scientific reports

OPEN

LogSum + L_2 penalized logistic regression model for biomarker selection and cancer classification

Xiao-Ying Liu✉, Sheng-Bing Wu, Wen-Quan Zeng, Zhan-Jiang Yuan & Hong-Bo Xu

Biomarker selection and cancer classification play an important role in knowledge discovery using genomic data. Successful identification of gene biomarkers and biological pathways can significantly improve the accuracy of diagnosis and help machine learning models have better performance on classification of different types of cancer. In this paper, we proposed a *LogSum* + L_2 penalized logistic regression model, and furthermore used a coordinate decent algorithm to solve it. The results of simulations and real experiments indicate that the proposed method is highly competitive among several state-of-the-art methods. Our proposed model achieves the excellent performance in group feature selection and classification problems.

With the development of DNA microarray technology[1,2], the biological researchers can analyze the expression levels of thousands of genes simultaneously. Many studies have shown that microarray data can be used to classify the different types of cancer, which includes how long the incubation period is, and what drugs are effective in the diagnosis and treatment processes.

From a biological point of view[3], only a small number of genes (biomarkers) strongly indicate the target cancer, while other genes are not related to disease. Therefore, the data with unrelated genes may bring noise, and make the machine learning approaches less easy to find pathogenic genes that cause the disease. Moreover, from a machine learning perspective, the large number of genes (features) with few samples in the datasets may cause overfitting[4], and have negative impact on classification performance. Due to the importance of these issues, effective gene (biomarker) selection methods are needed to help classify different cancer types and improve prediction accuracy.

In recent years, many methods for gene selection in microarray datasets have been developed and generally can be divided into three categories: filters, wrappers, and embedded methods. Filter methods[5–8] evaluate genes based on discriminative power without considering their regulation correlations with other genes. The main disadvantage of the filtering methods is that it examines each gene separately, and makes each gene independent, thereby ignores the possibility that the genes have combined and grouping effects. This is a common problem with statistical methods, such as *t*-test, which can also examine each gene individually.

Wrapper methods[9–11] utilize feature assessment measures based on the learning performance to select subsets of genes. Generally, they can acquire a small number of related genes to notable promote the learning ability. In some cases, the results of the wrapper methods are better than those of the filter methods. However, the main fault of wrapper methods is their computational cost is high.

A third set of feature selection approaches is the embedded methods[12–26] that perform feature selection as part of the learning procedure of a single process. Under similar learning performance, the computational efficiency of embedded methods is more efficient than wrapper approaches. Hence, embedded methods have recently attracted a lot of attention in the literature. The regularization methods are important embedded technologies, which can perform feature selection and model training simultaneously. Many regularization methods have been proposed, such as Lasso[12], SCAD[13], adaptive Lasso[14], MCP[15], L_q $(0 < q < 1)$[16], $L_{1/2}$[17,18], *LogSum*[19], etc. These methods perform well with the independent feature selection. When the features are highly correlated, some regularization methods which pay attention to the grouping effect can be used to select the groups of the relevant features, such as group Lasso[20], Elastic net[21], Fused Lasso[22], OSCAR[23], adaptive Elastic net[24], SCAD-L_2[25], $L_{1/2} + L_2$[26].

On the other hand, many machine learning models have been used to analyze microarray gene expression data for cancer classification. For example, Furey et al. used support vector machines (SVMs) to classify cell and tissue types[27]. Medjahed et al. applied the K-nearest neighbors (K-NN) to the diagnosis and classification of breast cancer[28]. Meanwhile, some researchers used the logistic regressions with optimization methods for

Computer Engineering Technical College, Guangdong Polytechnic of Science and Technology, Zhuhai 519090, Guangdong, China. ✉email: 631218194@qq.com

binary cancer classification[29–33]. However, the traditional logistic regression model has two obvious shortcomings, mainly in the following two aspects:

1. Feature selection problem.
   All or most of the feature coefficients obtained by fitting the logistic regression model are not zero, i.e. all most of the features are related to the classification target and not sparse. However, the key factors affecting the model are often only a few in many practical problems. This non-sparseness of the logistic models increases the computational complexity on the one hand and is not conducive to the actual interpretation of the practical problems.
2. Overfitting problem.
   The logistic regression models can often obtain good precision for the training data, but for the test data outside the training set, the classification accuracy rate is not ideal. In fact, not only logistic regression, many other data analysis models will also be affected by overfitting. It has become one of the hot research topics in statistics, machine learning and other fields.

In recent years, there is growing interesting to apply the regularization techniques in the logistic regression models to solve the above mentioned two shortcomings. For example, Tibshirani and Friedman[34,35] proposed the sparse logistic regression based on the Lasso regularization and the coordinate descent methods. Algamal et al.[36,37] proposed the adaptive Lasso and the adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Like sparse logistic regression with the L_1 regularization method, Cawley and Talbot[30] investigated sparse logistic regression with Bayesian regularization. Liang et al.[38] investigated the sparse logistic regression model with the $L_{1/2}$ penalty for gene selection in cancer classification.

Inspired by above mentioned methods, in this paper, we proposed a $LogSum + L_2$ penalized logistic regression model. The main contributions of this paper include.

1. Our proposed method can not only select sparse features (biomakers), but also identify the groups of the relevant features (gene pathways). The coordinate decent algorithm is used to solve the $LogSum + L_2$ penalized logistic regression model.
2. We also evaluate the capability of our proposed method and compare its performance with other regularization methods. The results of simulations and real experiments indicate that the proposed method is highly competitive among several state-of-the-art methods.

The rest of this paper is organized as follows. In "Related works" section, we introduce the related work. "Methods" section represents the $LogSum + L_2$ penalized logistic regression model and its optimization algorithm. "Experiments experimental results and discussion" section analyzes the results of the simulated data. "Discussion and conclusion" section analyzes the results of real data. Section 6 concludes this paper.

Related works

Sparse penalized logistic regression.

We focused on binary classification using logistic regression (LR), which is a statistical method for modeling a binary classification problem. Suppose we have n samples and p genes. Datasets X and y are the genes matrix and the dependent variable, respectively. So, the n samples mean the set D, x_{ij} denotes the value of gene j for the ith samples, y_i is a corresponding variable that takes a value of 0 or 1, $y_i = 0$ indicates the ith sample in Class 1 and $y_i = 1$ indicates the ith sample is in Class 2. Then, we define a classifier $f(x) = \frac{e^x}{(1+e^x)}$ such that for any input x with class label y, $f(x)$ predicts y correctly. The LR is given as follows:

$$P(y_i = 1|X_i) = f(X_i'\beta) = \frac{e^{(X_i'\beta)}}{1 + e^{(X_i'\beta)}} \tag{1}$$

In Eq. (1), $\beta = (\beta_0, \beta_1, ..., \beta_p)$ are the coefficients need to be estimated. We should notice that β_0 is the intercept. The log-likelihood function of the transformation of Eq. (1) is defined as:

$$l(\beta) = -\sum_{i=1}^{n} \{y_i \log[f(X_i'\beta)] + (1 - y_i) \log[1 - f(X_i'\beta)]\} \tag{2}$$

Then we can obtain the coefficients β when Eq. (2) is minimized. In the cancer classification problem with high-dimensional and low-sample size data ($p \gg n$),directly solving the logistic model (2) will make overfitting. Therefore, to solve this problem, we need add a regularization term to (2), the sparse logistic regression can be modelled as:

$$\beta = argmin\left\{ l(\beta) + \lambda \sum_{j=1}^{p} p(\beta_j) \right\} \tag{3}$$

where $l(\beta)$ is the loss function, $p(\beta)$ is the penalty function, and $\lambda > 0$ is a control parameter.

A coordinate decent algorithm for different thresholding operators.

The coordinate decent algorithm is a "one-at-a-time" approach[40], and before considering the coordinate descent algorithm for the nonlinear logistic regularization, we first introduce a linear regression case. The objective function of the linear regression is as follow:

$$min\left\{\frac{1}{2n}||y - X\beta||^2 + P_\lambda(\beta)\right\} \tag{4}$$

where $y = (y_1, \dots, y_n)^T$ is the vector of n response variables,$X_i = (x_{i1}, x_{i2}, \dots, x_{ij})$ is ith input variables with dimensionality p and y_i is the corresponding response variable. $||.||$ denotes the L_2-norm.

The coordinate decent algorithm "one-at-a-time" is to solve β_j and other $\beta_{k\neq j}$(represent the coefficients $\beta_{k\neq j}$ remained after jth element β_j is removed) are fixed. The Eq. (4) can be rewritten as:

$$R(\beta) = argmin\left\{\frac{1}{2n}\left(y_i - \left(\sum_{k\neq j} x_{ik}\beta_k + x_{ij}\beta_j\right)\right)^2 + \lambda\sum_{k\neq j} P(\beta_k) + \lambda P(\beta_j)\right\} \tag{5}$$

In Eq. (5), kth represents other features than the jth feature.

The first order derivative at β_j can be estimated as:

$$\frac{\partial R}{\partial \beta_j} = \sum_{i=1}^{n}\left(-x_{ij}\left(y_j - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\right)\right) + \lambda P(\beta_j) = 0 \tag{6}$$

We define $\tilde{y}_i^{(j)} = \sum_{k\neq j} x_{ik}\beta_k$ as a part of fitting β_j, $\tilde{r}_i^{(j)} = y_i - \tilde{y}_i^{(j)}$, and $w_j = \sum_{i=1}^{n} x_{ij}\tilde{r}_i^{(j)}$, where $\tilde{r}_i^{(j)}$ represents the partial residuals with respect to the jth feature.

