Computers in Biology and Medicine 67 (2015) 136–145

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification

Zakariya Yahya Algamal, Muhammad Hisyam Lee*

Department of Mathematical Sciences, Universiti Teknologi Malaysia 81310 Skudai, Johor, Malaysia

ARTICLE INFO

Article history:
Received 14 June 2015
Received in revised form
7 October 2015
Accepted 8 October 2015

Keywords:
Adaptive elastic net
Oracle property
Regularized logistic regression
Cancer classification
Gene selection

ABSTRACT

Cancer classification and gene selection in high-dimensional data have been popular research topics in genetics and molecular biology. Recently, adaptive regularized logistic regression using the elastic net regularization, which is called the adaptive elastic net, has been successfully applied in high-dimensional cancer classification to tackle both estimating the gene coefficients and performing gene selection simultaneously. The adaptive elastic net originally used elastic net estimates as the initial weight, however, using this weight may not be preferable for certain reasons: First, the elastic net estimator is biased in selecting genes. Second, it does not perform well when the pairwise correlations between variables are not high. Adjusted adaptive regularized logistic regression (AAElastic) is proposed to address these issues and encourage grouping effects simultaneously. The real data results indicate that AAElastic is significantly consistent in selecting genes compared to the other three competitor regularization methods. Additionally, the classification performance of AAElastic is comparable to the adaptive elastic net and better than other regularization methods. Thus, we can conclude that AAElastic is a reliable adaptive regularized logistic regression method in the field of high-dimensional cancer classification.

1. Introduction

Recently, molecular biology and genetics research has been transformed from the study of individual genes to the exploration of the whole genome. DNA microarrays technology is one such technique to measure the expression levels of thousands of genes in a single experiment [1–4]. Cancer classification based on microarray gene expression data has become one of the most active research topics in biomedical research, which is suitable for comparing the gene expression levels in tissues under different conditions, such as normal versus abnormal [5,6].

However, cancer classification with DNA microarray data is a challenging issue because of its high dimensionality and the small samples size. Typically, the number of genes is more than thousands from a hundred or less tissue samples [7,8]. Due to the high dimensionality and the small sample size, gene selection is an important issue for cancer classification and has been extensively studied in recent years. The application of gene selection methods allows the identification of a small number of important genes that can be used as biologically relevant genes of the appropriate cancer [9–11]. From the viewpoint of biologists, gene selection can increase the classification accuracy of the classification method by removing irrelevant and noisy genes [12–14].

Many gene selection methods have been proposed to select a subset of genes that can have high classification accuracy for cancer classification. Recently, regularization methods, which are capable of conducting efficient gene selection and model estimation simultaneously, have gained popularity [15,16]. From the statistical perspective, regularization methods can control the effects of the overfitting and multicollinearity [17]. Numerous statistical methods have been successfully applied in the area of cancer classification. Among them, logistic regression (LR) is considered to be a powerful discriminative method. LR provides predicted probabilities of class membership and easy interpretation of the gene coefficients [17]. However, LR is neither applicable nor suitable for high-dimensional cancer classification because the design matrix is singular. Thus, the iteration methods, such as Newton–Raphson's method cannot work [18]. Regularized logistic regression (RLR) has been successfully applied in high-dimensional cancer classification [6,19–23]. The benefits of RLR are that (a) the classification accuracy can often be improved by shrinking the regression coefficients, and (b) selecting a small subset of genes that exhibits the strongest effects provides a classification model with easy interpretation.

* Corresponding author. Tel.: +60 7 5534236; fax: +60 7 556 6162.
E-mail addresses: zak.sm_stat@yahoo.com (Z.Y. Algamal), mhl@utm.my (M.H. Lee).

http://dx.doi.org/10.1016/j.compbiomed.2015.10.008
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

An RLR with different regularization terms can be applied. The most widely and popular regularized term is the least absolute shrinkage and selection operator (LASSO) [24]. LASSO imposes the ℓ_1-norm regularization to the loss function. Because of the ℓ_1-norm property, LASSO can perform variable selection by assigning some genes coefficients to zero. For this reason, LASSO has gained popularity in high-dimensional data.

Despite the advantage of LASSO, it has three shortcomings [25,26]. First, LASSO has a biased gene selection, which means it is an inconsistent gene selection method because it regularizes all gene coefficients equally [27]. In other words, LASSO does not have the oracle property, which refers to the probability of selecting the right set of genes (with nonzero coefficients) converges to one, and that the estimators of the nonzero coefficients have asymptotically normal distribution with the same means and covariances as if the zero coefficients are known in a prior [28,29]. Related to this limitation of LASSO, concerning the oracle property, Zou [30] proposed the adaptive LASSO in which adaptive weights are used for regularizing different coefficients in the ℓ_1-norm regularization. Second, it cannot select more genes than the number of samples. Last, in the microarray gene data, there is grouping among genes, where genes that share a common biological pathway have a high pairwise correlation with each other. LASSO tries to select only one gene or a few of them among a group of correlated genes. To overcome the last two limitations, Zou and Hastie [26] proposed the elastic net regularization, for which the regularization is a linear combination of ℓ_1-norm and ℓ_2-norm. Similar to LASSO, elastic net lacks the oracle property even though it outperforms LASSO. Zou and Zhang [31] proposed adaptive elastic net to handle grouping effects and enjoy the oracle property simultaneously.

In high-dimensional classification data, however, the adaptive elastic net faces practical problems where a maximum likelihood estimate (MLE), which is usually proposed as an initial weight, is simply infeasible, and, hence, the adaptive elastic net is no longer applicable. Zou and Zhang [31] proposed using the elastic net estimates as an initial weight in adaptive elastic net; however, using this weight may not be preferable for three reasons: First, it is well known that gene selection by elastic net can be inconsistent [31,32]. In other words, this initial weight is biased in selecting genes. Second, elastic net exhibits difficulties when a group of genes is nearly linearly dependent, because it does not take into account the correlation structure among genes [33]. Last, the elastic net does not perform well when the pairwise correlations between genes are not extremely high; El Anbari and Mkhadri [34] stated that if the absolute correlation between genes is slightly less than 0.95, the elastic net may be slightly less reliable.

In this study, a new initial weight inside ℓ_1-norm regularization in adaptive elastic regularized logistic regression is proposed, which 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 regularization in regularized logistic regression by improving the gene selection consistency while still maintaining the grouping effects. To evaluate the effectiveness of the new initial weight, we applied three DNA microarray datasets of cancer classification. Moreover, a comparison is made with other regularization terms and initial weights.

The rest of this paper is arranged as follows: Section 2 displays the regularized logistic regression, the adaptive regularized logistic regression, and the proposed method. While Section 3 covers the real data application results. Finally, the conclusion is covered by Section 4.

