Informatics in Medicine Unlocked 24 (2021) 100622

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Gene selection and classification of microarray gene expression data based on a new adaptive L_1-norm elastic net penalty

Aiedh Mrisi Alharthi [a,b], Muhammad Hisyam Lee [a,*], Zakariya Yahya Algamal [c]

[a] Department of Mathematical Sciences, Universiti Teknologi Malaysia, Skudai, Malaysia
[b] Department of Mathematics, Taif University, Taif, Saudi Arabia
[c] Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

ARTICLE INFO

Keywords:
Adapted elastic net
Penalized logistic regression
Gene selection
Cancer diagnosis

ABSTRACT

The removal of irrelevant and insignificant genes has always been a major step in microarray data analysis. The application of gene selection methods in biological datasets has greatly increased, supporting expert systems in cancer diagnostic capability with high classification accuracy. Penalized logistic regression (PLR) using the elastic net (EN) has been widely used in high-dimensional cancer classification in recent years to estimate the gene coefficients and perform gene selection simultaneously. However, the EN estimator does not satisfy the oracle properties. This paper proposes the PLR using the adaptive elastic net (AEN), abbreviated as PLRAEN, to address the inconsistency. Our method employs the ratio (BWR) as an initial weight inside the L_1-norm of the EN model. Several experiments were performed on a simulation study for a different number of predictor variables, sample sizes, and correlation coefficients and also on three public gene expression datasets to evaluate the effectiveness. Experimental results demonstrate that the proposed method consistently outperforms two other contemporary penalized methods regarding classification accuracy and the number of selected genes. Therefore, we conclude that PLRAEN is a better method to implement gene selection in the high-dimensional cancer classification field.

1. Introduction

New technologies address the immense growth of data. These technologies help researchers transfer huge chunks of information into organized data. Big data might have irrelevant or redundant features (gene expressions). Therefore, researchers prefer to pick important genes by selecting a small subset of significant features from available datasets. Gene selection speeds up the learning process and improves the work of the model [1,2]. Using microarray technology, researchers can classify both cancerous and normal tissues, depending on gene expression profiles. Recently, many studies were conducted on gene expression datasets to determine the varieties of cancer. They also forecast clinical results to diagnose patients with cancer [3–5].

Microarray datasets of gene expression have many properties that obstruct the evolution of these techniques. One of these properties is the high dimensionality of the datasets. The gene expression dataset involves several genes, p, with only a limited number of observations, n. This means, in the matrix representing gene expressions, the number of columns is much larger than the number of rows, $p \gg n$ [6]. Another problem is that microarray data usually suffers from a high level of technical noise. Therefore, it is crucial to overcome these two problems to reasonably increase the Classification Accuracy (CA) associated with microarray data [7].

In the last three decades, statisticians have developed many selection methods to select important genes. These methods fall into three main categories: First, the filter category. It involves the most popular feature selection methods, where each gene is independently examined regardless of its group performance. The second is the wrapper category. It uses various algorithms to evaluate the process of selecting gene groups. Although the wrapper methods are more efficient in feature selection than the filter methods, they are computationally expensive, such as forward gene selection and backward gene elimination. The third is the embedded category, which combines the advantages of the filter and wrapper categories. It includes regularization (penalizing) methods that can simultaneously perform modeling and gene selection [8–10].

Penalized logistic regression (PLR) is one of the most widely used penalty-based regularization methods. It is used to select genes and

* Corresponding author.
E-mail addresses: aiedh.harthi@gmail.com (A.M. Alharthi), mhl@utm.my (M.H. Lee), zakariya.algamal@uomosul.edu.iq (Z.Y. Algamal).

https://doi.org/10.1016/j.imu.2021.100622
Received 12 April 2021; Received in revised form 18 May 2021; Accepted 25 May 2021
Available online 29 May 2021

classify them. Penalized methods belong to the class of embedded methods, efficient in selecting and classifying genes. In recent years, logistic regression (LR) has received tremendous consideration. Penalized methods add a kind of penalty term to the LR to perform selection and classification altogether. Many LR models can be used with different penalties. One of these penalties is called "Least Absolute Shrinkage and Selection Operator" (also known as LASSO). LASSO is based on L_1-norm [11]. Another penalty method is the so-called "Smoothly Clipped Absolute Deviation" (SCAD) [12]. Other penalties are the elastic net [13], the adaptive L_1-norm [14] and the adaptive elastic net (AEN) methods [15,16].

Although LASSO is capable of selecting features, it has three shortcomings [17,18]. First, it is related to the number of features that LASSO selects. When the dataset is high dimensional, LASSO cannot select more genes than the sample size. The sample size selected by LASSO is bound above by n. Second, the LASSO fails to consider the grouping effect when selecting genes. LASSO is expected either to select the whole group of highly correlated genes (if they are related to the disease) or to leave it all (if they are unrelated). However, LASSO selects only one or a few members of each highly correlated group of genes related to the study. Hastie and Zou [13] proposed the EN to overcome some of these shortcomings. The EN method employs a penalty that is linearly composed of L_1-norm and L_2-norm. Third, the LASSO method is biased in gene selection because it penalizes all gene coefficients on an equal basis. As a result of this weakness, LASSO lacks the oracle properties [12]. To solve the challenges due to the lack of oracle properties, Zou [14] developed a new regularization technique called the adaptive LASSO technique (ALASSO). Some weights were used to penalize each coefficient inside the L_1-norm-based penalty. In the ALASSO, modified weights are used to penalize the coefficients in the L_1-norm-based penalty.

The L_1-norm penalty model is one of the most common approaches in penalized methods. A drawback of the L_1-norm penalty model is that it equally penalizes all genes, causing the selection procedure to be inconsistent [12,14]. In this study, a penalized logistic regression model with adaptive elastic net (PLRAEN) is proposed to improve the gene selection performance. This is done by employing a ratio (BWR) as an initial weight inside the L_1-norm with the EN model to classify people concerning catching cancer correctly. This weight, in some sense, reflects the importance of genes individually. Experiments demonstrate that our method, compared to other similar methods, has the highest selection accuracy.

The main contributions of this paper are summarized as follows.

- This paper proposes PLRAEN to address inconsistencies in gene selection and classification.
- The ratio BWR is employed as an initial weight inside the L_1-norm of the EN model.
- PLRAEN has the adaptability advantage over the other used penalized methods in encouraging grouping effect and selecting genes consistently in high dimensional data with logistic regression models.
- The proposed method can be seen effectively under a different range of correlation values.

Besides this introduction, the previous related work is reviewed in Section 2. Section 3 provides a brief introduction to related works concerning penalized LR models. The PLRAEN method is presented in Section 4. Some evaluation metrics are presented in Sections 5. The results and the experimental study intended to evaluate the efficiency of PLRAEN compared to the EN and the AEN methods are presented and debated in Section 6. This paper is then concluded in Section 7.

