applied sciences

MDPI

Article

Gene Selection in Cancer Classification Using Sparse Logistic Regression with $L_{1/2}$ Regularization

Shengbing Wu *, Hongkun Jiang, Haiwei Shen and Ziyi Yang

Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China; jiang.hk.xm@gmail.com (H.J.); yykjthaiwei@buu.edu.cn (H.S.); yangziyi091100@163.com (Z.Y.)
* Correspondence: shengbing.wu@163.com

Received: 8 August 2018; Accepted: 4 September 2018; Published: 6 September 2018

Abstract: In recent years, gene selection for cancer classification based on the expression of a small number of gene biomarkers has been the subject of much research in genetics and molecular biology. The successful identification of gene biomarkers will help in the classification of different types of cancer and improve the prediction accuracy. Recently, regularized logistic regression using the L_1 regularization has been successfully applied in high-dimensional cancer classification to tackle both the estimation of gene coefficients and the simultaneous performance of gene selection. However, the L_1 has a biased gene selection and dose not have the oracle property. To address these problems, we investigate $L_{1/2}$ regularized logistic regression for gene selection in cancer classification. Experimental results on three DNA microarray datasets demonstrate that our proposed method outperforms other commonly used sparse methods (L_1 and L_{EN}) in terms of classification performance.

Keywords: gene selection; cancer classification; regularized logistic regression; $L_{1/2}$ regularization

1. Introduction

With the development of DNA microarray technology, biological researchers can pay more attention to simultaneously studying the expression levels of thousands of genes [1,2]. Cancer classification based on gene expression levels is one of the most active topics in genome research, which is appropriate for gene expression levels in different situations (e.g., normal and abnormal) [3,4]. However, cancer classification using DNA microarray data is a challenge because of the data's high dimension and small sample size [5]. Generally, the number of genes ranges in the thousands from a hundred or fewer tissue samples, and so gene selection has recently emerged as important technology for cancer classification [6]. Gene selection is applied because only a small subset of genes is strongly indicative of a targeted disease. From the biological perspective, effective gene selection methods can be desirable to help to classify different types of cancer and improve the accuracy of prediction [7–9].

Many gene selection methods have been proposed for selection of the subset of meaningful and important genes that can achieve high cancer classification performance. Recently, there has been growing interest in applying regularization techniques in gene selection. Regularization methods are an important embedded technique [10–13]. From the statistical perspective, regularization methods can prevent over-fitting. Many statistical methods have been successfully applied to cancer classification. Among them, logistic regression [14–17] is a powerful discriminative method, and has a direct probabilistic interpretation that can obtain classification probabilities apart from the class label information. However, logistic regression is not suitable for solving the high-dimensional and small sample size problem because the design matrix is singular. Thus, Newton–Raphson's method cannot work. Regularized logistic regression has been successfully applied in cancer classification in order to be suitable for high dimension and small sample size [7,8]. The advantages of regularized logistic regression can improve the classification accuracy by shrinking the regression coefficients and selecting a small subset of

genes. Different regularization terms are applied to regularized logistic regression. The widely popular regularization term is L_1 penalty, which is the least absolute shrinkage and selection operator (lasso) [18]. Meanwhile, there are various of versions of L_1, such as smoothly clipped absolute deviation (SCAD) [19], maximum concave penalty (MCP) [20], group lasso [21], and so on. The L_1 regularization can assign some genes' coefficients to zero for variable selection. Thus, the L_1 regularization has been widely applied to data with high dimension and small sample size.

Although a well-known regularization method is the L_1 penalty, it has some limitations [22]. The L_1 regularization does not have oracle property [19], which means the aim-listed probability of selecting the right set of genes (with nonzero coefficients) converges to one, and the estimators of the nonzero coefficients have asymptotically normal distribution with the same means and covariances as if the zero coefficients were known in the prior. Besides, there is grouping among genes in DNA microarray data. Related to this limitation, concerning the grouping property, Zhou and Hastie proposed the elastic net penalty (L_{EN}) [23], which is a linear combination of L_1 and L_2 penalties. In addition, L_1 regularization is not sparser. To overcome this limitation, Xu et al. proposed the $L_{1/2}$ penalty—a method that can be taken as a representative of L_q $(0 < q < 1)$ penalty in both sparsity and computational efficiency, and has demonstrated many attractive properties, such as unbiasedness and oracle properties [24–26]. Therefore, we investigated $L_{1/2}$ regularized logistic regression for gene selection in cancer classification. The approach is suitable for DNA data with high dimension and small sample size. To evaluate the effectiveness of the approach, three public datasets were applied to cancer classification. Additionally, we compared other commonly used sparse methods (L_1 and L_{EN}) to our methods.

Our research can be summarized as follows are given as follows:

- identification of gene biomarkers will help to classify different types of cancer and improve the prediction accuracy.
- The $L_{1/2}$ penalized logistic regression is used as a gene selection method for cancer classification to overcome the over-fitting problem with high-dimensional data and small sample size.
- Experimental results on three GEO lung cancer datasets corroborate our ideas and demonstrate the correctness and effectiveness of $L_{1/2}$ penalized logistic regression.