2. Methods

2.1. Regularized 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. Assume that we have n observations and p genes. Let $y_i \in \{0,1\}$ be the response variable value for observation i, $i=1,2,...,n$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})^T$ be the i^{th} gene vector of the gene matrix $\mathbf{X}$. Then, the response variable is related to genes by

$$\pi_i = p(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}, \quad i=1,2,...,n \tag{1}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ is a $p \times 1$ vector of unknown gene coefficients. The log-likelihood function of the logit transformation of Eq. (1) is defined as

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\}. \tag{2}$$

Regularized logistic regression adds a nonnegative regularization term to the negative log-likelihood function, $\ell(\beta)$, such that the size of gene coefficients in high-dimension can be controlled. Several regularization terms have been discussed in the literature [23,24,26,35]. The ℓ_1-norm regularization, proposed by Tibshirani [36], is one of the popular regularization terms. The ℓ_1-norm regularization performs gene selection and estimation simultaneously by constraining the negative log-likelihood function of gene coefficients. Thus, the RLR is defined as:

$$RLR = -\ell(\beta) + \lambda P(\beta). \tag{3}$$

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

$$\hat{\beta}_{RLR} = \text{argmin}_{\beta}\left[-\sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i\} + \lambda\, P(\beta)\right], \tag{4}$$

where $\lambda\, P(\beta)$ is the regularization term that regularized 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 regularization. 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 regularization term on the coefficient estimate 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 $(n^{-1})\sum_{i=1}^{n} x^2{}_{ij} = 1, \quad \forall j \in \{1,2,...,p\}$. As a result, the intercept β_0 is not regularized. The estimation of the vector β using the LASSO (ℓ_1-norm regularization) is defined as:

$$\hat{\beta}_{LASSO} = \text{argmin}_{\beta}\left[-\sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i\} + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right], \tag{5}$$

where λ is a tuning parameter. It reduces to the MLE estimator when $\lambda = 0$. On the other hand, if $\lambda \to \infty$, the regularization term forces all the gene coefficients to be zero. In practice, the value of λ is often chosen by a cross-validation procedure. Eq. (5) can be efficiently solved by the coordinate descent algorithm [37,38].

Elastic net is a regularization method for gene selection, which is introduced by Zou and Hastie [26] to deal with the first two drawbacks of LASSO. Elastic net tries to combine the ℓ_2-norm with ℓ_1-norm to deal with the highly correlated genes and to perform gene selection simultaneously. The RLR using elastic net penalty is defined by

$$\hat{\beta}_{Elastic} = \mathrm{argmin}_{\beta}\left[-\sum_{i=1}^{n}\{y_i \log(\pi_i)+(1-y_i)\log(1-\pi_i\} + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}{\beta_j}^2\right]. \tag{6}$$

As we observe from Eq. (6), elastic net estimator depends on two non-negative tuning parameters λ_1 and λ_2 which leads to regularized logistic regression solution.

2.2. Adaptive regularized logistic regression

According to Fan and Li [28], LASSO does not attain the oracle property. This is because LASSO is equally regularizing all the coefficients, leading the estimation to be biased. To overcome this drawback, Zou [30] proposed the adaptive LASSO where adaptive weights are assigned for regularizing different coefficients in the ℓ_1-norm penalty. By assigning the small coefficients with large weight and the large coefficients with low weight, it could be possible to reduce the selection bias, and, therefore, it can consistently select the relevant coefficients.

The regularized logistic regression using the adaptive LASSO of β is defined by:

$$\hat{\beta}_{ALASSO} = \mathrm{argmin}_{\beta}\left[-\sum_{i=1}^{n}\{y_i \log(\pi_i)+(1-y_i)\log(1-\pi_i\} + \lambda\sum_{j=1}^{p} w_j\left|\beta_j\right|\right], \tag{7}$$

where $\mathbf{w}=(w_1,\ldots,w_p)^T$ is $p\times 1$ weight vector and $w_j=(|\hat{\beta}_j|)^{-\gamma}$, where $\gamma>0$, and $\hat{\beta}$ is a root n-consistent initial value, which means that it converges to the true estimate β with $O_p(n^{-1/2})$.

In a similar way to LASSO, the elastic net does not enjoy the oracle property even though it performs much better in classification accuracy [31,32]. Additionally, Zou and Zhang [31] pointed out that the adaptive LASSO outperforms LASSO in terms of achieving the oracle property, even though the grouping effect problem for adaptive LASSO remains. As a result, the adaptive elastic net was introduced by Zou and Zhang [31] and Ghosh [32], which combines the ℓ_2-norm regularization with the adaptive LASSO. The improved regularization method, adaptive elastic net, outperforms adaptive LASSO in terms of gene selection consistency and grouping effect simultaneously. For fixed λ_2, the regularized logistic regression using the adaptive elastic net (AElastic) of β is defined by:

$$\hat{\beta}^{*}_{AElastic} = \mathrm{argmin}_{\beta}\left[\begin{array}{c}-\sum_{i=1}^{n}\{y_i^* \log(\pi_i)+(1-y_i^*)log(1-\pi_i)\} \\ +\lambda_1\sum_{j=1}^{p} w_j\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\beta_j^2\end{array}\right], \tag{8}$$

where $\mathbf{y}^* = \begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}_{(n+p)\times 1}$ is the augmented vector [26], and $w_j=(|\hat{\beta}_j|)^{-\gamma}$, $j=1,2,\ldots,p$ is the adaptive weight based on the initial estimator $\hat{\beta}$ for a positive constant γ.

2.3. The proposed method

In cancer classification, genes exhibit certain natural grouping structures; for example, gene expression profiles may be grouped according to their pathways, and it is often preferable that a group of genes are either kept or eliminated from the classification together. Furthermore, the regularization method that selects the correct subset of genes with probability tending to one is desired. The adaptive elastic net was successfully applied for gene selection in cancer classification [35,39,40].

Choosing the initial weight is crucial in the adaptive elastic net. Ghosh [32] studied the grouping effect in the adaptive elastic net by using the ordinary least squares (OLS) as the initial weight in low-dimension data. In logistic regression, MLE instead of OLS was proposed as an initial weight. In high-dimensional cancer classification, however, using MLE is simply infeasible and hence the adaptive elastic net is no longer applicable. Zou and Zhang [31], on the other hand, proposed using the elastic net as an initial weight either in low-dimensional data or high-dimensional data. Generally, the elastic net estimator is inconsistent in itself. In other words, this initial weight is biased in selecting genes. In addition, the elastic net performs well when the pairwise correlations between variables are very high. El Anbari and Mkhadri [34] stated that if the absolute correlation between genes is less than 0.95, the elastic net may be slightly less reliable. Moreover, the elastic net does not take into account the correlation structure among genes [33].