To consider the correlation of features, Elastic Net (L_{EN})[21] had been proposed, which emphasizes a grouping effect. The L_{EN} penalty function is given as follows:

$$P(\beta) = (1 - a)\frac{1}{2}||\beta||_{L_2}^2 + a||\beta||_{L_1} \tag{7}$$

The penalty function of L_{EN} is combination of L_1 penalty and ridge penalty which $a = 1$ and $a = 0$ respectively. Therefore, Eq. (6) is rewritten as follows:

$$\frac{\partial R}{\partial \beta_j} = \sum_{i=1}^{n}\left(-x_{ij}\left(y_j - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\right)\right) + \lambda(1 - a)\beta_j + \lambda a = 0 \tag{8}$$

Zou and Hastie have proposed the univariate solution[21] for a L_{EN} penalized regression coefficient as follows:

$$\beta_j = f_{L_{EN}}(w_j, \lambda, a) = \frac{S(w_j, \lambda a)}{1 + \lambda(1 - a)} \tag{9}$$

where $S(w_j, \lambda a)$ is soft thresholding operator for the L_1 penalty if a is equal to 1, so Eq. (9) can be divided into three situations as follows:

$$\beta_j = Soft(w_j, \lambda) = \begin{cases} w_j + \lambda & \text{if } w_j < -\lambda \\ w_j - \lambda & \text{if } w_j > \lambda \\ 0 & \text{if } -\lambda \leq w_j \leq \lambda \end{cases} \tag{10}$$

Fan et al. have proposed the SCAD penalty[13], which can produce sparse set of solutions and approximately unbiased coefficients for large coefficients. Its penalty function is shown as follows:

$$p_{\lambda,a}(\beta) = \begin{cases} \lambda\beta & \text{if } \beta \neq \lambda \\ \frac{a\lambda\beta - \frac{1}{2}(\beta^2 + \lambda^2)}{a-1} & \text{if } \lambda < \beta < a\lambda \\ \frac{\lambda(a^2-1)}{2(a-1)} & \text{if } \beta > a\lambda \end{cases} \tag{11}$$

Additionally, the SCAD thresholding operator is given as follows:

$$\beta_j = f_{SCAD}(w_j, \lambda, a) = \begin{cases} S(w_j, \lambda) & \text{if } \left|w_j\right| < 2\lambda \\ \frac{S\left(w_j, \frac{a\lambda}{a-1}\right)}{1 - \frac{1}{a-1}} & \text{if } 2\lambda < \left|w_j\right| \leq a\lambda \\ w_j & \text{if } \left|w_j\right| > a\lambda \end{cases} \tag{12}$$

Like the SCAD penalty, Zhang et al. have proposed the maximum concave penalty (MCP)[15]. The formula of its penalty function is shown as:

$$p_{\lambda,a}(\beta) = \begin{cases} \lambda\beta & \text{if } \beta \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if } \beta > \gamma\lambda \end{cases} \tag{13}$$

And the MCP thresholding operator is given as follows:

$$\beta_j = f_{MCP}(w_j, \lambda, \gamma) = \begin{cases} \frac{S(w_j,\lambda)}{1-\frac{1}{\lambda}} & \text{if } \left|w_j\right| \leq \gamma\lambda \\ w_j & \text{if } \left|w_j\right| > \gamma\lambda \end{cases} \tag{14}$$

In Eq. (14), γ is the experience parameter.

Xu et al. have proposed $L_{1/2}$ regularization[17], and its penalty function can be written:

$$min\left\{\frac{1}{2n}\left\|y - X\beta\right\|^2 + \lambda\sum_{j}^{p}\left|\beta_j\right|^{\frac{1}{2}}\right\} \tag{15}$$

Then the univariate half thresholding operator for a $L_{1/2}$ penalized linear regression coefficient is given as follows:

$$\beta_j = Half(w_j, \lambda) = \begin{cases} \frac{2}{3}w_j\left(1 + \cos\frac{2(\pi - \phi_\lambda(w_j))}{3}\right) & \text{if } \left|w_j\right| > \frac{3}{4}(\lambda)^{\frac{2}{3}} \\ 0 & \text{if } otherwise \end{cases} \tag{16}$$

in Eq. (16), $\phi_\lambda(w) = \frac{\lambda}{8}\left(\frac{|w|}{3}\right)^{-\frac{3}{2}}$.

To consider the correlation of genes, Huang et al. have proposed *HLR* regularization[26]. Equation (15) can be rewritten:

$$min\left\{\frac{1}{2n}\left\|y - X\beta\right\|^2 + \lambda\left(\sum_{j}^{p}\left(a\left|\beta_j\right|^{\frac{1}{2}} + (1-a)\left|\beta_j\right|^2\right)\right)\right\} \tag{17}$$

And the univariate half thresholding operator for the *HLR* penalized linear regression coefficient is as follows:

$$\beta_j = HLR(w_j, \lambda) = \frac{Half(w_j, \lambda a)}{1 + \lambda(1-a)} \tag{18}$$

Theoretically, the L_0 regularization produces the better solutions with more sparsity, but it is *NP* problem. Therefore, Candes et al. have[19] proposed *LogSum* penalty, which approximates much better the L_0 regularization. We could rewrite the penalty function of the *LogSum* regularization as follows:

$$min\left\{\frac{1}{2n}\left\|y - X\beta\right\|^2 + \lambda\sum_{j}^{p}log(\left|\beta_j\right| + \varepsilon)\right\} \tag{19}$$

where $\varepsilon > 0$ should be set arbitrarily small, to closely make the *LogSum* penalty resemble the L_0-norm. Equation (19) has a local minimal[39].

$$f_{Logsum}(w_j, \lambda, \varepsilon) = D(w_j, \lambda, \varepsilon) = \begin{cases} sign(w_j)\frac{c_1 + \sqrt{c_2}}{2} & \text{if } c_2 > 0 \\ 0 & \text{if } c_2 \leq 0 \end{cases} \tag{20}$$

where $\lambda > 0, 0 < \varepsilon < \sqrt{\lambda}, c_1 = w_j - \varepsilon, c_2 = c_1^2 - 4(\lambda - w_j\varepsilon)$.

Methods

$LogSum + L_2$ penalized logistic regression model.

In this paper, we proposed the $LogSum + L_2$ penalized logistic regression model for feature group selection. We could write the $LogSum + L_2$ penalty as follows:

$$min\left\{\frac{1}{2n}\left\|y - X\beta\right\|^2 + \lambda\left(\sum_{j}^{p}\left(\lambda_1 log(\left|\beta_j\right| + \varepsilon) + \lambda_2\left|\beta_j\right|^2\right)\right)\right\} \tag{21}$$

where $\left\|y - X\beta\right\|^2$ is the loss function, (y, X) is a data set, $\varepsilon > 0$ is a constant, $\lambda > 0$, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are regularization parameters that control the complexity of the penalty function.

Figure 1 describes the contour plots on two-dimensional for the penalty functions of L_1, L_{EN}, *HLR* and $LogSum + L_2$ approaches. It is demonstrated that the $LogSum + L_2$ penalty is non-convex for the given parameters λ_1 and λ_2 in Eq. (21).

The $LogSum + L_2$ thresholding operator is given as follows:

$$\beta_j = f_{Logsum+L_2}\left(w_j, \lambda, \varepsilon\right) = \begin{cases} sign\left(w_j\right)\frac{\left(\left|w_j\right| - (1+2\lambda_2)\varepsilon\right) + \sqrt{\left(\left|w_j\right| + (1+2\lambda_2)\varepsilon\right)^2 - 4\lambda_1(1+2\lambda_2)}}{2(1+2\lambda_2)} & if\ \left|w_j\right| > \lambda \\ 0 & if\ otherwise \end{cases} \tag{22}$$

where $\lambda = 2\sqrt{\lambda_1(1 + 2\lambda_2)} - (1 + 2\lambda_2)\varepsilon, \lambda_1 + \lambda_2 = 1$.

