BMC Bioinformatics

RESEARCH ARTICLE Open Access

Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification

Yong Liang[1*], Cheng Liu[1], Xin-Ze Luan[1], Kwong-Sak Leung[2], Tak-Ming Chan[2], Zong-Ben Xu[3] and Hai Zhang[3]

Abstract

Background: Microarray technology is widely used in cancer diagnosis. Successfully identifying gene biomarkers will significantly help to classify different cancer types and improve the prediction accuracy. The regularization approach is one of the effective methods for gene selection in microarray data, which generally contain a large number of genes and have a small number of samples. In recent years, various approaches have been developed for gene selection of microarray data. Generally, they are divided into three categories: filter, wrapper and embedded methods. Regularization methods are an important embedded technique and perform both continuous shrinkage and automatic gene selection simultaneously. Recently, there is growing interest in applying the regularization techniques in gene selection. The popular regularization technique is Lasso (L_1), and many L_1 type regularization terms have been proposed in the recent years. Theoretically, the Lq type regularization with the lower value of q would lead to better solutions with more sparsity. Moreover, the $L_{1/2}$ regularization can be taken as a representative of Lq ($0 < q < 1$) regularizations and has been demonstrated many attractive properties.

Results: In this work, we investigate a sparse logistic regression with the $L_{1/2}$ penalty for gene selection in cancer classification problems, and propose a coordinate descent algorithm with a new univariate half thresholding operator to solve the $L_{1/2}$ penalized logistic regression. Experimental results on artificial and microarray data demonstrate the effectiveness of our proposed approach compared with other regularization methods. Especially, for 4 publicly available gene expression datasets, the $L_{1/2}$ regularization method achieved its success using only about 2 to 14 predictors (genes), compared to about 6 to 38 genes for ordinary L_1 and elastic net regularization approaches.

Conclusions: From our evaluations, it is clear that the sparse logistic regression with the $L_{1/2}$ penalty achieves higher classification accuracy than those of ordinary L_1 and elastic net regularization approaches, while fewer but informative genes are selected. This is an important consideration for screening and diagnostic applications, where the goal is often to develop an accurate test using as few features as possible in order to control cost. Therefore, the sparse logistic regression with the $L_{1/2}$ penalty is effective technique for gene selection in real classification problems.

Keywords: Gene selection, Sparse logistic regression, Cancer classification

Background

With the development of DNA microarray technology, the biology researchers can analyze the expression levels of thousands of genes simultaneously. Many studies have demonstrated that microarray data are useful for classification of many cancers. However, from the biological perspective, only a small subset of genes is strongly indicative of a targeted disease, and most genes are irrelevant to cancer classification. The irrelevant genes may introduce noise and decrease classification accuracy. Moreover, from the machine learning perspective, too many genes may lead to overfitting and can negatively influence the classification performance. Due to the significance of these problems, effective gene selection methods are desirable to help to classify different cancer types and improve prediction accuracy.

In recent years, various approaches have been developed for gene selection of microarray data. Generally, they are divided into three categories: filter, wrapper and embedded methods. Filter methods evaluate a gene based on discriminative power without considering its correlations with other genes [1-4]. The drawback of filter methods is that it examines each gene independently, ignoring the possibility that

* Correspondence: yliang@must.edu.mo
[1]Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau, China
Full list of author information is available at the end of the article

groups of genes may have a combined effect which is not necessarily reflected by the individual performance of genes in the group. This is a common issue with statistical methods such as *T*-test, which examine each gene in isolation.

Wrapper methods utilize a particular learning method as feature evaluation measurement to select the gene subsets in terms of the estimated classification errors and build the final classifier. Wrapper approaches can obtain a small subset of relevant genes and can significantly improve classification accuracy [5,6]. For example, Guyon et al. [7] proposed a gene selection approach utilizing support vector machines (SVM) based on recursive feature elimination. However, the wrapper methods greatly require extensive computational time.

The third group of gene selection procedures is embedded methods, which perform the variable selection as part of the statistical learning procedure. They are much more efficient computationally than wrapper methods with similar performance. Embedded methods have drawn much attention recently in the literature. The embedded methods are less computationally expensive and less prone to over fitting than the wrapper methods [8].

Regularization methods are an important embedded technique and perform both continuous shrinkage and automatic gene selection simultaneously. Recently, there is growing interest in applying the regularization techniques in the logistic regression models. Logistic regression is a powerful discriminative method and has a direct probabilistic interpretation which can obtain probabilities of classification apart from the class label information. In order to extract key features in classification problems, a series of regularized logistic regression methods have been proposed. For example, Shevade and Keerthi [9] proposed the sparse logistic regression based on the Lasso regularization [10] and Gauss-Seidel methods. Glmnet is the general approach for the L_1 type regularized (including Lasso and elastic net) linear model using a coordinate descent algorithm [11,12]. Similar to sparse logistic regression with the L_1 regularization method, Gavin C. C. and Nicola L. C. [13] investigated sparse logistic regression with Bayesian regularization. Inspired by the aforementioned methods, we investigate the sparse logistic regression model with a $L_{1/2}$ penalty, in particular for gene selection in cancer classification. The $L_{1/2}$ penalty can be taken as a representative of Lq ($0 < q < 1$) penalty and has demonstrated many attractive properties, such as unbiasedness, sparsity and oracle properties [14].

In this paper, we develop a coordinate descent algorithm to the $L_{1/2}$ regularization in the sparse logistic regression framework. The approach is applicable to biological data with high dimensions and low sample sizes. Empirical comparisons with sparse logistic regressions with the L_1 penalty and the elastic net penalty demonstrate the effectiveness of the proposed $L_{1/2}$ penalized logistic regression for gene selection in cancer classification problems.

Methods

Sparse logistic regression with the $L_{1/2}$ penalty

In this paper, we focus on a general binary classification problem. Suppose we have n samples, $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is i^{th} input pattern with dimensionality p and y_i is a corresponding variable that takes a value of 0 or 1; $y_i = 0$ indicates the i^{th} sample in Class 1 and $y_i = 1$ indicates the i^{th} sample is in Class 2. The vector X_i contains p features (for all p genes) for the i^{th} sample and x_{ij} denotes the value of gene j for the i^{th} sample. Define a classifier $f(x) = e^x/(1 + e^x)$ such that for any input x with class label y, $f(x)$ predicts y correctly. The logistic regression is expressed as:

$$P(Y_i = 1|X_i) = f\left(X'_i\beta\right) = \frac{\exp\left(X'_i\beta\right)}{1 + \exp\left(X'_i\beta\right)} \tag{1}$$

Where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are the coefficients to be estimated, note that β_0 is the intercept. The log-likelihood is:

$$l(\beta|D) = -\sum_{i=1}^{n}\left\{y_i \log\left[f\left(X'_i\beta\right)\right] + (1-y_i)\log\left[1-f\left(X'_i\beta\right)\right]\right\} \tag{2}$$

We can obtain β by minimizing the log-likelihood (2). In high dimensional application with $p >> n$, directly solving the logistic model (2) is ill-posed and may lead to overfitting. Therefore, the regularization approaches are applied to address the overfitting problem. When adding a regularization term to (2), the sparse logistic regression can be modelled as:

$$\beta = arg\,min\left\{l(\beta|D) + \lambda\sum_{j=1}^{p}P(\beta_j)\right\} \tag{3}$$

Where $\lambda > 0$ is a tuning parameter and $P(B)$ is a regularization term. The popular regularization technique is Lasso (L_1) [10], which has the regularization term $P(\beta) = \Sigma\,|\beta|$. Many L_1 type regularization terms have been proposed in the recent years, such as SCAD [15], elastic net [16], and MC+ [17].