From these aforementioned drawbacks, using the elastic net estimator in adaptive elastic regularized logistic regression in high-dimensional cancer classification may not be preferable. The ratio of the standard error of the ridge regression estimator to the ridge regression estimator is proposed as the initial weight in the adaptive elastic net. According to the nature of the ℓ_2-norm, the ridge penalty tries to force the estimated coefficients of highly correlated genes to be close to each other. In particular, this property in the elastic net may help to select or omit the highly correlated genes together if their coefficients are close to each other. However, this property loses the capability for estimating the coefficients of highly correlated genes with different magnitudes, especially with different signs [41]. The advantage of using the standard error of the ridge estimator $s_{\hat{\beta}_{Ridge}}$ is to adjust the regularized logistic regression using the adaptive elastic net

Table 1
The detail information for the used datasets.

Data set	# Samples	# Genes	Classes
Prostate	102	5966	Tumor/Non-tumor
DLBCL	77	7129	DLBCL/FL
Colon	62	2000	Tumor/Normal

Table 2
Evaluation performance (on average) of the methods used according to the testing dataset over 50 partitions. The number in parenthesis is the standard error.

	# Selected genes	CA	Sen.	Sep.
Prostate				
Elastic	44(1.13)	90.64(0.51)	90.84(0.38)	90.71(0.37)
AElastic	44(1.07)	91.22(0.47)	90.90(0.37)	91.33(0.36)
AERidge	42(1.03)	90.35(0.49)	90.60(0.37)	90.31(0.35)
AAElastic	48(0.87)	93.04(0.38)	91.52(0.32)	92.80(0.35)
DLBCL				
Elastic	54(1.25)	92.35(0.46)	89.20(0.38)	93.64(0.43)
AElastic	55(1.12)	93.84(0.41)	91.07(0.36)	94.27(0.42)
AERidge	49(1.11)	91.90(0.40)	88.83(0.37)	92.68(0.44)
AAElastic	61(1.04)	95.04(0.31)	92.14(0.36)	95.08(0.42)
Colon				
Elastic	24(1.17)	93.55(0.91)	91.58(0.70)	95.84(0.73)
AElastic	24(1.08)	94.24(0.90)	91.87(0.61)	96.31(0.65)
AERidge	23(1.09)	91.74(0.66)	90.91(0.57)	93.34(0.63)
AAElastic	28(0.94)	96.40(0.64)	92.21(0.54)	96.91(0.63)

(AAElastic) when using ridge regression estimates or elastic net estimates as an initial weight. As a result, AAElastic is able to improve genes selection consistently and maintain the grouping effects simultaneously. Cule and De Iorio [42] proposed a procedure to calculate the $s_{\hat{\beta}_{Ridge}}$ depending on principal component analysis.

Table 3
Two-way ANOVA for area under the curve over 50 partitions.

Source	df	SS	MS	F	*p*-value
Methods	3	0.10563	0.03521	53.83805	0.00000
Datasets	2	0.12015	0.06007	91.86491	0.00000
Error	594	0.38847	0.00065		
Total	599	0.61426			

Table 4
P-value of Duncan's multiple range test for area under the curve.

	Elastic	AElastic	AERidge	AAElastic
Elastic		0.034	0.027	0.001
AElastic			0.013	0.012
AERidge				0.007
AAElastic				

Let $\hat{\beta}_{Ridge} = (\hat{\upbeta}_{1(Ridge)}, \ldots, \hat{\upbeta}_{p(Ridge)})^T$ be the vector of ridge regression estimate, $\mathbf{s}_{\hat{\beta}_{Ridge}} = (\mathrm{s}_{1(\hat{\beta}_{Ridge})}, \ldots, s_{p(\hat{\beta}_{Ridge})})^T$ be the vector of the standard error of the ridge regression, then $\mathbf{w}_{Ratio} = (w_{1(ratio)}, \ldots, \mathrm{w}_{\mathrm{p(ratio)}})^T$ be the ratio weight vector where$w_j = (s_{j(\hat{\beta}_{Ridge})}/|\hat{\beta}_{j(Ridge)}|)^{-\gamma}, \quad j = 1, 2, \ldots, p$. Furthermore, let $\mathbf{x}_j{}^{**} = \mathbf{x}^*{}_j / w_{Ratio,j}, j = 1, 2, \ldots, p$, then, the regularized logistic regression using AAElastic is defined as:

$$\hat{\beta}^{**}{}_{AAElastic} = \mathrm{argmin}_{\beta} \left[\begin{array}{c} -\sum_{i=1}^{n} \{ y_i^* \log(\pi_i) + (1 - y_i^*) \ln(1 - \pi_i) \} \\ + \lambda_1 \sum_{j=1}^{p} w_{Ratio,j} \left| \beta_j \right| + \lambda_2 \sum_{j=1}^{p} \beta_j{}^2 \end{array} \right]. \qquad (9)$$

Eq. (9) can be effectively solved by the coordinate descent method in *glmnet* package [38]. After solving Eq. (9), the true vector estimator $\hat{\beta}$ is calculated as:

$$\hat{\beta}_{AAElastic} = (1 + \lambda_2) \hat{\beta}^{**}{}_{AAElastic} / w_{Ratio,j}, \quad j = 1, 2, \ldots, p. \qquad (10)$$

