Applied Mathematical Modelling 105 (2022) 179–196

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Regression with adaptive lasso and correlation based penalty

Yadi Wang[a,b], Wenbo Zhang[a,b], Minghu Fan[a,b], Qiang Ge[a,b], Baojun Qiao[a,b], Xianyu Zuo[a,b,*], Bingbing Jiang[c]

[a] Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, 475004, China
[b] Institute of Data and Knowledge Engineering, School of Computer and Information Engineering, Henan University, Kaifeng, 475004, China
[c] School of Information Science and Engineering, Hangzhou Normal University, Hangzhou, 311121, China

ARTICLE INFO

Article history:
Received 7 June 2021
Revised 8 December 2021
Accepted 11 December 2021
Available online 22 December 2021

Keywords:
Feature selection
Classification
Mutual information
Correlation based penalty
Adaptive logistic regression

ABSTRACT

Feature selection for high-dimensional data is an important issue in machine learning, pattern recognition and bioinformatics fields. Feature selection algorithms are proposed to select the relevant feature subset from the original features. To adaptively identify the important highly correlated features from high-dimensional data which often beneficial to improve classification accuracy is a challenge. In this paper, we propose a regularized logistic regression with adaptive Lasso and correlation based penalty model to select informative highly correlated features adaptively. To incorporate significance of features into regression model, we first measure significance of each feature based on mutual information, and propose an adaptive weight construction strategy. Based on the adaptive weight construction strategy, the proposed adaptive logistic regression can impose a large amount of penalty on irrelevant features, and thus noise features are easily removed from the model and remain the informative features. The experimental results on the simulation and real-world datasets demonstrate the effectiveness and the superiority the proposed model by comparing it to existing competing regularized logistic regression models.

© 2021 Elsevier Inc. All rights reserved.

1. Introduction

Feature selection plays a critical role in data mining [1], pattern recognition [2] and bioinformatics applications [3]. Feature selection is conducive to reduce dimensionality, remove irrelevant data, and improve resultant learning accuracy of the high-dimensional data [4], which includes large p (features) and small n (samples). In recent years, even though a large number of methods have been proposed for feature selection. Many feature selection methods confront severe challenges in terms of effectiveness and efficiency, because of the recent increase in data dimensionality. Generally, the feature selection methods are divided into three categories: filter, wrapper and embedded methods [5]. Filter methods [6] evaluate a feature based on an evaluation function, and are independent of classifier [7]. The computational complexity of the filter methods is low, but their accuracy is not guaranteed [8]. Wrapper methods [9] calculate the score of a feature subset based on a specific classifier, which can obtain high classification. However, since they need to constantly build models, the computational complexity is relatively large. Embedded methods [10] select a feature subset in the learning stage. Embedded methods have been widely used because of their less computational complexity and less prone to over-fitting than filter and wrapper

* Corresponding author.
E-mail address: xianyu_zuo@163.com (X. Zuo).

https://doi.org/10.1016/j.apm.2021.12.016
0307-904X/© 2021 Elsevier Inc. All rights reserved.

methods. Regularization models are important embedded methods and perform both continuous shrinkage and automatic feature selection simultaneously.

Various feature selection approaches have been proposed based on the framework of regularization [11], which are closely related to the case study. Most of the regularization models are focus on the individual feature selection. As a representative of the individual regularization model, Lasso [12] enjoys certain advantages such as computational simplicity and satisfactory numerical performance. However, there exist grouping structures in gene expression data which describe the inherent inter-connections among genes. Lasso fails to select the highly correlated genes, which may lead to inefficient models. Toward this end, many regularization models based on structured sparsity are the effective methods for performing grouped feature selection [13]–[22]. Feature selection algorithms based on structured sparsity have received widespread attention, and a large number of algorithms have been proposed, Gui et al. [13] systematically studied the relationship between these structured sparsity feature selection algorithms. Yuan et al. [14] proposed group Lasso model by using $L_{2,1}$ norm penalty, which is extended to logistic regression [15]. Simon et al. [16] presented a sparse group lasso model to achieve group-wise sparsity and within group sparsity simultaneously. However, the effectiveness of group Lasso and sparse group Lasso models relies greatly on the group division method [17]. Another similar couple model is the elastic net [18] and penalized regression with correlation-based penalty (CP) [19] which is extended to L1CP [20], both of which deal with the grouped feature selection problem. Chen et al. [21] proposed a matrix elastic net regularized multivariate Huber regression model, which can reduce the negative effect of outliers. However, most of these regularization models may produce inefficient estimation and inconsistent feature selection results because to uniformly impose penalty on all features overlooks considering the significance of each feature [22].