Theoretically, the Lq type regularization $P(\beta) = \Sigma\,|\beta|^q$ with the lower value of q would lead to better solutions with more sparsity. However when q is very close to zero, difficulties with convergence arise. Therefore, Xu et al. [14] further explored the properties of Lq ($0 < q < 1$) regularization and revealed the extreme importance and special role of the $L_{1/2}$ regularization. They proposed that when $1/2 < q < 1$, the $L_{1/2}$ regularization can yield most sparse results and its difficulty with convergence is not very high compared with that of the L_1 regularization, while when $0 < q < 1/2$, the performance of Lq penalties makes no significant difference and solving the $L_{1/2}$ regularization is much simpler than solving the L_0 regularization. Hence, the $L_{1/2}$ regularization can

be taken as a representative of Lq ($0 < q < 1$) regularizations. In this paper, we apply the $L_{1/2}$ penalty to the logistic regression model. The sparse logistic regression model based on the $L_{1/2}$ penalty has the form:

$$\beta_{1/2} = arg\,min\left\{ l(\beta|D) + \lambda\sum_{j=1}^{p}|\beta_j|^{1/2} \right\} \quad (4)$$

The $L_{1/2}$ regularization has been demonstrated many attractive properties, such as unbiasedness, sparsity and oracle properties. The theoretical and experimental analyses show that the $L_{1/2}$ regularization is a competitive approach. Our work in this paper also reveals the effectiveness of the $L_{1/2}$ regularization to solve the nonlinear logistic regression problems with a small number of predictive features (genes).

A coordinate descent algorithm for the $L_{1/2}$ penalized logistic regression

The coordinate descent algorithm [11,12] is a "one-at-a-time" approach, and its basic procedure can be described as follows: for each coefficients, to partially optimize the target function with respect to $\beta_j (j = 1, 2, \ldots, p)$ with the remaining elements of β fixed at their most recently updated values.

Before introducing the coordinate descent algorithm for the nonlinear logistic regularization, we first consider a linear regularization case. Suppose the dataset D has n samples, $D = \{(X_1, y_1), (X_2, y_2), \ldots, (Xn, y_n)\}$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is i^{th} input variables with dimensionality p and y_i is the corresponding response variable. The variables are standardized: $\sum_{i=1}^{n} x_{ij}^2 = 1$ and $\sum_{i=1}^{n} y_i = 0$. Therefore, The linear regression with the regularization term can be expressed as:

$$R(\beta) = arg\,min\left\{ \frac{1}{n}\sum_{i=1}^{n}\left(y_i - X^{'}\beta\right)^2 + \lambda\sum_{j=1}^{p}P\left(\beta_j\right) \right\} \quad (5)$$

Where $P(B)$ is the regularization term. The coordinate descent algorithm solves β_j and other $\beta_{k\neq j}$ ($k \neq j$ represent the parameters remained after j^{th} element is removed) are fixed. The equation (5) can be rewritten as:

$$R(\beta) = argmin\left\{ \frac{1}{n}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k + x_{ij}\beta_j \right)^2 + \lambda\sum_{k\neq j}P(\beta_k) + \lambda P\left(\beta_j\right) \right\} \quad (6)$$

The first order derivative at β_j can be estimated as:

$$\frac{\partial R}{\partial \beta_j} = \sum_{i=1}^{n}\left(-x_{ij}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j \right) \right) + \lambda P\left(\beta_j\right)' = 0 \quad (7)$$

Define $\tilde{y}_i^{(j)} = \sum_{k\neq j} x_{ik}\beta_k$ as the partial residual for fitting β_j and $\omega_j = \sum_{i=1}^{n} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right)$, the univariate soft thresholding operator of the coordinate descent algorithm [11] for the L_1 regularization (Lasso) can be defined as:

$$\beta_j = S(\omega_j, \lambda) = \begin{cases} \omega_j + \lambda & if\,\omega_j < -\lambda \\ \omega_j - \lambda & if\,\omega_j > \lambda \\ 0 & if\,|\omega_j| < \lambda \end{cases} \quad (8)$$

Similarly, for the L_0 regularization, the thresholding operator of the coordinate descent algorithm can be defined as:

$$\beta_j = Hard(\omega_j, \lambda) = \omega I(|\omega_j| > \lambda) \quad (9)$$

where I is the indicator function. This formula is equivalent to the hard thresholding operator [17].

According to equations (8) and (9), we can know that the different penalties are associated with different thresholding operators. Therefore, Xu et al. [18] proposed a half thresholding operator to solve the $L_{1/2}$ regularization for linear regression model. It is an iterative algorithm and can be seen as multivariate half thresholding approach. In this paper, we propose the univariate half thresholding operator of the coordinate descent algorithm for the $L_{1/2}$ regularization. Based on equation (7), the gradient of the $L_{1/2}$ regularization at β_j can be expressed as:

$$\frac{\partial R}{\partial \beta_j} = \beta_j - \omega_j + \lambda\frac{sign\left(\beta_j\right)}{4\sqrt{\left|\beta_j\right|}} = 0 \quad (10)$$

Firstly, we consider the $\beta_j > 0$ statement, and let, $\sqrt{\left|\beta_j\right|} = \mu$, $\beta_j = \mu^2$. When $\beta_j > 0$, the equation (10) can be redefined as:

$$\mu^3 - \omega_j\mu + \frac{\lambda}{4} = 0 \quad (11)$$

There are three cases of $\omega_j < 0$, $0 < \omega_j < \frac{3}{4}\lambda^{\frac{2}{3}}$, and $\omega_j > \frac{3}{4}\lambda^{\frac{2}{3}}$ respectively.

(i) If $\omega_j < 0$, the three roots of equation (11) can be expressed as follows:

$\mu_1 = -2\ r\sin\ \frac{\phi}{3}$, $\mu_2 = r\sin\ \frac{\phi}{3} + i\sqrt{3}\ r\ \cos\ \frac{\phi}{3}$ and $\mu_3 = r\sin\ \frac{\phi}{3} - i\ \sqrt{3}\ r\ \cos\ \frac{\phi}{3}$,

where $r = \sqrt{\frac{|\omega_j|}{3}}$, $\phi = \arccos\left(\frac{\lambda}{8r^3}\right)$. When $r > 0$, none of the roots satisfices $\mu_1 > 0$. Thus, there is no solution to equation (11) when $\omega_j < 0$.

(ii) If $0 < \omega_j < \frac{3}{4}\lambda^{\frac{2}{3}}$, the three roots of equation (11) are:

$\mu_1 = -2\ r\ cos\ \frac{\phi}{3}$, $\mu_2 =\ r\ cos\ \frac{\phi}{3} + i\ \sqrt{3}r\ \sin\ \frac{\phi}{3}$ and $\mu_3 = r\ cos\ \frac{\phi}{3} - i\ \sqrt{3}r\ \sin\ \frac{\phi}{3}$.

There is still no solution to equation (11) in this case.