The proof of Eq. (22) is given as follows:

Considering the regression model has the following form

$$y = X\beta + e \tag{23}$$

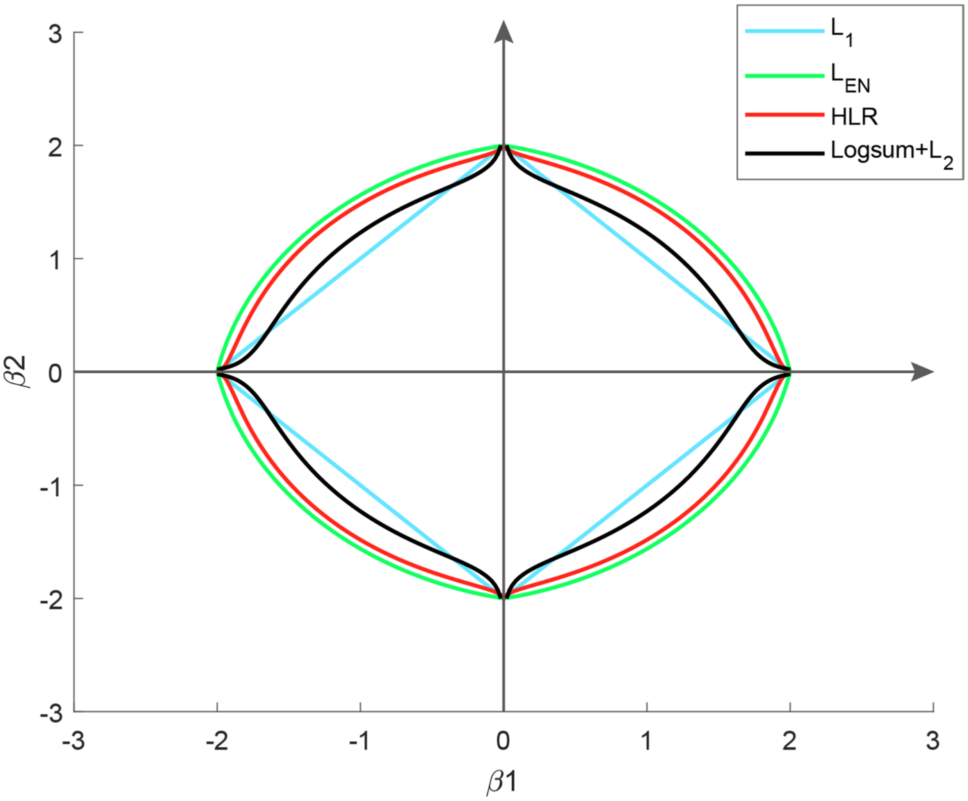


Figure 1. Contour plots (two-dimensional) for the regularization methods.

where the response $y \in R^n$, the predictors $X = (x_1, x_2, ..., x_p), X \in R^{n\times p}$ and the error term $e = (e_1, e_2, ..., e_n)$ are i.i.d. with mean 0 variance σ^2.

The *Logsum* $+ L_2$ regularization can be expressed as:

$$l_{Logsum+L_2}(\beta; \lambda_1, \lambda_2) = \frac{1}{2}\|y - X\beta\|^2 + \sum_{j=1}^{p} (\lambda_1 \log(|\beta_j + \varepsilon|) + \lambda_2\beta_j^2) \tag{24}$$

Its first partial derivative with respect to β_k is given by follows:

$$\begin{aligned} \frac{\partial l_{Logsum+L_2}}{\partial \beta_k} &= \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)(-x_{ik}) + \frac{\lambda_1 sign(\beta_k)}{|\beta_k| + \varepsilon} + 2\lambda_2\beta_k \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j\neq k}^{p} x_{ij}\beta_j - x_{ik}\beta_k\right)(-x_{ik}) + \frac{\lambda_1 sign(\beta_k)}{|\beta_k| + \varepsilon} + 2\lambda_2\beta_k \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j\neq k}^{p} x_{ij}\beta_j\right)(-x_{ik}) + \sum_{i=1}^{n} x_{ik}^2\beta_k + \frac{\lambda_1 sign(\beta_k)}{|\beta_k| + \varepsilon} + 2\lambda_2\beta_k \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)(-x_{ik}) + \beta_k + \frac{\lambda_1 sign(\beta_k)}{|\beta_k| + \varepsilon} + 2\lambda_2\beta_k \end{aligned} \tag{25}$$

Equation (25) is obtained from condition that the design matrix X is orthonormal. By setting the first partial derivative equal to zero, we obtain the estimator with its kth element $\hat{\beta}_k$.

We first considers the situation $\beta_j > 0$, let $r_i^{(k)} = y_i - \sum_{j\neq k}^{p} x_{ij}\beta_j, w_k = \sum_{i=1}^{n} r_i^{(k)}(-x_{ik})$. Set the first partial derivative $\frac{\partial l_{Logsum+L_2}}{\partial \beta_k} = 0$, we have:

$$-w_k + \beta_k + \frac{\lambda_1}{\beta_k + \varepsilon} + 2\lambda_2\beta_k = 0 \tag{26}$$

and Eq. (26) is equivalent to follows:

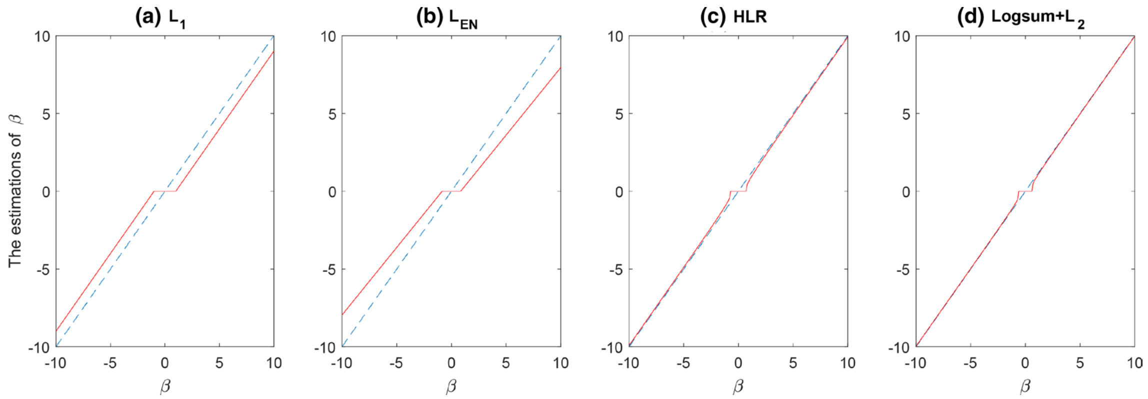


Figure 2. Exact solution of (**a**) L_1 (**b**) L_{EN} (**c**) HLR (**d**) $LogSum + L_2$ in an orthogonal design.

$$(1+2\lambda_2)\beta_k^2 - (w_k - (1+2\lambda_2)\varepsilon)\beta_k - w_k\varepsilon + \lambda_1 = 0 \tag{27}$$

Let

$$\begin{aligned}\Delta &= (w_k - (1+2\lambda_2)\varepsilon)^2 - 4(1+2\lambda_2)(\lambda_1 - w_k\varepsilon)\\ &= (w_k + (1+2\lambda_2)\varepsilon)^2 - 4\lambda_1(1+2\lambda_2)\end{aligned}$$

We discuss the solutions of Eq. (27) according to the value of Δ.

1. if $\Delta < 0$, Eq. (27) has no solution, that is no real root.
2. if $\Delta = 0$, Eq. (27) has unique root, that is $\widehat{\beta}_k = \frac{w_k-(1+2\lambda_2)\varepsilon}{2(1+2\lambda_2)}$.
3. if $\Delta > 0$, Eq. (27) has two roots, we have

$$(w_k + (1+2\lambda_2)\varepsilon)^2 > 4\lambda_1(1+2\lambda_2)$$

$$w_k + (1+2\lambda_2)\varepsilon > 2\sqrt{\lambda_1(1+2\lambda_2)}$$

Therefore, when $w_k \geq 2\sqrt{\lambda_1(1+2\lambda_2)} - (1+2\lambda_2)\varepsilon$, we obtain the estimator

$$\widehat{\beta}_k = \frac{w_k - (1+2\lambda_2)\varepsilon + \sqrt{(w_k + (1+2\lambda_2)\varepsilon)^2 - 4\lambda_1(1+2\lambda_2)}}{2(1+2\lambda_2)} \tag{28}$$

For $\beta_k < 0$, we can obtain the estimator in a similar way. Finally, we obtain the thresholding function of the $Logsum + L_2$ regularization as Eq. (22).

According to different thresholding operators, we also discuss three properties to satisfy the coefficient estimator as shown in Fig. 2:

(a) *Unbiasedness* the resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias;
(b) *Sparsity* the resulting estimator is a thresholding rule, which automatically sets a small estimated coefficient to zero to reduce model complexity;
(c) *Continuity* the resulting estimator is continuous to avoid instability in model prediction.

Figure 2 shows four regularization methods:L_1, L_{EN}, HLR and $LogSum + L_2$ penalties with an orthogonal design matrix in the regression model. The estimators of L_1 and L_{EN} are biased, whereas the HLR penalty is asymptotically unbiased. Similar to the HLR method, the $LogSum + L_2$ approach also performs better than L_1 and L_{EN} in the property of unbiasedness. All of these four regularization methods fulfil requirements of sparsity and continuity.