Adaptively selecting important features is a challenge for feature selection. Some regularization models have been developed to achieve adaptively feature selection by constructing weight coefficients of features. The weight construction strategy in [23] is based on the initial consistent estimator and the weight construction strategies in [24] and partly adaptive weight construction strategies in [25] are constructed by initial elastic net estimator which face a practical problem in finding the type of initial weight. Wang et al. [26] and Fang et al. [27] proposed adaptive group Lasso models in terms of a least squares estimator and a group bridge estimator, respectively. A modified adaptive lasso with weights adopting the ranking-based feature selection method is proposed by Patil et al. [28]. Based on the correlation-based penalty in [19], Algamal et al. [29] proposed a adaptive penalized logistic regression model. The weight construction method in [30] is proposed based on Wilcoxon rank sum test. Liu et al. [31] constructed feature weights by using Pearson correlation coefficient. Li et al. [32] proposed a adaptive weight strategy based on t-test measure. The above three kinds of the weight are constructed by the statistical measurements. However, since most of the aforementioned weight construction strategies are very sensitive to outliers or noise in the dataset, some features irrelevant to classification might be identified, which leads to a reduction in model performance. Many information theoretic feature selection methods [33] based on entropy, mutual information [34] and other information theory measure [35] have been proposed to select the optimal feature subset successfully. Some information measures such as mutual information (MI) [36] and normalized mutual information, i.e., symmetrical uncertainty (SU) [37] are widely used as a measure of the relevance of features. In addition, the information measures rely only on the probability distribution of a random variable instead of on its actual values, so they are more effective to assess the feature-class relevance [38]. By introducing the normalized mutual information into the adaptive weight construction, the above-mentioned disadvantages could be avoided to some extent and the existing efficient regularization models such as in [29] and [30] will be improved in this report.

Inspired by the above ideas, we measure the significance of each feature based on mutual information, and then construct an adaptive weight construction strategy. Since the general Lasso-type penalty does consider the importance degree of each feature. In order to overcome this, an adaptive Lasso type penalty based on the proposed adaptive weight construction strategy is discriminately imposed on each feature. Furthermore, a regularized logistic regression with adaptive Lasso and correlation based penalty (ALCP) model is proposed to adaptively select informative and highly correlated features. The optimization problem of ALCP model can be solved by a regularized solution path algorithm based on the coordinate descent algorithm. In this paper, we focus on researching the selection of desirable grouped features of binary-class classification. The main technical contributions of this paper are summarized as follows:

- A new regularized logistic regression with adaptive Lasso and correlation based penalty is proposed for feature selection and encouraging grouping effects simultaneously.
- A new adaptive weight construction method is proposed based on the significance of each feature in terms of mutual information.
- Based on the coordinate descent algorithm, a solving algorithm is further developed for the proposed ALCP model. The experiments demonstrate that the classification performance and feature selection performance of the proposed ALCP model are superior to the other competing regression models.

The rest of this paper is organized as follows. Section 2 gives a brief description of the research problems. The regularized logistic regression with adaptive Lasso and correlation based penalty (ALCP) model is presented in Section 3, and the corresponding algorithm is developed in Section 4. Experimental results obtained from the simulation and real-world datasets are presented in Section 5. Finally, the work is summarized in Section 6.

2. Problem description

The regularized logistic regression models try to remove the irrelevant and redundant features while selecting the most relevant features, i.e., important information in the data set is kept, which is desired for high-dimensional data. Given a data set $(X, \boldsymbol{y}) = \{(\boldsymbol{x}_i, y_i)|i = 1, \ldots, n\}$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ is the input vector. n and p are the number of samples and features, respectively. Let $X = (\boldsymbol{x}_1; \ldots; \boldsymbol{x}_n) = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(p)})$ be the model matrix in which $\boldsymbol{x}_{(j)} = (x_{1j}, \ldots, x_{nj})^T$ is the j^{th} predictor (feature). $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the $n \times 1$ response variable vector. For binary classification, the response variable $\boldsymbol{y}$ includes output labels $y_i \in \{0, 1\}$. We also assume that the response vector $\boldsymbol{y}$ is centered and the predictors are standardized, i.e.,