(iii) If $\omega_j > \frac{3}{4}\lambda^{\frac{2}{3}}$, the three roots of equation (11) are given by:

$$\cdot\mu_1 = -2\ r\ cos\ \frac{\phi}{3},\ \cdot\mu_2 = \ 2r\ cos\ \left(\frac{\pi-\phi}{3}\right) \text{and}\ \mu_3 = 2r\ cos\ \left(\frac{\pi+\phi}{3}\right).$$

In this case, the μ_2 is a unique solution of equation (10). Thus, the equation (11) has non-zero roots only when $\omega_j > \frac{3}{4}\lambda^{\frac{2}{3}}$. The unique solution of equation (10) is as follow:

$$\beta_j = (\mu_2)^2 = \frac{2}{3}|\omega_j|\left(1 + \cos\left(\frac{2\left(\pi - \phi\left(\omega_j\right)\right)}{3}\right)\right)$$

On the other hand, in the $\beta_j < 0$ statement, we denoted $\sqrt{|\beta_j|} = \mu$ and $\beta_j = -\mu^2$. The equation (10) can be transformed into the equation:

$$\mu^3 - \omega_j\mu - \frac{\lambda}{4} = 0 \quad (12)$$

The equation (12) also has a unique solution when $\omega_j < -\frac{3}{4}\lambda^{\frac{2}{3}}$:

$$\mu_2 = 2r\cos\left(\tfrac{\pi-\phi}{3}\right) \quad \text{and} \quad \beta_j = -(\mu_2)^2 = -\tfrac{2}{3}|\omega_j| \times \left(1 + \cos\left(\frac{2\left(\pi-\phi\left(\omega_j\right)\right)}{3}\right)\right).$$

In conclusion, the univariate half thresholding operator can be expressed as:

$$\beta_j = Half\left(\omega_j, \lambda\right) = \begin{cases} \frac{2}{3}\omega_j\left(1 + \cos\left(\frac{2\left(\pi - \phi_\lambda\left(\omega_j\right)\right)}{3}\right)\right) & if\ |\omega_j| > \frac{3}{4}(\lambda)^{\frac{2}{3}} \\ 0 & otherwise \end{cases} \quad (13)$$

where $\phi_\lambda(\omega)$ satisfies:

$$\cos(\phi_\lambda(\omega)) = \frac{\lambda}{8}\left(\frac{|\omega|}{3}\right)^{-\frac{3}{2}}$$

The coordinate descent algorithm for the $L_{1/2}$ regularization makes repeated use of the univariate half thresholding operator. The details of the algorithm will be described later. This coordinate descent algorithm for the regularization can be extended to the sparse logistic regression model. Based on the objective function (3) of the sparse logistic regression, one-term Taylor series expansion for $l(B)$ has the form of

$$L(\beta,\lambda) \approx \frac{1}{2n}\sum_{i=1}^{n}(Z_i - X_i\beta)' W_i (Z_i - X_i\beta) + \sum_{j=1}^{p} P\left(\beta_j\right) \quad (14)$$

Where $Z_i = X_i\tilde{\beta} + \frac{Y_i - f\left(X_i\tilde{\beta}\right)}{f\left(X_i\tilde{\beta}\right)\left(1-f\left(X_i\tilde{\beta}\right)\right)}$ is an estimated response, $W_i = f\left(X_i\tilde{\beta}\right)\left(1-f\left(X_i\tilde{\beta}\right)\right)$ is a weight and $f\left(X_i\tilde{\beta}\right) = \exp\left(X_i\tilde{\beta}\right)/\left(1 + \exp\left(X_i\tilde{\beta}\right)\right)$ is a evaluated value at current parameters. Redefine the partial residual for fitting current $\tilde{\beta}_j$ as $\tilde{Z}_i^{(j)} = \sum_{i=1}^{n} W_i\left(\tilde{Z}_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k\right)$ and $\sum_{i=1}^{n} x_{ij}\left(Z_i - \tilde{Z}_i^{(j)}\right)$, we can directly apply the coordinate descent algorithm with the $L_{1/2}$ penalty for sparse logistic regression and the details are given follows:

Algorithm: The coordinate descent algorithm for sparse logistic with the $L_{1/2}$ penalty

Step 1: Initialize all $\beta_j\ (m) = 0$ (j =1,2,…,p) and λ, set $m = 0$;

Step 2: Compute $Z\ (m)$ and $W(m)$ and approximate the loss function(14) based on the Current $\beta\ (m)$;

Step 3: Update each $\beta_j\ (m)$, and cycle over j=1,…,p, until $\beta_j\ (m)$ does not change;

Step 3.1: Calculate $Z_i^{(j)}(m) = \sum_{i=1}^{n} W_i(m)\left(Z_i(m) - \sum_{k\neq j} x_{ik}\beta_k(m)\right)$ and $\omega_j(m) = \sum_{i=1}^{n} x_{ij}\left(Z_i(m) - Z_i^{(j)}(m)\right)$;

Step 3.2: Update $\beta_j(m) = Half(\omega_j(m), \lambda)$;

Step 4: Let $m = m + 1$, $\beta(m+1) \leftarrow \beta(m)$,

repeat Steps 2, 3 until $\sum_{i=1}^{p}\left(|\beta_i(m+1)| - |\beta_i(m)|\right) < 10^{-8}$.

The coordinate descent algorithm for the $L_{1/2}$ penalized logistic regression works well in the sparsity problems, because the procedure does not need to change many irrelevant parameters and recalculate partial residuals for each update step.

Results

Analyses of simulated data

In this section, we evaluate the performance of the sparse logistic regression with the $L_{1/2}$ penalty in simulation study. We generate high-dimensional and low sample size data which contain many irrelevant features. Two methods are compared with our proposed approach: Sparse logistic regression with the Elastic Net penalty (L_{EN}) and Sparse logistic regression with the Lasso penalty (L_1).

We generated the vectors $\gamma_{i0},\gamma_{i1},\ldots,\gamma_{ip}$ (i = 1,...,n) independently from the standard normal distribution and the predictor vector(i=1,...,n) is generated by $x_{ij} = \gamma_{ij}\sqrt{1-\rho} + \gamma_{i0}\sqrt{\rho}$ (j=1,..., p), where ρ is the correlation coefficient of the predictor vectors [19]. The simulated data set generated from the logistic model:

$$\log\left(\frac{Y_i}{1-Y_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \sigma\cdot\varepsilon \qquad (15)$$

Where ε is the independent random error generated from $N(0,1)$ and σ is the parameter which controls the signal to noise. In every simulation, the dimension p of the predictor vector is 1000, and the first five true coefficients are nonzero: β_1 = 1, β_2 = 1, β_3 = -1, β_4 = -1, β_5 = 1, and $\beta_j = 0 (6 \le j \le 1000)$.

The estimation of the optimal tuning parameter λ in the sparse logistic regression models can be done in many ways and is often done by k-fold cross-validation (CV). Note that the choice of k will depend on the size of the training set. In our experiments, we use 10-fold cross-validation (k=10). The elastic net method has two tuning parameters, we need to cross-validate on a two-dimensional surface [16].