2. Related work

Conventionally, statistical learning methods have been used to select genes independently. Among these methods, a general hybrid adaptive classifier ensemble [19], Nested cross-validation with ensemble feature selection and classification model [20], as well as support vector machine (SVM) and its extensions [21,22], have been commonly used in cancer classification for gene selection. The L_1-penalized logistic regression is becoming increasingly relevant and popular when dealing with high data and focusing on feature selection and classification performance. However, when the penalties of various coefficients are all the same and unrelated to the data, the LASSO estimates can be problematic. Several previous studies have proposed methods to select the genes more efficiently by adding various penalty techniques. LASSO and its extensions [23–27] have been used to select genes using the L_1 norm penalty in logistic regression. Penalized logistic regression has been constructed using a Bayesian regularization term [28,29]. A few such methods use multi-stage sparse logistic regression models with L_1 norm penalty [5], while others use AEN [15,30], SCAD penalty [31] and weighted L_1 penalty. These approaches have been successfully applied to gene selection and improve classification accuracy. However, none of the previous works propose a ratio (BWR) as an initial weight inside the L_1-norm with the EN model for gene selection in cancer classification.

3. Penalized logistic regression models

LR is one of the most popular machines learning algorithms for binary classification, where the response variable values are coded as zero (0) and one (1). For example, while classifying cancer, the response variable takes either (1) for cancerous cases or (0) for non-cancerous cases. In various classification fields, classical LR with a penalty is incorporated to perform gene selection and classification simultaneously. In this study, the PLR model is used to address the gene expression classification problems. It penalizes the model because there are too many genes. In LR, the regression equation is non-linearly related to the linear combination of the predictor variables.

For illustration purposes, gene expression profiles are often represented as a matrix $\mathbf{X} \in R^{n\times p}$ $(n \ll p)$, where each column denotes a gene, and each row indicates a sample. The entry $x_{i,j}$ indicates the expression value of the j^{th} gene of the i^{th} sample and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is the i^{th} input sample. Let $\mathbf{y} = (y_1, ..., y_n)^T$ be the response vector, where y_i is the corresponding classification label that takes values of (0) or (1). The response variable $\mathbf{y}$ is classified according to a linear combination of the $n \times p$ matrix with real entries; written $\mathbf{X}^T\boldsymbol{\beta}$. The symbol $\mathbf{X}^T$ denotes the transpose of the design matrix $\mathbf{X}$ and $\boldsymbol{\beta}$ is a vector of the dimension p of the unknown coefficients $(\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T)$. In general, in LR, the response variable y has a Bernoulli distribution, and the probability that y is equal to 1 given the value of $x$is denoted as $\pi(x)$ is

$$p(y_i = 1 | x_{ij}) = \pi(x_i) = \frac{\exp\left(x_j^T \beta_j\right)}{1 + \exp\left(x_j^T \beta_j\right)}, j = 1, 2, ..., p \tag{1}$$

$$f(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, i = 1, 2, ..., n \tag{2}$$

The likelihood function of LR is given as

$$L(\beta, y_i) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \tag{3}$$

Then, the log-likelihood function is:

$$\ell\left(\beta, y_i\right) = \sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)\} \tag{4}$$

LR is a powerful discriminative method used in classification (variable selection). Despite its efficiency in linear models related to regular data, the LR is not applicable as a classification tool when the dataset is high dimensional because the design matrix is not invertible. Consequently, it fails to provide reliable estimates for the regression

coefficients. Additionally, when datasets of genes are high dimensional, for example, when there are several genes (or features in general), the overfitting problem arises. Moreover, its estimators may also suffer from multicollinearity [32].

From a statistical viewpoint, the other (unrelated) genes might generate noise and lower the classification performance. Therefore, statisticians usually prefer to apply gene selection methods that can remove unrelated and redundant genes to improve CA. Besides LR, the classification methods available include the penalizing logistic regression method used to eliminate high dimensionality and improve the CA [33]. Although penalization methods are commonly used in high dimensionality, Doerken et al. [34] demonstrated that the methods could also perform well in low dimensional data.

In PLR, a positive penalty term is added to the log-likelihood function forcing some coefficients to become zero to obtain a sparse solution. PLR imposes a penalty term on the equation of the logistic model that has too many genes. Accordingly, under some constraint on the coefficients, the coefficients of less contributive genes become either very close to zero or exactly zero. This process is also known as regularizations. The setting of the method is as follows.

The penalized log-likelihood equation is expressed as

$$\mathrm{PLR} = -\ell(\boldsymbol{\beta}, y_i) + \lambda g(\boldsymbol{\beta}) \tag{5}$$

where, $\ell(\boldsymbol{\beta}, y_i)$ denotes the log-likelihood as Eq. (4), $g(\boldsymbol{\beta})$ denotes the penalty term, and λ is a regulation factor used to tune the penalty amount. Then the PLR of Eq. (5) is minimized with respect to the λ to find the coefficients estimates. The penalty is used to decrease the estimates' variances and force them to be biased, resulting in improved prediction accuracy [35]. These penalizing methods are from a class of embedded selection methods frequently used in classification and feature selection in high-dimensional datasets [36].

Before solving the PLR minimization problem, let the response vector $\mathbf{y}$ is centered and the columns of $\mathbf{X}$ (genes) are usually standardized so that $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 1$, and $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, 2, \ldots, p$. Standardization sets the intercept term (β_0) to zero. $\boldsymbol{\beta}$ is estimated using LASSO (L_1-norm penalization) as follows.

$$\widehat{\beta}_{LASSO} = \mathrm{argmin}_{\beta}\left[-\sum_{i=1}^{n}\{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{6}$$

where, λ is the tuning parameter. When $\lambda = 0$, Eq. (6) decreases to the usual minimum likelihood estimator. As $\lambda \to \infty$, penalization forces all predictor variables to be zero.

Another important penalized method that is used in gene selection is the EN. It was invented by Hastie and Zou [13] to address the weaknesses of LASSO. EN combines L_2, and L_1 norms to address genes with high correlation and select relative genes at once. PLR, based on EN penalty, is given in the following equation:

$$\widehat{\beta}_{Elastic} = \mathrm{argmin}_{\beta}\left[-\sum_{i=1}^{n}\{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right] \tag{7}$$

Eq. (7) indicates the EN estimator depends on two regulation factors that assume only non-negative values, λ_1 and λ_2. Eq. (7) gives a PLR solution.