2. Methods

2.1. Regularized Logistic Regression

In this paper, we only consider a general binary classification problem and get a predictor vector X and a response variable y, which consists of genes and corresponding tissue samples, respectively. Suppose we have n samples, $D = (X_1, y_1), (X_2, y_2), ..., (X_n, y_n)$, where $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is ith input pattern with dimensionality p, which means the X_i has p descriptors and x_{ij} denotes the value of gene j for the ith sample. y_i is a corresponding variable that takes a value of 0 or 1. Define a classifier $f(x) = e^x/(1 + e^x)$, and the logistic regression is shown as follows:

$$P(y_i = 1|X_i) = f(X_i^{'}\beta) = \frac{exp(X_i^{'}\beta)}{1 + exp(X_i^{'}\beta)}. \quad (1)$$

Additionally, the log-likelihood can be expressed as follows:

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

We can get the value of vector β from Equation (2). However, solving Equation (2) can result in over-fitting with data of high dimension and small sample size. Therefore, in order to address the problem, we add the regularization terms to Equation (2):

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

where $l(\beta)$ and $p(\beta)$ are loss function and penalty function, respectively, and $\lambda > 0$ is a tuning parameter. Note that $p(\beta) = \sum |\beta|^q$. When q is equal to 1, the L_1 has been proposed. Moreover, there are various of versions of L_1, such as SCAD, MCP, group lasso, and so on. We add the L_1 regularization to Equation (2). The formula is expressed as follows:

$$\beta = argmin\{l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|\}. \tag{4}$$

From a biologist's point of view, there is a grouping property among genes, which is a limitation of L_1 regularization. To overcome this limitation, Zou et al. proposed the elastic net (L_{EN}) regularization method for gene selection. The L_{EN} regularization tries to combine L_1 with L_2 in order to search for highly correlated genes and perform gene selection simultaneously. The regularized logistic regression using L_{EN} is exhibited as follows:

$$\beta = argmin\{l(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2\}. \tag{5}$$

As we observe from Equation (5), λ_1 and λ_2 control the sparsity and group effect, respectively. The coefficient β depends on two non-negative tuning parameters λ_1 and λ_2. In order to simplify Equation (5), let λ_1 plus λ_2 equal to 1. Thus, we can rewrite Equation (5) as:

$$\beta = argmin\{l(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + (1 - \lambda_1) \sum_{j=1}^{p} |\beta_j|^2\}. \tag{6}$$

2.2. $L_{1/2}$ Regularized Logistic Regression

Despite the advantages of L_1 and L_{EN}, there are some limitations. L_1 and L_{EN} have a biased gene selection, and they do not have an oracle property. Besides, theoretically, the L_q-type regularization $p(\beta) = \sum |\beta|^q$ with the lower value of q would lead to better solutions with more sparsity. However, difficulties with convergence arise when q is very close to zero. Therefore, Xu et al. proposed $L_{1/2}$ regularization. When $\frac{1}{2} < q < 1$, comparing with L_1, the convergence of $L_{1/2}$ regularization is not high, while when $0 < q < \frac{1}{2}$, comparing with L_0, solving the $L_{1/2}$ regularization is much simpler. Thus, the $L_{1/2}$ regularization can be taken as a representative of $L_q(0 < q < 1)$ regularization. The $L_{1/2}$ regularized logistic regression is as follows:

$$\beta = argmin\{l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^{\frac{1}{2}}\}, \tag{7}$$

where the value of β can be obtained by calculating Equation (7).

In this paper, we apply the coordinate descent algorithm to solve Equation (7). The algorithm is a "one-at-a-time" algorithm and solves β_j, and other $\beta_{j \neq k}$ (representing the parameters remaining after the jth element is removed) are fixed [7,8]. Suppose that we have n samples, $D = (X_1, y_1), (X_2, y_2), ..., (X_n, y_n)$, where $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is the ith input pattern with dimensionality p, which means the X_i has p genes and x_{ij} denotes the value of genes j for the ith sample. y_i is a corresponding variable that takes a value of 0 or 1. $y_i = 0$ indicates that the ith sample is in Class 1 and $y_i = 1$ indicates that the ith

sample is in Class 2. Inspired by Friedman et al. [27], Xu et al. [26], and Xia et al. [28], the univariate half thresholding operator for a $L_{1/2}$-penalized logistic regression coefficient is as follows:

$$\beta_j = Half(w_j, \lambda) = \begin{cases} \frac{2}{3} w_j (1 + \cos \frac{2(\pi - \phi_\lambda(w_j))}{3}) & \text{if } |w_j| > \frac{3}{4}(\lambda)^{\frac{2}{3}}, \\ 0 & \text{if } otherwise. \end{cases} \tag{8}$$

Besides, the univariate thresholding operator of the coordinate descent algorithm for the L_{EN} regularization can be defined as:

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

where $S(w_j, \lambda a)$ is a soft thresholding operator for the L_1 if a is equal to 1, as follows:

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

Inspired by Reference [7], Equation (7) is linearized by one-term Taylor series expansion:

$$L(\beta, \lambda) \approx \frac{1}{2n} \sum_{i=1}^{n} (Z_i - X_i\beta)' W_i (Z_i - X_i\beta) + \lambda \sum_{j=1}^{n} |\beta|^{\frac{1}{2}}, \tag{11}$$

where $Z_i = X_i\widetilde{\beta} + \frac{Y_i - f(X_i\widetilde{\beta})}{f(X_i\widetilde{\beta})(1 - f(X_i\widetilde{\beta}))}$, $W_i = f(X_i\widetilde{\beta})(1 - f(X_i\widetilde{\beta}))$, and $f(X_i\widetilde{\beta}) = \frac{exp(X_i\widetilde{\beta})}{(1 + exp(X_i\widetilde{\beta}))}$. Redefine the partial residual for fitting $\widetilde{\beta}_j$ as $\widetilde{Z}_i^{(j)} = \sum_{i=1}^{n} W_i(\widetilde{Z}_i - \sum_{k \neq j} x_{ik}\widetilde{\beta}_k)$ and $\sum_{i=1}^{n} x_{ij}(Z_i - \widetilde{Z}_i^{(j)})$. A pseudocode of coordinate descent algorithm for $L_{1/2}$ penalized logistic regression is described in Algorithm 1 [7].

Algorithm 1: A coordinate descent algorithm for $L_{1/2}$ penalized logistic regression.

Input: X, y, and λ are chosen by 5-fold cross-validation
Output: β
while $\beta(m)$ *does not change* **do**
  Initialize all $\beta_j(m) = 0 (j = 1, 2, 3, ..., p)$, set $m = 0$
  Calculate $Z(m)$ and $W(m)$ and the loss function Equation (11) based on $\beta(m)$
  Update each $\beta_j(m)$ and cycle $j = 1, 2, 3, ..., p$
    $\widetilde{z}_i^{(j)}(m) \leftarrow \sum_{k \neq j} x_{ik}\beta_k(m)$
    and $w_j(m) \leftarrow w_j(m) x_{ij}(Z_i(m) - \widetilde{Z}_i^{(j)}(m))$
  Update $\beta_j(m)$ by Equation (8)
  Let $m \leftarrow (m + 1)$, $\beta(m + 1) \leftarrow \beta(m)$
end

2.3. Classification Evaluation Criteria

In order to evaluate the cancer classification performance of the proposed method, accuracy, sensitivity, and specificity were applied to three public DNA microarray data. The formulas of accuracy, sensitivity, and specificity are shown as follows [29]:

$$Sensitivity = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP}, \quad Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP refers to true positives, TN refers to true negatives, FP refers to false positives, and FN refers to false negatives.

3. Datasets

In this section, three public QSAR datasets were obtained online, including GSE10072 [30], GSE19804 [31], and GSE4115 [32]. A brief description of these datasets is given in Table 1.

Table 1. Three publicly available cancer datasets used in the experiments

Datasets	No. of Samples	No. of Genes	Class
GSE10072	107	22283	Normal/Tumor
GSE19804	120	54675	Normal/Tumor
GSE4115	187	22215	Normal/Tumor

3.1. GSE10072

The dataset is provided by the National Cancer Institute (NIH). There are 107 samples, of which 58 are lung tumor, and the other 49 are normal lung. Each sample contained 22,283 genes.

3.2. GSE19804

We obtained this dataset online. For data preprocessing, we utilized 120 samples, which consisted of 60 lung cancer and 60 lung normal samples, with 54,675 genes for the model as input.

3.3. GSE4115

This cancer dataset is from the Boston University Medical Center. After preprocessing, the number of lung cancer and normal lung samples was 97 and 90, respectively. Each sample contained 22,215 descriptors.

4. Results

In this section, two methods are compared to our proposed method, including L_{EN} and L_1. To evaluate the prediction accuracy of the three logistic regression models, we first used random partition to divide the samples. That is to say, the samples were divided into training samples (70%) and testing samples (30%). The detailed information of the three publicly available datasets used in the experiments are shown in Table 2. Secondly, in order to obtain the tuning parameter λ, we applied 5-fold cross validation to the training set. Thirdly, the classification evaluation criteria were the corresponding average number at 50 runs.

Table 2. Detailed information of the three publicly available datasets used in the experiments.

Datasets	No. of Training (Class 1/Class 2)	No. of Testing (Class 1/Class 2)
GSE10072	75 (35 Normal/40 Tumor)	32 (14 Normal/18 Tumor)
GSE19804	84 (46 Normal/38 Tumor)	36 (14 Normal/22 Tumor)
GSE4115	131 (67 Normal/64 Tumor)	56 (31 Normal/25 Tumor)

Table 3 shows that the results of the training set and testing set were obtained by L_1, L_{EN}, and $L_{1/2}$. The results obtained by $L_{1/2}$ were better those of L_1 and L_{EN}. For example, for the training set in the dataset GSE10072, the values of sensitivity, specificity, and accuracy of $L_{1/2}$ were the same as for L_1. Besides, the values of sensitivity and accuracy of L_{EN} were 0.98, and 0.99 lower than those of $L_{1/2}$. For the testing set in dataset GSE4115, $L_{1/2}$ and L_{EN} ranked first and second, respectively. L_1 was the last. For instance, the value of accuracy of $L_{1/2}$ was 0.80, higher than the 0.77 and 0.78 of L_1 and L_{EN}, respectively. Moreover, $L_{1/2}$ was more sparse than L_1 and L_{EN}. As shown in Figure 1, In dataset GSE17084, the number of selected genes of $L_{1/2}$ was 8, lower than the respective 33 and 82 of L_1 and L_{EN}. In a word, $L_{1/2}$ was superior to L_1 and L_{EN}.