We consider the cases with the training sample size n = 50, 80, 100, the correlation coefficient ρ =0.1, 0.4 and the noise control parameter σ =0.2, 0.6 respectively. Each classifier was evaluated on a test data set including 100 samples. The experiments were repeated 30 times and we report the average test errors in Table 1. As shown in Table 1, when the sample size n increases, the prediction performances of all the three methods are improved. For example when ρ =0.1, and σ =0.2, the average test errors of the $L_{1/2}$ method are 28.2%, 10.7% and 8.1% with the sample sizes n=50, 80, and 100 respectively. When the correlation parameter ρ and the noise parameter σ increase, the prediction performances of all the three methods are decreased. For example, when ρ =0.4 and n =100, the average test errors from the $L_{1/2}$ method increased from 9.1% to 15.1%, in which σ increased from 0.2 to 0.6. When σ =0.6 and n =80, the average test error from the $L_{1/2}$ method increase from 18.4% to 20.5%, in which ρ increased from 0.1 to 0.4. Moreover, in our simulation, the influence of the noise may be larger than that of the variable correlation for the prediction performance of all the three methods. On the other hand, at the same parameter setting case, the prediction performance of the $L_{1/2}$ method is consistent and better than the results of the L_{EN} and L_1 methods. For example, when ρ =0.1, σ =0.2 and n=100, the predictive error of the $L_{1/2}$ method is 8.1% much better than 16.9% and 15.7% got by the L_{EN} and L_1 methods respectively.

Table 1 The average errors (%) for the test data sets obtained by the sparse logistic regressions with the $L_{1/2}$, L_{EN} and L_1 penalties in 30 runs

	Sample size	$L_{1/2}$	L_{EN}	L_1
$\rho = 0.1$, $\sigma = 0.2$	n=50	28.2	31.8	31.2
	n=80	10.7	23.1	22.2
	n=100	8.1	16.9	15.7
$\rho = 0.1$, $\sigma = 0.6$	n=50	31.4	33.1	33.3
	n=80	18.4	27.1	26.6
	n=100	14.2	22.4	21.3
$\rho = 0.4$, $\sigma = 0.2$	n=50	30.1	32.6	33.0
	n=80	11.1	23.3	22.9
	n=100	9.1	19.0	16.4
$\rho = 0.4$, $\sigma = 0.6$	n=50	35.1	35.5	36.3
	n=80	20.5	27.2	26.9
	n=100	15.1	22.7	22.9

Table 2 The average number of variables selected by the sparse logistic regressions with the $L_{1/2}$, L_{EN} and L_1 penalties in 30 runs

	Sample size	$L_{1/2}$	L_{EN}	L_1
$\rho = 0.1$, $\sigma = 0.2$	n=50	7.5	31.6	27.1
	n=80	8.8	43.1	40.3
	n=100	8.9	49.7	45.7
$\rho = 0.1$, $\sigma = 0.6$	n=50	8.3	33.6	29.2
	n=80	10.6	45.7	41.9
	n=100	10.8	54.4	50.1
$\rho = 0.4$, $\sigma = 0.2$	n=50	7.8	33.5	28.3
	n=80	8.9	44.5	41.8
	n=100	9.0	51.2	46.6
$\rho = 0.4$, $\sigma = 0.6$	n=50	8.6	41.3	29.9
	n=80	10.7	45.9	44.1
	n=100	11.2	56.4	53.4

Table 2 shows the average number of the variables selected in 30 runs for each method. Since the simulation datasets have x_1-x_5 relevant features, the idealized average number of variables selected by each method is 5. In Table 2, the results obtained by the $L_{1/2}$ penalized method are obviously closed to 5 and 3–10 times smaller than those of the L_{EN} and L_1 penalties at the same parameter setting. For example, when ρ =0.1, σ =0.2 and n=100, the average numbers from the L_{EN} and L_1 methods are 49.7 and 45.7 respectively, and the result of $L_{1/2}$ method is 8.9. Moreover, when the sample size n, the correlation parameter ρ, and the noise parameter σ increase, the average numbers from all the three methods increase, but the values of the L_{EN} and L_1 methods increase faster than those of the $L_{1/2}$ method. This means that the $L_{1/2}$ penalized method consistently outperforms than other two methods in term of variable selection.

To further evaluate the performance of the $L_{1/2}$ penalized method, we report the frequency with which each relevant variable was selected among 30 runs for each method in Table 3. When the sample size is small (n=50), the $L_{1/2}$ penalty selects the relevant variables slightly less frequently than the other two methods and all the three methods select true nonzero coefficients with difficulties, especially when ρ and σ are relatively large. For example, when ρ =0.4, σ =0.6, n=50, and for β_5, the selected frequencies of the $L_{1/2}$, L_{EN} and L_1 methods are 12, 14 and 13 respectively in 30 runs. As n increases, all the three methods tend to select the true nonzero coefficients more accurately and the $L_{1/2}$ penalty method performs slightly better, in terms of variable frequencies, than the other two methods under the different parameter settings of ρ and σ. To sum up, Tables 1, 2 and 3 clearly show that the $L_{1/2}$ method is winner among the competitors in terms of both prediction accuracy and variable selection in the different variable correlation and noise situations.

Table 3 The frequencies of the relevant variables obtained by the sparse logistic regressions with the $L_{1/2}$, L_{EN} and L_1 penalties in 30 runs

	Sample size	Method					
$\rho = 0.1$, $\sigma = 0.2$	n=50	$L_{1/2}$	21	22	19	15	15
		L_{EN}	24	25	21	17	17
		L_1	22	24	20	15	17
	n=80	$L_{1/2}$	30	30	30	30	30
		L_{EN}	30	29	30	30	30
		L_1	30	29	30	30	30
	n=100	$L_{1/2}$	30	30	30	30	30
		L_{EN}	30	30	30	30	30
		L_1	30	30	30	30	30
$\rho = 0.1$, $\sigma = 0.6$	n=50	$L_{1/2}$	17	17	17	14	14
		L_{EN}	18	19	17	16	14
		L_1	18	18	18	16	15
	n=80	$L_{1/2}$	30	29	30	28	28
		L_{EN}	30	28	30	28	27
		L_1	30	28	30	27	26
	n=100	$L_{1/2}$	30	30	30	30	30
		L_{EN}	30	30	30	30	30
		L_1	30	30	30	28	30
$\rho = 0.4$, $\sigma = 0.2$	n=50	$L_{1/2}$	19	18	18	16	15
		L_{EN}	21	22	21	17	17
		L_1	18	21	19	16	17
	n=80	$L_{1/2}$	30	30	30	30	30
		L_{EN}	30	28	30	29	29
		L_1	30	27	30	29	29
	n=100	$L_{1/2}$	30	30	30	30	30
		L_{EN}	30	30	30	30	30
		L_1	30	30	30	29	29
$\rho = 0.4$, $\sigma = 0.6$	n=50	$L_{1/2}$	14	16	15	12	12
		L_{EN}	17	17	17	12	14
		L_1	17	15	14	9	13
	n=80	$L_{1/2}$	29	25	26	28	29
		L_{EN}	28	24	24	27	24
		L_1	27	24	24	23	23
	n=100	$L_{1/2}$	30	29	30	30	30
		L_{EN}	30	27	28	28	30
		L_1	29	27	27	28	30

Analyses on microarray data

In this section, we compare our proposed $L_{1/2}$ penalized method with the L_{EN} and L_1 methods on 4 publicly available gene expression datasets: Leukaemia, Prostate, Colon and DLBCL. A brief description of these datasets is given below and summarized in Table 4.

Leukaemia dataset

The original dataset was provided by Golub et al. [7], and contains the expression profiles of 7,129 genes for 47 patients of acute lymphoblastic leukaemia (ALL) and 25 patients of acute myeloid leukaemia (AML). For data preprocessing, we followed the protocol detailed in the supplementary information to Dudoit et al. [1]. After thresholding, filtering, applying a logarithmic transformation and standardizing each expression profile to zero mean and unit variance, a dataset comprising 3,571 genes remained.