The ALASSO technique was first introduced by Zou [14] to solve the overestimation problem of LASSO by replacing the L_1 penalty with weighted penalty [37]. Zou amended the L_1-penalty by assigning different weights to different coefficients. The assigned weights could be based on Ridge, LASSO, or other shrinkage techniques. The penalized logistic model associated with ALASSO is defined as follows:

$$\widehat{\beta}_{LASSO} = \mathrm{argmin}_{\beta}\left[\sum_{i=1}^{n}\{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\left(\left|\widehat{\beta}_j^{initial}\right|\right)^{\gamma}} \right], \tag{8}$$

where, $\lambda, \gamma \geq 0$ and $\widehat{\beta}_j^{initial}$ is an initial estimate for each β_j estimated using the LASSO technique or other shrinkage techniques. Here we set $\gamma = 1$, for simplicity.

Like the EN method, other penalized regression methods can achieve grouping effect, such as AEN methods proposed by Refs. [15,16], who proposed two AEN estimators. They added the adaptive weight into the L_1-norm penalty within the EN. The two AEN approaches are different in their adaptive weights. Zou and Zhang [15] construct the adaptive weight using the EN estimator. However, Ghosh [16] used the least-squares estimator to construct the adaptive weight. For fixed λ_2, the PLR using AEN of β is given by:

$$\widehat{\beta}_{AElastic} = \mathrm{argmin}_{\beta}\left[-\sum_{i=1}^{n}\{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda_1 \sum_{j=1}^{p} w_j|\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right] \tag{9}$$

where $w_j = (|\widehat{\beta}|)^{-\gamma}$, $j = 1, 2, \ldots, p$ is the modified weight produced by the $\widehat{\beta}$ initial estimator. Here γ is a non-negative constant. Eqs. (6)–(9) are solved by an algorithm called coordinate descent [35].

4. The proposed method

Regarding gene expression classification, classification efficiency should be enhanced to provide a reliable gene selection process and a deeper understanding of the classification question. High dimensionality may negatively impact a classifier's classification efficiency by raising the possibility of overfitting and extending the computation time. Furthermore, specific classification approaches are not explicitly applicable to the study of microarray gene expression data. When implementing classification methods to analyze data-on-data sets of gene expressions, it is important to exclude unrelated genes from the datasets to guarantee accuracy.

It is noticed that when the correlations between each pair of genes are very high, the EN method works efficiently. The authors of [38] noticed that if the genes are not highly correlated ($|r|$ is less than 0.95), the reliability of EN somewhat decreases. Another problem is that EN fails to consider the correlation nature of genes [39]. Moreover, Zou and Zhang [15] noticed that EN does not satisfy the oracle property and that the grouping effect problem remains. To solve the EN problems, the AEN was established by Zou and Zhang [15], and Ghosh [16] by adding an L_2-norm penalty to ALASSO.

Initially, to select genes, Dudoit et al. [40] performed it based on the ratio (BWR) of the sum of squares between gene groups (BSS) to the sum of squares for each gene within groups (WSS), defined as

$$BWR(j) = \frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)\left(\bar{x}_{kj} - \bar{x}_{.j}\right)^2}{\sum_i \sum_k I(y_i = k)\left(x_{ij} - \bar{x}_{kj}\right)^2}, \tag{10}$$

where, $I(.)$ is an indicator function, $\bar{x}_{.j}$ is the mean of column j that represents the expression level of gene j across all samples and $\bar{x}_{kj}$ is the mean of the values of gene j across samples of class k, where $k = 2$ [40,41]. In this study, we also have exactly two classes.

Selecting the initial weight is critical for AEN. Therefore, to improve

Table 1
Confusion matrix of classification.

	Prediction (+)	Prediction (−)
Actual (+)	True Positive	(TP) False Negative (FN)
Actual (−)	False Positive (FP)	True Negative (TN)

Table 2
Classification and variable selection performance of the PLRAEN and the competitor methods over 100 partitions when $\rho = 0.55$.

Methods	Model 1			Model 2		
	CA%	TP	FP	CA%	TP	FP
$n = 100, p = 1,000$						
EN	91.12 (0.08)	5	28	92.32 (0.11)	6	27
AEN	92.00 (0.09)	5	22	92.00 (0.09)	5	25
Proposed	96.00 (.007)	6	17	94.03 (0.07)	6	18
$n = 100, p = 5,000$						
EN	92.00 (0.06)	6	23	92.00 (0.07)	7	36
AEN	92.00 (0.09)	5	22	92.00 (0.14)	7	34
Proposed	94.21 (0.05)	6	13	96.30 (0.11)	8	20
$n = 100, p = 10,000$						
EN	86.20 (0.07)	8	29	89.11 (0.06)	8	38
AEN	89.58 (0.11)	8	28	91.84 (0.12)	7	39
Proposed	92.00 (0.05)	8	21	96.04 (0.05)	8	22
$n = 200, p = 1,000$						
EN	86.02 (0.07)	5	29	88.23 (0.06)	6	26
AEN	88.31 (0.09)	5	21	91.10 (0.16)	6	25
Proposed	94.41 (0.05)	6	17	95.00 (0.07)	8	18
$n = 200, p = 5,000$						
EN	92.00 (0.07)	6	23	92.00 (0.14)	7	36
AEN	93.24 (0.09)	6	22	93.18 (0.17)	7	33
Proposed	95.00 (0.06)	6	14	96.04 (0.12)	8	21
$n = 200, p = 10,000$						
EN	86.00 (0.07)	8	29	88.17 (0.06)	8	38
AEN	91.05 (0.09)	7	30	91.00 (0.16)	7	38
Proposed	95.00 (0.07)	8	22	94.08 (0.08)	8	22
$n = 300, p = 1,000$						
EN	86.77 (0.07)	5	29	88.64 (0.06)	6	27
AEN	90.20 (0.09)	5	21	91.26 (0.16)	6	25
Proposed	96.00 (0.04)	6	17	94.57 (0.07)	7	19
$n = 300, p = 5,000$						
EN	91.08 (0.04)	6	23	95.00 (0.11)	7	36
AEN	92.16 (0.09)	6	25	94.00 (0.12)	7	34
Proposed	96.00 (0.07)	7	14	98.00 (0.11)	8	21
$n = 300, p = 1,0000$						
EN	88.07 (0.06)	8	29	95.21 (0.14)	8	38
AEN	92.00 (0.09)	8	28	92.00 (0.11)	8	37
Proposed	98.00 (0.07)	8	21	98.00 (0.09)	8	23

the selection of genes and ensure classification accuracy, we propose a new weight based on the ratio (BWR) as an initial weight inside L_1-norm with the EN.

The j^{th} component of the $p-$ dimensional weight vector $\mathbf{w} = (w_1, w_2, ..., w_p)^T$ is given by:

$$w_j = \frac{1}{|BWR(j)|}, \quad j = 1, 2, ..., p, \tag{11}$$

where, BWR(j) is the ratio of the gene j, in defined as Eq. (10).