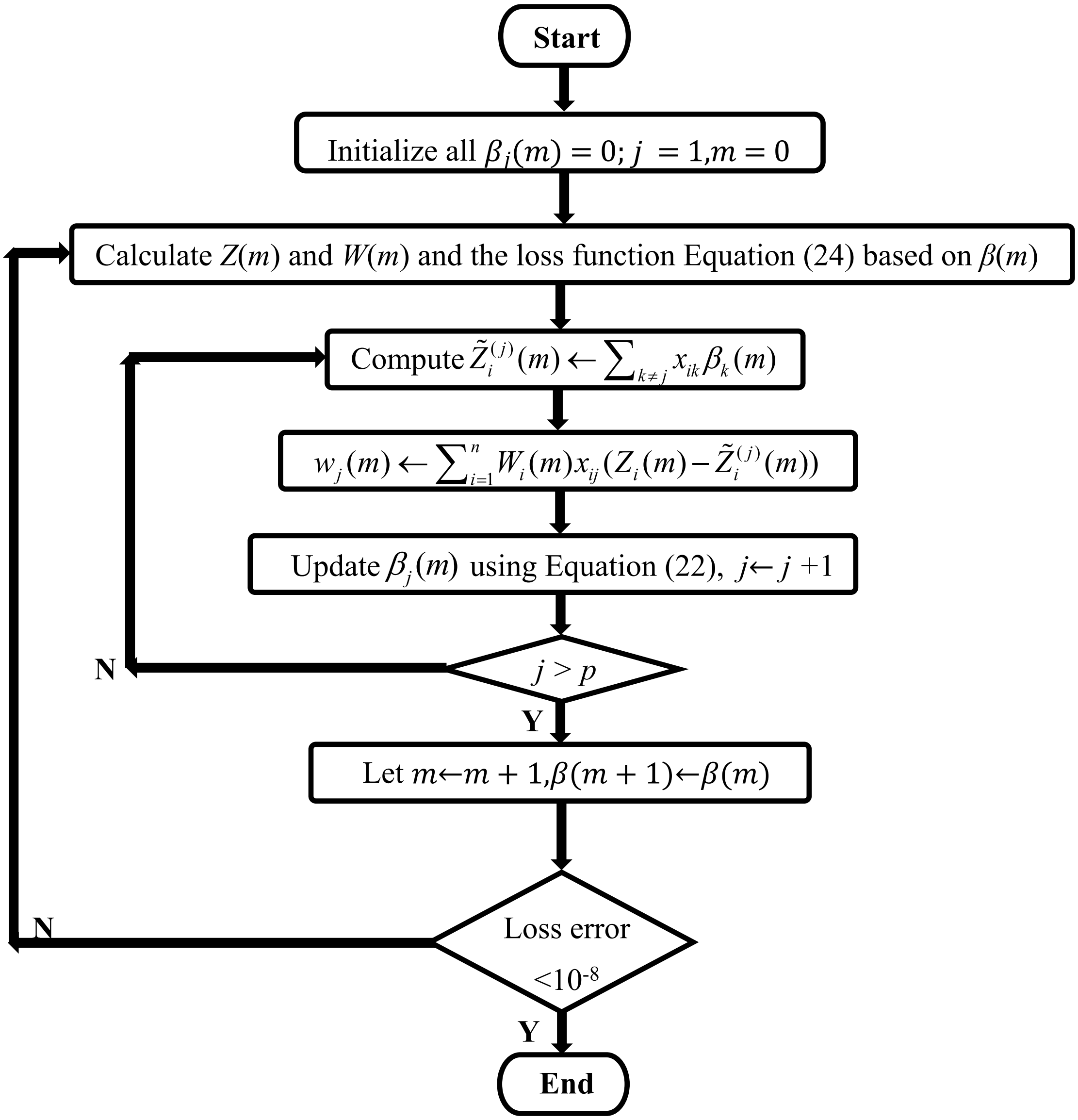
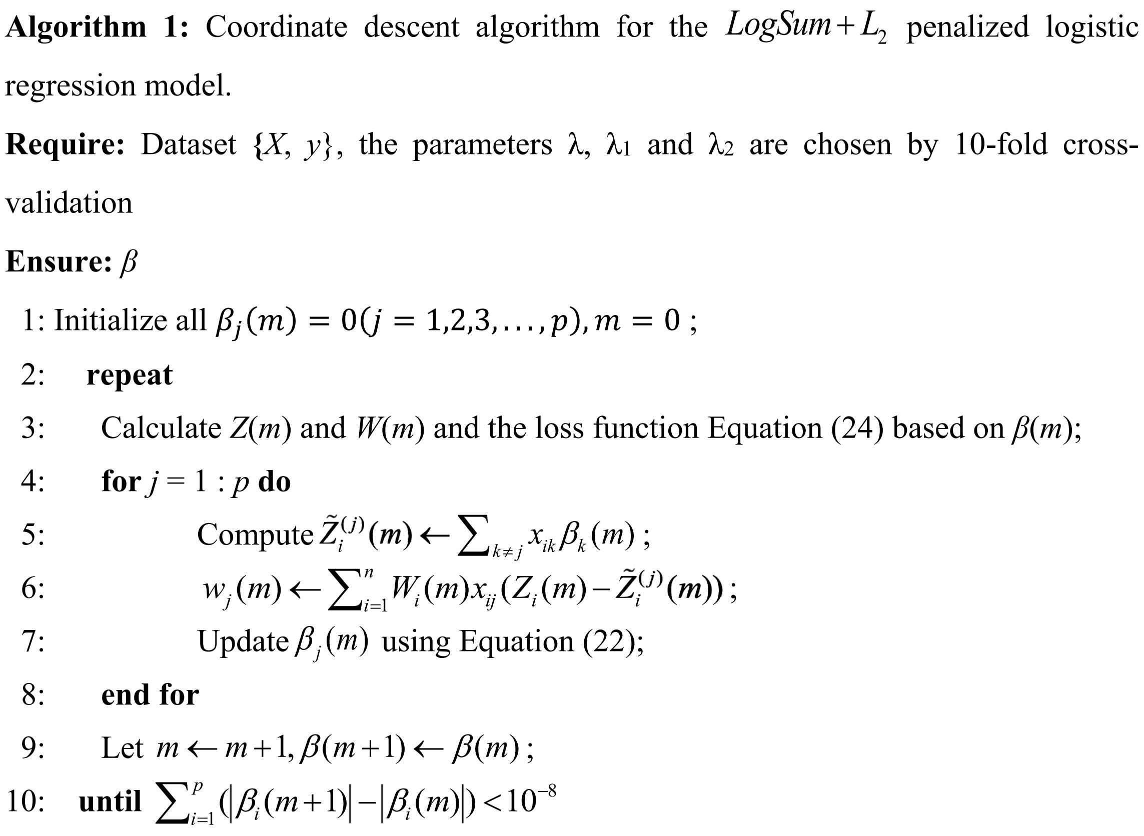

Start
Initialize all $\beta_j(m) = 0; j = 1, m = 0$
Calculate $Z(m)$ and $W(m)$ and the loss function Equation (24) based on $\beta(m)$
Compute $\tilde{Z}_i^{(j)}(m) \leftarrow \sum_{k \neq j} x_{ik}\beta_k(m)$
$w_j(m) \leftarrow \sum_{i=1}^{n} W_i(m)x_{ij}(Z_i(m) - \tilde{Z}_i^{(j)}(m))$
Update $\beta_j(m)$ using Equation (22), $j \leftarrow j+1$
N
$j > p$
Y
Let $m \leftarrow m+1, \beta(m+1) \leftarrow \beta(m)$
N
Loss error
$<10^{-8}$
Y
End

Figure 3. Flowchart of the coordinate descent algorithm for the $LogSum + L_2$ penalized logistic regression model.

A coordinate decent algorithm for the *LogSum* + L_2 model. Inspired by Liang et al.[38], Eq. (3) is linearized by one-term Taylor series expansion:

$$L(\beta, \lambda) \approx \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Z_i - X_i\beta)' W_i (Z_i - X_i\beta) + \lambda \left(\sum_{j}^{p} \left(\lambda_1 log\left(|\beta_j| + \right) + \lambda_2 |\beta_j|^2 \right) \right) \right\} \tag{30}$$

where $\varepsilon > 0$, $Z_i = X_i\tilde{\beta} + \frac{Y_i - f(X_i\tilde{\beta})}{f(X_i\tilde{\beta})(1-f(X_i\tilde{\beta}))}$ is the estimated response, $W_i = f(X_i\tilde{\beta})(1 - f(X_i\tilde{\beta}))$ is the weight and $f(X_i\tilde{\beta}) = \frac{exp(X_i\tilde{\beta})}{1+exp(X_i\tilde{\beta})}$. Redefine the partial residual for fitting current $\tilde{\beta}_j$ as $\tilde{Z}_i^{(j)} = \sum_{k \neq j} x_{ik}\tilde{\beta}_k$ and $w_j = \sum_{i=1}^{n} W_i x_{ij}(Z_i - \tilde{Z}_i^{(j)})$. A pseudocode of the coordinate descent algorithm for the $Logsum + L_2$ penalized logistic regression model is shown in Algorithm 1 (Fig. 3).



Algorithm 1: Coordinate descent algorithm for the $LogSum + L_2$ penalized logistic regression model.

Require: Dataset $\{X, y\}$, the parameters λ, λ1 and λ2 are chosen by 10-fold cross-validation

Ensure: β

```
Initialize all $\beta_j(m) = 0 (j = 1,2,3,\ldots,p), m = 0$ ;
repeat
  Calculate $Z(m)$ and $W(m)$ and the loss function Equation (24) based on $\beta(m)$;
  for $j = 1 : p$ do
      Compute $\tilde{Z}_i^{(j)}(\boldsymbol{m}) \leftarrow \sum_{k \neq j} x_{ik}\beta_k(m)$ ;
      $w_j(m) \leftarrow \sum_{i=1}^{n} W_i(m)x_{ij}(Z_i(m) - \tilde{Z}_i^{(j)}(\boldsymbol{m}))$ ;
      Update $\beta_j(m)$ using Equation (22);
  end for
  Let $m \leftarrow m+1, \beta(m+1) \leftarrow \beta(m)$ ;
until $\sum_{i=1}^{p}(|\beta_i(m+1)| - |\beta_i(m)|) < 10^{-8}$
```

Experiments experimental results and discussion

Analysis on simulated data. In this section, we analyze the performance of the proposed method (the $LogSum + L_2$ penalized logistic regression model) by simulation analysis. We compare the proposed method with other three methods, which are logistic regression with L_1, L_{EN}, *HLR* regularizations. We simulate data from the true model.

$$\log\left(\frac{y}{1-y}\right) = X\beta + \sigma\varepsilon,\ \varepsilon \sim N(0,\ 1)$$

where $X \sim N(0,\ 1)$, ε is the independent random error and σ is the parameter that controls the signal to noise. Two scenarios are presented here. In each example, the dimension of features is 1000. Here are the details of the two scenarios.