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1. \tag{1}$$

Based on the loss + penalty criterion, the regularized logistic regression models can be formulated into a generic regularized problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \Big\{ -l(\boldsymbol{\beta}) + \lambda R(\boldsymbol{\beta}) \Big\}, \tag{2}$$

where $l(\boldsymbol{\beta})$ is a negative log-likelihood function, λ is the regularization parameter, and $R(\boldsymbol{\beta})$ is a nonnegative regularization term. Features are selected according to the estimated coefficient vector $\hat{\boldsymbol{\beta}}$, i.e., the features with non-zero estimated coefficients in $\hat{\boldsymbol{\beta}}$ are selected. The regularization parameter λ is the tradeoff between the loss term and the regularization term (or penalty term).

3. Regularized logistic regression with adaptive lasso and correlation based penalty

3.1. Statistical learning model

The general logistic regression model which represents the class conditional probability is expressed as:

$$\pi(\boldsymbol{x}_i) = p(y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}, \quad i = 1, 2, \ldots, n \tag{3}$$

where $\boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. β_0 is the intercept and $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients. The logistic transformation of the vector of probability estimates $\pi(\boldsymbol{x}_i)$ is modeled by a linear function:

$$\log \frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)} = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{4}$$

The log-likelihood function of Eqn. (4) is defined as:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [y_i \log \pi(\boldsymbol{x}_i) + (1 - y_i) \log \{1 - \pi(\boldsymbol{x}_i)\}] \\ &= \sum_{i=1}^{n} \left[y_i \boldsymbol{x}_i^T \beta - \log \left\{ 1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) \right\} \right]. \end{aligned} \tag{5}$$

Based on the regularization framework in Eqn. (2), we add a new non-negative penalty term to the negative log-likelihood function $l(\boldsymbol{\beta})$ in (5), so that the size of feature coefficients in high-dimension data can be controlled. We combine the adaptive Lasso penalty and the correlation based penalty and propose an ALCP penalty, i.e.,

$$R\big(\boldsymbol{\beta}\big) = \alpha \sum_{j=1}^{p} w_j \big|\beta_j\big| + (1 - \alpha) \sum_{j=1}^{p-1} \sum_{i>j} \left\{ \frac{\big(\beta_i - \beta_j\big)^2}{1 - \rho_{ij}} + \frac{\big(\beta_i + \beta_j\big)^2}{1 + \rho_{ij}} \right\}, \tag{6}$$

where parameter $\alpha \in [0, 1]$, $\rho_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j$ indicates the correlation coefficients between the i^{th} and j^{th} features. The $w_j, j = 1, \cdots, p$ is the weight of feature based on mutual information which will be described in Section 3.2 in detail. The adaptive Lasso penalty $\sum_{j=1}^{p} w_j|\beta_j|$ encourages sparsity in the coefficients, which can select the features adaptively. The correlation based penalty $\sum_{j=1}^{p-1} \sum_{i>j} \{ \frac{(\beta_i - \beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\rho_{ij}} \}$ ($\rho_{ij}^2 \neq 1$ for $i \neq j$) as in [19] aims to encourage grouping effect of highly correlated features. For strong positive correlation features ($\rho_{ij} \approx 1$), the first term becomes dominant having the effect that estimates for β_i and β_j are similar ($\hat{\beta}_i \approx \hat{\beta}_j$). For strong negative correlation features ($\rho_{ij} \approx -1$), the second term becomes dominant and $\hat{\beta}_i$ will be close to $-\hat{\beta}_j$.

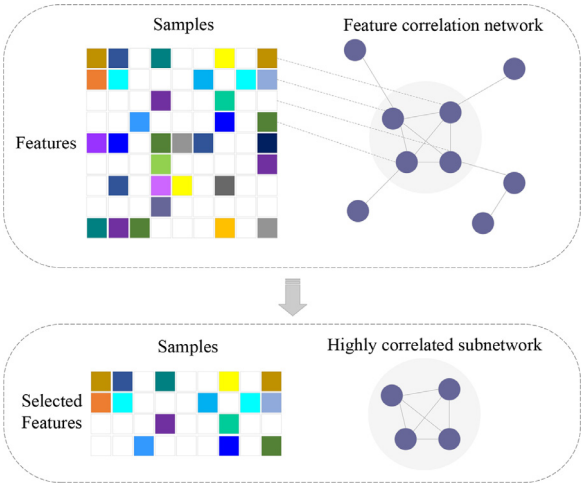

Fig. 1. The feature selection process of the proposed ALCP model.