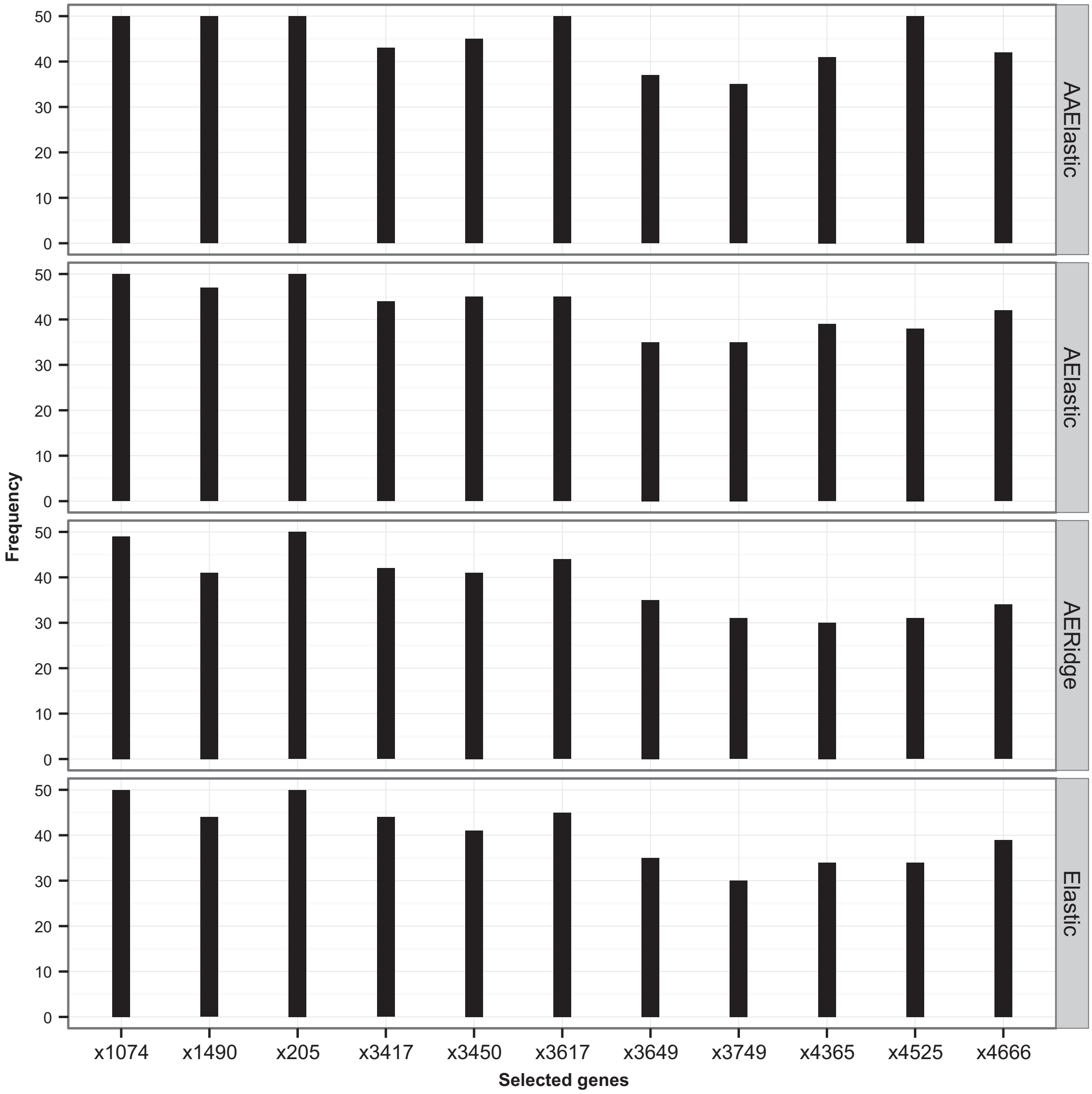

Fig. 1. The 11 most frequently selected genes from the prostate dataset.

In order to prove that our proposed method has the oracle property, the theoretical results were covered in the Supplementary file.

Theorem 1 (Oracle property). *: Suppose that* $A=\left\{j:\ \beta_j \neq 0\right\}$ *and* $\hat{A}(\lambda_1,\lambda_2)=\left\{j:\ \hat{\beta}_{AAElastic}(\lambda_1,\lambda_2)\neq 0\right\}$*. Under the regularity conditions (R1) – (R6), the adjusted adaptive elastic net (AAElastic) has the oracle property by satisfying the following:*

1. Consistency in variable selection: $\lim_{n\to\infty} p(\hat{A}(\lambda_1,\lambda_2)=A)=1$
2. Asymptotic normality: $\eta^{T\frac{(1+\lambda_2)\Sigma_A^{-1}}{2+\lambda_2}}\sqrt{\Sigma_A}\,\hat{\beta}_{AAElastic}(\lambda_1,\lambda_2)\sim N(0,\sigma^2)$, where η is a vector of norm 1, and $\Sigma_A = X_A^T X_A$.

2.4. Tuning parameter selection

For practical applications, one has to decide the values of the tuning parameters. Classically, cross-validation (CV) has been widely used. However, it is computationally intensive for AAElastic, simply because there are three tuning parameters λ_1, λ_2 and γ. For simplicity, $\gamma=1$ was used for the real data application. The λ_2 is typically assumed to take values from a range between 0 and 100. For each λ_2, the coordinate descent algorithm produces the entire solution path. Then the optimal pair of (λ_1,λ_2) is obtained using k-fold CV.

3. DNA microarray datasets application

To evaluate our proposed method, AAElastic, in the field of cancer classification, three publicly well-known binary cancer classification datasets were used. The first was the prostate cancer dataset published by [43]. It consisted of 102 samples of 52 prostate tumor samples and 50 non-tumor tissues, where each sample has 12600 genes. According to Yang et al. [44], a subset of 5966 genes was adapted in the classification by setting the intensity thresholds at 100–16000 units, then filtering out the genes with either max/min ≤ 5 or max–min ≤ 50.

The second was the diffuse large B-cell lymphoma (DLBCL) dataset published by [45]. The DLBCL dataset consisted of the gene