The proposed weight assigns a relatively larger weight to the gene with a low-value ratio and a smaller weight to the gene with a high ratio. Moreover, the adjusted L_1 part of the penalty performs gene selection by setting some coefficients to exactly 0, and the L_2 part of the penalty encourages the group selection by shrinking the coefficients of correlated genes toward each other. Accordingly, the L_1-norm can reduce inconsistency. After assigning weight to genes, the PLRAEN can select

Table 3
Classification and variable selection performance of the PLRAEN and the competitor methods over 100 partitions when $\rho = 0.95$.

Methods	Model 1			Model 2		
	CA%	TP	FP	CA%	TP	FP
$n = 100, p = 1,000$						
EN	86.06 (0.07)	7	26	88.31 (0.06)	8	25
AEN	88.00 (0.09)	7	24	90.00 (0.11)	8	29
Proposed	92.10 (0.05)	8	16	96.00 (0.07)	9	20
$n = 100, p = 5,000$						
EN	91.00 (0.08)	6	34	92.04 (0.11)	7	37
AEN	90.27 (0.09)	6	35	92.07 (0.11)	7	43
Proposed	94.17 (0.07)	7	18	96.00 (0.11)	9	22
$n = 100, p = 10,000$						
EN	88.13 (0.06)	8	38	92.00 (0.14)	8	38
AEN	90.00 (0.09)	8	37	92.00 (0.11)	8	41
Proposed	94.08 (0.07)	8	22	96.12 (0.10)	9	21
$n = 200, p = 1,000$						
EN	91.82 (0.08)	8	26	92.65 (0.11)	8	25
AEN	92.00 (0.09)	6	24	92.10 (0.12)	7	24
Proposed	95.04 (0.07)	8	16	95.00 (0.11)	9	21
$n = 200, p = 5,000$						
EN	86.72 (0.09)	6	34	85.21 (0.17)	7	37
AEN	90.11 (0.11)	7	34	88.00 (0.11)	8	33
Proposed	94.20 (0.07)	8	20	93.12 (0.09)	9	22
$n = 200, p = 10,000$						
EN	86.88 (0.08)	8	38	86.42 (0.17)	8	38
AEN	88.16 (012)	7	37	90.00 (0.11)	8	32
Proposed	93.18 (0.07)	8	21	94.32 (0.11)	9	21
$n = 300, p = 1,000$						
EN	86.14 (0.07)	8	26	88.00 (0.09)	8	25
AEN	88.00 (0.09)	7	24	88.00 (0.11)	7	21
Proposed	94.10 (0.05)	8	16	94.00 (0.07)	9	18
$n = 300, p = 5,000$						
EN	95.08 (0.07)	6	34	95.12 (0.14)	7	37
AEN	95.16 (0.09)	7	34	95.00 (0.11)	7	33
Proposed	96.14 (0.06)	8	19	96.23 (0.06)	9	19
$n = 300, p = 1,0000$						
EN	86.12 (0.07)	8	38	88.17 (0.06)	8	38
AEN	88.00 (0.09)	8	35	88.00 (0.12)	8	37

(continued on next page)

Table 3 (*continued*)

Methods	Model 1			Model 2		
	CA%	TP	FP	CA%	TP	FP
Proposed	92.00 (0.07)	8	21	92.30 (0.09)	9	22

Table 4
The characteristics of the three used datasets.

Datasets	Samples(n)	Genes(p)	Classes
Bip	61	22,283	31 control/30 bipolar disorder
Sco	54	22,283	15 normal/39 sick
Aut	146	54,613	64 healthy/82 autism

important genes with higher accuracy. The details of the PLRAEN algorithm are presented here. The PLRAEN equation has a global maximum as it has a convex form. Therefore, the coordinate descent algorithm is implemented to solve PLRAEN.

Step 1. Split each gene $x_{.j}$ based on the value of y into two classes x_{1j} and x_{2j}

Step 2. Find mean of $x_{.j}$, x_{1j} and x_{2j}

Step 3. Compute $BSS(j) = (\bar{x}_{1j} - \bar{x}_{.j})^2 + (\bar{x}_{2j} - \bar{x}_{.j})^2$

Step 4. Compute $WSS(j) = \sum (x_{1j} - \bar{x}_{1j})^2 + \sum (x_{2j} - \bar{x}_{2j})^2$

Step 5. Find $BWR(j) = \frac{BSS(j)}{WSS(j)}$

Step 6. Find $w_j, j = 1, 2, \ldots, p$

Step 7. Define $\tilde{}$

Step 8. Solve the PLRAEN

$$\hat{\beta}_{PLRAEN} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{ y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) \} + \lambda_1 \sum_{j=1}^{p} w_j |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right] \quad (12)$$

5. Method evaluation

Here, we evaluate the performance of the proposed method. The three common evaluation metrics of a predictive model, particularly in the healthcare setting, are the classification accuracy (CA), sensitivity (SEN), and specificity (SPE) [42]. The formulae for computing these metrics indicate the confusion matrix predicted vs. actual results shown in Table 1.

CA is a key efficiency measurement and is computed using Eq. (13). Average accuracy is an average of the accuracy results obtained from many cross-validation experiments. As CA does not differentiate between false positives and false negatives, SEN and SPE measurements are considered. SEN computes the TP rate, while SPE computes the TN rate. The SEN and SPE formulae are given by Eqs. (14) and (15), respectively.

Table 6
The averaged criteria over 100 times for the testing dataset.

Dataset	Methods	Testing set					
		% CA		% SEN		% SPE	
Bip	EN	87.97	(0.05)	88.77	(0.05)	86.58	(0.04)
	AEN	88.46	(0.07)	90.72	(0.06)	88.41	(0.06)
	Proposed	93.45	(0.04)	93.69	(0.05)	92.80	(0.04)
Aut	EN	90.07	(0.05)	91.64	(0.04)	90.55	(0.03)
	AEN	90.14	(0.04)	90.46	(0.04)	90.64	(0.06)
	Proposed	93.78	(0.04)	94.57	(0.04)	92.64	(0.05)
Sco	EN	89.58	(0.05)	88.36	(0.07)	89.63	(0.03)
	AEN	87.78	(0.03)	90.23	(0.02)	90.40	(0.02)
	Proposed	91.72	(0.03)	93.97	(0.03)	92.87	(0.02)

These metrics (criteria) are defined as:

$$CA = \frac{TN + TP}{FP + TP + TN + FN} \times 100\% \quad (13)$$

$$SEN = \frac{TP}{FN + TP} \times 100\% \quad (14)$$

$$SPE = \frac{TN}{TN + FP} \times 100\% \quad (15)$$

Here, *TP, FP, TN,* and *FN* denote the number of true positives, false positives, true negatives, and false negatives, respectively. The higher values of the evaluation criteria indicate better classification performance.

The one-way analysis of variance (ANOVA) was performed to prove the stability of the results for the proposed method. This was performed in addition to Tukey's honestly significant difference (HSD) test to evaluate the proposed method's classification results compared to the other two methods.