By further introducing the ALCP penalty (6) into the log-likelihood function (5), we propose the ALCP model:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}\in R^p}\left\{-l(\boldsymbol{\beta})+\lambda R(\boldsymbol{\beta})\right\}\\ &= \arg\min_{\boldsymbol{\beta}\in R^p}\Bigg\{-\sum_{i=1}^{n}\left[y_i\boldsymbol{x}_i^T\boldsymbol{\beta}-\log\left\{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\}\right]+\lambda\alpha\sum_{j=1}^{p}w_j\left|\beta_j\right|\\ &\quad+\lambda(1-\alpha)\sum_{j=1}^{p-1}\sum_{i>j}\left\{\frac{(\beta_i-\beta_j)^2}{1-\rho_{ij}}+\frac{(\beta_i+\beta_j)^2}{1+\rho_{ij}}\right\}\Bigg\}.\end{aligned} \tag{7}$$

The ALCP model imposes a large amount of penalty on irrelevant features, and thus a small assessment is made for unimportant features, or coefficients of irrelevant features may be assessed to exactly zero. This implies that ALCP model can select important features adaptively and also allow to select or remove highly correlated features together, i.e., grouping effect. Moreover, the correlation based penalty can be re-expressed as a quadratic form $\boldsymbol{\beta}^TQ\boldsymbol{\beta}$, where Q is a positive matrix with the element as:

$$q_{ij}=\begin{cases}2\sum_{s\neq i}\frac{1}{1-\rho_{is}^2}, & i=j\\ -2\frac{\rho_{ij}}{1-\rho_{ij}^2}, & i\neq j\end{cases} \tag{8}$$

Therefore, the ALCP model in (7) can also be represented as:

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in R^p}\left\{-\sum_{i=1}^{n}\left[y_i\boldsymbol{x}_i^T\boldsymbol{\beta}-\log\left\{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\}\right]+\lambda\alpha\sum_{j=1}^{p}w_j|\beta_j|+\lambda(1-\alpha)\boldsymbol{\beta}^TQ\boldsymbol{\beta}\right\}, \tag{9}$$

In fact, the penalty term $\boldsymbol{\beta}^TQ\boldsymbol{\beta}$ is a network penalty that can infer the correlation network of all the features, based on which ALCP can select the highly correlated features in the complete network. The detailed feature selection process is shown in Fig. 1.

3.2. Adaptive weight construction

In this section, we propose a new adaptive weight construction method based on mutual information. Before exploring the details of adaptive weight construction strategy, we review the several basic concepts about information theory. Let $\hat{\mathcal{X}}=(\hat{x}_1,\cdots,\hat{x}_n)^T$, $\mathcal{Y}=(y_1,\cdots,y_n)^T$ and $\mathcal{Z}=(z_1,\cdots,z_n)^T$ be three random variables. The information entropy [35] of variable $\hat{\mathcal{X}}$ can be expressed as:

$$H(\hat{\mathcal{X}})=-\sum_{\hat{x}\in\hat{\mathcal{X}}}p(\hat{x})\log p(\hat{x}), \tag{10}$$

where $p(\hat{x})$ is the probability density function of the random variable $\hat{\mathcal{X}}$. The entropy $H(\hat{\mathcal{X}})$ denotes an average estimation of the uncertainty of $\hat{\mathcal{X}}$. Mutual information [35] is adopted to measure the amount of information shared by $\hat{\mathcal{X}}$ and $\mathcal{Y}$. Mutual information is applied to describe the degree of correlation between the two variables and is defined as follows:

$$I(\hat{\mathcal{X}};\mathcal{Y})=\sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{y\in\mathcal{Y}}p(\hat{x},y)\log\frac{p(\hat{x},y)}{p(\hat{x})p(y)}, \tag{11}$$