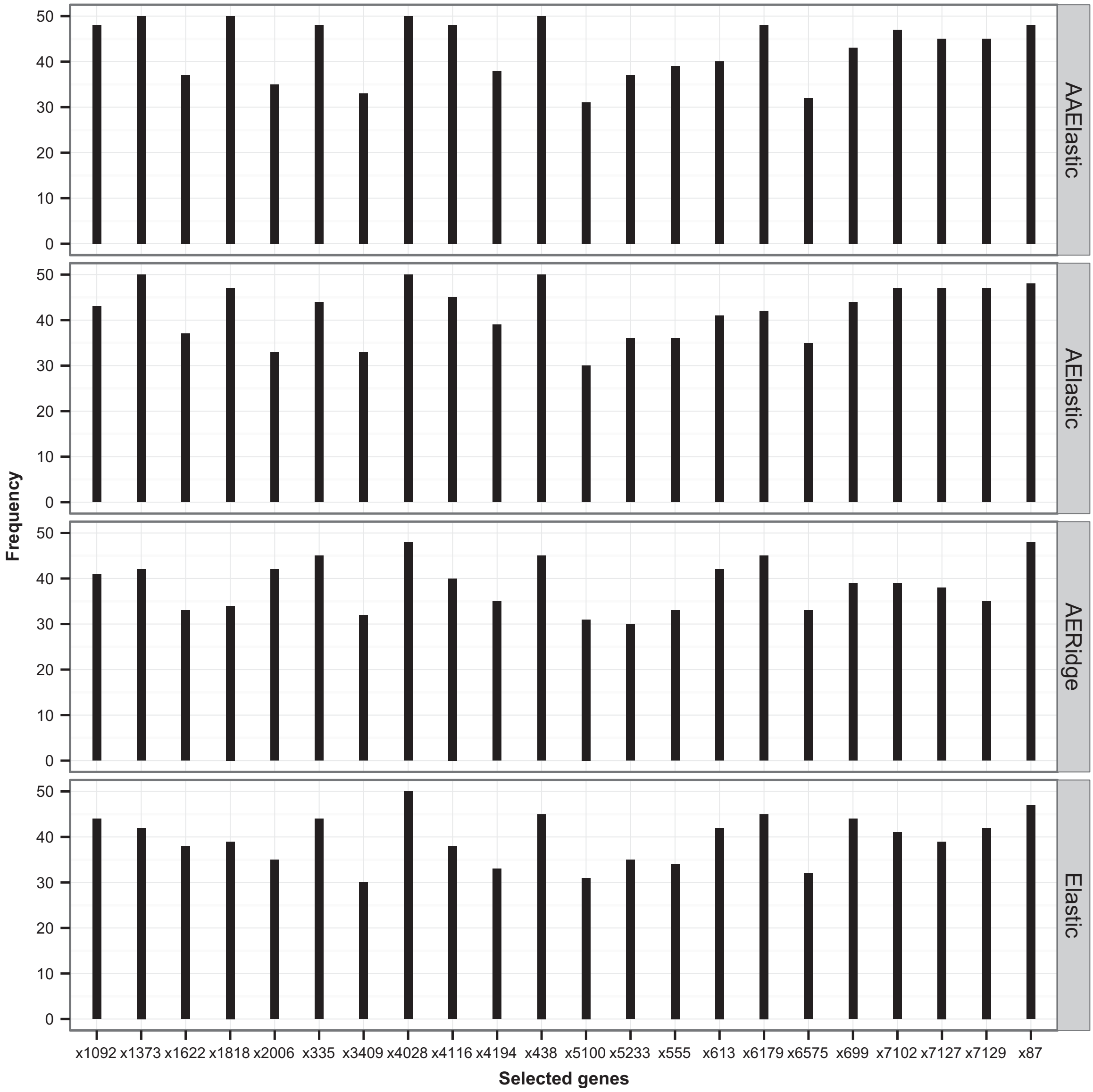

Fig. 2. The 23 most frequently selected genes from the DLBCL dataset.

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 7129 gene expression values.

The last was the colon cancer dataset published by [46]. It 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. The detailed information of these datasets is summarized in Table 1.

3.1. Performance evaluation criteria

In order to evaluate the performance of our proposed AAElastic method and to compare it with the Elastic, AElastic, and AERidge, three evaluation criteria were used depending on the testing dataset:

Classification accuracy (%) (CA)

$$\text{Classification accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \times 100\% \tag{11}$$

Sensitivity (%) (Sen.)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \tag{12}$$

Specificity (%) (Spe.)

$$\text{Specificity} = \frac{\text{FN}}{\text{FP}+\text{TN}} \times 100\% \tag{13}$$

where TP is the number of true positive, FP is the number of false positive, TN is the number of true negative, and FN is the number of false negative. Furthermore, we also performed a two-way analysis of variance (ANOVA), to show the statistical difference in the area under the curve (AUC) of the methods in the training dataset.

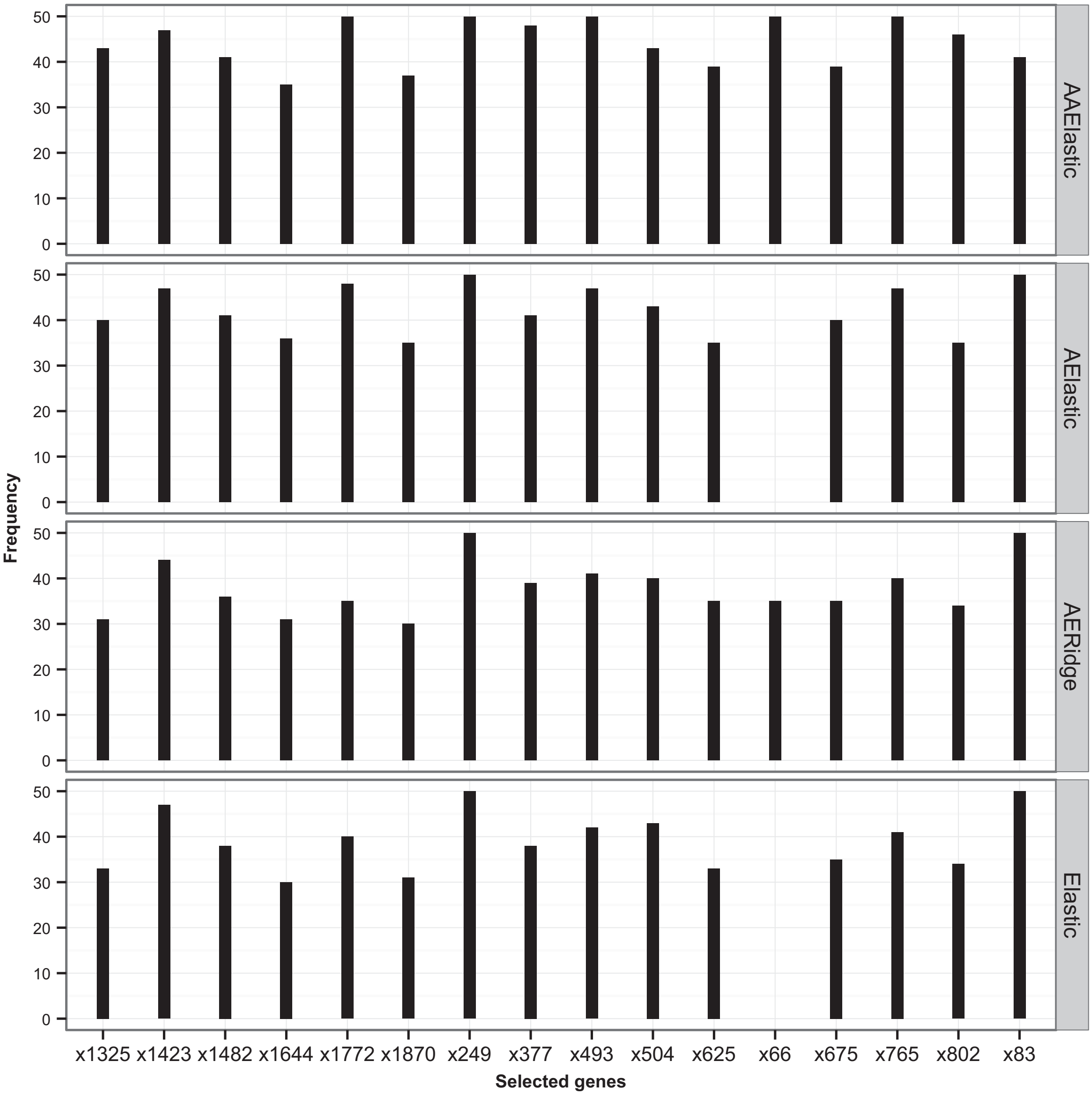

Fig. 3. The 16 most frequently selected genes from the colon dataset.

3.2. Application results

In order to enable a fair comparison, typically, each dataset was randomly partitioned into a training dataset, which comprised 70% of the samples, and a test dataset, which consisted of 30% of the samples. The partition repeated 50 times for each of the datasets. In order to obtain the best value of the pair (λ_1, λ_2), the 10-fold CV was employed using the training dataset. All the applications were conducted in R using the *glmnet* package. The average number of selected genes, the average classification accuracy, the average sensitivity, and the average specificity are presented in Table 2.

