Adjusted Adaptive Sparse Logistic Regression for Gene Selection in High-dimensional Microarray Data Classification

Dr. Zakariya Yahya Algamal[1] Dr. Hassan S. Uraibi[2]

[1](Assist. Prof.) Department of Statistics and Informatics, College of Computer Science and Mathematics, Mosul, University of Mosul, Iraq.

[2](Lecturer) Department of Statistics, College of Administration and Economics, Al-Qadisiyah University, Al Diwaniyah, Iraq.

Abstract

Gene selection in high-dimensional microarray data has become increasingly important in cancer classification. To tackle both estimating the gene coefficients and performing gene selection simultaneously, sparse logistic regression using the least absolute shrinkage and selection operator (LASSO) was successfully applied in high-dimensional microarray data. However, the LASSO has two major limitations. First, it does not encourage grouping effects. Second, it is biased in gene selection. The adaptive LASSO was originally proposed to overcome the selection bias. Similar to the LASSO, the adaptive LASSO does not encourage grouping effects. To address these issues, adjusted adaptive sparse logistic regression (AASLR) is proposed. Extensive applications using high-dimensional gene expression data show that our proposed method has high classification accuracy. Furthermore, it is able to select genes consistently, and, simultaneously, it is effective in selecting highly correlated genes. Thus, we can conclude that AASLR is a reliable sparse logistic regression method in the field of high-dimensional microarray data classification.

Keywords: Adaptive LASSO, microarray data classification, gene selection, grouping effects.

1 Introduction

One of the major advancement made in the field of bioinformatics is the emergence of DNA microarray technology. In cancer research, this technology facilitates the determination of the expression values of thousands of genes simultaneously. The gene expression data is used for various analyses to understand the biological significance of the tissue from which the genes were extracted for the experiment [1,2]. In most applications of microarray technology, the number of genes, p, is greater than the number of patients (tissues), n [3]. Dealing with the situation $p > n$, which is commonly known as high-dimensional data, poses a challenging task in the application of the statistical methods [4]. Overfitting and multicollinearity are the most common problems that arise in high-dimensional data when applying statistical classification methods.

In general, cancer classification analysis, based on microarray gene data, is a task of constructing a decision rule based on the dataset of genes and tissues, which is able to automatically assign new tissue to one of two categories [5,6]. High-dimensional cancer classification analysis has attracted much attention in both bioinformatics and computational biology, because the classical classification methods suffer from the curse of dimensionality [7].

Using all genes in the high-dimensional microarray data often results in model overfitting, particularly if there are irrelevant and noisy genes [8]. Consequently, removing irrelevant and noisy genes is an important target when dealing with high-dimensional cancer classification. In principle, gene selection aims to select a relatively small set of genes from a high-dimensional gene dataset, and, therefore, achieves high classification accuracy. Furthermore, selecting important genes can also help in early diagnosis and drug discovery for cancer patients [9,10]. Numerous statistical methods have been successfully applied in the area of cancer classification. Among them, logistic regression (LR) is considered as a powerful discriminative method. LR provides predicted probabilities of class membership and easy interpretation of the

gene coefficients [8]. However, LR is neither applicable nor suitable for the high-dimensional microarray data classification, because the design matrix is singular. Thus, iteration methods, such as Newton-Raphson's method, cannot work [11].

Recently, there has been growing interest in applying the sparse methods in high-dimensional cancer classification [8,12,11]. To tackle both estimating the gene coefficients and performing gene selection simultaneously, sparse logistic regression (SLR) has been successfully applied in high-dimensional cancer classification [13-17]. A SLR with different penalties can be applied. The most widely and popular penalty is the least absolute shrinkage and selection operator (LASSO) [18]. The LASSO imposes the l_1-norm penalty to the loss function. Because of the l_1-norm property, the LASSO can perform variable selection by assigning some gene coefficients to zero. For this reason, the LASSO has gained popularity in high-dimensional data. SLR with l_1-norm gives a sparse solution with high classification accuracy.

Despite the advantage of the LASSO, it has three shortcomings [19]. First, it cannot select more genes than the number of tissues. Second, in microarray gene data, there is grouping among genes, where genes that share a common biological pathway have a high pairwise correlation with each other. The LASSO tries to select only one gene or a few of them among a group of correlated genes. To overcome the first two limitations, Zou and Hastie [20] proposed the elastic net penalty, for which the penalty is a linear combination of l_1-norm and l_2-norm. Last, the LASSO has a bias gene selection, because it penalizes all gene coefficients equally [21,22]. In other words, the LASSO does not have the oracle properties, which refer to the probability of selecting the right set of genes (with nonzero coefficients) converged to one, and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients are known in advance [23].