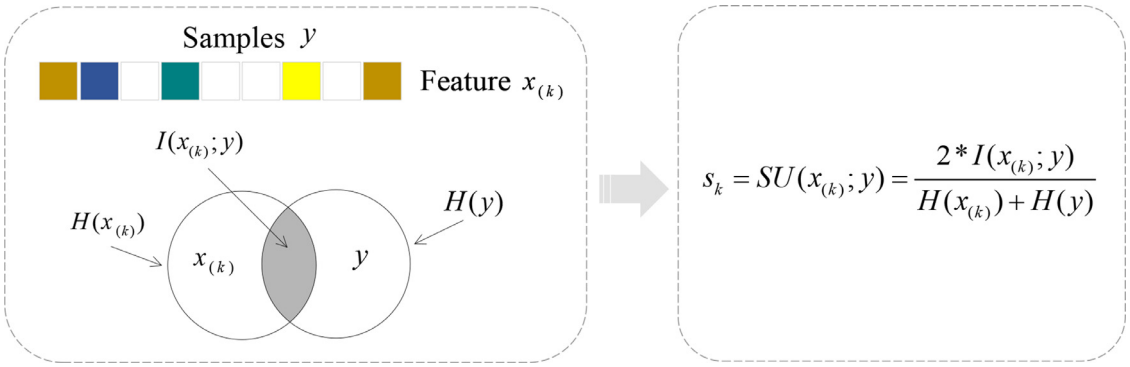

Fig. 2. The individual significance of the feature $x_{(k)}$.

In the following, we present a mechanism to evaluate the importance of the k^{th} feature by adopting the normalized mutual information which is also called symmetrical uncertainty (SU) [37]. Let s_k denote the individual significance of the k^{th} feature:

$$s_k = SU(\boldsymbol{x}_{(k)}, \boldsymbol{y}) = \frac{2 * I(\boldsymbol{x}_{(k)}; \boldsymbol{y})}{H(\boldsymbol{x}_{(k)}) + H(\boldsymbol{y})}, \tag{12}$$

where SU is obtained by normalizing the mutual information to the entropies of the features and class, and its value is restricted to the interval of $[0, 1]$. $SU(\boldsymbol{x}_{(k)}, \boldsymbol{y})$ indicates the correlation between feature $\boldsymbol{x}_{(k)}$ and class $\boldsymbol{y}$, and the construction procedure of the $SU(\boldsymbol{x}_{(k)}, \boldsymbol{y})$ is shown in Fig. 2. If $SU(\boldsymbol{x}_{(i)}, \boldsymbol{y})$ is larger than $SU(\boldsymbol{x}_{(j)}, \boldsymbol{y})$, it implies that the i^{th} feature $\boldsymbol{x}_{(i)}$ contains more information with respect to the class $\boldsymbol{y}$ than that of the j^{th} feature $\boldsymbol{x}_{(j)}$ does. Therefore, s_k can be used as a quantitative index to measure how significant a feature is, i.e., the higher the value of the s_k is, the more significant the feature $\boldsymbol{x}_{(k)}$ will be. In particular, $s_k = 0$ if the feature $\boldsymbol{x}_{(k)}$ can not provide any useful information for the class label. We use MIToolbox [33] to compute the symmetrical uncertainty in Eqn. (12), which is available at http://www.cs.man.ac.uk/~pococka4/MIToolbox.html.

To impose discriminative penalty on each feature in terms of importance degree in classification, we construct the weight coefficient of the k^{th} feature based on s_k in Eqn. (12), which can be defined as:

$$w_k = \begin{cases} \frac{1}{s_k}, & \text{if } s_k > \tau \\ \frac{1}{\tau}, & \text{otherwise} \end{cases} \tag{13}$$

which can be expressed as following:

$$w_k = \begin{cases} 1 \Big/ \left[\frac{2*I(x_{(k)};\boldsymbol{y})}{H(\boldsymbol{x}_{(k)})+H(\boldsymbol{y})}\right]^{\iota}, & \text{if } s_k > \tau \\ \frac{1}{\tau}, & \text{otherwise} \end{cases} \tag{14}$$

where $k = 1, \cdots, p$, the controllable parameter $0 < \tau \ll 1$ is a given threshold and ι is a given integer. From Eqn. (13), we can obtain that irrelevant features are imposed relatively large amount of weight, while small amount of weight is imposed on important features. This means that we can directly incorporate the significance of features in classification into the logistic regression modeling based on an adaptive Lasso penalty.

Remark 1. The initial consistent estimator is adopted to construct the weights for the adaptive lasso [23] and the initial elastic net estimator is adopted to construct the weights for the adaptive elastic net [24] and partly adaptive elastic net [25]. Though the aforementioned two kinds of weights have clear statistical meanings and could be roundly applied to assess the importance of genes, they can not indicate the obvious biological significance. In addition, these weights highly depend on the actual values or outliers of the original data. The constructed weights in this paper ensure the proposed ALCP model robust because they depend only on the probability distribution of the random variables rather than on their actual values. In addition, the proposed ALCP model for feature selection in classification can be directly incorporated into linear regression models, e.g., integrating the weighted L_1 norm penalty into adaptive Lasso models.