Table 4 Four publicly available gene expression datasets used in the experiments

Dataset	No. of genes	No. of samples	classes
Leukaemia	3571	72	ALL/AML
Prostate	5966	102	Normal/Tumor
Colon	2000	62	Normal/Tumor
DLBCL	6285	77	DLBCL/FL

Prostate dataset

This original dataset contains the expression profiles of 12,600 genes for 50 normal tissues and 52 prostate tumor tissues. For data preprocessing, we adopt the pretreatment method [20] to obtain a dataset with 102 samples. And each sample contains 5966 genes.

Colon dataset

The colon microarray data set in Alon et al. [21] has 2000 genes per sample and 62 samples which consist of 22 normal tissues and 40 cancer tissues. The Colon dataset are available at http://microarray.princeton.edu/oncology.

DLBCL dataset

This dataset contains 77 microarray gene expression profiles of the 2 most prevalent adult lymphoid malignancies: 58 samples of diffuse large B-cell lymphomas (DLBCL) and 19 observations of follicular lymphoma (FL). Each sample contains 7,129 gene expression values. More information on these data can be found in Shipp MA et al. [22]. For data preprocessing, we followed the protocol detailed in the supplementary information to Dudoit et al. [1], and a dataset comprising 6,285 genes remained.

We evaluate the prediction accuracy of the three penalized logistic regression models using random partition. This means that we divide the datasets at random such that approximate 70-80% of the datasets becomes training samples and the other 20-30% test samples. More information on these data is given in Table 5. For selecting the tuning parameter λ, we employ the ten-fold cross validation scheme using the training set. We repeat this procedure 30 times and the averaged misclassification errors were reported in Table 6. Here the denominators of the ten-fold cross validation errors and the test errors describe the sample size of training and test

Table 5 The detail information of 4 microarray datasets used in the experiments

Dataset	No.of Training(class1/class2)	No.of Testing(class1/class2)
Leukaemia	50(32 ALL/18 AML)	22 (15 ALL/7 AML)
Prostate	71(35 ALL/36 AML)	31(15 ALL/16 AML)
Colon	42(14 Normal/28 Tumor)	20(8 Normal/12 Tumor)
DLBCL	60(45 DLBCL/15FL)	17(13 DLBCL/4 FL)

Table 6 The classification performances of different methods for 4 gene expression datasets

Dataset	Method	Cross-validation error	Test error	No. of selected genes
Leukaemia	$L_{1/2}$	2/50	1/22	2
	L_{EN}	1/50	1/22	9
	L_1	1/50	1/22	6
Prostate	$L_{1/2}$	5/71	3/31	5
	L_{EN}	5/71	4/31	34
	L_1	5/71	3/31	25
Colon	$L_{1/2}$	4/42	3/20	5
	L_{EN}	5/42	4/20	13
	L_1	5/42	4/20	7
DLBCL	$L_{1/2}$	3/60	2/17	14
	L_{EN}	2/60	1/17	38
	L_1	3/60	3/17	23

datasets respectively. The fractions of the ten-fold cross validation errors and the test errors and the number of gene selected are the approximated integers of the corresponding average number at 30 runs. As shown in Table 6, for Leukaemia dataset, the classifier with the $L_{1/2}$ penalty gives the average ten-fold cross validation error of 2/50 and the average test error of 1/22 with about 2 genes selected. The classifiers with L_{EN} and L_1 methods give the average ten-fold cross validation errors of 1/50 and the average test errors of 1/22 with about 9 and 6 genes selected respectively. This means that all three methods can be successfully applied to high-dimensional classification problems and classify the Leukaemia dataset with same accuracies. Note that, the $L_{1/2}$ method achieved its success using only about 2 predictors (genes), compared to about 9 and 6 for the L_{EN} and L_1 methods. For Prostate and Colon datasets, it can be seen the $L_{1/2}$ method achieves the best classification performances with the highest accuracy rates using much fewer genes compared with those of the L_{EN} and L_1 methods. For DLBCL dataset, the $L_{1/2}$ logistic regression achieves better classification performance than that of the L_1 method and worse than that of the L_{EN} method. However, as well as other three datasets, the $L_{1/2}$ method achieved its success using much less predictors (about 14 genes), compared to about 38 and 23 for the L_{EN} and L_1 methods. This is an important consideration for screening and diagnostic applications, where the goal is often to develop an accurate test using as few features as possible in order to control cost.

Figures 1, 2 and 3 display the solution paths and the gene selection results of the three methods for the Prostate dataset in one sample run. Here the x-axis displays the number of running steps, the y-axis in the left sub-

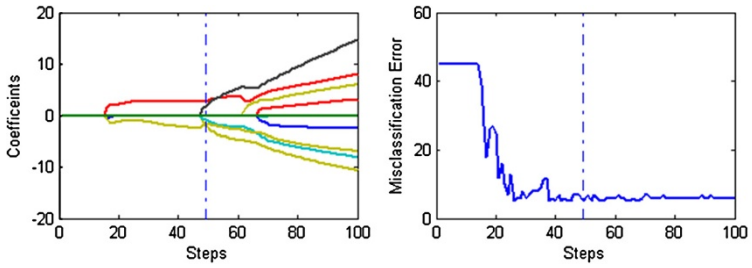

Figure 1 The results of the sparse logistic regression with the $L_{1/2}$ penalty on Prostate dataset. The solution paths and the gene selection results of the sparse logistic $L_{1/2}$ penalty methods for the Prostate dataset in one sample run.

figure is the coefficients measured gene importance and the y-axis in the right sub-figure is the misclassification errors based on the ten-fold cross validation. The optimal results of three methods are shown as vertical dotted lines. Figure 1 indicates that the number of nonzero coefficients (selected genes) of the optimal results obtained by the $L_{1/2}$ method is 5. In contrast, Figures 2 and 3 indicate that the numbers of nonzero coefficients (selected genes) of optimal results obtained by the L_{EN} and L_1 methods are 37 and 26 respectively. Generally speaking, the penalized logistic regression methods can be successfully applied to the cancer classification problems with high dimensional and low samples microarray data, and our proposed $L_{1/2}$ method achieves better performance especially in gene selection.

Brief biological analyses of the selected genes

The summaries of the 10 top-ranked informative genes found by the three sparse logistic regression methods for 4 gene expression datasets are shown in Tables 7, 8, 9 and 10 respectively. The genes with star(*) are the most frequently selected genes to construct the classifiers according to the last column of Table 6, and the common genes obtained by each classifier are emphasized with bold. The biologically experimental results proved some genes included in the frequently selected gene sets that produce high classification accuracy rate are mostly and functionally related to carcinogenesis or tumor histogenesis. For example, in Table 7, the most frequently selected gene set of each sparse logistic method for leukemia classification, including cystatin C (CST3) and myeloperoxidase (MPO) genes, that achieve high classification accuracy by the $L_{1/2}$ method, are experimentally proved to be correlated to leukemia of ALL or AML. The cystatin C gene is located at the extracellular region of the cell and has role in invasiveness of human glioblastoma cells. Decrease of cystatin C in the CSF might contribute to the process of metastasis and spread of the cancer cells in the leptomeningeal tissues [23]. The myeloperoxidase gene is taking role in anti-apoptosis process where cancer cells kill themselves [24]. For the colon dataset (Table 9), the most frequently selected gene set of each sparse logistic method includes genes such as guanylate cyclase activator 2B (GUCA2B), myosin, light chain 6, alkali, smooth muscle and non-muscle (MYL6) and Human desmin (DES) genes. These genes are the top 3 significant informative genes ranked by our proposed $L_{1/2}$ method and also selected by Ben-Dor et al. [25], Yang and Song [26] and Li et al. [27]. On the top of these genes lists is guanylate cyclase activator