6. Results and discussion

This section uses both simulation data and real microarray datasets

Table 7
One-way ANOVA for the classification accuracy over 50 partitions in the training set.

Datasets	Source	Df	SS	MS	F	P-value
Aut	Methods	2	0.02935	0.014675	35.18	0.000 (*)
	Error	147	0.01411	0.000523		
	Total	149	0.04346			
Bip	Methods	2	0.03858	0.019288	9.516	0.0001 (*)
	Error	147	0.29797	0.002027		
	Total	149	0.33655			
Sco	Methods	2	0.0845	0.04225	88.02	0.0001 (*)
	Error	147	0.07057	0.00048		
	Total	149	0.15507			

(*) Significant at. $\alpha = 0.05$

Table 5
The averaged criteria over 100 times for the training dataset.

Dataset	Methods	# Genes	Training set					
			% CA		% SEN		% SPE	
Bip	EN	48	93.69	(0.05)	93.70	(0.05)	93.97	(0.05)
	AEN	44	95.62	(0.06)	94.31	(0.07)	94.51	(0.06)
	Proposed	52	97.70	(0.04)	96.55	(0.06)	95.98	(0.05)
Aut	EN	72	95.39	(0.001)	95.38	(0.002)	95.82	(0.02)
	AEN	76	94.28	(0.002)	95.48	(0.003)	94.94	(0.02)
	Proposed	76	97.64	(0.04)	98.87	(0.02)	97.28	(0.03)
Sco	EN	248	92.63	(0.02)	91.39	(0.04)	91.84	(0.02)
	AEN	247	91.39	(0.02)	93.86	(0.02)	92.18	(0.04)
	Proposed	264	96.93	(0.03)	95.80	(0.02)	95.55	(0.03)

Table 8
P-value of Tukey HSD test for classification accuracy in the training set.

Datasets	Proposed vs EN	Proposed vs AEN	EN vs AEN
Aut	0.000 (*)	0.000 (*)	0.0184 (*)
Bip	0.000 (*)	0.047 (*)	0.1049
Sco	0.000 (*)	0.000 (*)	0.0147 (*)

(*) Significant at $\alpha = 0.05$

to illustrate the effectiveness of the proposed method, PLRAEN.

6.1. Simulation study

The data is simulated under the following framework. Two simulation models are considered for the logistic regression model to cover two practical scenarios: the correlation among predictor variables and the correlation between a group of predictor variables. The sample size, n, takes three values: 100, 200, and 300, where each n was randomly split into two parts: 50% for the training and 50% for the testing dataset. In addition, we considered the number of the predictor variables $p = 1,000,\ 5,000$, and $10,000$ because the magnitude of this number affects the resulting estimator in terms of variable selection; particularly, the value of FP [43–45]. Further, because we are interested in the groping effect, in which the pairwise correlation value is considered more important, three values of the pairwise correlation are considered $\rho = \{0.55,\ 0.95\}$. According to create low and high correlations among variables, these values are chosen, respectively [45–47]. In total, we have two models that consider the logistic regression model as follows:

Model 1: The data was generated according to the logistic regression model as

$$Y \sim B\left(\frac{\exp(\mathbf{X}\boldsymbol{\beta}_{true})}{1+\exp(\mathbf{X}\boldsymbol{\beta}_{true})}\right), \tag{16}$$

for both the training and the testing datasets. In this model, we set the following: The true vector $\boldsymbol{\beta}_{true} = (1.5,\ 1,\ 0.8,\ 0.7,\ -0.6,\ 9,\ -3,\ 2,\ 0,\ \ldots,\ 0)^T$, with nonzero variables $q = 8$, and zero variables $= p - q$. The predictor variables matrix $\mathbf{X}$ is generated from a multivariate normal distribution $N(0,\ \Sigma)$, where Σ is the covariance matrix with $\Sigma_{i,j} = \rho^{|i-j|}$ $(i,\ j = 1,\ 2,\ \ldots,\ p)$ and, therefore, the predictor variables are correlated.

Model 2: The data was generated from Eq. (16). In this model, we set the following: The true vector $\boldsymbol{\beta}_{true} = (1.5,\ 1,\ 0.8,\ 0.7,\ -0.6,\ 9,\ -3,\ 2,\ 1,\ 0,\ \ldots,\ 0)^T$, with nonzero variables $q = 9$ and zero variables $= p - q$. The nonzero predictor variables are generated as

$$\begin{aligned}
&\text{Group } 1 : \mathbf{x}_j = \mathbf{v}_1 + \varepsilon,\ \mathbf{v}_1 \sim N(0,\ 1),\ j = 1,\ 2,\ 3;\\
&\text{Group } 2 : \mathbf{x}_j = \mathbf{v}_2 + \varepsilon,\ \mathbf{v}_2 \sim N(0,\ 1),\ j = 4,\ 5,\ 6;\\
&\text{Group } 3 : \mathbf{x}_j = \mathbf{v}_3 + \varepsilon,\ \mathbf{v}_3 \sim N(0,\ 1),\ j = 7,\ 8,\ 9,
\end{aligned}$$

while the zero predictor variables are generated as $\mathbf{x}_j \sim N(0,\ 1), j = 10,\ 11,\ \ldots,\ p - q$. Therefore, the predictor variables within each group were correlated, while the predictor variables from different groups were uncorrelated. To ensure that the correlations among variables within each group are 0.55, and 0.95, the ε was generated according to $\varepsilon \sim N(0,\ 0.8)$, and $\varepsilon \sim N(0,\ 0.01)$, respectively.

In an elastic net, there are two tuning parameters λ_1 and λ_2, and, therefore, two-dimensional surface cross-validation (CV) is need. Following [13,38], first, we fix $\lambda_2 = \{0,\ 0.01,\ 0.1,\ 1,\ 10,\ 100\}$, then for each λ_2 value, 10-fold CV was employed to find the best value of λ_1.

The simulation process was repeated 100 times for each model. The median of CA with different values of n, p and ρ are presented in Tables 2 and 3, respectively. The values in the parentheses denote the corresponding standard deviation. The number of truly relevant variables selected (TP) and the number of irrelevant variables not selected (TN) are recorded to quantify variable selection performance. For comparison purposes, the proposed method's performance was compared with other existing methods, namely EN and AEN.

Tables 2 and 3 summarize the classification and variable selection performance of the PLRAEN and the competitor methods over 100 partitions for $\rho = 0.55$ and $\rho = 0.95$, respectively. As is shown in Tables 2 and 3, in all cases, our proposed method consistently attained the highest CA for the logistic regression simulation models; thus, it gave the best predictive performance. For the number of TP and TN, the proposed method performed well in selecting the true nonzero correlated variables. It reduced the model selecting of the true zero variables in all cases. This implies that our proposed method can select the true relevant variables.