3.3. Adaptive grouping effect

The process of the complex diseases are generally caused by the important genes in pathways instead of individual genes. The elastic net models [18,24] are widely known for its performance of encouraging grouping effect. Generally, if the regression coefficients of the group with highly correlated variables incline to be equal, then the regression model can encourage the grouping effect. It should be noted that the important features may be highly correlated with some inessential features, the redundant noise variables could be included in these models. The following theorem shows that

the ALCP model can select the important highly correlated features adaptively, which take turns encourages an adaptive grouping effect, the ability of an model to automatically identify the features in groups into the model if one feature among them is identified.

Theorem 1. *Let* $\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)$ *be the solution of ALCP (9). If* $\hat{\beta}_j\hat{\beta}_l \neq 0$ *then*

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{1}{\lambda(1-\alpha)}\|\gamma_j \boldsymbol{x}_{(j)} - \gamma_l \boldsymbol{x}_{(l)}\|_1 = \frac{1}{\lambda(1-\alpha)}\sum_{i=1}^{n}\left|\gamma_j x_{ij} - \gamma_l x_{il}\right|. \tag{15}$$

Moreover, if $\boldsymbol{x}_{(l)}, \boldsymbol{x}_{(2)}, \ldots, \boldsymbol{x}_{(p)}$ *have been standardized to have mean 0 and unit length, then*

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{\sqrt{n}}{\lambda(1-\alpha)}\sqrt{1-\varphi\rho_{jl}}\sqrt{\gamma_j^2+\gamma_l^2}, \tag{16}$$

where $\gamma_j = \frac{w_l}{w_j+w_l}$, $\gamma_l = \frac{w_j}{w_j+w_l}$, $\rho_{jl} = \boldsymbol{x}_{(j)}^T\boldsymbol{x}_{(l)} = \sum_{i=1}^{n} x_{ij}x_{il}$, *and* $\varphi = \frac{2\gamma_j\gamma_l}{\gamma_j^2+\gamma_l^2}$.

Proof. Based on the estimated coefficient vector $\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)$, we construct the following bias $\hat{\beta}_0^*$ and coefficient vector $\boldsymbol{\beta}^*$, where

$$\hat{\beta}_0^* = \hat{\beta}_0$$

$$\hat{\beta}_{j^{'}}^* = \begin{cases} \frac{w_j}{w_j+w_l}\hat{\beta}_j + \frac{w_l}{w_j+w_l}\hat{\beta}_l & \text{if} \quad j^{'} = j, l \\ \hat{\beta}_{j^{'}} & \text{otherwise} \end{cases}$$

Let

$$L\left(\lambda, \alpha, \beta_0, \boldsymbol{\beta}\right) = \sum_{i=1}^{n} l\left(\boldsymbol{\beta}\right) + \lambda\alpha\sum_{j=1}^{p} w_j|\beta_j| + \lambda(1-\alpha)\boldsymbol{\beta}^T Q\boldsymbol{\beta}.$$

By the definition of $\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\beta}_0^*$ and $\hat{\boldsymbol{\beta}}^*$, we have $0 \leq L(\lambda, \alpha, \hat{\beta}_0^*, \hat{\boldsymbol{\beta}}^*) - L(\lambda, \alpha, \hat{\beta}_0, \hat{\boldsymbol{\beta}})$, i.e.,

$$\begin{aligned} 0 \leq & \sum_{i=1}^{n} l\left(\hat{\beta}_0^*, \hat{\boldsymbol{\beta}}^*\right) + \lambda\alpha\sum_{j=1}^{p} w_j|\hat{\beta}_j^*| + \lambda(1-\alpha)\hat{\boldsymbol{\beta}}^{*T}Q\hat{\boldsymbol{\beta}}^* \\ & - \sum_{i=1}^{n} l\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right) - \lambda\alpha\sum_{j=1}^{p} w_j|\beta_j| - \lambda(1-\alpha)\hat{\boldsymbol{\beta}}^T Q\hat{\boldsymbol{\beta}}. \end{aligned}$$

Note that $l\left(\beta_0, \boldsymbol{\beta}\right)$ is an Lipschitz continuous function. Hence,

$$\left|l\left(\hat{\beta}_0^*, \hat{\boldsymbol{\beta}}^*\right) - l\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)\right| \leq \left|\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)^T\boldsymbol{x}_i\right|. \tag{17}$$