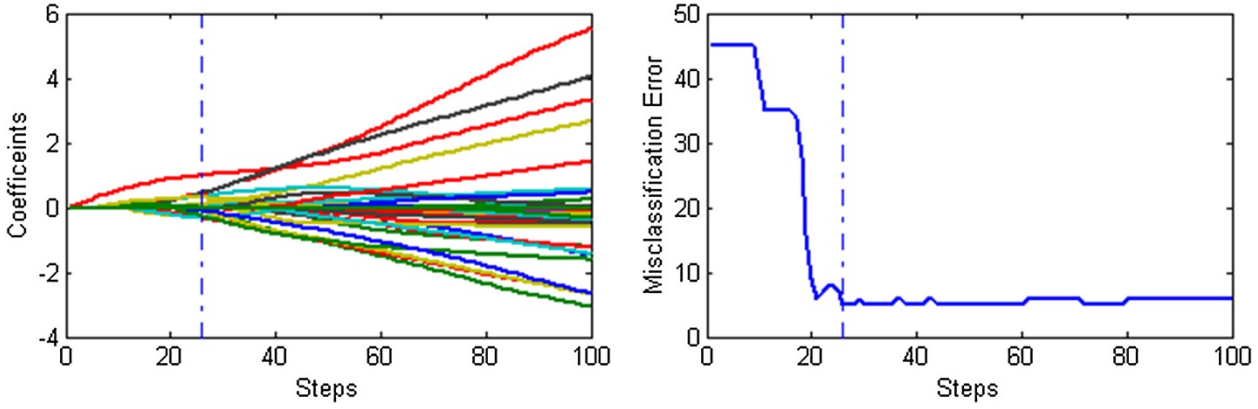

Figure 2 The results of the sparse logistic regression with the L_{EN} penalty on Prostate dataset. The solution paths and the gene selection results of the sparse logistic elastic net penalty methods for the Prostate dataset in one sample run.

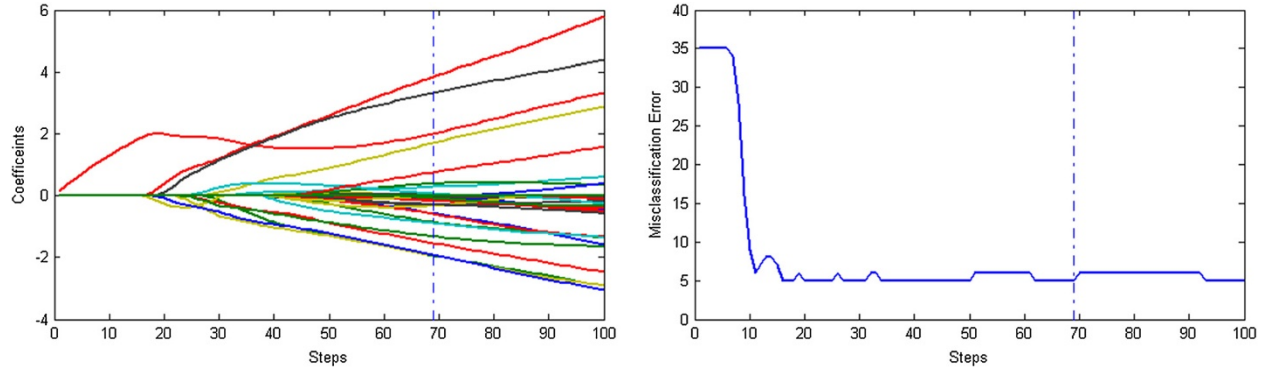

Figure 3 The results of the sparse logistic regression with the L_1 penalty on Prostate dataset. The solution paths and the gene selection results of the sparse logistic L_1 penalty methods for the Prostate dataset in one sample run.

2B (GUCA2B) gene. Notterman et al. [28] showed that a reduction of uroguanylin might be an indication of colon tumors, and Shailubhai et al. [29] reported that treatment with uroguanylin has a positive therapeutic significance to the reduction in pre-cancerous colon ploys.

In Tables 7, 8, 9 and 10, some genes are only frequently selected by the L1/2 method, but not discovered by the L_{EN} and L_1 methods. The evidence from the literatures showed that they are cancer related genes. For example, for the colon dataset, the genes cholinergic receptor, nicotinic, delta polypeptide (CHRND) and platelet/endothelial cell adhesion molecule-1 (PECAM1) were also selected by Maglietta R. et al. [30], Wiese A.H. et al. [31], Wang S. L. et al. [32], and Dai J. H. and Xu Q. [33]. These genes can significantly discriminate between non-dissected tumors and micro dissected invasive tumor cells. It is remarkable that apparently (to our knowledge) some discovered genes that have not been seen in any past studies.

On the other hand, from Tables 7, 8, 9 and 10, we found that the most frequently selected genes and their ranking orders by the LEN and L1 methods are much similar compared with those of the L1/2 method. The main reasons are that the classification hypothesis needs not be unique as the samples in gene expression data lie in a high-dimensional space, and both of the LEN and L1 methods are based on the L1 type penalty.

Table 7 The 10 top-ranked informative genes found by the three sparse logistic regression methods from the Leukaemia dataset

Rank	Gene description		
	$L_{1/2}$	L_{EN}	L_1
1	**CST3 cystatin C** *	CFD complement factor D (adipsin) *	**CST3 cystatin C** *
2	**MPO myeloperoxidase** *	**CST3 cystatin C** *	CFD complement factor D (adipsin) *
3	**IL8 interleukin 8**	**MPO myeloperoxidase** *	**MPO myeloperoxidase** *
4	GYPB glycophorin B (MNS blood group)	**DNTT deoxynucleotidyltransferase, terminal** *	**IL8 interleukin 8** *
5	**IGL immunoglobulin lambda locus**	TCL1A T-cell leukemia/lymphoma 1A *	**DNTT deoxynucleotidyltransferase, terminal** *
6	**DNTT deoxynucleotidyltransferase, terminal**	**IGL immunoglobulin lambda locus** *	TCL1A T-cell leukemia/lymphoma 1A *
7	LOC100437488 interleukin-8-like	**IL8 interleukin 8** *	**IGL immunoglobulin lambda locus**
8	**LTB lymphotoxin beta (TNF superfamily, member 3)**	ZYX zyxin *	**LTB lymphotoxin beta (TNF superfamily, member 3)**
9	TCRB T cell receptor beta cluster	**LTB lymphotoxin beta (TNF superfamily, member 3)** *	CD79A CD79a molecule, immunoglobulin-associated alpha
10	S100A9 S100 calcium binding protein A9	CD79A CD79a molecule, immunoglobulin-associated alpha	HBB hemoglobin, beta

The genes with star(*) are the most frequently selected genes to construct the classifiers according to the last column of Table 6, and the common genes obtained by $L_{1/2}$, L_{EN} , L_1 classifiers are emphasized with bold.