In relation to the last limitation of the LASSO, oracle properties, Zou [24] proposed the adaptive LASSO in which adaptive weights are used for penalizing different coefficients in the l_1-norm penalty. In high-dimensional classification data, however, the adaptive LASSO faces two practical problems: (1) a maximum likelihood estimates (MLE) is usually proposed as an initial weight. In high-dimensional cancer classification, the MLE is not available, and, hence, the adaptive LASSO is no longer applicable. (2) The adaptive LASSO still has poor performance when there is grouping among genes [25].

In this study, a new initial weight inside the l_1-norm penalty in adaptive sparse logistic regression is proposed. It is defined as the ratio of the standard error of the ridge regression estimator to the ridge regression estimator. The main objective behind this new initial weight is to adjust the l_1-norm penalty in sparse logistic regression by improving consistent genes selection (oracle property) and encouraging the l_1-norm penalty to select more correlated genes inside a group (grouping effects). To evaluate the effectiveness of the new initial weight, we apply four public cancer classification datasets. Moreover, a comparison is done with other penalties and initial weights.

The rest of this paper is arranged as follows: Section 2 displays the sparse logistic regression and the proposed method. Section 3 describes the results and discussion of the real data analysis. The conclusion is covered by section 4.

2 Methods

2.1 Sparse logistic regression

Logistic regression is a statistical method to model a binary classification problem. The regression function has a nonlinear relation with the linear combination of the genes. In cancer classification, the response variable of the logistic regression has two values either 1 for the

tumor class or 0 for the normal class. Let $\mathbf{y}_i \in \{0,1\}$ be a vector of size $n \times 1$ of tissues, and let $\mathbf{x}_j$ be a $p \times 1$ vector of genes. The logistic transformation of the vector of probability estimates $\pi_i = p(y_i = 1 | \mathbf{x}_j)$ is modeled by a linear function, logit transformation:

$$\ln[\pi_i / 1 - \pi_i] = \beta_0 + \sum_{j=1}^{p} \mathbf{x}_j^T \beta_j, \; i = 1, 2, ..., n, \tag{1}$$

where β_0 is the intercept and β_j is a $p \times 1$ vector of unknown gene coefficients. The log-likelihood function of (1) is defined as:

$$\mathrm{l}(\beta_0, \beta) = \sum_{i=1}^{n} \left\{ \mathbf{y}_i \ln \pi(\mathbf{x}_{ij}) + (1 - \mathbf{y}_i) \ln(1 - \pi(\mathbf{x}_{ij}) \right\}. \tag{2}$$

Logistic regression offers the advantage of simultaneously estimating the probabilities $\pi(\mathbf{x}_{ij})$ and $1 - \pi(\mathbf{x}_{ij})$ for each class and classifying subjects. The probability of classifying the ith sample in class 1 is estimated by $\hat{\pi}_i = \exp(\beta_0 + \sum_{j=1}^{p} \mathbf{x}_j^T \beta_j) / 1 + \exp(\beta_0 + \sum_{j=1}^{p} \mathbf{x}_j^T \beta_j)$. The predicted class is then obtained by $I\{\hat{\pi}_i > 0.5\}$, where $I(\mathrm{g})$ is an indicator function.

SLR adds a nonnegative penalty term to (1), such that the size of gene coefficients in high-dimension can be controlled. Several penalty terms have been discussed in the literature [26,8,14,18]. The l_1-norm penalty, proposed by Tibshirani [18], is one of the popular penalty terms. The l_1-norm penalty performs genes selection and estimation simultaneously by constraining the log-likelihood function of gene coefficients. The sparse method for the logistic regression is obtained by adding the penalty term to the log-likelihood function:

$$SLR = \sum_{i=1}^{n} \left\{ \mathbf{y}_i \ln \pi(\mathbf{x}_{ij}) + (1 - \mathbf{y}_i) \ln(1 - \pi(\mathbf{x}_{ij}) \right\} + \lambda P(\beta). \tag{3}$$