1. In Scenario 1, the dataset consists of 200 observations, we set σ = 0.3 and simulate the group feature situation.

$$\beta = \left(\underbrace{2,2,2,2,2,}_{5}\underbrace{0,\ldots,0}_{995}\right);$$

$$x_i = \rho \times x_1 + (1-\rho) \times x_i,\ i = 2,3,4,5;$$

where ρ is the correlation coefficient of the group features.

In this example, there is one set of related features. The ideal sparse regression method should select 5 real features and set other 995 features as noise features to zero.

2. In Scenario 2, we set σ = 0.4 and the dataset consists of 400 observations, and defined two group features.

$$\beta = \left(\underbrace{2,2,2,2,2,1.5,-2,1.7,3,-2.5,}_{10}\underbrace{3,\ldots,3}_{10}\underbrace{0,\ldots,0}_{980}\right);$$

$$x_i = \rho \times x_1 + (1-\rho) \times x_i, i = 2,3,\ldots,10;$$

ρ	Method	Scenario							
		Accuracy		Sensitivity		Specificity		AUC	
		1	2	1	2	1	2	1	2
0.2	L_1	90.00% (1.85%)	98.78% (0.37%)	91.30% (1.12%)	99.82% (0.01%)	88.73%(2.14%)	97.45% (0.49%)	97.12% (0.53%)	98.12% (0.32%)
	L_{EN}	87.14% (2.28%)	99.16% (0.12%)	86.96% (2.97%)	99.78% (0.03%)	87.32% (3.17%)	97.33% (0.52%)	95.08% (0.93%)	97.89% (0.35%)
	HLR	94.29% (0.59%)	98.65% (0.35%)	95.65% (0.62%)	99.82% (0.01%)	92.96% (1.26%)	98.31% (0.38%)	98.84% (0.21%)	98.53% (0.33%)
	$LogSum + L_2$	**100%** (0%)	**99.50%** (0.01%)	**100%** (0%)	**99.96%** (0.01%)	**100%** (0%)	**99.42%** (0.01%)	**100%** (0%)	**99.27%** (0.06%)
0.6	L_1	91.43% (1.35%)	98.65% (0.26%)	87.69% (2.61%)	98.76% (0.13%)	94.67% (0.92%)	97.24% (0.29%)	97.37% (0.31%)	98.16% (0.28%)
	L_{EN}	85.71% (2.03%)	97.76% (0.31%)	69.23% (2.84%)	97.84% (0.20%)	100% (0%)	98.86% (0.22%)	96.04% (0.48%)	98.07% (0.34%)
	HLR	90.71% (1.76%)	98.65% (0.23%)	87.69% (1.48%)	99.12% (0.04%)	93.33% (0.81%)	98.21% (0.26%)	97.58% (0.40%)	98.54% (0.32%)
	$LogSum + L_2$	**97.86%** (0.21%)	**99.23%** (0.02%)	**95.38%** (0.62%)	**99.30%** (0.02%)	**100%** (0%)	**99.10%** (0.02%)	**100%** (0%)	**98.97%** (0.09%)

Table 1. Training results of different methods on the simulated datasets. Numbers in parentheses are the standard deviations and the best results are highlighted in bold.

$$x_i = \rho \times x_{11} + (1 - \rho) \times x_i, i = 12, 13, \ldots, 20;$$

In this example, there are two sets of related group features. The ideal penalized logistic regression method should select 20 real features and set other 980 features as noise features to zero.

In this experiment, we initialize the coefficient ρ of features' correlation as 0.2, 0.6 respectively, and hope to observe the accuracy of testing under different correlations by running different correlation values. The L_1 and L_{EN} approaches were executed by Glmnet (http://web.stanford.edu/~hastie/glmnet_matlab/, MATLAB version 2014-a). We use the tenfold cross-validation (CV) approach to optimize the regularization parameters or tuning parameters (balance the tradeoff between data fit and model complexity) of the L_1, L_{EN}, *HLR* and $LogSum + L_2$ approaches.

At the beginning, we divided the datasets at random into the training sets and the test sets. In our experiment, the approximate 70% of samples are proposed as training sets, and the rest are used as test sets. We repeated the simulations 30 times for each penalty method and computed the mean classification accuracy, mean classification sensitivity, and mean classification specificity on the training and test datasets respectively. To evaluate the quality of the selected features for the regularization approaches, the sensitivity and specificity of the feature selection performance[39] were defined as the follows:

$$\text{True Negative} \ \ (TN) := \left|\overline{\beta}.*\overline{\hat{\beta}}\right|_0, \quad \text{False Positive} \ \ (FP) := \left|\overline{\beta}.*\hat{\beta}\right|_0$$

$$\text{False Negative} \ \ (FN) := \left|\beta.*\overline{\hat{\beta}}\right|_0, \quad \text{True Positive} \ \ (TP) := \left|\beta.*\hat{\beta}\right|_0$$

$$\beta\text{-Sensitivity} := \frac{TP}{TP + FN}, \quad \beta\text{-Specificity} := \frac{TN}{TN + FP}$$

where the $.*$ is the element-wise product, and $|.|_0$ calculates the number of non-zero elements in a vector, $\overline{\beta}$ and $\overline{\hat{\beta}}$ are the logical "not" operators on the vector β and $\hat{\beta}$.

The training results of different methods on simulate datasets are reported in Table 1. As it can be seen, for all scenarios, our proposed $LogSum + L_2$ procedure generally achieves higher or comparable classification performance than the L_1, L_{EN} and *HLR* methods. For example, in the Scenario 1 with $\rho = 0.6$, our proposed method gained the 97.86% of accuracy, 95.38% of sensitivity and 100% of specificity, all of this data has increased by 6% for other methods. And whatever Scenario 1 or 2, the $LogSum + L_2$ methods always show the highest accuracy of training set, both $\rho = 0.2$ and $\rho = 0.6$. In summary, in the case of different scenarios and different values ρ, the $LogSum + L_2$ penalized logistic regression model is always the best.

Table 2 shows test results of different methods on simulate datasets. We can find that the performance of the $LogSum + L_2$ penalized logistic regression model is still the best one among the four methods. And in Scenario 1, whatever $\rho = 0.2$ or $\rho = 0.6$, the $LogSum + L_2$ approach shows similar values, but in Scenario 2, the sensitivity of the $LogSum + L_2$ model is far apart, and its accuracy and specificity are not much different compared with other three methods.

Table 3 shows the feature selection of all competing regularization methods. As shown in Table 3, these are the β-Sensitivity and β-Specificity. The approximate results are similar to the previous two Tables. In the

		Scenario							
		Accuracy		Sensitivity		Specificity		AUC	
ρ	Method	1	2	1	2	1	2	1	2
0.2	L_1	75.00% (3.82%)	71.67% (3.19%)	78.31% (2.83%)	78.57% (2.88%)	74.19% (3.64%)	63.50% (4.31%)	86.76% (1.63%)	80.80% (2.37%)
	L_{EN}	78.33% (3.15%)	66.67% (4.18%)	79.54% (2.68%)	75.00% (3.07%)	87.10% (2.63%)	53.13% (6.16%)	84.43% (1.87%)	76.23% (3.49%)
	HLR	80.00% (1.86%)	65.00% (4.03%)	79.63% (2.66%)	71.43% (3.92%)	87.10% (2.52%)	58.76% (5.83%)	88.65% (1.31%)	77.23% (3.46%)
	$LogSum + L_2$	**85.00%** (1.55%)	**76.67%** (3.17%)	**86.21%** (1.43%)	**81.00%** (3.17%)	**89.87%** (1.96%)	**78.13%** (3.05%)	**93.99%** (0.62%)	**83.82%** (2.29%)
0.6	L_1	68.33% (4.01%)	58.33% (4.92%)	62.07% (4.65%)	59.09% (4.83%)	70.97% (3.62%)	57.89% (5.52%)	82.76% (1.94%)	65.67% (4.46%)
	L_{EN}	71.67% (3.61%)	56.67% (5.32%)	55.17% (5.18%)	63.64% (4.78%)	77.42% (2.95%)	44.74% (8.03%)	81.76% (2.43%)	59.93% (5.04%)
	HLR	73.33% (3.33%)	55.00% (5.57%)	58.62% (4.96%)	59.09% (5.02%)	80.65% (2.31%)	52.63% (5.24%)	86.76% (1.88%)	52.27% (5.71%)
	$LogSum + L_2$	**85.00%** (1.73%)	**70.00%** (2.83%)	**82.76%** (2.04%)	**69.09%** (3.11%)	**87.10%** (1.78%)	**76.32%** (2.71%)	**92.10%** (0.50%)	**71.77%** (4.32%)

Table 2. Test results of different methods on the simulated datasets. Numbers in parentheses are the standard deviations and the best results are highlighted in bold.