As can be seen from Table 2, AAElastic selected more genes than the other three methods for all the datasets. In DLBCL, for instance, AAElastic selected 61 genes compared to 54, 55, and 49 genes for Elastic, AElastic, and the AERidge, respectively. Importantly, AAElastic had the potential to select more genes than the other three methods, indicating that most of these additionally selected genes were probably not highly correlated.

In terms of classification accuracy, AAElastic achieved a maximum accuracy of 93.04%, 95.04% and 96.40% for prostate, DLBCL, and colon datasets, respectively. Furthermore, it is clear from the results that AAElastic outperformed the AERidge in terms of classification accuracy for all datasets. This improvement in classification accuracy is mainly due to the AAElastic ability in taking into account the standard error of the ridge regression. Moreover, AAElastic slightly improved the classification accuracy compared to AElastic. The improvements were 2.00%, 1.27%, and 2.29% for the prostate, DLBCL, and colon datasets.

It can also be seen from Table 2 that AAElastic has the best results in terms of the sensitivity and specificity. AAElastic has the largest sensitivity of 91.52%, 92.14%, and 92.21% for the prostate, DLBCL, and colon datasets, respectively. This indicated that AAElastic significantly succeeded in identifying the patients who in fact have the cancer with a probability of 0.915, 0.921, and 0.922, respectively.

On the other hand, the results for the specificity represent the probability of an adaptive regularized logistic regression method in identifying the patients who are normal. In terms of the specificity, AAElastic significantly outperformed the AElastic, AERidge, and Elastic for all datasets. In the prostate cancer dataset, for example, AAElastic has the largest probability of 0.928 in identifying the normal patients compared to 0.913, 0.903, and 0.907 for AElastic, AERidge, and Elastic, respectively.

To further highlight the classification stability for the proposed method, the AAElastic seeks to prove that it can classify high-dimensional cancer data with a high degree of accuracy compared to the other three methods used. Depending on the training dataset, a two-way ANOVA was used to check whether the AAElastic, AElastic, AERidge, and the Elastic were statistically significant, and if there was any significant difference between the three datasets used in terms of AUC. Table 3 reports the two-way ANOVA results. From Table 3, the results showed statistically significant differences between the AAElastic and the three other methods used in terms of AUC. In addition, we can see that the prostate, DLBCL, and colon datasets had different area under the curve values.

Furthermore, Duncan's multiple range test was used to obtain more detailed information about the differences between the AAElastic and the three other methods used. Table 4 lists the *p*-value of each compared pair of methods. It is apparent from Table 4 that the AAElastic showed statistical differences compared to the AElastic, AERidge, and Elastic in terms of AUC.

Besides classification accuracy, gene selection consistency is another aspect associated with adaptive regularized logistic regression. To measure the consistency of gene selection, Figs. 1–3 display the frequency of the top selected genes by the AAElastic, AElastic, AERidge, and the Elastic for prostate, DLBCL, and colon datasets, respectively.

As we can see from Fig. 1, only 11 genes were frequently selected by all methods. It is clearly seen that AAElastic showed more consistency in selecting the top genes. For example, it successfully

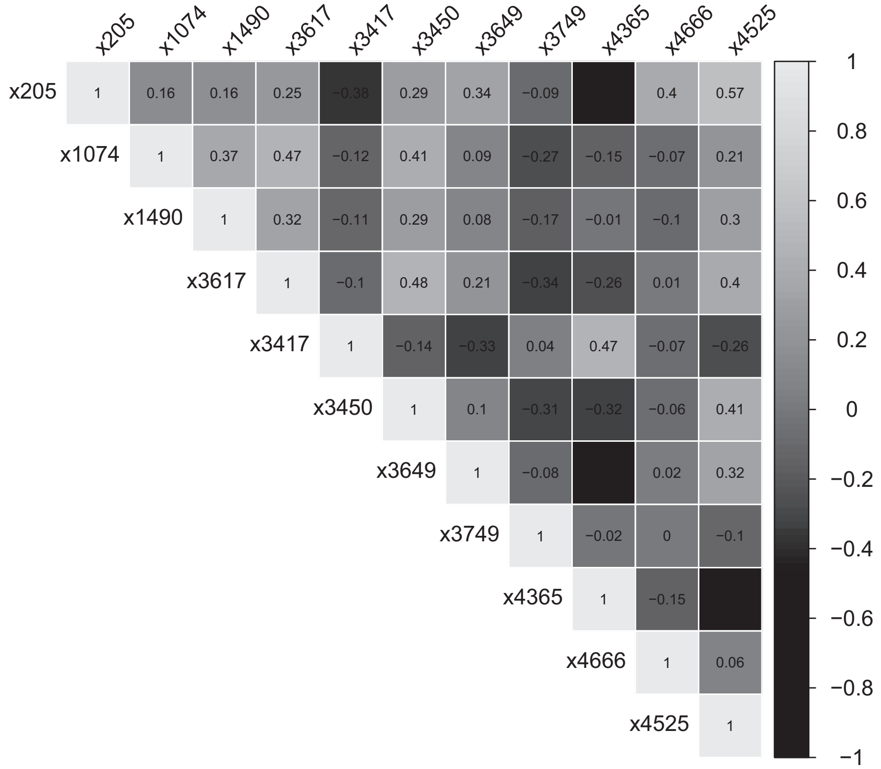

Fig. 4. The correlation matrix among the most frequently selected genes of the prostate dataset.

selected the gene index (name) × 1074 (H.sapiens cDNA), × 1490 (H. sapiens ABC), × 205 (H.sapiens mRNA for RET), × 3617 (H.sapiens GSTA4 mRNA), and × 4525 (hepatoma mRNA for serine protease hepsin) with probability equal to 1, while the other three methods only selected × 1074 and × 205 with a probability equal to 1. By looking at the correlation among the 11 top selected genes from Fig. 4, the correlation between × 205 and × 4525 was 0.568, which is not very high, but AAElastic selected these two genes together with 100%.