The estimation of the vector β is obtained by minimizing (3):

$$\hat{\beta}_{SLR} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left\{ \mathbf{y}_i \ln \pi(\mathbf{x}_{ij}) + (1 - \mathbf{y}_i) \ln(1 - \pi(\mathbf{x}_{ij}) \right\} + \lambda P(\beta) \right], \tag{4}$$

where $\lambda P(\beta)$ is the penalty term that sparse the estimates. The penalty term depends on the positive tuning parameter, λ, which controls the tradeoff between fitting the data to the model and the effect of the penalty. In other words, it controls the amount of shrinkage. For the $\lambda = 0$, we obtain the MLE solution. In contrast, for large values of λ the influence of the penalty term on the coefficient estimates increases. Choosing the tuning parameter is an important part of the model fitting. If the focus is on classification, the tuning parameter should find the right balance between the bias and variance to minimize the misclassification error. Without loss of generality, it is assumed that the genes are standardized, $\sum_{i=1}^{n} x_{ij} = 0$ and $(n^{-1})\sum_{i=1}^{n} x_{ij}^{2} = 1, \quad \forall j \in \{1, 2, ..., p\}$. The estimation of the vector β using the LASSO (l_1-norm penalty) is defined as:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left\{ \mathbf{y}_i \ln \pi(\mathbf{x}_{ij}) + (1 - \mathbf{y}_i) \ln(1 - \pi(\mathbf{x}_{ij}) \right\} + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right], \qquad (5)$$

where λ is a tuning parameter. It reduces to the MLE estimator when $\lambda = 0$. On the other hand, if $\lambda \to \infty$, the penalty forces all the genes to be zeros. In practice, the value of λ is often chosen by a cross-validation procedure. To solve (5), the traditional numerical methods are through MLE or the Newton-Raphson algorithm. However, the computation of these methods is prohibitive when the number of genes is large [15]. Equation (5) can be efficiently solved by the coordinate descent algorithm [27,28].

The LASSO has an advantage in that it is computationally feasible in high-dimensional classification data. On the other hand, the LASSO has three main drawbacks. First of all, if $p > n$, the LASSO selects at most n genes because of the nature of the convex optimization problem. In addition, the LASSO cannot handle the effect of grouping. When the pairwise correlations among a group of genes are very high, then the LASSO tends to select only one gene from the whole group and does not take into account which one is selected [20]. Lastly, the LASSO lacks the oracle properties, as stated in Fan and Li [23].

2.2 Adjusted adaptive sparse logistic regression

According to Fan and Li [23], a good penalty term should result in an estimator with three properties: unbiasedness, sparsity and continuity. Unbiasedness means the resulting estimator has no over penalization for large parameters to avoid unnecessary modeling biases. Sparsity is another property that an estimator enjoys. In other words, the resulting estimator automatically sets insignificant parameters to zero. Lastly, continuity is the third property, meaning that the resulting estimator is continuous in data in order to avoid instability in model prediction.

One of the main reasons for the LASSO not to be consistent, i.e., lacking the oracle property [23,29,22] is that it equally penalizes all the coefficients, which over-penalizes the irrelevant genes leading it to be a biased estimator. To alleviate this drawback, Zou [24] proposed the adaptive LASSO in which adaptive weights are used for penalizing different coefficients in the l_1-norm penalty. The basic idea behind the adaptive LASSO is that by assigning a higher weight to the small coefficients and lower weight to the large coefficients, it is possible to reduce the selection bias; therefore, it can consistently select the model. Furthermore, the adaptive LASSO solution is continuous from its definition, which enables it to enjoy oracle properties. The sparse logistic regression using the adaptive LASSO (ASLR) of β is defined by:

$$\hat{\beta}_{ASLR} = \arg\min_{\beta}\left[\sum_{i=1}^{n}\left\{\mathbf{y}_i \ln\pi(\mathbf{x}_{ij}) + (1-\mathbf{y}_i)\ln(1-\pi(\mathbf{x}_{ij})\right\} + \lambda\sum_{j=1}^{p} w_j \left|\beta_j\right|\right], \qquad (6)$$

where $\mathbf{w}_j = (\mathrm{w}_1,...,w_p)^T$ is $p\times 1$ data-driven weight vector. It depends on the root n-consistent initial values of $\hat{\beta}$ and $\mathbf{w}_j = (|\hat{\beta}_j|)^{-\gamma}$, where γ is a positive constant. The adaptive LASSO originally used MLE estimates for the initial weight [24]. This is no longer valid in high-dimensional data. Several researchers have used the LASSO estimates as an alternative initial weight [30]. However, using a LASSO estimator in an adaptive LASSO

sparse logistic when $p > n$ may not be preferable for three reasons. First, the LASSO estimator is inconsistent in itself. In other words, this initial weight is biased in selection genes. Second, it does not take into account the weights for all the genes in any implantation which means that some genes will be selected and the others will be set to zero. Last, when there is a group of correlated genes, the LASSO fails to select the grouped genes together.