		Scenario			
		β-Sensitivity		β-Specificity	
ρ	Method	1	2	1	2
0.2	L_1	73.45% (2.95%)	71.53% (2.94%)	99.90% (0.01%)	95.81% (0.52%)
	L_{EN}	73.16% (2.26%)	71.31% (2.32%)	99.95% (0.01%)	76.65% (2.73%)
	HLR	74.62% (3.05%)	73.15% (2.89%)	99.95% (0.01%)	95.45% (1.92%)
	$LogSum + L_2$	**82.57%** (2.58%)	**80.11%** (2.74%)	**99.95%** (0.01%)	**99.60%** (0.01%)
0.6	L_1	64.18% (3.56%)	62.43% (4.62%)	99.70% (0.01%)	95.00% (0.73%)
	L_{EN}	65.36% (3.63%)	63.34% (4.13%)	99.95% (0.01%)	76.00% (3.04%)
	HLR	65.41% (3.81%)	63.62% (4.51%)	99.90% (0.01%)	95.96% (0.65%)
	$LogSum + L_2$	**73.50%** (2.92%)	**72.31%** (3.86%)	**99.85%** (0.01%)	**99.24%** (0.01%)

Table 3. Results of β-sensitivity, β-specificity obtained by four methods. (Numbers in parentheses are the standard deviations and the best results are highlighted in bold).

same ρ value, the $LogSum + L_2$ penalized logistic regression model contains the greatest number of features and highest sensitivity and specificity. And in different ρ value, the performance of $\rho = 0.6$ always greater than the performance of $\rho = 0.2$.

Analysis of real data. We use three publicly available lung cancer microarray datasets, which download from GEO (https://www.ncbi.nlm.nih.gov/geo/). Some detail information and introduction will be shown below:

1. GSE10072: Series GSE10072 is a gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. Tobacco smoking can cause 90% of lung cancer cases, but the changes in the level of the molecules that lead to cancer development and affect survival are still unclear.
2. GSE19188: Series GSE19188 is a dataset about gene expression for early stage Non-small-cell lung carcinoma (NSCLC). 156 tumors and normal samples are aggregated into the expected group. The prognostic characteristics of 17 genes showed the best correlation with the survival time after surgery.
3. GSE19804: Series GSE19804 is a dataset about Genome-wide screening of transcriptional modulation in non-smoking female lung cancer in Taiwan. Although smoking is a major risk factor for lung cancer, only 7% of women with lung cancer in Taiwan have a history of smoking, which is much lower than that of white women. Researchers extracted RNA from paired tumors and normal tissues for gene expression analysis to

Dataset	No. of probes	Classes (Class1/Class2)	No. of sample (Class1/Class2)
GSE10072	22,284	Normal/Lung Cancer	107 (49/58)
GSE19188	54,675	Normal/Lung Cancer	156 (88/91)
GSE19804	54,675	Normal/Lung Cancer	120 (60/60)

Table 4. Three publicly available lung cancer gene expression datasets.

Data	Method	Training accuracy	Test accuracy	No. selected genes
GSE10072	L_1	98.32% (0.14%)	95.12% (0.31%)	23 (1.97)
	HLR	99.04% (0.04%)	98.4% (0.17%)	72 (8.45)
	LEN	98.21% (0.16%)	92.1% (0.94%)	11 (1.32)
	$LogSum + L_2$	**99.43%** (0.02%)	**99.15%** (0.08%)	7 (0.82)
GSE19188	L_1	97.11% (0.21%)	51.46% (6.05%)	72 (9.33)
	HLR	98.33% (0.09%)	47.56% (7.41%)	121 (10.34)
	LEN	96.3% (0.28%)	46.19% (5.23%)	17 (2.03)
	$LogSum + L_2$	**99.25%** (0.01%)	**75%** (3.44%)	10 (1.21)
GSE19804	L_1	99.05% (0.02%)	95.2% (0.61%)	37 (4.32)
	HLR	99.05% (0.02%)	94.6% (0.64%)	70 (7.73)
	LEN	97.14% (0.22%)	96.6% (0.58%)	9 (1.03)
	$LogSum + L_2$	**99.41%** (0.01%)	**98.45%** **(0.23%)**	6 (0.82)

Table 5. Training and test accuracy and number of selected genes of three lung cancer datasets in four methods. Numbers in parentheses are the standard deviations and the best results are highlighted in bold.

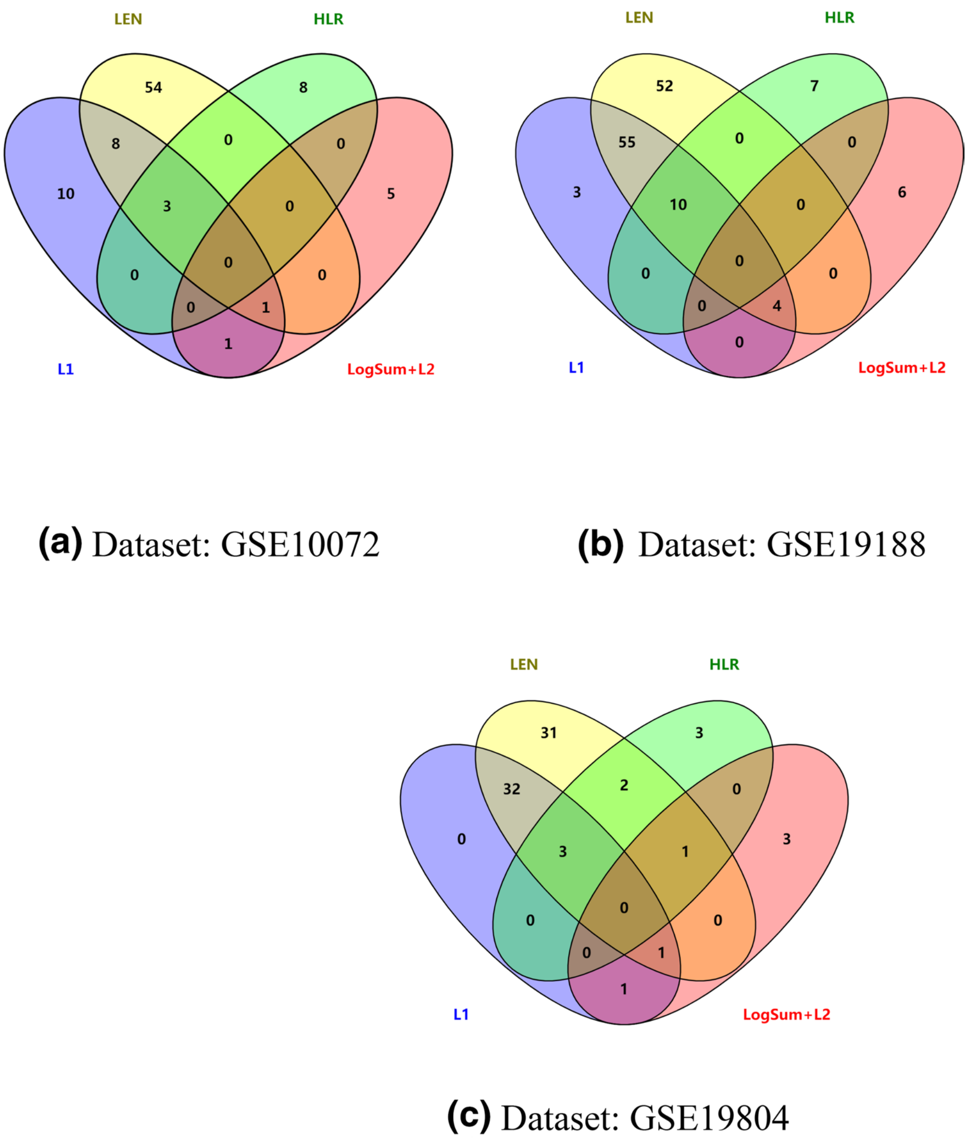
explain this phenomenon. This dataset and its reports comprehensively analyze the molecular characteristics of lung cancer in non-smoking women in Taiwan.

The GSE10072 dataset contains 22,284 microarray gene expression profiles and GSE19188 and GSE 19,804 both have 54,675 microarray gene expression profiles. As same as simulation data, we randomly divide the datasets such that 70% of the datasets become training samples and 30% become test samples. A brief introduction of these datasets is summarized in Table 4.

Table 5 describes the average training and test accuracies are obtained by different variable selection methods in the three datasets. It is easy to find that the performance of the $LogSum + L_2$ penalized logistic regression model is better than other three approaches. For example, in terms of training accuracy, the $LogSum + L_2$ approach reached 99.43%, and other three methods are 98.32%, 99.04% and 98.21% respectively in GSE10072 dataset. In GSE19188 dataset, we observe the test accuracy of the $LogSum + L_2$ method is 75%, and other three methods are 51.46%, 47.56% and 46.19% respectively. From the number of selected genes, we can find the $LogSum + L_2$ penalized logistic regression model always select the lowest number of genes and the L_{EN} approach select the highest number of genes.