To sum up, the simulation results seem to indicate that the performance of PLRAEN is superior to the EN and AEN in terms of variable selection and CA. It has the adaptability advantage over the other used penalized methods in encouraging grouping effect and selecting variables consistently in high dimensional data. Moreover, PLRAEN can be used successfully in various correlation values.

Table 9
One-way ANOVA for the classification accuracy over 50 partitions in the testing set.

Datasets	Source	Df	SS	MS	F	P-value
Aut	Methods	2	0.045	0.022501	50.13	0.000 (*)
	Error	147	0.06598	0.000449		
	Total	149	0.11098			
Bip	Methods	2	0.09209	0.04604	42.33	0.000 (*)
	Error	147	0.15991	0.00109		
	Total	149	0.252			
Sco	Methods	2	0.03045	0.015224	9.559	0.0001 (*)
	Error	147	0.23412	0.001593		
	Total	149	0.26457			

(*) Significant at $\alpha = 0.05$

Table 10
P-value of Tukey HSD test for classification accuracy in the testing set.

Datasets	Proposed vs. EN	Proposed vs. AEN	EN vs. AEN
Aut	0.000 (*)	0.000 (*)	0.9888
Bip	0.000 (*)	0.000 (*)	0.7407
Sco	0.000 (*)	0.0487 (*)	0.0677

(*) Significant at $\alpha = 0.05$

6.2. Real data studies

The proposed method (PLRAEN) is applied to three well-known gene expression datasets with different numbers of genes and different sample sizes to evaluate its performance and demonstrate its advantages over the other competitive methods. These datasets are publicly available and previously used by many researchers. Three public datasets have been used in this study to evaluate the performance of our method. First, the Bipolar disorder (Bip) dataset. Its sample size was 61, including 31 control observations and 30 observations with bipolar disorder. Gene expressions of 22,283 human genes are captured using Affymetrix technology [48,49]. The second is the Sarcoma (Sco) dataset. It involved the expression profiles of 22,283 human genes measured on 54 patients; where 15 people were normal, and 39 people had the disease [49,50]. The third is the Autism (Aut) dataset, which represented the gene expressions of 146 children from peripheral blood lymphocytes (PBL). The complete RNA was obtained using Affymetrix Human U133 Plus 2.0, including 39 expression arrays for microarray experiments. This dataset contained 54,613 genes, 82 with autism and 64 were healthy. Furthermore, this dataset has been recently analyzed by Refs. [18,51,52]. The main characteristics of the three datasets are summarized in Table 4.

The proposed PLRAEN method is effective through comparative experiments with two other methods, namely EN and AEN, where they

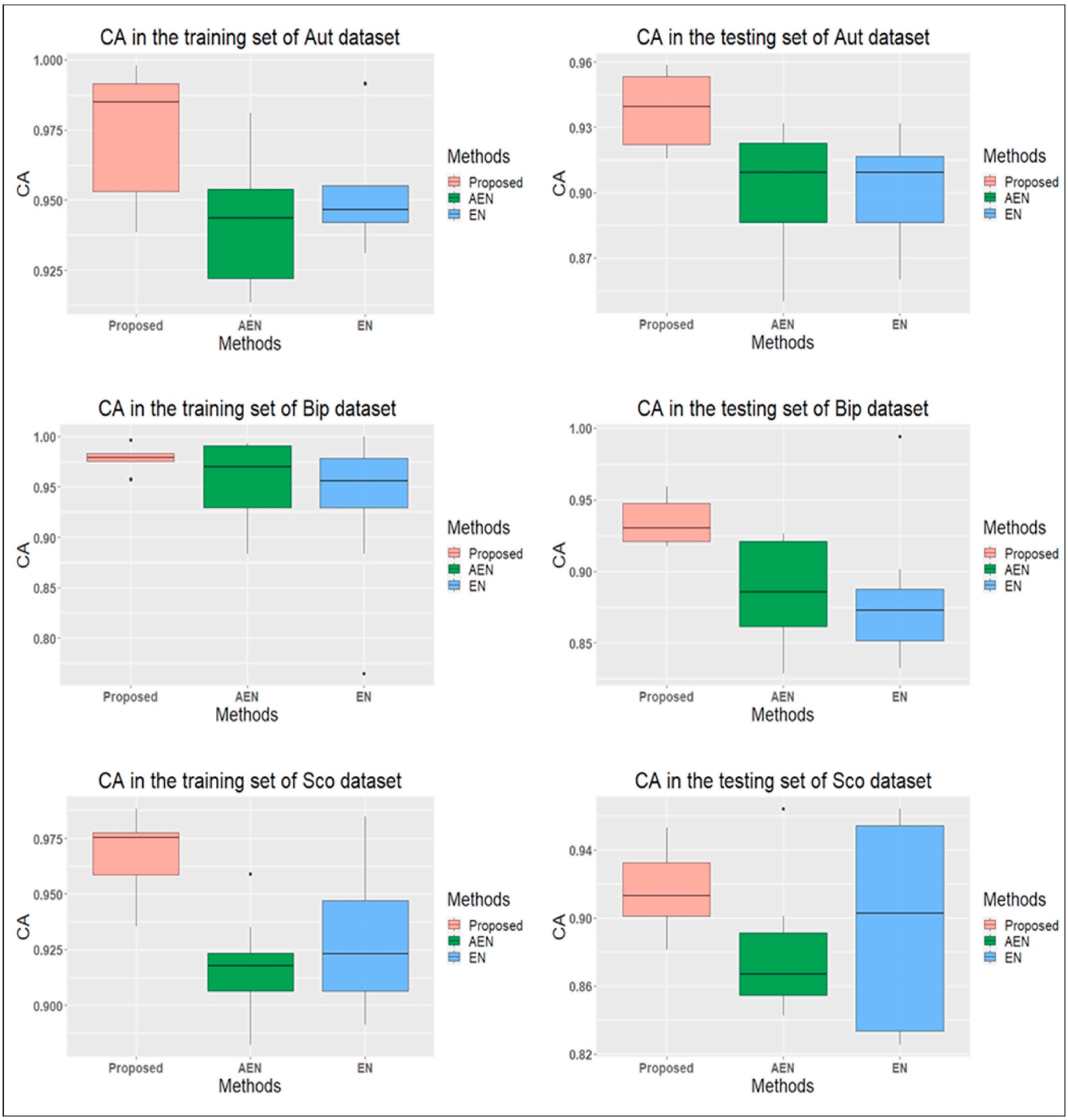

Fig. 1. The CA in the training and testing parts for the three datasets by the three methods.

and the proposed method were applied to the datasets considered. Each dataset was randomly subdivided into two parts, the training (70%) and the testing (30%) parts to perform CV. The cross-validation was conducted ten times using the training subset to select the optimal values of λ_1 and λ_2. The result was an average of 100 replications of the experiment. The value of the tuning parameters for each method was allowed a value in the interval $[0, 100]$. The implementations of these methods were done in R using the Glmnet package.

