reach a classification accuracy of only 91.61%. In this paper, we selected 15 tumor feature genes and used SVM as a classifier to categorize tumors. The classification accuracy achieved 100%, which is significantly higher than the extended VizRank method by 8.39%. For the Stjude tumor dataset, Sun et al. (2014) used ReliefF, Maximum Relevance Minimum Redundancy, and Kernel Partial Least Squares methods to obtain 1000 feature genes, which were classified by the SVM classifier and used the 5-fold cross-validation method. Classification accuracy rates are 86.20%, 88.90%, and 89.90%, respectively. In this paper, only 28 feature genes were selected and classified by SVM as a classifier. The classification accuracy rate was 91.67% with 10-fold cross-validation. The proposed method is 5.47%, 2.68%, and 1.68% higher than ReliefF, Maximum Relevance Minimum Redundancy and Kernel Partial Least Squares, respectively. For the SRBCT tumor dataset, Li et al. (2017) utilized the MRSRC method with a classification accuracy of 98.8%. Our method not only considers the sparseness between the feature genomes but also notes the sparseness within the feature genome. Therefore, when 15 feature genes were selected, the classification accuracy rate reached 100%, which is 1.2% higher than the MRSRC method. For the Lymphoma tumor dataset, the proposed method only selected eight feature genes, and classification accuracy reached 100%. Piao et al. (2012) utilized the ECBGS method to select nine feature genes, and the classification accuracy also reached 100%. This is the result of considering the correlation between the signature genes in both the ECBGS method and the methods herein. For Lung tumor dataset, Lv et al. (2016) proposed a multi-objective optimization algorithm (MOEDA) for tumor classification and selected 25 feature genes with a classification accuracy of 96.00%. In this paper, only 20 genes were chosen as the

distinguishing genes, and the classification accuracy reached 95.73%.

In summary, the above results fully demonstrated the analysis of sparseness of intra- and inter-group genes using the tumor classification model SGL-SVM we constructed. The model can significantly improve the classification accuracy of ALLAML, Prostate, DLBCL, MLL, Stjude, SRBCT, Lymphoma, and Lung datasets. The tumor classification results are satisfactory.

3.5 Effectiveness of SGL-SVM on NGS datasets

To validate the effectiveness of the SGL-SVM method, we use two different types of tumor datasets from TCGA (The Cancer Genome Atlas) (Weinstein et al., 2013; Akbani et al., 2014; Xu et al., 2019). Two types of tumor include 104 cases of breast invasive carcinoma (BRCA) and 213 cases of glioblastoma multiforme (GBM). For each tumor, Level 3 datasets containing DNA methylation, miRNA expression, and gene expression data are used. These two types of tumor data undergo three pretreatment steps: deletion of outliers, input of missing data and normalization (Xu et al., 2019). Table 10 shows the two datasets from TCGA.

Table 10

Information about the BRAC and GBM datasets.

Tumor types	DNA methylation	miRNA expression	Gene expression	No. of Sample	No. of Class	#instance per class
BRAC	23094	354	17814	104	4	18/51/12/23
GBM	1305	534	12042	213	4	56/34/58/65

First, we perform a feature gene primary selection on the normalized datasets using the Kruskal-Wallis rank sum test. Second, using sparse group Lasso for further feature gene selection. Finally, the classification accuracy of NGS tumor dataset is tested by using the characteristic gene and the support vector machine (SVM) with RBF as the kernel function. Among them, for the BRAC dataset, the number of characteristic genes of DNA methylation, miRNA expression and Gene expression data are selected as 17, 13 and 9 respectively. For GBM dataset, the number of characteristic genes of DNA methylation, miRNA expression and Gene expression data are selected as 17, 12 and 19 respectively. The overall accuracy of the proposed method compared to other methods on the BRAC and GBM data sets is shown in Table 11.

Table 11

Comparison proposed method with some other methods in the BRAC and GBM datasets.