To overcome these limitations, the ratio of the standard error of the ridge regression estimator to the ridge regression estimator has been proposed as an initial weight in the adaptive LASSO sparse logistic regression. According to the nature of the l_2-norm, the ridge penalty tries to force the estimated gene coefficients of highly correlated genes to be close to each other. However, this property loses the capability of estimating the coefficients of highly correlated genes with different magnitude, especially with different signs [19,31]. The advantage of using the standard error of the ridge estimator $s_{\hat{\beta}_{Ridge}}$ is to adjust the sparse logistic regression using the adaptive LASSO (AASLR) when using ridge regression estimates as an initial value. As a result, the AASLR is able to improve gene selection consistently (oracle property) and encourage the l_1-norm penalty in selecting more correlated genes inside a group (grouping effects). Cule and De Iorio [32] proposed a procedure to calculate the $s_{\hat{\beta}_{Ridge}}$ depending on the principal component analysis. Let $\hat{\beta}_{Ridge} = (\hat{\beta}_{1(Ridge)}, ..., \hat{\beta}_{\mathrm{p}(Ridge)})^T$ be the vector of ridge regression estimate, $\mathbf{s}_{\hat{\beta}_{Ridge}} = (s_{1(\hat{\beta}_{Ridge})}, ..., s_{\mathrm{p}(\hat{\beta}_{Ridge})})^T$ be the vector of the standard error of the ridge regression, and $\mathbf{w}_{Ratio} = (w_{1(ratio)}, ..., w_{p(ratio)})^T$ be the ratio weight vector where $\mathbf{w}_j = (\mathbf{s}_{j(\hat{\boldsymbol{\beta}}_{Ridge})} / |\hat{\beta}_{j(Ridge)}|)^{-\gamma}, \quad j = 1, 2, ..., p$. Then a coordinate descent method can be used to solve AASLR. The computation details are given in algorithm 1.

Algorithm 1 The coordinate descent method for AASLR

Step 1: Input $\mathbf{x}_{ij}$ and $\mathbf{w}_{Ratio}$.

Step 2: Define $\mathbf{x}_{ij}^{**} = \mathbf{x}_{ij} / \mathbf{w}_{Ratio}, j = 1, 2, ..., p$.

Step 3: Solve the sparse logistic regression using the LASSO for all λ and γ values,

$$\hat{\beta}^{**}_{AASLR} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left\{ \mathbf{y}_i \ln \pi(\mathbf{x}_{ij}^{**}) + (1-\mathbf{y}_i) \ln(1-\pi(\mathbf{x}_{ij}^{**}) \right\} + \lambda \sum_{j=1}^{p} \mathbf{w}_{j\,(Ratio)} \left| \boldsymbol{\beta}_j \right| \right\}.$$

Step 4: Output $\hat{\beta}^{*}_{j\,(AASLR)} = \hat{\beta}_j^{**} / \mathbf{w}_{Ratio}$.

2.3 Tuning parameter selection

For practical applications, one has to decide the values of λ. Classically, cross-validation (CV) has been widely used. However, it is computationally intensive for AASLR, simply because there are two tuning parameters: λ and γ. For simplicity, $\gamma = 1$ was used for the simulation study and the real data application. Then, the AASLR tuning parameters were reduced to only λ.

3 Results and discussion

To evaluate our proposed method AASLR in the field of cancer classification, four publicly well-known binary cancer classification datasets were used: diffuse large B-cell lymphoma (DLBCL) [33], prostate cancer [34], leukemia cancer [35], and colon cancer [36]. The detailed information of these datasets is summarized in Table 1. The DLBCL dataset consisted of the gene expression values of 77 samples that were measured by high-density oligonucleotide microarrays of the two most prevalent adult lymphoid malignancies, which comprised 58 samples of diffuse large B-cell lymphomas (DLBCL) and 19 samples of follicular lymphoma (FL). Each sample contained 7,129 gene expression values. The original prostate dataset contained 12600 genes for 52 prostate tumor samples and 50 non-tumor tissues. A subset of

5966 genes was adapted in the classification. In the leukemia dataset, there were two types of patients: 47 patients of acute lymphoblastic leukemia (ALL) and 25 patients of acute myeloid leukemia (AML). The total expression profiles were 7129 genes. The colon cancer dataset, contained gene expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes obtained with an Affymetrix oligonucleotide array. A subset of 2000 genes with the highest minimal intensity across the samples was used.