According to Eqn. (17) and the definition of $\hat{\beta}_0^*$ and $\hat{\boldsymbol{\beta}}^*$, we have

$$\begin{aligned} \left|\sum_{i=1}^{n} l\left(\hat{\beta}_0^*, \hat{\boldsymbol{\beta}}^*\right) - \sum_{i=1}^{n} l\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)\right| &\leq \sum_{i=1}^{n}\left|\left(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\right)^T\boldsymbol{x}_i\right| \\ &= \frac{1}{w_j+w_l}\sum_{i=1}^{n}\left|(\hat{\beta}_j - \hat{\beta}_l)\left(w_j x_{il} - w_l x_{ij}\right)\right| \\ &= \frac{1}{w_j+w_l}\left|\hat{\beta}_j - \hat{\beta}_l\right|\sum_{i=1}^{n}\left|w_l x_{ij} - w_j x_{il}\right| \\ &= \frac{1}{w_j+w_l}\left|\hat{\beta}_j - \hat{\beta}_l\right|\left\|w_l\boldsymbol{x}_{(j)} - w_j\boldsymbol{x}_{(l)}\right\|_1. \end{aligned} \tag{18}$$

For the weighted L_1-norm penalty, we have

$$\begin{aligned} \sum_{j=1}^{p} w_j|\hat{\beta}_j^*| - \sum_{j=1}^{p} w_j|\hat{\beta}_j| &= w_j\left(|\hat{\beta}_j^*| - |\hat{\beta}_j|\right) + w_l\left(|\hat{\beta}_l^*| - |\hat{\beta}_l|\right) \\ &= w_j\left(\left|\frac{w_j}{w_j+w_l}\hat{\beta}_j + \frac{w_l}{w_j+w_l}\hat{\beta}_l\right| - |\hat{\beta}_j|\right) \\ &+ w_l\left(\left|\frac{w_j}{w_j+w_l}\hat{\beta}_j + \frac{w_l}{w_j+w_l}\hat{\beta}_l\right| - |\hat{\beta}_l|\right) \leq 0, \end{aligned} \tag{19}$$

Based on the reference [19], we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{*T}Q\hat{\boldsymbol{\beta}}^{*}-\hat{\boldsymbol{\beta}}^{T}Q\hat{\boldsymbol{\beta}} &= \|\sqrt{Q}\hat{\boldsymbol{\beta}}^{*}\|_2^2-\|\sqrt{Q}\hat{\boldsymbol{\beta}}\|_2^2\\ &= \sum_{j'=1}^{p} q_{j'j'}\left(\hat{\beta}_{j'}^{*}+\hat{\beta}_{j'}\right)\left(\hat{\beta}_{j'}^{*}-\hat{\beta}_{j'}\right)+\sum_{j',l'=1}^{p} 2q_{j'l'}\left(\hat{\beta}_{j'}^{*}\hat{\beta}_{l'}^{*}-\hat{\beta}_{j'}\hat{\beta}_{l'}\right)\\ &\le -\frac{2(p-1)}{1-\rho_{jl}^{2}}\left(\hat{\beta}_j-\hat{\beta}_l\right)^2.\end{aligned} \tag{20}$$

Based on Eqns. (17)-(20), we can obtain

$$\begin{aligned}0 &\le \frac{1}{w_j+w_l}\left|\hat{\beta}_j-\hat{\beta}_l\right|\|w_l\boldsymbol{x}_{(j)}-w_j\boldsymbol{x}_{(l)}\|_1\\ &\quad -\lambda(1-\alpha)\frac{2(p-1)}{1-\rho_{jl}^{2}}\left(\hat{\beta}_j-\hat{\beta}_l\right)^2,\end{aligned}$$

$$|\hat{\beta}_j-\hat{\beta}_l|\le\frac{1-\rho_{jl}^{2}}{\lambda(1-\alpha)2(p-1)}\|\frac{w_l}{w_j+w_l}\boldsymbol{x}_{(j)}-\frac{w_j}{w_j+w_l}\boldsymbol{x}_{(l)}\|_1\le\frac{1}{\lambda(1-\alpha)}\|\frac{w_l}{w_j+w_l}\boldsymbol{x}_{(j)}-\frac{w_j}{w_j+w_l}\boldsymbol{x}_{(l)}\|_1,$$