Table 8 The 10 top-ranked informative genes found by the three sparse logistic regression methods from the Prostate dataset

Rank	Gene description		
	$L_{1/2}$	L_{EN}	L_1
1	**SLC43A3 solute carrier family 43, member 3** *	**AMOTL2 angiomotin like 2** *	**USP4 ubiquitin specific peptidase 4 (proto-oncogene)** *
2	**CD22 CD22 molecule** *	**USP4 ubiquitin specific peptidase 4 (proto-oncogene)** *	**CD22 CD22 molecule** *
3	KHDRBS1 KH domain containing, RNA binding, signal transduction associated 1 *	**EIF4EBP2 eukaryotic translation initiation factor 4E binding protein 2** *	**EIF4EBP2 eukaryotic translation initiation factor 4E binding protein 2** *
4	ZNF787 zinc finger protein 787 *	PRAF2 PRA1 domain family, member 2 *	Gene symbol:AA683055, probe set: 34711_at *
5	GMPR guanosine monophosphate reductase *	CACYBP calcyclin binding protein *	**AMOTL2 angiomotin like 2** *
6	**AMOTL2 angiomotin like 2**	Gene symbol:AA683055, probe set: 34711_at *	VSNL1 visinin-like 1 *
7	**EIF4EBP2 eukaryotic translation initiation factor 4E binding protein 2**	VSNL1 visinin-like 1 *	FLNC filamin C, gamma *
8	USP2 ubiquitin specific peptidase 2	**SLC43A3 solute carrier family 43, member 3** *	PRAF2 PRA1 domain family, member 2 *
9	**USP4 ubiquitin specific peptidase 4 (proto-oncogene)**	**CD22 CD22 molecule** *	CACYBP calcyclin binding protein *
10	ACTN4 actinin, alpha 4	TMCO1 transmembrane and coiled-coil domains 1 *	**SLC43A3 solute carrier family 43, member 3** *

The genes with star(*) are the most frequently selected genes to construct the classifiers according to the last column of Table 6, and the common genes obtained by $L_{1/2}$, L_{EN} , L_1 classifiers are emphasized with bold.

Construct KNN classifier with the most frequently selected relevant genes

In this section, to further evaluate the performance and prediction generality of the sparse logistic regression with $L_{1/2}$ penalty, we constructed KNN (k =3, 5) classifiers using the relevant genes which were most frequently selected by the $L_{1/2}$ penalized logistic regression method. In this experiment, we use the random leave-one-out cross validation (LOOCV) to evaluate the predictive ability and repeat 50 runs.

Table 9 The 10 top-ranked informative genes found by the three sparse logistic regression methods from the colon dataset

Rank	Gene description		
	$L_{1/2}$	L_{EN}	L_1
1	**GUCA2B guanylate cyclase activator 2B (uroguanylin)** *	**GUCA2B guanylate cyclase activator 2B (uroguanylin)** *	**GUCA2B guanylate cyclase activator 2B (uroguanylin)** *
2	**MYL6 myosin, light chain 6, alkali, smooth muscle and non-muscle** *	**MYH9 myosin, heavy chain 9, non-muscle** *	**ATPsyn-Cf6 ATP synthase-coupling factor 6, mitochondrial** *
3	**DES desmin** *	**DES desmin** *	**MYH9 myosin, heavy chain 9, non-muscle** *
4	CHRND cholinergic receptor, nicotinic, delta polypeptide *	**MYL6 myosin, light chain 6, alkali, smooth muscle and non-muscle** *	GSN gelsolin *
5	PECAM1 platelet/endothelial cell adhesion molecule-1 *	GSN gelsolin *	**MYL6 myosin, light chain 6, alkali, smooth muscle and non-muscle** *
6	**ATPsyn-Cf6 ATP synthase-coupling factor 6, mitochondrial**	COL11A2 collagen, type XI, alpha 2 *	COL11A2 collagen, type XI, alpha 2 *
7	ATF7 activating transcription factor 7	**ATPsyn-Cf6 ATP synthase-coupling factor 6, mitochondrial** *	MXI1 MAX interactor 1, dimerization protein *
8	PROBABLE NUCLEAR ANTIGEN (Pseudorabies virus)[accession number:T86444]	ssb single-strand binding protein *	UQCRC1 ubiquinol-cytochrome c reductase core protein I *
9	**MYH9 myosin, heavy chain 9, non-muscle**	Sept2 septin 2 *	**DES desmin** *
10	MYH10 myosin, heavy chain 10, non-muscle	MXI1 MAX interactor 1, dimerization protein *	ZEB1 zinc finger E-box binding homeobox 1*

The genes with star(*) are the most frequently selected genes to construct the classifiers according to the last column of Table 6, and the common genes obtained by $L_{1/2}$, L_{EN} , L_1 classifiers are emphasized with bold.

Table 10 The 10 top-ranked informative genes found by the three sparse logistic regression methods from the DLBCL dataset

Rank	Gene description		
	$L_{1/2}$	L_{EN}	L_1
1	**CCL21 chemokine (C-C motif) ligand 21** *	MTH1 metallothionein 1H *	MTH1 metallothionein 1H *
2	HLA-DQB1 major histocompatibility complex, class II, DQ beta 1 *	**MT2A metallothionein 2A** *	**MT2A metallothionein 2A** *
3	**MT2A metallothionein 2A** *	**SFTPA1 surfacant protein A1** *	**CCL21 chemokine (C-C motif) ligand 2** *
4	THRSP thyroid hormone responsive *	TCL1A T-cell leukemia/lymphoma 1A *	**SFTPA1 surfacant protein A1** *
5	**Igj immunoglobulin joining chain** *	ZFP36L2 ZFP36 ring finger protein-like 2 *	POLD2 polymerase (DNA directed), delta 2, accessory subunit *
6	TCL1A T-cell leukemia/lymphoma 1A *	FCGR1A Fc fragment of IgG, high affinity Ia, receptor (CD64) *	**Igj immunoglobulin joining chain** *
7	GOT2 glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2) *	**Igj immunoglobulin joining chain** *	MELK maternal embryonic leucine zipper kinase *
8	Plod procollagen lysyl hydroxylase *	TRB2 Homeodomain-like/winged-helix DNA-binding family protein *	CKS2 CDC28 protein kinase regulatory subunit 2 *
9	STXBP2 syntaxin binding protein 2 *	MELK maternal embryonic leucine zipper kinase *	EIF2A eukaryotic translation initiation factor 2A, 65kDa *
10	**SFTPA1 surfacant protein A1** *	**CCL21 chemokine (C-C motif) ligand 2** *	AQP4 aquaporin 4 *

The genes with star(*) are the most frequently selected genes to construct the classifiers according to the last column of Table 6, and the common genes obtained by $L_{1/2}$, L_{EN} , L_1 classifiers are emphasized with bold.

Table 11 summarizes classification accuracies of four datasets with KNN classifiers with selected genes by our proposed methods. From Table 11, we can see that all the classification accuracies are high than 90%, especially the classification accuracy on the Leukaemia dataset is 98.3%. The KNN classifiers with relevant genes which were selected by the sparse logistic regression with the $L_{1/2}$ penalty can achieve high classification accuracy. The results indicate that the sparse logistic regression with the $L_{1/2}$ penalty can select power discrimination genes.

Conclusions

In cancer classification application based on microarray data, only a small subset of genes is strongly indicative of a targeted disease. Thus, feature selection methods play an important role in cancer classification. In this paper, we propose and model sparse logistic regression with the $L_{1/2}$ penalty, and develop the corresponding coordinate descent algorithm as a novel gene selection approach. The proposed method utilizes a novel univariate half thresholding to update the estimated coefficients.