Table 1 The detail information for the used data sets

Data set	# samples	# genes	Classes
DLBCL	77	7129	DLBCL / FL
Prostate	102	5966	Tumor / Non-tumor
Colon	62	2000	Tumor / Normal
Leukemia	72	7129	ALL / AML

In order to enable a fair comparison, we randomly partitioned each dataset into a training dataset, which comprised 70% of the samples, and a test dataset, which consisted of 30% of the samples. In order to get the best value of λ, the 10-fold CV was employed using the training dataset with 25 times. All the applications were conducted in R using the *glmnet* package. The averaged number of selected genes, the averaged classification accuracy (%) (CA), and Youden's index (YI) in both the training and testing datasets are reported in Table 2. For comparison purposes, the performance of the LASSO, adaptive sparse logistic regression with LASSO as an initial weight ($ASLR_{LASSO}$), and adaptive sparse logistic regression with ridge regression as an initial weight ($ASLR_{Ridge}$) was also evaluated.

As can be seen from Table 2, AASLR selected more genes than the other three methods. In leukemia, for instance, AASLR selected 25 genes compared to 18, 15, and 14 genes for $ASLR_{Ridge}$, $ASLR_{LASSO}$, and the LASSO, respectively. Importantly, AASLR has the potential to select more genes than the other three methods, indicating that most of these additionally selected genes were probably highly correlated.

Furthermore, AASLR has average classification accuracy in both the training and testing sets, and is much better than $ASLR_{Ridge}$, $ASLR_{LASSO}$, and the LASSO in the DLBCL, prostate, and leukemia datasets, respectively. For the colon dataset, AASLR has slightly better classification accuracy. For example, in the DLBCL data, the classification accuracy of AASLR in the training (testing) set was 99.583 (96.741), which was greater than 97.736 (93.674) for the $ASLR_{Ridge}$, 96.287 (92.035) for $ASLR_{LASSO}$, and 96.011 (91.731) for the LASSO. In terms of Youden's index, the averaged values in all the datasets were considerably higher for AASLR in both the training and testing datasets, where the maximal Youden's index is 1 [37].

On the other hand, the LASSO generally performed slightly worse than the other three methods in terms of classification accuracy and Youden's index for either the training or the testing dataset, although it did select less genes. Besides, $ASLR_{Ridge}$ performed slightly better than $ASLR_{LASSO}$. This is because ASLR is dependent on the LASSO weight, which was biased in gene selection.

Table 2 The averaged evaluation criteria over 25 time for the used data sets

	Training set			Testing set	
	# genes	CA	YI	CA	YI
DLBCL					
LASSO	12	96.011	0.895	91.731	0.859
$ASLR_{LASSO}$	13	96.287	0.905	92.035	0.907
$ASLR_{Ridge}$	19	97.736	0.912	93.674	0.912
AASLR	24	99.583	0.937	96.741	0.940
Prostate					
LASSO	14	98.441	0.894	88.749	0.877
$ASLR_{LASSO}$	16	98.718	0.903	88.782	0.883
$ASLR_{Ridge}$	18	98.872	0.910	89.107	0.891
AASLR	25	99.014	0.955	93.317	0.917
Colon					
LASSO	10	93.551	0.743	78.882	0.721

$ASLR_{LASSO}$	10	93.803	0.754	79.107	0.738
$ASLR_{Ridge}$	10	94.325	0.769	79.438	0.741
AASLR	15	94.484	0.784	80.107	0.757
Leukemia					
LASSO	14	98.891	0.947	95.148	0.907
$ASLR_{LASSO}$	15	98.904	0.951	95.471	0.918
$ASLR_{Ridge}$	18	98.974	0.955	96.076	0.924
AASLR	25	99.638	0.981	98.085	0.977

3.1 The consistency of the proposed method

To further evaluate the ability of AASLR in consistent gene selection, Fig. 1 depicts the boxplots of the number of selected genes of AASLR, $ASLR_{Ridge}$, $ASLR_{LASSO}$, and the LASSO in all the data over the 25 times. It is clear that AASLR gave much more consistent results than the other three methods. For instance, using the whiskers of the boxplots as a reference, the AASLR is likely to choose a subset of genes of size 23 to 28 genes, as compared to a subset of size 11 to 22, 10 to 22, and 5 to 22 genes for the $ASLR_{Ridge}$, $ASLR_{LASSO}$, and the LASSO in the prostate dataset, respectively. This clearly demonstrated that the size of the selected genes obtained from AASLR was consistent each time.