In order to search the common gene signatures selected by the different methods, we used VENNY software to generate Venn diagrams. As show in Fig. 4, we consider the common gene signatures selected by the logistic regression model with L_1, L_{EN}, *HLR* and $LogSum + L_2$ regularizations, which are the most relevant signatures of lung cancer. Many genes selected by the $LogSum + L_2$ penalized logistic regression model do not appear in the results of the other three regularization methods. For example, the $LogSum + L_2$ approach selects 5, 6, and 3 unique genes from GSE10072, GSE19188 and GSE19804 datasets respectively. This means that the $LogSum + L_2$ penalized logistic regression model can find the different genes and pathways related to lung cancer compared with other three regularization methods.

Figures 5, 6 and 7 show the interactive networks of all the features selected by the $LogSum + L_2$ penalized logistic regression model. The integrative networks among these selected features are represented by the cBioPortal from publicly lung cancer datasets. The circles with thick border represent the selected genes, and the rest circles with gradient color-coded represent genes according to their alteration frequencies in databases. The



(a) Dataset: GSE10072

(b) Dataset: GSE19188

(c) Dataset: GSE19804

Figure 4. Venn diagram analysis of the results of L_1, L_{EN}, HLR and $LogSum + L_2$ regularization methods.

hexagons represent target drugs, and among of them some with yellow color represent the drugs approved by FDA. The links connected some selected genes represent that they have regulation correlations with group effect.

In GSE10072 dataset, from Fig. 5, we find a gene named EGFR, which has been conformed as the important target gene of NSCLC[40]. It belongs to ERBB receptor tyrosine kinase family, which include some other genes like HER2, HER3 and HER4. Due to observed patterns of oncogenic mutation of EGFR and HER2, many research works report their attractive option for targeted therapy in patients with NSCLC.

As shown in Fig. 6, three important genes TUBB1, PRKD1 and STK11 have been selected, and genes PRKD1 and STK11 have the regulation correlation with group effect from GSE19188 dataset. In fact, there are many drugs have been developed to target the gene TUBB1. And many research works report that genes PRKD1 and STK11 significantly influence the patients' survival rates across all tumors[41].

As shown in Fig. 7, four important genes EPCAM, SMC3, HIST1H2BL, and LMNA and their regulation correlations with group effect have been selected from GSE19804 dataset. Many research works report that the epithelial cell adhesion molecule (EPCAM) represents true oncogenes as the tumor-associated calcium signal transducer, and study the relationship between gene EPCAM and NSCLC[42].

Table 6 summarizes that the genes were selected by the $LogSum + L_2$ penalized logistic regression model. At the beginning of the experiments, the attribute of genes is prob set ID. Thus, we could transform prob set ID to gene symbol by using the website DAVID (https://david.ncifcrf.gov). According to the experimental results, the $LogSum + L_2$ penalized logistic regression model can find some unique genes, which cannot be identified by other regularization models but are significantly related to the disease. Therefore, we believe that the $LogSum + L_2$ penalized logistic regression model can accurately and efficiently identify cancer-related genes.

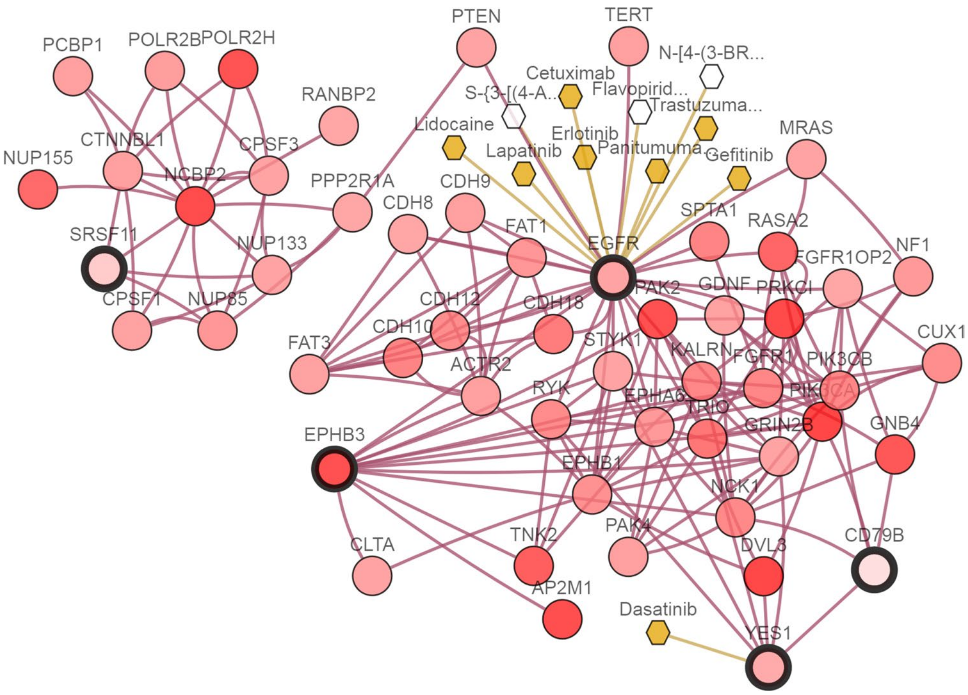


Figure 5. Maximum Integrative Network of features selected by the $LogSum + L_2$ penalized logistic regression model in GSE10072 dataset.

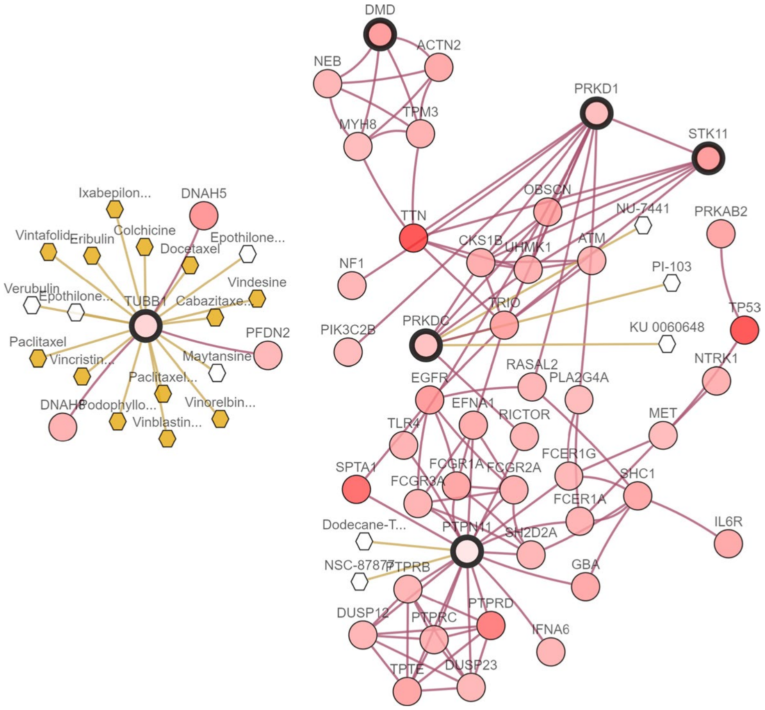


Figure 6. Maximum Integrative Network of features selected by the $LogSum + L_2$ penalized logistic regression model in GSE19188 dataset.

Discussion and conclusion

Successful identification of gene biomarkers and biological pathways can significantly improve the accuracy of diagnosis and help machine learning models have better performance on classification of different types of cancer. Many researchers used the logistic regressions with optimization methods for binary cancer classification. However, the traditional logistic regression model has two obvious shortcomings: feature selection and

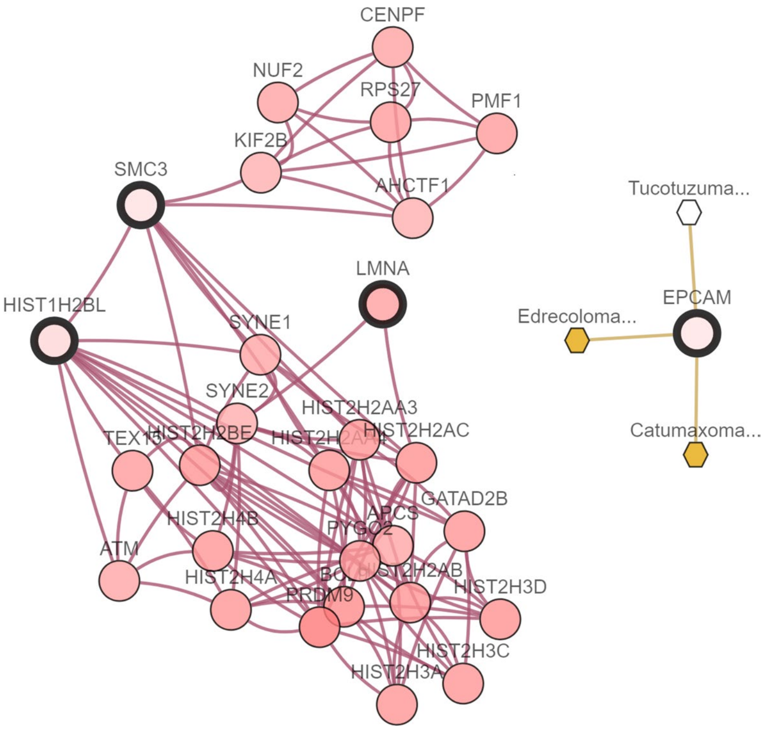


Figure 7. Maximum Integrative Network of features selected by the $LogSum + L_2$ penalized logistic regression model in GSE19804 dataset.