Both simulation and microarray data studies show that the sparse logistic regression with the $L_{1/2}$ penalty achieve higher classification accuracy than those of ordinary L_1 and elastic net regularization approaches, while fewer but informative genes are selected. Therefore, the sparse logistic regression with the $L_{1/2}$ penalty is the effective technique for gene selection in real classification problem.

In this paper, we use the proposed method to solve binary cancer classification problem. However, many cancer classification problems involve multi-category microarray data. We plan to extend our proposed method to solve multinomial penalized logistic regression for multiclass cancer classification in our future work.

Table 11 Summary of the results of KNN classifiers using the most frequently selected genes by our proposed $L_{1/2}$ penalized logistic regression method

Methods	K-NN(k=3)	K-NN(k=5)
Leukaemia	98.3%	94.4%
Prostate	95.1%	94.2%
Colon	95.1%	90.6%
DLBCL	94.8%	91.2%

Competing interests

All authors declare that they have no competing interests.

Authors' contributions

YL, CL and XZL developed the gene selection methodology, designed and carried out the comparative study, wrote the code, and drafted the manuscript. KSL, TMC, ZBX and HZ brought up the biological problem that prompted the methodological development and verified and provided discussion on the methodology, and co-authored the manuscript. The authors read and approved the manuscript.

Acknowledgements
This research was supported by Macau Science and Technology Develop Funds (Grant No. 017/2010/A2) of Macau SAR of China and the National Natural Science Foundations of China (Grant No. 2013CB329404, 11131006, 61075054, and 11171272).

Author details
[1]Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau, China. [2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China. [3]Faculty of Science, Xi'an Jiaotong University, Xian, China.

Received: 4 July 2012 Accepted: 30 May 2013
Published: 19 June 2013

References

1. Dudoit S, Fridlyand S, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *J Am Stat Assoc* 2002, **97**(457):77–87.
2. Li T, Zhang C, Ogihara M: **A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.** *Bioinformatics* 2004, **20:**2429–2437.
3. Lee JW, Lee JB, Park M, Song SH: **An extensive evaluation of recent classification tools applied to microarray data.** *Com Stat Data Anal* 2005, **48:**869–885.
4. Ding C, Peng H: **Minimum redundancy feature selection from microarray gene expression data.** *J. Bioinform. Comput* 2005, **3**(2):185–205.
5. Monari G, Dreyfus G: **Withdrawing an example from the training set: an analytic estimation of its effect on a nonlinear parameterized model.** *Neurocomputing Letters* 2000, **35:**195–201.
6. Rivals I, Personnaz L: **MLPs (mono-layer polynomials and multi-layer perceptrons) for nonlinear modeling.** *J Mach Learning Res* 2003, **3:**1383–1398.
7. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531–537.
8. Guyon I, Elisseff A: **An Introduction to variable and feature selection.** *J Mach Learning Res* 2003, **3:**1157–1182.
9. Shevade SK, Keerthi SS: **A simple and efficient algorithm for gene selection using sparse logistic regression.** *Bioinformatics* 2003, **19:**2246–2253.
10. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J. R. Statist. Soc. B* 1996, **58:**267–288.
11. Fiedman J, Hastie T, Hofling H, Tibshirani R: **Path wise coordinate optimization.** *Ann. Appl. Statist.* 2007, **1:**302–332.
12. Fiedman J, Hastie T, Hofling H, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *J. Statist. Softw.* 2010, **33:**1–22.
13. Gavin CC, Talbot LC: **Gene selection in cancer classification using sparse logistic regression with Bayesian regularization.** *Bioinformatics* 2006, **22:**2348–2355.
14. Xu ZB, Zhang H, Wang Y, Chang XY, Liang Y: **$L_{1/2}$ regularization.** *Sci China Series F* 2010, **40**(3):1–11.
15. Fan J, Li R: **Variable selection via nonconcave penalized likelihood and its oracle properties.** *J. Amer. Statist. Assoc.* 2001, **96:**1348–1361.
16. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J Royal Stat Soc Series B* 2005, **67**(2):301–320.
17. Zhang CH: **Nearly unbiased variable selection under minimax concave penalty.** *Ann. Statist.* 2010, **38:**894–942.
18. Xu ZB, Chang XY, Xu FM, Zhang H: **$L_{1/2}$ Regularization: a thresholding representation theory and a fast solver.** *IEEE Transact Neural Networks Learn Syst* 2012, **23**(7):1013–1027.
19. Sohn I, Kim J, Jung SH, Park C: **Gradient lasso for Cox proportional hazards model.** *Bioinformatics* 2009, **25**(14):1775–1781.
20. Yang K, Cai ZP, Li JZ, Lin GH: **A stable gene selection in microarray data analysis.** *BMC Bioinformatics* 2006, **7:**228.
21. Alon U, Barkai N, Notterman D, Gish K, Ybarra S, Mack D, Levine A: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Nat Acad Sci USA* 1999, **96**(12):6745–6750.
22. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Amgel M, Reich M, Pinkus GS, Ray TS, Kovall MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: **Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning.** *Nat Med* 2002, **8:**68–74.
23. Nagai A, Terashima M, Harada T, Shimode K, Takeuchi H, Murakawa Y, *et al*: **Cathepsin B and H activities and cystatin C concentrations in cerebrospinal fluid from patients with leptomeningeal metastasis.** *Clin Chim Acta* 2003, **329:**53–60.
24. Moroz C, Traub L, Maymon R, Zahalka MA: **A novel human ferritin subunit from placenta with immunosuppressive activity.** *J Biol Chem* 2002, **277:**12901–12905.
25. Ben-Dor A, *et al*: **Tissue classification with gene expression profiles.** *J Comput Biol* 2000, **7:**559–583.
26. Yang AJ, Song XY: **Bayesian variable selection for disease classification using gene expression data.** *Bioinformatics* 2010, **26:**215–222.
27. Li HD, Xu QS, Liang YZ: **Random frog: an efficient reversible jump Markov chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification.** *Anal Chim Acta* 2012, **740:**20–26.
28. Notterman DA, Alon U, Sierk AJ, Levine AJ: **Minimax probability machine. Advances in neural processing systems.** *Cancer Res* 2001, **61:**3124–3130.
29. Shailubhai K, Yu H, Karunanandaa K, Wang J, Eber S, Wang Y, Joo N, Kim H, Miedema B, Abbas S, Boddupalli S, Currie M, Forte L: **Uroguanylin treatment suppeesses polyp formation in the Apc(Min/+) mouse and indices apoptosis in human colon adenocarcinoma cells via cyclic GMP.** *Cancer Res* 2000, **60:**5151–5157.
30. Maglietta R, Addabbo A, Piepoli A, Perri F, Liuni S, Pesole G, Ancona N: **Selection of relevant genes in cancer diagnosis based on their prediction accuracy.** *Art Intell Med* 2007, **40:**29–44.
31. Wiese AH J, Lassmann S, Nahrig J, Rosenberg R, Hofler H, Ruger R, Werner M: **Identification of gene signatures for invasive colorectal tumor cells.** *Cancer Detect Prev* 2007, **31:**282–295.
32. Wang SL, Li XL, Zhang SW, Gui J, Huang DS: **Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction.** *Comp Biol Med* 2010, **40:**179–189.
33. Dai JH, Xu Q: **Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification.** *App Soft Comp* 2013, **13:**211–221.

doi:10.1186/1471-2105-14-198
Cite this article as: Liang *et al.*: Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. *BMC Bioinformatics* 2013 **14**:198.