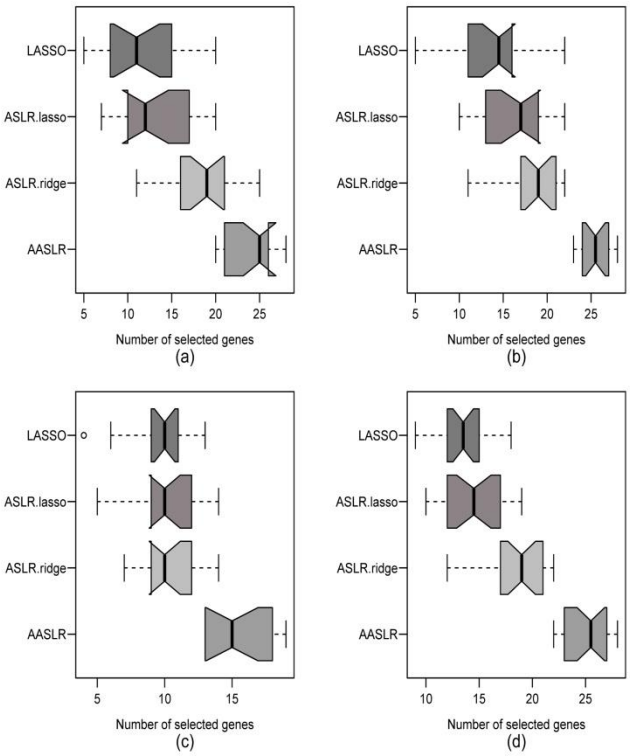

Fig. 1 Number of selected genes over 25 times for the used methods. (a) DLBCL. (b) prostate. (c) colon. (d) leukemia

3.2 Grouping effects

To focus on the capability of AASLR in encouraging grouping effects, we listed the most frequently highly correlated selected genes in the leukemia dataset in Table 3. The correlation matrix of these selected genes is given in Fig. 2. We can observe that the AASLR successfully selected the most highly correlated genes. For example, the highest correlation among the selected genes was 0.918 between gene index 2348 and 4535. These two correlated genes were selected together by AASLR with 100% compared to 72% for $ASLR_{Ridge}$, 36% for $ASLR_{LASSO}$, and 20% for the LASSO.

Furthermore, AASLR selected the most important highly correlated genes 21 times out of 25 times, with the percentage equal to 84%. On the other hand, it can be observed that the $ASLR_{Ridge}$ did not perform well in selecting highly correlated genes, although it selected more

genes compared to the LASSO and $ASLR_{LASSO}$. In contrast, the LASSO and $ASLR_{LASSO}$ failed to select the highly correlated genes together; their percentages were 16% and 20%, respectively. The success of AASLR in selecting more correlated genes than the other methods, especially $ASLR_{LASSO}$, is due to its ability to adjust the adaptive weight.

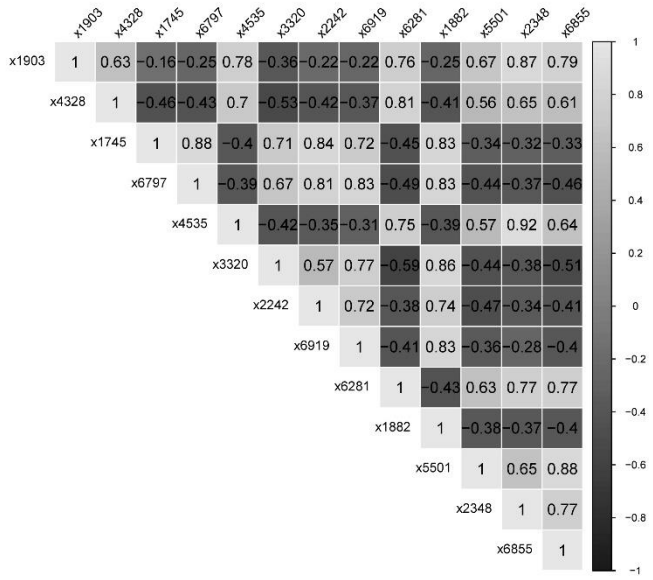

Fig. 2 The correlation matrix between the top 13 selected genes for leukemia dataset

Table 3 Frequencies of the most 13 selected genes in leukemia data set over 25 times