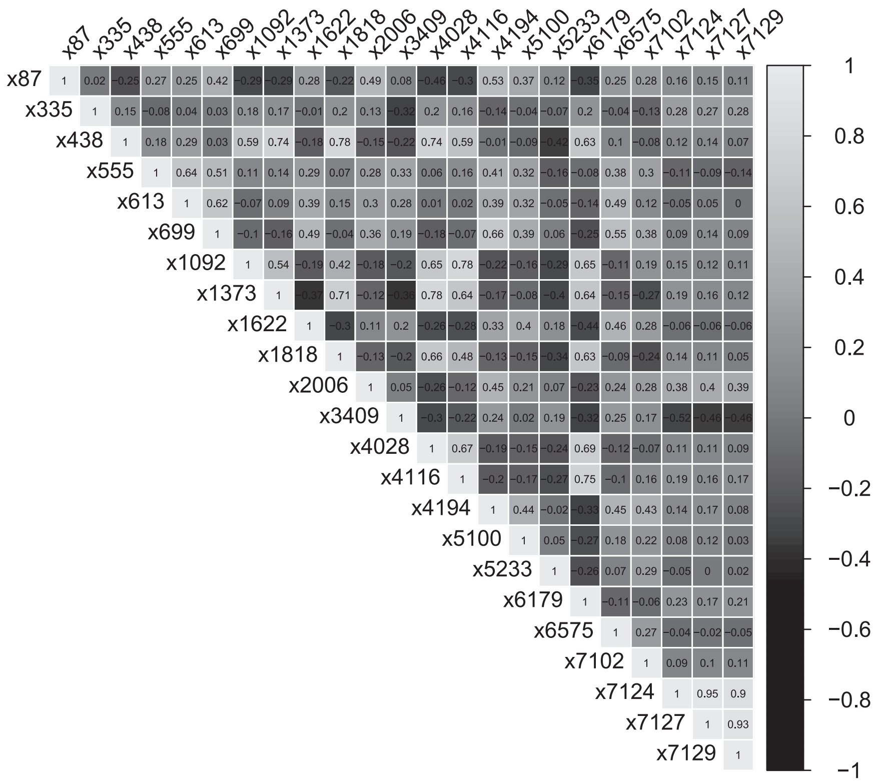

Fig. 5. The correlation matrix among the most frequently selected genes of the DLBCL dataset.

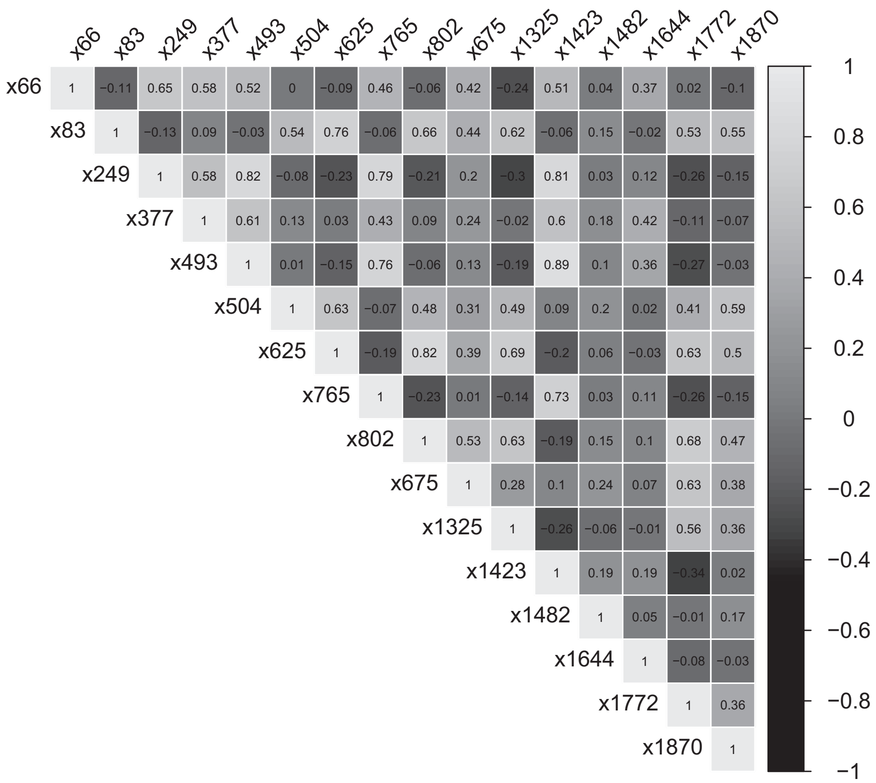

Fig. 6. The correlation matrix among the most frequently selected genes of the colon dataset.

Compared to the AElastic and AERidge, they selected these genes with a percentage of 74% and 62%, respectively. Hence, the ability of AAElastic in selected correlated genes with no high correlation can be inferred.

Similar to the prostate dataset, AAElastic provided consistent gene selection for the DLBCL dataset. Among the top 23 frequently selected genes (Fig. 2), six genes: × 1373 (Macrophage migration inhibitory factor (MIF)), × 1818 (heat shock 60 kDa protein 1 (chaperonin)), × 4028 (lactate dehydrogenase A), × 4116 (ALDOA Aldolase A), × 438 (T-COMPLEX PROTEIN 1), and × 6179 (enolase 1, (alpha)), frequently appeared together in all the methods. It is apparent that AAElastic consistently selected them with a probability of 0.96 compared to 0.84, 0.78, and 0.78 of AElastic, AERidge, and Elastic, respectively. By checking the correlation matrix from Fig. 5, we can observe that the correlations among these six genes range between 0.53 and 0.78. This could be the reason why the AAElastic selected these six genes together more frequently compared to the other three methods.

Again, from Fig. 4, we can see that the AAElastic is more consistent than the other three methods. It selected genes × 1772 (Homo sapiens), × 249 (Human desmin gene, complete cds), × 493 (MYOSIN HEAVY CHAIN, NOUMUSCLE), × 66 (HUMAN), and × 765 (SMOOTH MUSCLE) together with a percentage of 100%. However, the AAElastic was selected × 83 (Human mRNA) less than AElastic, AERidge, and Elastic, where × 83 achieved correlation with the range between 0.53 and 0.75 with some selected genes (Fig. 6). In contrast, AAElastic selected gene x66 with a percentage of 100%, while both AElastic and Elastic failed to selected it, although gene × 66 has a correlation equal to 0.52, 0.51, and 0.65 with × 493, × 1432 (H.sapiens mRNA for p cadherin), and × 249, respectively.

Overall, it is obvious that the microarray real datasets results demonstrated the use of AAElastic in terms of classification accuracy for both the training and testing datasets, sensitivity, and specificity. Additionally, it outperformed the AElastic, AERidge, and Elastic in terms of gene selection consistency. Furthermore, it is clear from the application results that for the values of the pairwise correlations, AAElastic dominates the other three methods via grouped selection.

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 computational methods have failed to identify a small subset of important genes. To tackle both estimating the gene coefficients and performing gene selection simultaneously, adaptive regularized logistic regression was successfully applied in high-dimensional cancer classification. In this research, we proposed AAElastic for consistent gene selection and encouraging the grouping effect simultaneously in high-dimensional cancer classification. From the results, which were based on three microarray real datasets, it was proved that AAElastic was competitive, effective, and yielded positive results in terms of (a) classification accuracy, sensitivity, and specificity, and (b) consistency in gene selection. Furthermore, it is clear from the application results that for the values of the pairwise correlations, AAElastic dominates the other three methods via grouped selection. Therefore, we can conclude the effectiveness of the proposed AAElastic method in practice.

Summary

A proposed penalized method as a tool for gene selection, adjusted adaptive regularized logistic regression (AAElastic), is employed in high-dimensional cancer classification. AAElastic can perform consistency selection and deal with grouping effects simultaneously. Compared to other commonly used regularization methods, the results show that not only does AAElastic obtain the best classification ability by consistency selection, but also by encouraging the grouping effects in selecting more correlated genes.