Datasets	Experiments	Methods	BRAC ACC (%)	GBM ACC (%)

DNA	Xu et.al. 2019	DFNForest	0.731	0.596
methylation	This Paper	SGL-SVM	0.808	0.788
miRNA	Xu et.al. 2019	DFNForest	0.769	0.539
expression	This Paper	SGL-SVM	0.779	0.694
Gene	Xu et.al. 2019	DFNForest	0.808	0.865
expression	This Paper	SGL-SVM	0.856	0.826

As can be shown in Table 11, for the BRAC dataset, Xu et al. (2019) used DFNForest method and through 5-fold cross-validation, the classification accuracy of DNA methylation was 73.1%, the classification accuracy of miRNA expression was 76.9%, and the classification accuracy of Gene expression was 80.8%. However, in this paper, through 5-fold cross-validation, the classification accuracy of DNA methylation reached 80.8%, the classification accuracy of miRNA expression reached 77.9%, and the classification accuracy of Gene expression reached 85.6%, all of which were higher than DFNForest. For GBM dataset, Xu et al. (2019) used DFNForest method and through 5-fold cross-validation, the classification accuracy of DNA methylation was 59.6%, the classification accuracy of miRNA expression was 53.9%, and the classification accuracy of Gene expression was 86.5%. In this paper, the classification accuracy of DNA methylation was 78.8% using SGL-SVM method, 19.2% higher than DFNForest method. The classification accuracy of miRNA expression reached 69.4%, 15.5% higher than DFNForest method. The classification accuracy of Gene expression reached 82.6%, which was 3.9% lower than the DFNForest method. This is mainly because the DFNForest method proposed by Xu, J. et.al. (2019) takes into account the inherent statistical characteristics of each data type and different omics data correlation, and this article mainly considers the sparseness of genes within and between groups. Therefore, in summary, the SGL-SVM method proposed in this paper has good performance on two types of NGS datasets, which has certain research significance.

4. Conclusion

With the advent of the big data era, tumor classification based on gene expression data has quickly become a hot topic in bioinformatics. This paper presents a novel method for tumor classification—SGL-SVM. First, on normalized tumor datasets, we used the Kruskal-Wallis rank sum test as the primary selection to remove some redundant genes. Second, SGL further selected feature genes. Finally, SVM served as the classifier. To test the effectiveness of the SGL-SVM method, eight microarray tumor datasets were selected for the experiment, and we obtained

satisfactory classification accuracy by 10-fold cross-validation. To validate the effectiveness of our method in NGS data, this paper used SGL-SVM method to test two kinds of tumor datasets from TCGA, and obtained high classification accuracy through 5-fold cross validation. The experimental results show that for the two-class and multi-class microarray datasets and NGS datasets, SGL-SVM has better performance, selecting fewer genes and achieving better classification accuracy. Compared with the Elastic Net, Group Lasso, and ReliefF, SGL considers the group effect of variables and selects more effective feature genes. Therefore, in the future work, we will choose a more suitable primary method for SGL, optimize the parameters of SVM and implement it on other tumor datasets. Considering the forefront regularization method, improving the proposed method of feature selection to further reduce the number of feature genes and improve the classification accuracy will become our future research direction.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (Nos. 61863010 and 11771188), the Key Research and Development Program of Shandong Province of China (2019GGX101001), the Natural Science Foundation of Shandong Province of China (Nos. ZR2017MA014 and ZR2018MC007), and the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159). This work used the Extreme Science and Engineering Discovery Environment, which is supported by the National Science Foundation (No. ACI-1548562).

Conflict of interest

The authors declare no conflict of interest.

References

Akbani, R., Ng, K. S., Werner, H. M., Zhang, F., Ju, Z., Liu, W., Yang, J-Y., Lu, Y., Weinstein, J. N., & Mills, G. B., 2014. A pan-cancer proteomic analysis of The Cancer Genome Atlas (TCGA) project. Cancer Research. 74(19):4262.

Algamal, Z. Y., Lee, M. H., 2015. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Syst. Appl. 42, 9326-9332.

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.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Scott, A.A., Jane, E.S., Lewis, B.S., Rob, P., Monique, L.B., Mark, D.M., Stephen, E.S., Eric, S.L., Todd, R.G., Stanley, J., 2002. Korsmeyer MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30, 41-47.

Behrmann, J., Etmann, C., Boskamp, T., Casadonte, R., Kriegsmann, J., Maass, P., 2017. Deep learning for tumor classification in imaging mass spectrometry. Bioinformatics 34, 1215-1223.

Bharat, S., Vyas, O. P., 2014. A meta-heuristic regression-based feature selection for predictive analytics. Data. Sci. J. 13, 106-118.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M.A., Loda, M., Weber, G.M., Mark, E.J., Lander, E.S., Wong, W.H., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., Meyerson, M., 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA 98, 13790-13795.

BolonCanedo, V., SanchezMarono, N., AlonsoBetanzos, A., Benitez, J.M., Herrera, F., 2014. A review of microarray datasets and applied feature selection methods. Inf. Sci. 282, 111-135.

Borgi, M. A., Labate, D., Arbi, M. E., Amar, C. B., 2015. Sparse multi-stage regularized feature learning for robust face recognition. Expert Syst. Appl. 42, 269-279.