		Frequency			
Gene Index	Gene name	LASSO	$ASLR_{LASSO}$	$ASLR_{Ridge}$	AASLR
4535	SSR2 Signal sequence receptor, beta	21	22	18	25
4328	MCP Membrane cofactor protein	7	5	20	25
2348	ACADM acyl-coenzyme A dehydrogenase, C-4 to C-12 straight chain	5	9	17	25
1745	C-yes-1 mRNA	19	18	25	25
2242	INTEGRAL MEMBRANE PROTEIN E16	16	11	10	22
6919	Skeletal beta-tropomyosin	13	17	16	21
1882	CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)	7	7	20	22
6797	GYPB Glycophorin B	5	10	24	25
3320	Guanine nucleotide exchange factor p532 mRNA	4	5	17	21
5501	TOP2B topoisomerase (DNA) II b (180 kDa)	16	15	18	25
1903	Recombination activating protein (RAG-1) gene	9	6	17	23
6855	TCF3 transcription factor 3 (E2A immunoglobulin enhancer-binding factors E12/E	6	7	14	25
6281	MYL1 myosin light chain (alkali)	9	8	17	23

3.3 Stability test

In the stability test for the proposed method, the AASLR seeks to prove that it can classify high-dimensional cancer data with a high degree of accuracy compared to the other three used

methods. Depending on the training dataset, a two-way analysis of variance (ANOVA) was used as a statistical test to check whether the AASLR, $ASLR_{Ridge}$, $ASLR_{LASSO}$, and the LASSO were statistically significant and if there was any significant difference between the four datasets used in terms of classification accuracy. Table 4 reports the two-way ANOVA results. From Table 4, the results showed statistically significant differences between the AASLR and the three other used methods in terms of classification accuracy. In addition, we can see that the DLBCL, prostate, colon, and leukemia datasets had different classification accuracy values. Furthermore, Duncan's multiple range test was used to obtain more detailed information about the differences between the AASLR and the two three used methods. Table 5 lists the p-value of each compared pair of methods. It is apparent from Table 5 that the AASLR showed statistical differences compared to the $ASLR_{Ridge}$, $ASLR_{LASSO}$, and LASSO in terms of classification accuracy.

Table 4 Two-way ANOVA for average classification accuracy over 25 times

Source	df	SS	MS	F	*p*-value
Methods	3	7362.3	2454.1	103.1	0.000
Datasets	3	1826.7	608.9	25.5	0.008
Error	396	9426.8	23.8		
Total	399	18615.8			

Table 5 P-value of Duncan’s multiple range test for average classification accuracy

	LASSO	$ASLR_{LASSO}$	$ASLR_{Ridge}$	AASLR
LASSO		0.036	0.000	0.000
$ASLR_{LASSO}$			0.007	0.000
$ASLR_{Ridge}$				0.004
AASLR				

To summarize, it is obvious that the microarrays real datasets results demonstrated the use of AASLR in terms of classification accuracy, Youden’s index for both the training and testing

sets. In addition, it outperformed the other competitor methods in terms of consistent selection, selection of highly correlated genes, and stability test.

4 Conclusion

Cancer classification is one of the most important applications in gene expression data. However, due to the high-dimensionality problem of genes, many supervised computational methods have failed to identify a small subset of important genes. To tackle both estimating the gene coefficients and performing gene selection simultaneously, sparse logistic regression was successfully applied in high-dimensional microarray data classification. In this research, we have proposed AASLR for consistent gene selection and selecting highly grouped genes simultaneously in high-dimensional tumor classification. The results, which were based on four microarray real datasets, proved that AASLR yielded positive and useful results in terms of (a) classification accuracy and Youden's index for both the training and testing datasets, (b) consistency in gene selection, (c) selecting highly correlated genes, and (d) stability test. Therefore, we can conclude the effectiveness of the proposed AASLR method in practice. We restricted our attention to the binary case, but AASLR can be extended to cover high-dimensional multi-classification microarray data.

References

1. Cui Y, Zheng C-H, Yang J, Sha W (2013) Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. Comput Biol Med 43 (7):933-941. doi:http://dx.doi.org/10.1016/j.compbiomed.2013.04.018

2. Liao JG, Chin K-V (2007) Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics 23 (15):1945-1951. doi:10.1093/bioinformatics/btm287