Conflict of Interest Statement

None declared.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2015.10.008.

References

[1] J. Kalina, Classification methods for high-dimensional genetic data, Biocybern, Biomed. Eng. 34 (2014) 10–18.

[2] S. Ma, J. Huang, Penalized feature selection and classification in bioinformatics, Brief. Bioinform. 9 (2008) 392–403.

[3] A. Kastrin, B. Peterlin, Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data, Expert. Syst. Appl. 37 (2010) 5178–5185.

[4] B. Chandra, M. Gupta, An efficient statistical feature selection approach for classification of gene expression data, J. Biomed. Inform. 44 (2011) 529–535.

[5] E. Lotfi, A. Keshavarz, Gene expression microarray classification using PCA–BEL, Comput. Biol. Med. 54 (2014) 180–187.

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

[7] S. Zheng, W. Liu, An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification, Comput. Biol. Med. 41 (2011) 1033–1040.

[8] C.-H. Zheng, Y.-W. Chong, H.-Q. Wang, Gene selection using independent variable group analysis for tumor classification, Neural. Comput. Appl. 20 (2011) 161–170.

[9] Y. Cui, C.-H. Zheng, J. Yang, W. Sha, Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data, Comput. Biol. Med. 43 (2013) 933–941.

[10] S. Kar, K. Das Sharma, M. Maitra, Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique, Expert. Syst. Appl. 42 (2015) 612–627.

[11] D. Du, K. Li, X. Li, M. Fei, A novel forward gene selection algorithm for microarray data, Neurocomputing 133 (2014) 446–458.

[12] I. Kamkar, S.K. Gupta, D. Phung, S. Venkatesh, Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso, J. Biomed. Inform. (2014).

[13] Y. Lei, H. Yue, M.E. Berens, Stable gene selection from microarray data via sample weighting, IEEE Trans. Comput. Biol. Bioinform. 9 (2012) 262–272.

[14] H. Peng, Y. Fu, J. Liu, X. Fang, C. Jiang, Optimal gene subset selection using the modified SFFS algorithm for tumor classification, Neural. Comput. Appl. 23 (2013) 1531–1538.

[15] X. Nan, N. Wang, P. Gong, C. Zhang, Y. Chen, D. Wilkins, Biomarker discovery using 1-norm regularization for multiclass earthworm microarray gene expression data, Neurocomputing 92 (2012) 36–43.

[16] S. Winham, C. Wang, A. Motsinger-Reif Alison, A comparison of multifactor dimensionality reduction and L1-penalized regression to identify gene-gene interactions in genetic association studies, Stat. Appl. Genet. Mol. Biol. 10 (2011) 1–25.

[17] Y. Liang, C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z.-B. Xu, H. Zhang, Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification, BMC Bioinform. 14 (2013) 198.

[18] C. Bielza, V. Robles, P. Larrañaga, Regularized logistic regression without a penalty term: an application to cancer classification with microarray data, Expert. Syst. Appl. 38 (2011) 5110–5118.

[19] G.C. Cawley, N.L.C. Talbot, Gene selection in cancer classification using sparse logistic regression with Bayesian regularization, Bioinformatics 22 (2006) 2348–2355.

[20] S.K. Shevade, S.S. Keerthi, A simple and efficient algorithm for gene selection using sparse logistic regression, Bioinformatics 19 (2003) 2246–2253.

[21] J. Zhu, T. Hastie, Classification of gene microarrays by penalized logistic regression, Biostatistics 5 (2004) 427–443.

[22] S. Li, T. Eng Chong, Dimension reduction-based penalized logistic regression for cancer classification using microarray data, IEEE Trans. Comput. Biol. Bioinform. 2 (2005) 166–175.

[23] L. Zhenqiu, J. Feng, T. Guoliang, W. Suna, S. Fumiaki, T. Ming, Sparse logistic regression with Lp penalty for biomarker identification, Stat. Appl. Genet. Mol. Biol. 6 (2007) 1–22.

[24] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.

[25] Z.F. Zeny, Y. Xiaojian, S. Sanjeena, D.M. Paul, The LASSO and sparse least squares regression methods for SNP selection in predicting quantitative traits, IEEE Trans. Comput. Biol. Bioinform. 9 (2012) 629–636.

[26] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–320.

[27] J. Fan, Y. Fan, E. Barut, Adaptive robust variable selection, Ann. Stat. 42 (2014) 324–351.

[28] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc. 96 (2001) 1348–1360.

[29] R. Alhamzawi, K. Yu, D.F. Benoit, Bayesian adaptive Lasso quantile regression, Stat Model. 12 (2012) 279–297.

[30] H. Zou, The adaptive lasso and its oracle properties, J. Am. Stat. Assoc. 101 (2006) 1418–1429.

[31] H. Zou, H.H. Zhang, On the adaptive elastic-net with a diverging number of parameters, Ann. Stat. 37 (2009) 1733–1751.

[32] S. Ghosh, On the grouped selection and model complexity of the adaptive elastic net, Stat. Comput. 21 (2011) 451–462.

[33] P. Bühlmann, P. Rütimann, S. van de Geer, C.-H. Zhang, Correlated variables in regression: Clustering and sparse estimation, J. Stat. Plan. Inference 143 (2013) 1835–1858.

[34] M. Anbari, A. Mkhadri, Penalized regression combining the L 1 norm and a correlation based penalty, Sankhya B 76 (2014) 82–102.

[35] J. Li, Y. Jia, Z. Zhao, Partly adaptive elastic net and its application to microarray classification, Neural Comput. Appl. 22 (2013) 1193–1200.

[36] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser B (Methodological) 58 (1996) 267–288.

[37] M.Y. Park, T. Hastie, Penalized logistic regression for detecting gene interactions, Biostatistics 9 (2008) 30–50.

[38] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.

[39] X. Chen, Adaptive elastic-net sparse principal component analysis for pathway association testing, Stat. Appl. Genet. Mol. Biol. 10 (2011) 1–23.

[40] J.-T. Li, Y.-M. Jia, An improved elastic net for cancer classification and gene selection, Acta Automat. Sin. 36 (2010) 976–981.

[41] S. Wang, B. Nan, S. Rosset, J. Zhu, Random lasso, Ann. Appl. Stat. 5 (2011) 468–485.

[42] E. Cule, M. De Iorio, Ridge regression in prediction problems: Automatic choice of the ridge parameter, Genet. Epidemiol. 37 (2013) 704–714.

[43] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.

[44] K. Yang, Z. Cai, J. Li, G. Lin, A stable gene selection in microarray data analysis, BMC Bioinform. 7 (2006) 228–243.

[45] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med., 8, (2002) 68–74.

[46] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. 96 (1999) 6745–6750.