Table 3. Mean results of empirical datasets. The results of our proposed method are given in bold.

Methods	Datasets	Training Set (5-CV)			Testing Set		
		Sensitivity	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy**
L_1	GSE10072	1.00	1.00	1.00	0.92	0.98	0.95
	GSE19084	1.00	0.98	0.99	0.87	0.72	0.81
	GSE4115	0.83	0.97	0.91	0.77	0.74	0.73
	Mean	0.94	0.98	0.97	0.85	0.81	0.83
L_{EN}	GSE10072	0.98	1.00	0.99	0.93	0.94	0.94
	GSE19084	1.00	0.98	0.99	0.90	0.68	0.81
	GSE4115	0.94	0.98	0.96	0.78	0.85	0.78
	Mean	0.97	0.99	0.98	0.87	0.82	0.84
$L_{1/2}$	GSE10072	**1.00**	**1.00**	**1.00**	**0.94**	**1.00**	**0.97**
	GSE19084	**1.00**	**1.00**	**1.00**	**0.92**	**0.75**	**0.87**
	GSE4115	**0.98**	**0.99**	**0.98**	**0.78**	**0.93**	**0.83**
	Mean	**0.99**	**1.00**	**0.99**	**0.88**	**0.89**	**0.89**

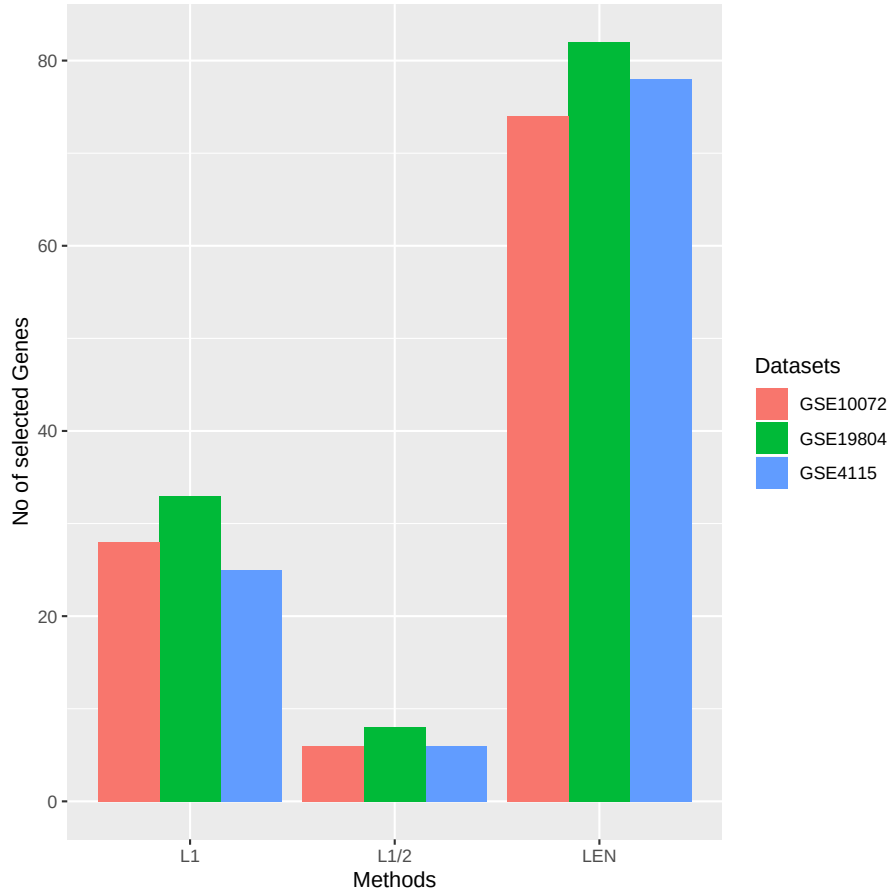
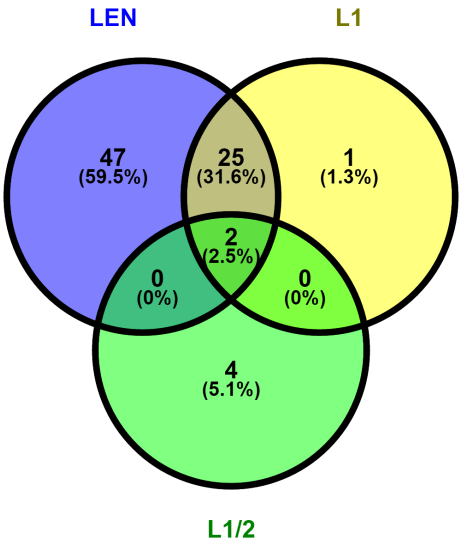
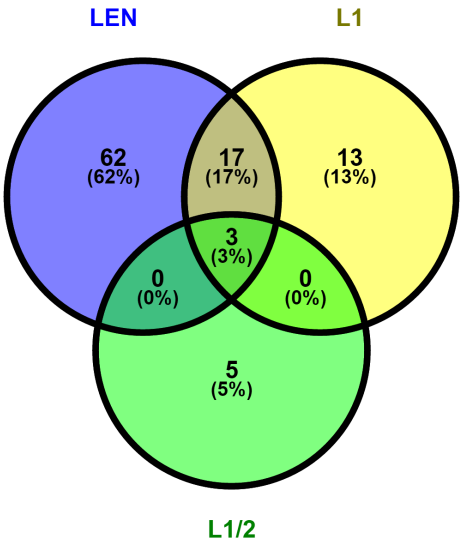