Tables 5 and 6 summarize the average number of important genes chosen by the corresponding method (# genes), CA, SEN, and SPE for the training and the testing subsets of the original dataset when applying the proposed method, EN, and AEN. The corresponding standard deviation is written in parenthesis.

Table 5 shows that our proposed PLRAEN method has the highest average number of selected genes among all other methods, where it selected 52 genes for the Bip dataset. In comparison, EN and AEN selected 48 and 44 genes, respectively.

Tables 5 and 6 show that in each dataset, the mean of CA, SEN, and SPE in the training and testing parts produced by our method are higher than the measures produced by EN and AEN. For example, the training (testing) CA of the proposed method was 97.64% (93.78%) in the Aut dataset, greater than 95.39% (90.07%) for EN and 94.28% (90.14%) for

AEN. We also observe that PLRAEN in the testing set had the highest sensitivities of 93.69%, 94.57%, and 93.97% for the Bip, Aut, and Sco datasets, respectively. Furthermore, the highest specificities of the training set were 95.98% (Bip), 97.28 (Aut), and 95.55% (Sco) for PLRAEN.

In addition to the Tukey HSD test, one-way ANOVA was performed to evaluate our method's obtained classification results. Upon rejecting the null hypothesis, a Tukey HSD test provides us more details about the differences between each pair of the three methods. Tables 7 and 9 summarize the ANOVA results for the CA in the training and the testing parts. The results indicate significant differences among the three methods, for all datasets, regarding the CA. Moreover, the Tukey HSD test was implemented to acquire details about the differences between the PLRAEN and the other applied methods. Tables 8 and 10 lists the *p*-value of each pair of methods. The PLRAEN demonstrated significant CA performance regarding the EN and AEN.

Further examining our method's CA performance, Fig. 1 indicates that the mean of CA in all three datasets (Aut, Bip, and Sco) resulting from our method is higher than the corresponding average of CA from other methods. The box plot related to the proposed methods shows that the CA distribution is more symmetric and more stable as its spread is the least among the other methods. This shows that our method performs better than others.

To further highlight the performance of the PLRAEN, we compared the obtained results for the same dataset (Autism), regarding the number of selected genes and CA, with other three methods: the Bayesian lasso quantile regression (Blassou) reported by Ref. [26], the adaptive penalized logistic regression (APLR) proposed by Ref. [18], and SCAD-support vector machine using firefly algorithm (FFA1) presented by Ref. [31]. Our method selected more genes than the other three methods, where it selected 76 genes while Blassou, APLR, and FFA1, respectively, selected 13, 9, and 21 genes. Importantly, PLRAEN has the potential to select more genes than other methods, indicating that most of these additionally selected genes were probably highly correlated. Additionally, our method achieved a higher CA of 97.64% compared with 96.20% for Blassou, 93.27% for APLR, and 93.35% for FFA1.

The proposed method's superior classification performance was generally shown through three aspects: high CA, SEN, and SPE for both the training and testing datasets. Meeting these three aspects simultaneously nominates the proposed method as a promising gene selection method. Moreover, as a classification process, our adaptive penalized method is the best classification process compared to competitor methods. This demonstrates that our method considers the weights of the genes.

7. Conclusion

Comparing the results obtained by applying the proposed method to simulated datasets and three well-known datasets (Bip, Aut, and Sco.) with other methods (EN and AEN) applied to the same datasets, we confirm that the performance of our method as a classification and gene selection process is more efficient than the other methods concerning CA and selection of genes. This assures that our method is a significant classification and gene selection method, and it may be applied to other cancer-related datasets.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

A special thanks to the Taif University for its financial sponsorship. Also, Universiti Teknologi Malaysia for providing the facilities.

References

[1] Potharaju SP, Sreedevi M. Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. Clin. Epidemiol. Glob. Heal. 2019;7:171–6. https://doi.org/10.1016/j.cegh.2018.04.001.
[2] Liu X-Y, Liang Y, Wang S, Yang Z-Y, Ye H-S. A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. IEEE Access 2018;6:22863–74. https://doi.org/10.1109/ACCESS.2018.2818682.
[3] Kourou K, Exarchos TP, Exarchos KP, V Karamouzis M, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005.
[4] Tran QN, Arabnia HR. Emerging trends in applications and infrastructures for computational biology, bioinformatics, and systems biology. Elsevier; 2016. https://doi.org/10.1016/C2015-0-01779-8.
[5] Algamal ZY, Lee MH. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. Adv. Data Anal. Classif. 2019;13:753–71. https://doi.org/10.1007/s11634-018-0334-1.
[6] Ayyad SM, Saleh AI, Labib LM. Gene expression cancer classification using modified K-Nearest Neighbors technique. Biosystems 2019;176:41–51. https://doi.org/10.1016/j.biosystems.2018.12.009.
[7] Yang ZZ-Y, Liang Y, Zhang H, Chai H, Zhang B, Peng C. Robust sparse logistic regression with the lq ($0 < q < 1$) regularization for feature selection using gene expression data ZIYI. IEEE Access 2018;6:68586–95. https://doi.org/10.1109/ACCESS.2018.2880198.
[8] Min W, Liu J, Zhang S. Network-Regularized sparse logistic regression models for clinical risk prediction and biomarker discovery. IEEE ACM Trans Comput Biol Bioinf 2018;15:944–53. https://doi.org/10.1109/TCBB.2016.2640303.
[9] Dashtban M, Balafar M, Suravajhala P. Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics 2018;110:10–7. https://doi.org/10.1016/j.ygeno.2017.07.010.
[10] Nakariyakul S. A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification. PloS One 2019;14:1–17. https://doi.org/10.1371/journal.pone.0212333.
[11] Tibshirani R. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B. 1996;58:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[12] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 2001;96:1348–60. https://doi.org/10.1198/016214501753382273.
[13] Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodol. 2005;67:301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
[14] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006;101:1418–29. https://doi.org/10.1198/016214506000000735.
[15] Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Ann Stat 2009;37:1733–51. https://doi.org/10.1214/08-AOS625.
[16] Ghosh S. On the grouped selection and model complexity of the adaptive elastic net. Stat Comput 2011;21:451–62. https://doi.org/10.1007/s11222-010-9181-4.
[17] Wang S, Nan B, Rosset S, Zhu J. Random lasso. Ann Appl Stat 2011;5:468–85. https://doi.org/10.1214/10-AOAS377.
[18] Algamal ZY. Classification of gene expression autism data based on adaptive penalized logistic regression. Electron. J. Appl. Stat. Anal. 2017;10:561–71. https://doi.org/10.1285/i20705948v10n2p561.
[19] Yu Zhiwen, Le Li, Liu Jiming. Guoqiang han, hybrid adaptive classifier ensemble. IEEE Trans. Cybern 2015;45:177–90. https://doi.org/10.1109/TCYB.2014.2322195.
[20] Zhong Y, Chalise P, He J. Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. Commun Stat Simulat Comput 2020:1–18. https://doi.org/10.1080/03610918.2020.1850790.
[21] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422. https://doi.org/10.1023/A:1012487302797.
[22] Tian Yingjie, Qi Zhiquan, Ju Xuchan, Shi Yong, Liu Xiaohui. Nonparallel support vector machines for pattern classification. IEEE Trans. Cybern. 2014;44:1067–79. https://doi.org/10.1109/TCYB.2013.2279167.
[23] Huang H, Gao Y, Zhang H, Li B. Weighted Lasso estimates for sparse logistic regression: non-asymptotic properties with measurement errors. Acta Math Sci 2021;41:207–30. https://doi.org/10.1007/s10473-021-0112-6.
[24] Wang Y, Li X, Ruiz R. Weighted general group lasso for gene selection in cancer classification. IEEE Trans. Cybern. 2019;49:2860–73. https://doi.org/10.1109/TCYB.2018.2829811.
[25] Kwon S, Lee S, Na O. Tuning parameter selection for the adaptive LASSO in the autoregressive model. J Korean Surg Soc 2017;46:285–97. https://doi.org/10.1016/j.jkss.2016.10.005.
[26] Algamal ZY, Alhamzawi R, Mohammad Ali HT. Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression. Comput Biol Med 2018;97:145–52. https://doi.org/10.1016/j.compbiomed.2018.04.018.
[27] Algamal ZY, Lee MH. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Syst Appl 2015;42:9326–32. https://doi.org/10.1016/j.eswa.2015.08.016.
[28] Cawley GC, Talbot NLC. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics 2006;22:2348–55. https://doi.org/10.1093/bioinformatics/btl386.
[29] Alhamzawi R, Yu K, Benoit DF. Bayesian adaptive Lasso quantile regression. Stat Model An Int J 2012;12:279–97. https://doi.org/10.1177/1471082X1101200304.