3. Zheng S, Liu W (2011) An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification. Comput Biol Med 41 (11):1033-1040. doi:http://dx.doi.org/10.1016/j.compbiomed.2011.08.011
4. Piao Y, Piao M, Park K, Ryu KH (2012) An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics 28 (24):3306-3315. doi:10.1093/bioinformatics/bts602
5. Kalina J (2014) Classification methods for high-dimensional genetic data. Biocybern Biomed Eng 34 (1):10-18. doi:http://dx.doi.org/10.1016/j.bbe.2013.09.007
6. Benoit DF, Alhamzawi R, Yu K (2013) Bayesian lasso binary quantile regression. Computation Stat 28:2861–2873
7. Lotfi E, Keshavarz A (2014) Gene expression microarray classification using PCA–BEL. Comput Biol Med 54 (0):180-187. doi:http://dx.doi.org/10.1016/j.compbiomed.2014.09.008
8. Liang Y, Liu C, Luan X-Z, Leung K-S, Chan T-M, Xu Z-B, Zhang H (2013) Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. BMC Bioinformatics 14 (1):198-211
9. Chen K-H, Wang K-J, Wang K-M, Angelia M-A (2014) Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. Appl Soft Comput 24 (0):773-780. doi:http://dx.doi.org/10.1016/j.asoc.2014.08.032
10. Pang S, Havukkala I, Hu Y, Kasabov N (2007) Classification consistency analysis for bootstrapping gene selection. Neural Comput Applic 16 (6):527-539. doi:10.1007/s00521-007-0110-1
11. Bielza C, Robles V, Larrañaga P (2011) Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. Expert Syst Appl 38 (5):5110-5118. doi:http://dx.doi.org/10.1016/j.eswa.2010.09.140
12. Bootkrajang J, Kabán A (2013) Classification of mislabelled microarrays using robust sparse logistic regression. Bioinformatics 29 (7):870-877. doi:10.1093/bioinformatics/btt078
13. Shevade SK, Keerthi SS (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 19 (17):2246-2253. doi:10.1093/bioinformatics/btg308
14. Zhenqiu L, Feng J, Guoliang T, Suna W, Fumiaki S, Ming T (2007) Sparse logistic regression with Lp penalty for biomarker identification. Stat Appl Genet Mol Biol 6 (1):1-22
15. Zhu J, Hastie T (2004) Classification of gene microarrays by penalized logistic regression. Biostatistics 5 (3):427-443. doi:10.1093/biostatistics/kxg046
16. Cawley GC, Talbot NLC (2006) Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics 22 (19):2348-2355. doi:10.1093/bioinformatics/btl386
17. Li S, Eng Chong T (2005) Dimension reduction-based penalized logistic regression for cancer classification using microarray data. IEEE Trans Comput Bi 2 (2):166-175. doi:10.1109/TCBB.2005.22
18. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Statist Soc Ser B 58 (1):267-288. doi:10.2307/2346178
19. Wang S, Nan B, Rosset S, Zhu J (2011) Random lasso. Ann Appl Stat 5 (1):468-485. doi:10.1214/10-AOAS377
20. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Roy Statist Soc Ser B 67 (2):301-320. doi:10.1111/j.1467-9868.2005.00503.x
21. Fan J, Fan Y, Barut E (2014) Adaptive robust variable selection. Ann Statist 42 (1):324-351
22. Alhamzawi R, Yu K, Benoit DF (2012) Bayesian adaptive Lasso quantile regression. Stat Model 12 (3):279–297

23. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc 96 (456):1348-1360
24. Zou H (2006) The Adaptive Lasso and Its Oracle Properties. J Amer Statist Assoc 101 (476):1418-1429. doi:10.1198/016214506000000735
25. El Anbari M, Mkhadri A (2013) The adaptive gril estimator with a diverging number of parameters. Commun Stat Theor 42 (14):2634-2660. doi:10.1080/03610926.2011.615438
26. Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1):55-67
27. Park MY, Hastie T (2008) Penalized logistic regression for detecting gene interactions. Biostatistics 9 (1):30-50. doi:10.1093/biostatistics/kxm010
28. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33 (1):1-22
29. Li J, Jia Y, Zhao Z (2012) Partly adaptive elastic net and its application to microarray classification. Neural Comput Applic 22 (6):1193-1200. doi:10.1007/s00521-012-0885-6
30. Bühlmann P, Van De Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Heidelberg
31. Qian W, Yang Y (2013) Model selection via standard error adjusted adaptive lasso. Ann Ins Stat Math 65 (2):295-318. doi:10.1007/s10463-012-0370-0
32. Cule E, De Iorio M (2013) Ridge regression in prediction problems: Automatic choice of the ridge parameter. Genet Epidemiol 37 (7):704-714. doi:10.1002/gepi.21750
33. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8 (1):68-74
34. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1 (2):203-209. doi:http://dx.doi.org/10.1016/S1535-6108(02)00030-2
35. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286 (5439):531-537. doi:10.1126/science.286.5439.531
36. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96 (12):6745-6750. doi:10.1073/pnas.96.12.6745
37. Becker N, Toedt G, Lichter P, Benner A (2011) Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. BMC Bioinformatics 12 (1):138-151