Figure 1. The number of genes selected by L_1, L_{EN}, and $L_{1/2}$.

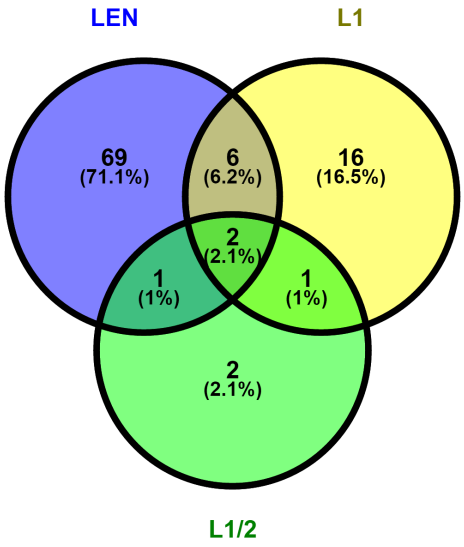
In order to search the common gene signatures selected by the different methods, we used VENNY software (2.1.0 Centro Nacional de Biotecnología, Madrid, Spain, 2015) [33] to generate Venn diagrams. As shown in Figure 2, we considered the common gene signatures selected by the logistic regression model with L_1, L_{EN}, and $L_{1/2}$ regularization methods, which are the most relevant signatures of lung cancer. Hence, 2, 3, and 2 common genes were found in these methods for different datasets.



(**a**) Dataset: GSE10072



(**b**) Dataset: GSE19804



(**c**) Dataset: GSE4115

Figure 2. Venn diagram analysis of the results of L_1, L_{EN}, and $L_{1/2}$.

Table 4 shows that the genes were selected by $L_{1/2}$. At the beginning of the experiments, the attribute of genes was prob set ID. Thus, we could transform prob set ID to gene symbol by using the software DAVID 6.8 [34]. The data distribution for the selected genes is displayed in Figures 3–5. From inspecting the figures, we can find that some genes facilitated the classification of lung tumor and normal lung, such as FAM107A, KDELR2, AASS, and SFRP1 for dataset GSE10072; and SOCS2 and EHD2 for dataset GSE19804. In addition, we found that a common gene in the three different datasets using $L_{1/2}$ was EGFR [35,36]. However, due to the distribution of the data of different datasets, we cannot use gene EFGR to classify different types of cancer and improve the prediction accuracy. Furthermore, the literature indicates that never-smokers with adenocarcinoma have the highest incidence of EGFR, HER2, ALK, RET, and ROS1 mutations [37]. Therefore, our proposed $L_{1/2}$ is an effective technique in gene selection and classification.

Table 4. The genes selected by $L_{1/2}$ for different datasets.

Dataset: GSE10072		
Prob_ID	Gene Symbol	Gene Name
209074_s_at	FAM107A	family with sequence similarity 107 member A (FAM107A)
200700_s_at	KDELR2	KDEL endoplasmic reticulum protein retention receptor 2 (KDELR2)
201983_s_at	EGFR	epidermal growth factor receptor (EGFR)
210852_s_at	AASS	aminoadipate-semialdehyde synthase (AASS)
202037_s_at	SFRP1	secreted frizzled related protein 1 (SFRP1)
203295_s_at	ATP1A2	ATPase Na^+/K^+ transporting subunit alpha 2 (ATP1A2)
Dataset: GSE19804		
Prob_ID	Gene Symbol	Gene Name
1555636_at	CD300LG	CD300 molecule like family member g (CD300LG)
206938_at	SRD5A2	steroid 5 alpha-reductase 2 (SRD5A2)
44654_at	G6PC3	glucose-6-phosphatase catalytic subunit 3 (G6PC3)
45297_at	EHD2	EH domain containing 2 (EHD2)
1552696_at	NIPA1	non-imprinted in Prader–Willi/Angelman syndrome 1 (NIPA1)
45687_at	prr14	proline-rich 14 (PRR14)
203373_at	SOCS2	suppressor of cytokine signaling 2 (SOCS2)
210984_x_at	EGFR	epidermal growth factor receptor (EGFR)
Dataset: GSE4115		
Prob_ID	Gene Symbol	Gene Name
205560_at	PCSK5	pro-protein convertase subtilisin/kexin type 5 (PCSK5)
200003_s_at	MIR680	microRNA 6805 (MIR6805)
201983_s_at	EGFR	epidermal growth factor receptor (EGFR)
210187_at	FKBP1A	FK506 binding protein 1A (FKBP1A)
205364_at	ACOX2	acyl-CoA oxidase 2 (ACOX2)
206628_at	SLC5A1	solute carrier family 5 member 1 (SLC5A1)