Prob_ID	Gene symbol	Gene name
Dataset: GSE10072		
201839_s_at	EPCAM	Epithelial cell adhesion molecule (EPCAM)
200685_at	SRSF11	Serine and arginine rich splicing factor 11(SRSF11)
204600_at	EPHB3	EPH receptor B3(EPHB3)
205297_s_at	CD79B	CD79b molecule (CD79B)
202932_at	YES1	YES proto-oncogene 1, Src family tyrosine kinase (YES1)
201983_s_at	EGFR	Epidermal growth factor receptor (EGFR)
201596_x_at	KRT18	Keratin 18(KRT18)
Dataset: GSE19188		
204292_x_at	STK11	Serine/threonine kinase 11(STK11)
205880_at	PRKD1	Protein kinase D1(PRKD1)
208694_at	PRKDC	Protein kinase, DNA-activated, catalytic polypeptide (PRKDC)
205868_s_at	PTPN11	Protein tyrosine phosphatase, non-receptor type 11(PTPN11)
214250_at	NUMA1	Nuclear mitotic apparatus protein 1(NUMA1)
231657_s_at	CCDC74A	Coiled-coil domain containing 74A(CCDC74A)
220939_s_at	DPP8	Dipeptidyl peptidase 8(DPP8)
210704_at	FEZ2	Fasciculation and elongation protein zeta 2(FEZ2)
208601_s_at	TUBB1	Tubulin beta 1 class VI(TUBB1)
207660_at	DMD	Dystrophin (DMD)
Dataset: GSE19804		
1553655_at	CDC20B	Cell division cycle 20B(CDC20B)
201839_s_at	EPCAM	Epithelial cell adhesion molecule (EPCAM)
1552370_at	C4ORF33	Chromosome 4 open reading frame 33(C4orf33)
1556925_at	SMC3	Structural maintenance of chromosomes 3(SMC3)
207611_at	HIST1H2BL	Histone cluster 1 H2B family member l(HIST1H2BL)
1554600_s_at	LMNA	Lamin A/C(LMNA)

Table 6. The genes are selected by the $LogSum + L_2$ penalized logistic regression model for different datasets.

overfitting problems. In this paper, we proposed the $LogSum + L_2$ penalized logistic regression model. Our proposed method can not only select sparse features (biomakers), but also identify the groups of the relevant features (gene pathways). The coordinate decent algorithm is used to solve the $LogSum + L_2$ penalized logistic regression model. We also evaluate the capability of our proposed method and compare its performance with other regularization methods. The results of simulations and real experiments indicate that the proposed method is highly competitive among several state-of-the-art methods. The disadvantage of the proposed method is its three regularization parameters need to be tuned by the *k*-fold cross-validation approach.

In recent years, increasing associations between of microRNAs (miRNAs) and human diseases have been identified. Based on accumulating biological data, many computational models for potential miRNA-disease associations inference have been developed[43–46]. We will apply the proposed $LogSum + L_2$ penalized logistic regression model to identify the non-coding RNA biomarker of human complex diseases as the future direction of our research.

Received: 21 May 2020; Accepted: 25 November 2020
Published online: 17 December 2020

References

1. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**(1–3), 389–422 (2002).
2. Heller, M. J. DNA microarray technology: Devices, systems, and applications. *Annu. Rev. Biomed. Eng.* **4**(1), 129–153 (2002).
3. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**(9), 1–8 (2003).
4. Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**(1), 1–12 (2004).
5. Dudoit, S., Fridlyand, J. & Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**(457), 77–87 (2002).
6. Li, T., Zhang, C. & Ogihara, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* **20**(15), 2429–2437 (2004).
7. Lee, J. W., Lee, J. B., Park, M. & Song, S. H. An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.* **48**(4), 869–885 (2005).
8. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **3**(02), 185–205 (2005).
9. Monari, G. & Dreyfus, G. Withdrawing an example from the training set: An analytic estimation of its effect on a non-linear parameterised model. *Neurocomputing* **35**(1–4), 195–201 (2000).
10. Rivals, I. & Personnaz, L. MLPs (mono-layer polynomials and multi-layer perceptrons) for nonlinear modeling. *J. Mach. Learn. Res.* **3**, 1383–1398 (2003).
11. Liu, X. Y., Liang, Y., Wang, S., Yang, Z. Y. & Ye, H. S. A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. *IEEE Access* **6**, 22863–22874 (2018).
12. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn Res.* **3**, 1157–1182 (2003).
13. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**(456), 1348–1360 (2001).
14. Zhang, H. H. & Lu, W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**(3), 691–703 (2007).
15. Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38**(2), 894–942 (2010).
16. Rosset, S. & Zhu, J. Piecewise linear regularized solution paths. *Ann. Stat.* **35**, 1012–1030 (2007).
17. Xu, Z., Zhang, H., Wang, Y., Chang, X. & Liang, Y. L1/2 regularization. *Sci. China Inf. Sci.* **53**(6), 1159–1169 (2010).
18. Xu, Z., Chang, X., Xu, F. & Zhang, H. L1/2 regularization: A thresholding representation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.* **23**(7), 1013–1027 (2012).
19. Candes, E. J., Wakin, M. B. & Boyd, S. P. Enhancing sparsity by reweighted L1 minimization. *J. Fourier Anal. Appl.* **14**(5–6), 877–905 (2008).
20. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **68**(1), 49–67 (2006).
21. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **67**(2), 301–320 (2005).
22. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Stat.* **32**(2), 407–499 (2004).
23. Fan, J. & Li, R. Variable selection for Cox's proportional hazards model and frailty model. *Ann. Stat.* **30**, 74–99 (2002).
24. Zou, H. & Zhang, H. H. On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* **37**(4), 1733 (2009).
25. Zeng, L. & Xie, J. Group variable selection via SCAD-L 2. *Statistics* **48**(1), 49–66 (2014).
26. Huang, H. H., Liu, X. Y. & Liang, Y. Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2+2 regularization. *PLoS ONE* **11**(5), e0149675 (2016).
27. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**(10), 906–914 (2000).
28. Medjahed, S. A., Saadi, T. A. & Benyettou, A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. *Int. J. Comput. Appl.* **62**(1), 1–5 (2013).
29. Zhou, X., Liu, K. Y. & Wong, S. T. Cancer classification and prediction using logistic regression with Bayesian gene selection. *J. Biomed. Inform.* **37**(4), 249–259 (2004).
30. Cawley, G. C. & Talbot, N. L. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics* **22**(19), 2348–2355 (2006).
31. Algamal, Z. Y. & Lee, M. H. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. *Adv. Data Anal. Classif.* **13**(3), 753–771 (2019).
32. Algamal, Z. An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. *Electron. J. Appl. Stat. Anal.* **10**(1), 242–256 (2017).
33. Shevade, S. K. & Keerthi, S. S. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19**(17), 2246–2253 (2003).
34. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B (Methodol.)* **58**(1), 267–288 (1996).
35. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**(1), 1 (2010).
36. Algamal, Z. Y. & Lee, M. H. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Syst. Appl.* **42**(23), 9326–9332 (2015).

37. Algamal, Z. Y. & Lee, M. H. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Comput. Biol. Med.* **67**, 136–145 (2015).
38. Liang, Y. *et al.* Sparse logistic regression with a L 1/2 penalty for gene selection in cancer classification. *BMC Bioinform.* **14**(1), 198 (2013).
39. Xia, L. Y. *et al.* Descriptor selection via log-sum regularization for the biological activities of chemical structure. *Int. J. Mol. Sci.* **19**(1), 30 (2018).
40. Jänne, P. A. *et al.* AZD9291 in EGFR inhibitor–resistant non–small-cell lung cancer. *N. Engl. J. Med.* **372**(18), 1689–1699 (2015).
41. Nath, A. & Chan, C. Genetic alterations in fatty acid transport and metabolism genes are associated with metastatic progression and poor prognosis of human cancers. *Sci. Rep.* **6**, 18669 (2016).
42. Pak, M. G., Shin, D. H., Lee, C. H. & Lee, M. K. Significance of EpCAM and TROP2 expression in non-small cell lung cancer. *World J. Surg. Oncol.* **10**(1), 53 (2012).
43. Chen, X., Wang, L., Qu, J., Guan, N. N. & Li, J. Q. Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* **34**(24), 4256–4265 (2018).
44. Chen, X., Xie, D., Zhao, Q. & You, Z. H. MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **20**(2), 515–539 (2019).
45. Chen, X., Yin, J., Qu, J. & Huang, L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* **14**(8), e1006418 (2018).
46. Chen, X., Yan, C. C., Zhang, X. & You, Z. H. Long non-coding RNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **18**(4), 558–576 (2017).

Author contributions

X.Y.L. and S.B.W. conceived the conception, designed and developed the method, acquired and analyzed the data and result. W.Q.Z., Z.J.Y. and H.B.X. wrote, reviewed and revised the manuscript. X.Y.L. is the correspondence author. All authors have read and approved the final manuscript.

Funding

Tis work is supported in part by the Key Project for University of Department of Education of Guangdong Province of China Funds (Natural) under Grant.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.-Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.