[30] Algamal ZY, Lee MH. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Comput Biol Med 2015;67:136–45. https://doi.org/10.1016/j.compbiomed.2015.10.008.
[31] Al-Thanoon NA, Qasim OS, Algamal ZY. Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification. Comput Biol Med 2018;103:262–8. https://doi.org/10.1016/j.compbiomed.2018.10.034.
[32] Algamal ZY, Lee MH, Al-Fakih AM, Aziz M. High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty. J Chemom 2017;31:e2889. https://doi.org/10.1002/cem.2889.
[33] El Guide M, Jbilou K, Koukouvinos C, Lappa A. Comparative study of L 1 regularized logistic regression methods for variable selection. Commun Stat Simulat Comput 2020:1–16. https://doi.org/10.1080/03610918.2020.1752379.
[34] Doerken S, Avalos M, Lagarde E, Schumacher M. Penalized logistic regression with low prevalence exposures beyond high dimensional settings. PloS One 2019;14:e0217057. https://doi.org/10.1371/journal.pone.0217057.
[35] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Software 2010;33:1–22. https://www.ncbi.nlm.nih.gov/pubmed/20808728.
[36] Liang Y, Liu C, Luan X-Z, Leung K-S, Chan T-M, Xu Z-B, Zhang H. Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. BMC Bioinf 2013;14:198. https://doi.org/10.1186/1471-2105-14-198.
[37] Bühlmann P, van de Geer S. Statistics for high-dimensional data. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. https://doi.org/10.1007/978-3-642-20192-9.
[38] El Anbari M, Mkhadri A. The adaptive gril estimator with a diverging number of parameters. Commun Stat Theor Methods 2013;42:2634–60. https://doi.org/10.1080/03610926.2011.615438.
[39] Bühlmann P, Rütimann P, van de Geer S, Zhang C-H. Correlated variables in regression: clustering and sparse estimation. J Stat Plann Inference 2013;143:1835–58. https://doi.org/10.1016/j.jspi.2013.05.019.
[40] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87. https://doi.org/10.1198/016214502753479248.
[41] Peng H, Fu Y, Liu J, Fang X, Jiang C. Optimal gene subset selection using the modified SFFS algorithm for tumor classification. Neural Comput Appl 2013;23:1531–8. https://doi.org/10.1007/s00521-012-1148-2.
[42] Gladstone E, Smolina K, Morgan SG, Fernandes KA, Martins D, Gomes T. Sensitivity and specificity of administrative mortality data for identifying prescription opioid-related deaths. CMAJ (Can Med Assoc J) 2016;188:E67–72. https://doi.org/10.1503/cmaj.150349.
[43] Androulakis E, Koukouvinos C, Mylona K. Tuning parameter estimation in penalized least squares methodology. Commun Stat Simulat Comput 2011;40:1444–57. https://doi.org/10.1080/03610918.2011.575507.
[44] Chen J, Chen Z. Extended BIC for small-n-large-P sparse GLM. Stat Sin 2012;22:555–74. https://doi.org/10.5705/ss.2010.216.
[45] Mkhadri A, Ouhourane M. A group VISA algorithm for variable selection. Stat. Methods Appt. 2015;24:41–60. https://doi.org/10.1007/s10260-014-0281-8.
[46] Fu G-H, Zhang W-M, Dai L, Fu Y-Z. Group variable selection with oracle property by weight-fused adaptive elastic net model for strongly correlated data. Commun Stat Simulat Comput 2014;43:2468–81. https://doi.org/10.1080/03610918.2012.752841.
[47] Zeng L, Xie J. Group variable selection via SCAD- L 2. Statistics (Ber). 2014;48:49–66. https://doi.org/10.1080/02331888.2012.719513.
[48] Ryan MM, Lockstone HE, Huffaker SJ, Wayland MT, Webster MJ, Bahn S. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Mol Psychiatr 2006;11:965–78. https://doi.org/10.1038/sj.mp.4001875.
[49] Shen Q, Mei Z, Ye B-X. Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. Comput Biol Med 2009;39:646–9. https://doi.org/10.1016/j.compbiomed.2009.04.008.
[50] Detwiller KY, Fernando NT, Segal NH, Ryeom SW, D'Amore PA, Yoon SS. Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A. Canc Res 2005;65:5881–9. https://doi.org/10.1158/0008-5472.CAN-04-4078.
[51] Latkowski T, Osowski S. Computerized system for recognition of autism on the basis of gene expression microarray data. Comput Biol Med 2015;56:82–8. https://doi.org/10.1016/j.compbiomed.2014.11.004.
[52] Latkowski T, Osowski S. Data mining for feature selection in gene expression autism data. Expert Syst Appl 2015;42:864–72. https://doi.org/10.1016/j.eswa.2014.08.043.