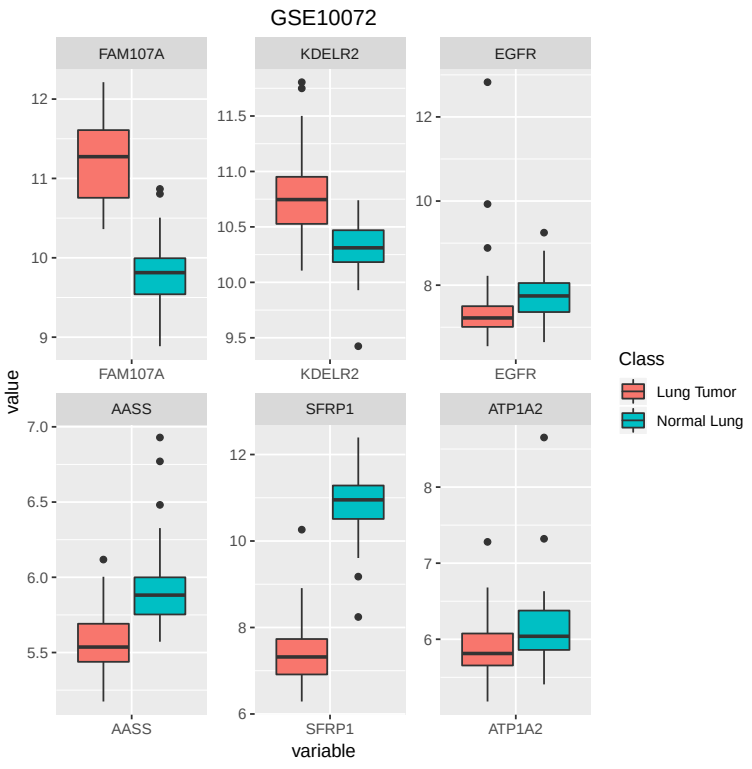


Figure 3. The box plots of selected genes by $L_{1/2}$ for dataset GSE10072.

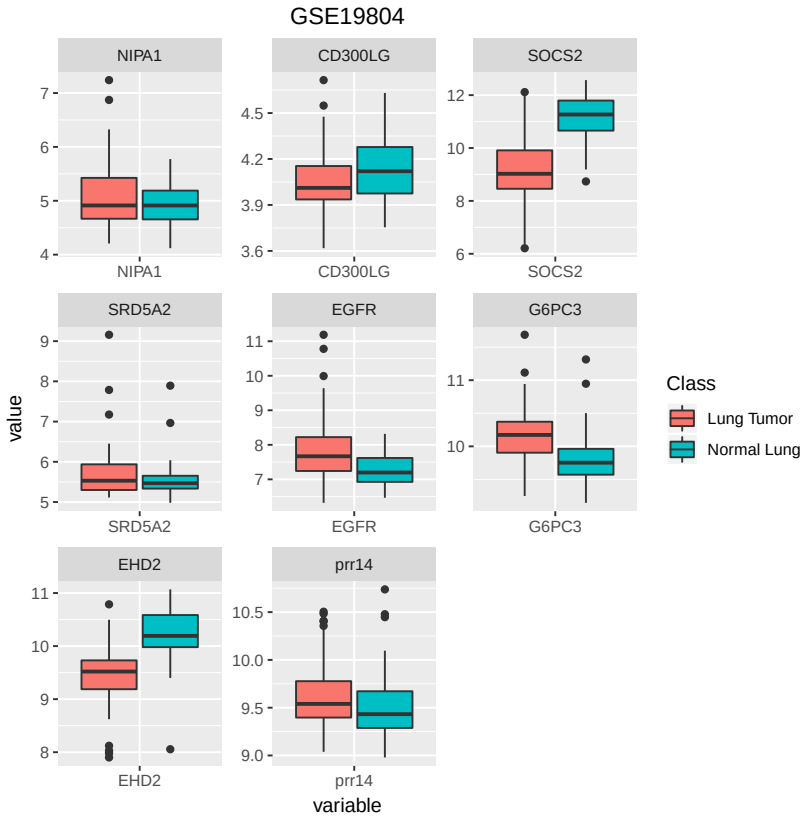


Figure 4. The box plots of selected genes by $L_{1/2}$ for dataset GSE19804.