Breiman, L., 1995. Better Subset Regression Using the Nonnegative Garrote. Technometrics 37, 373-384.

Castillo, D., Galvez, J. M., Herrera, L. J., & Rojas, I., 2017. Breast Cancer Microarray and RNASeq Data Integration Applied to Classification. In International Work-Conference on Artificial Neural Networks (pp. 123-131). Springer, Cham.

Chandra, B., Gupta, M., 2011. An efficient statistical feature selection approach for classification of gene expression data. J. Biomed. Inform. 44, 529-535.

Chan, H. P., Kim, S. B., 2015. Sequential random k-nearest neighbor feature selection for high-dimensional data. Expert Syst. Appl. 42, 2336-2342.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37-46.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273-297.

Dagliyan, O., Uneyyuksektepe, F., Kavakli, I. H., Turkay, M., 2011. Optimization based tumor classification from microarray gene expression data. PLos One 6, e14579.

Dettling, M., 2004. BagBoosting for tumor classification with gene expression data. Bioinformatics. 20, 3583-3593.

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

Elingaramil, S., Li, X., & He, N., 2013. Applications of nanotechnology, next generation sequencing and microarrays in biomedical research. J. Nanosci. Nanotechno. 13(7), 4539-4551.

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

Foygel, R., Drton, M., 2010. Exact block-wise optimization in group lasso and sparse group lasso for linear regression. arXiv. 1010.3320.

Frank, L. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109-135.

Fu, W.J., 1998. Penalized regressions: the bridge versus the lasso. J. Comput. Graph. Statist. 7, 397-416.

Gao, J., Kwan, P. W., Shi, D., 2010. Sparse kernel learning with lasso and bayesian inference algorithm. Neural Networks 23, 257-264.

Glaab, E., Bacardit, J., Garibaldi, J. M., Krasnogor, N., 2012. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. PLos One 7, e39932.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H.A., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537.

Guo, Y., Liu, S., Li, Z., Shang, X., 2018. Bcdforest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. BMC Bioinformatics. 19, 118.

Han, F., Yang, C., Wu, Y. Q., Zhu, J. S., Ling, Q. H., Song, Y. Q., Huang, D. S., 2017. A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information. IEEE/ACM. Trans. Comput. Biol. Bioinform. 14, 85-96.

Hewett, R., Kijsanayothin, P., 2008. Tumor classification ranking from microarray data. BMC Genomics. 9, S21.

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.

Kang, C., Huo, Y., Xin, L., Tian, B., Yu, B., 2019. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. J. Theor. Biol. 463, 77-91.

Kang, S., Song, J., 2017. Robust gene selection methods using weighting schemes for microarray data analysis. BMC Bioinformatics. 18, 389.

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.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673-679.

Kolali, K. M., Bazrafkan, M., 2016. A novel sparse coding algorithm for classification of tumors based on gene expression data. Med. Biol. Eng. Comput. 54, 869-876.

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

Latkowski, T., Osowski, S., 2015. Data mining for feature selection in gene expression autism data. Expert Syst. Appl. 42, 864-872.

Li, W., Bo, L., Wen, Z., Min, C., Li, P., 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.

Liu, J., Ye, J., 2010. Fast overlapping group lasso. arXiv. 1009.0306.

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.

Luo, X., Liu, F., Yang, S., Wang, X., Zhou, Z., 2015. Joint sparse regularization based Sparse Semi-Supervised Extreme Learning Machine (S3ELM) for classification. Knowl-Based. Syst. 73, 149-160.

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.

Ma, S., Song, X., Huang, J., 2007. Supervised group Lasso with applications to microarray data analysis. BMC Bioinform. 8, 60.

Margalit, O., Somech, R., Amariglio, N., Rechavi, G., 2005. Microarray-based gene expression profiling of hematologic malignancies: basic concepts and clinical applications. Blood Rev. 19, 223-234.

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

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., 2009. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Stat. Sci. 27, 538-557.

Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Warnatz, H.J., Grobner, S., Segurawang, M., Zichner, T., Rudneva, V.A., Warnatz, H., Sidiropoulos, N., Phillips, A.H., Schumacher, S.E., Kleinheinz, K., Waszak, S.M., Erkek, S., Jones, D.T., Worst, M., Kool, B.C., Zapatka, M., Jager, N., Chavez, L., Hutter, B., Bieg, M., Paramasivam, N., Heinold, M., Gu, Z., Ishaque, N., Jagerschmidt, C., Imbusch, C.D., Jugold, A., Hubschmann, D., Risch, T., Amstislavskiy, V., Gonzalez, F.G.R., Weber, U., Wolf, S., Robinson, G.W., Zhou, X., Wu, G., Finkelstein, D., Liu, Y., Cavalli, F.M.G., Luu, B., Ramaswamy, V., Wu, X., Koster, J., Ryzhova, M., Cho, Y., Pomeroy, S.L., Heroldmende, C., Schuhmann, M.U., Ebinger, M., Liau, L.M., Mora, J., Mclendon, R.E., Jabado, N., Kumabe, T., Chuah, E., Ma, Y., Moore, R.A., Mungall, A.J., Mungall, K., Thiessen, N., Tse, K., Wong, T., Jones, S.J.M., Witt, O., Milde, T., Deimling, A.V., Capper, D., Korshunov, A., Yaspo, M., Kriwacki, R.W., Gajjar, A., Zhang, J., Beroukhim, R., Fraenkel, E., Korbel, J.O., Brors, B., Schlesner, M., Eils, R., Marra, M.A., Pfister, S.M., Taylor, M.D., 2017. Peter Lichter The whole-genome landscape of medulloblastoma subtypes. Nature 547, 311-317.

Osborne, M. R., Presnell, B., Turlach, B. A., 2000. A new approach to variable selection in least squares problems. Ima. J. Numer. Anal. 20, 389-403.

Piao, Y., Piao, M., Park, K., Ryu, K. H., 2012. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics 28, 3306-3315.

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

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, A.J., 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.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group lasso. J. Comput. Graph. Stat. 22, 231-245.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., Damico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer cell 1, 203-209.

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

Sun, S. Q. Peng, Q. K., 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. B. 58, 267-288.

Tibshirani, R., Saunders, M. A., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Series. B. Stat. Methodol. 67, 91-108.

Wang, A., An, N., Chen, G., Liu, L., Alterovitz, G., 2018. Subtype dependent biomarker identification and tumor classification from gene expression profiles. Knowl-Based. Syst. 146, 104-117.

Wang, H., Leng, C., 2008. A note on adaptive group lasso. Comput. Stat. Data. Anal. 52, 5277-5286.

Wang, M., Song, L., Wang, X., 2016. Quadratic approximation via the SCAD penalty with a diverging number of parameters. Commun. Stat-Simul. C. 45, 1-16.

Wang, M., Wang, X., 2014. Adaptive lasso estimators for ultrahigh dimensional generalized linear models. Stat. Probabil. Lett. 89, 41-50.

Wang, X., Wang, M., 2016. Variable selection for high-dimensional generalized linear models with the weighted elastic-net procedure. J. Appl. Stat. 43, 796-809.

Wang, X., Wang, M., 2017. Adaptive group bridge estimation for high-dimensional partially linear models. J. Inequal. Appl. 2017, 158.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander. C., Stuart, J. M., & Cancer Genome Atlas Research Network, 2013. The cancer genome atlas pan-cancer analysis project. Nat Genet, 45(10), 1113.

Xu, X., Ghosh, M., 2015. Bayesian variable selection and estimation for group lasso. Bayesian Anal. 10, 909-936.

Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., Dawood, H., 2019. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC bioinformatics. 20, 527.

Yang, A., Cao, T., Li, R., Liao, B., 2012. A hybrid gene selection method for cancer classification based on clustering algorithm and euclidean distance. J. Comput. Theor. Nanosci. 9, 611-615.

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.H., Evans, W.E., Naeve, C., Wong, L., Downing, J.R., 2002. Classification subtype discovery and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell 1, 133-143.

Yu, B., Zhang, Y., 2013. The analysis of colon cancer gene expression profiles and the extraction

of informative genes. J. Comput. Theor. Nanosci. 10, 1097-1103.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Series. B. Stat. Methodol. 68, 49-67.

Zainuddin, Z., Ong, P., 2011. Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. Expert Syst. Appl. 38, 13711-13722.

Zhao, G., Wu, Y., 2016. Feature subset selection for cancer classification using weight local modularity. Sci. Rep. 6, 34759.

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

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series. B. Stat. Methodol. 67, 301-320.

Authors' contributions

Bin Yu, Yanhao Huo and Lihui Xin conceived the algorithm, prepared the datasets, carried out experiments, and drafted the manuscript. Yanhao Huo, Lihui Xin and Chuanze Kang carried out the data coding and data interpretation. Minghui Wang and Qin Ma designed, analyzed experiments and provided some suggestions for the manuscript writing. All authors read and approved the final manuscript.