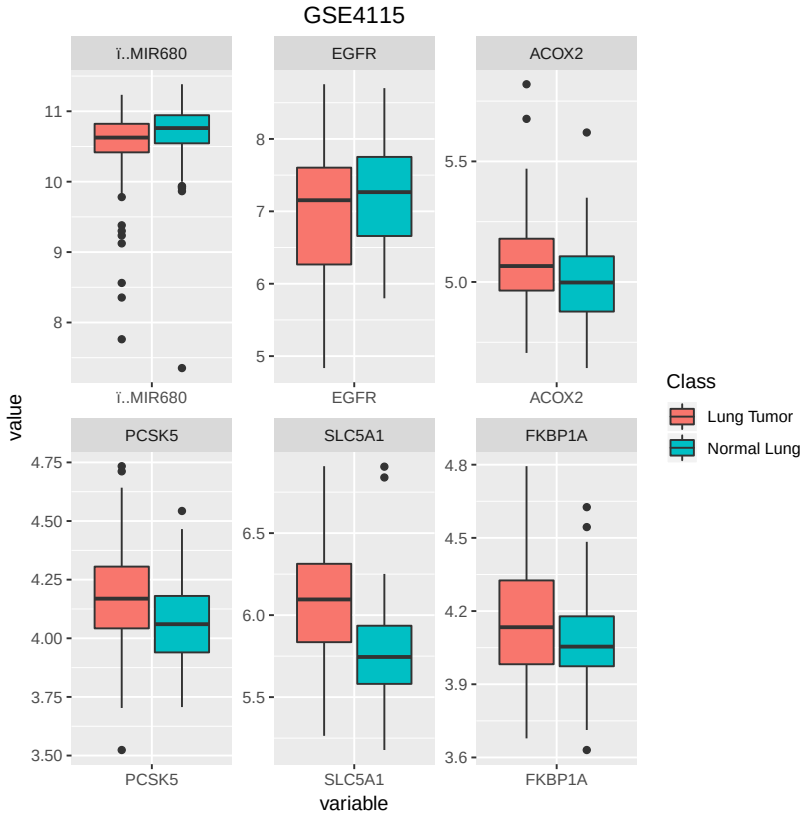


Figure 5. The box plots of selected genes by $L_{1/2}$ for dataset GSE4115.

5. Conclusions

In cancer classification with data of high dimension and small sample size, only a small number of genes strongly suggest specific diseases. Therefore, gene selection is widely popular in cancer classification. Especially, regularization methods have the capacity to select a small subset of meaningful and important genes. In this study, we applied $L_{1/2}$ to a logistic regression model to perform gene selection. Additionally, during the updating of the estimated coefficients, the proposed method utilizes a novel univariate half thresholding.

Experimental results on three cancer datasets demonstrated that our proposed method outperformed the other commonly used sparse methods (L_1 and L_{EN}) in terms of classification performance, while fewer but informative genes were selected—especially the gene EFGR. Therefore, $L_{1/2}$ regularization is a promising tool for feature selection in classification problems.

Author Contributions: S.W. contributed to collecting datasets and analyzing data. S.W. and H.J. designed and implemented the algorithm. H.S. and Z.Y. contributed to the interpretation of the results. S.W. took the lead in writing the manuscript. S.W., H.J., H.S., and Z.Y. revised the manuscript.

Funding: This research received no external funding

Acknowledgments: This work is supported by the Macau Science and Technology Development Funds Grand No. 003/2016/AFJ from the Macau Special Administrative Region of the People's Republic of China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kalina, J. Classification methods for high-dimensional genetic data. *Biocybern. Biomed. Eng.* **2014**, *34*, 10–18. [CrossRef]
2. Kastrin, A.; Peterlin, B. Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. *Expert Syst. Appl.* **2010**, *37*, 5178–5185. [CrossRef]
3. Lotfi, E.; Keshavarz, A. Gene expression microarray classification using PCA–BEL. *Comput. Biol. Med.* **2014**, *54*, 180–187. [CrossRef] [PubMed]
4. Algamal, Z.Y.; Lee, M.H. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Syst. Appl.* **2015**, *42*, 9326–9332. [CrossRef]
5. Chen, S.X.; Qin, Y.L. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Stat.* **2010**, *38*, 808–835. [CrossRef]
6. Yata, K.; Aoshima, M. Intrinsic dimensionality estimation of high-dimension, low sample size data with d-asymptotics. *Commun. Stat. Theory Methods* **2010**, *39*, 1511–1521. [CrossRef]
7. Liang, Y.; Liu, C.; Luan, X.Z.; Leung, K.S.; Chan, T.M.; Xu, Z.B.; Zhang, H. Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. *BMC Bioinform.* **2013**, *14*, 198. [CrossRef] [PubMed]
8. Huang, H.H.; Liu, X.Y.; Liang, Y. Feature selection and cancer classification via sparse logistic regression with the hybrid $L_{1/2+2}$ regularization. *PLoS ONE* **2016**, *11*, e0149675. [CrossRef] [PubMed]
9. Huang, H.H.; Liu, X.Y.; Liang, Y.; Chai, H.; Xia, L.Y. Identification of 13 blood-based gene expression signatures to accurately distinguish tuberculosis from other pulmonary diseases and healthy controls. *Bio-Med. Mater. Eng.* **2015**, *26*, S1837–S1843. [CrossRef] [PubMed]
10. Ma, S.; Huang, J. Regularized gene selection in cancer microarray meta-analysis. *BMC Bioinform.* **2009**, *10*, 1. [CrossRef] [PubMed]
11. Deng, H.; Runger, G. Gene selection with guided regularized random forest. *Pattern Recognit.* **2013**, *46*, 3483–3489. [CrossRef]
12. Allen, G.I. Automatic feature selection via weighted kernels and regularization. *J. Comput. Graph. Stat.* **2013**, *22*, 284–299. [CrossRef]
13. Zou, H.; Yuan, M. Regularized simultaneous model selection in multiple quantiles regression. *Comput. Stat. Data Anal.* **2008**, *52*, 5296–5304. [CrossRef]

14. Harrell, F.E. Ordinal logistic regression. In *Regression Modeling Strategies*; Springer: New York, NY, USA, 2015; pp. 311–325.
15. Menard, S. *Applied Logistic Regression Analysis*; Sage: Newcastle upon Tyne, UK, 2002; Volume 106.
16. Hayes, A.F.; Matthes, J. Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behav. Res. Methods* **2009**, *41*, 924–936. [CrossRef] [PubMed]
17. Wang, Q.; Dai, H.N.; Wu, D.; Xiao, H. Data analysis on video streaming QoE over mobile networks. *EURASIP J. Wirel. Commun. Netw.* **2018**, *2018*, 173. [CrossRef]
18. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288.
19. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
20. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
21. Meier, L.; Van De Geer, S.; Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 53–71. [CrossRef]
22. Feng, Z.Z.; Yang, X.; Subedi, S.; McNicholas, P.D. The LASSO and sparse least squares regression methods for SNP selection in predicting quantitative traits. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **2012**, *9*, 629–636. [CrossRef] [PubMed]
23. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
24. Xu, Z.B.; Guo, H.L.; Wang, Y.; Zhang, H. Representative of $L_{1/2}$ regularization among L_q $(0 < q \leq 1)$ regularizations: An experimental study based on phase diagram. *Acta Autom. Sin.* **2012**, *38*, 1225–1228.
25. Xu, Z.; Zhang, H.; Wang, Y.; Chang, X.; Liang, Y. $L_{1/2}$ regularization. *Sci. China Inf. Sci.* **2010**, *53*, 1159–1169. [CrossRef]
26. Xu, Z.; Chang, X.; Xu, F.; Zhang, H. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1013–1027. [PubMed]
27. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1. [CrossRef] [PubMed]
28. Xia, L.Y.; Wang, Y.W.; Meng, D.Y.; Yao, X.J.; Chai, H.; Liang, Y. Descriptor Selection via Log-Sum Regularization for the Biological Activities of Chemical Structure. *Int. J. Mol. Sci.* **2017**, *19*, 30. [CrossRef] [PubMed]
29. Sohal, H.; Eldridge, S.; Feder, G. The sensitivity and specificity of four questions (HARK) to identify intimate partner violence: A diagnostic accuracy study in general practice. *BMC Fam. Pract.* **2007**, *8*, 49. [CrossRef] [PubMed]
30. Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. Available online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10072 (accessed on 27 December 2017).
31. Genome-Wide Screening of Transcriptional Modulation in Non-Smoking Female Lung Cancer in Taiwan. Available online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19804 (accessed on 3 June 2018).
32. Airway Epithelial Gene Expression Diagnostic for the Evaluation of Smokers with Suspect Lung Cancer. Available online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4115 (accessed on 27 December 2017).
33. Oliveros, J. An Interactive Tool for Comparing Lists with Venn's Diagrams (2007–2015). Available online: http://bioinfogp.cnb.csic.es/tools/venny/index.html (accessed on 21 May 2018).
34. Da Wei Huang, B.T.S.; Stephens, R.; Baseler, M.W.; Lane, H.C.; Lempicki, R.A. DAVID gene ID conversion tool. *Bioinformation* **2008**, *2*, 428. [CrossRef]
35. Rosell, R.; Carcereny, E.; Gervais, R.; Vergnenegre, A.; Massuti, B.; Felip, E.; Palmero, R.; Garcia-Gomez, R.; Pallares, C.; Sanchez, J.M.; et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): A multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* **2012**, *13*, 239–246. [CrossRef]

36. Kobayashi, S.; Boggon, T.J.; Dayaram, T.; Jänne, P.A.; Kocher, O.; Meyerson, M.; Johnson, B.E.; Eck, M.J.; Tenen, D.G.; Halmos, B. EGFR mutation and resistance of non–small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **2005**, *352*, 786–792. [CrossRef] [PubMed]
37. Richards, E. Molecular Profiling of Lung Cancer. Ph.D. Thesis, Imperial College London, London, UK, 2013.