which can be easily transformed into Eqn. (15). Moreover, if $\boldsymbol{x}_{(l)},\boldsymbol{x}_{(2)},\ldots,\boldsymbol{x}_{(p)}$ have been standardized to have mean 0 and unit length, then

$$\begin{aligned}\|\gamma_j\boldsymbol{x}_{(j)}-\gamma_l\boldsymbol{x}_{(l)}\|_1 &\le \sqrt{n}\|\gamma_j\boldsymbol{x}_{(j)}-\gamma_l\boldsymbol{x}_{(l)}\|_2\\ &\le \sqrt{n}\sqrt{\gamma_j^2+\gamma_l^2-2\gamma_j\gamma_l\boldsymbol{x}_{(j)}^{T}\boldsymbol{x}_{(l)}}\\ &= \sqrt{n}\sqrt{\gamma_j^2+\gamma_l^2}\sqrt{1-\varphi\rho_{jl}}\end{aligned} \tag{21}$$

We substitute the Eqn. (21) into Eqn. (15) yields the Eqn. (21), which complete the proof. □

From the Theorem 1, we can conclude that for highly correlated predictors, ALCP model can encourage their coefficient parameters to be equal. The highly correlated features in groups can be selected or removed adaptively, which implies the proposed ALCP model has the adaptive group effecting of feature selection.

4. A learning algorithm for the proposed ALCP model

An learning algorithm for the proposed ALCP model according to the coordinate descent algorithm [39] is developed in this section. Friedman et al. [40] have been proved that the Newton algorithm for maximizing the log-likelihood (5) amounts to iteratively reweighted least squares. Hence, the quadratic approximation of the log-likelihood (5) based on the current estimator $\tilde{\beta}$ (i.e., Taylor expansion about current estimates) can be expressed as follows:

$$l(\beta)\approx l(\tilde{\beta})+(\beta-\tilde{\beta})\nabla l(\tilde{\beta})+\frac{1}{2}(\beta-\tilde{\beta})H(\beta-\tilde{\beta}), \tag{22}$$

where $\nabla l(\tilde{\beta})$ is the gradient of $l(\beta)$, $H=\frac{1}{n}X^{T}TX$ is the Hessian matrix of $l(\beta)$, and T is a diagonal matrix with elements:

$$t_i=\tilde{\pi}(\boldsymbol{x}_i)(1-\tilde{\pi}(\boldsymbol{x}_i)), \tag{23}$$

where $\tilde{\pi}(\boldsymbol{x}_i)$ is assessed in the current parameter estimates. Hence, the log-likelihood (5) can be approximated as follows:

$$l(\beta)=\frac{1}{2}\sum_{i=1}^{n}\left[t_i\left(z_i-\boldsymbol{x}_i^{T}\boldsymbol{\beta}\right)^2\right], \tag{24}$$

where $z_i=\boldsymbol{x}_i^{T}\tilde{\boldsymbol{\beta}}+\frac{y_i-\tilde{\pi}(\boldsymbol{x}_i)}{\tilde{\pi}(\boldsymbol{x}_i)(1-\tilde{\pi}(x_i))}$ is an estimated pseudo response in terms of the current parameters. Therefore, the ALCP model in Eqn. (7) is approximated as follows:

$$\hat{\boldsymbol{\beta}}=\arg\min\left\{\frac{1}{2}\sum_{i=1}^{n}\left[t_i\left(z_i-\boldsymbol{x}_i^{T}\boldsymbol{\beta}\right)^2\right]+\lambda R(\boldsymbol{\beta})\right\}, \tag{25}$$

The approximation of the ALCP model in Eqn. (25) can be expressed as following:

$$f(\tilde{\boldsymbol{\beta}})=\frac{1}{2}\sum_{i=1}^{n}t_i\left(z_i-\tilde{\beta}_0-x_{ij}\beta_j-\sum_{k\neq j}^{p}x_{ik}\tilde{\beta}_k\right)^2+\lambda R(\boldsymbol{\beta}), \tag{26}$$

where

$$\begin{aligned}R(\boldsymbol{\beta}) &= \lambda\alpha w_j|\beta_j|+\lambda(1-\alpha)\sum_{k\neq j}^{p}\left(\frac{(\beta_j-\beta_k)^2}{1-\rho_{jk}}+\frac{(\beta_j+\beta_k)^2}{1+\rho_{jk}}\right)+\\ &\quad \lambda\alpha\sum_{k\neq j}w_k|\tilde{\beta}_k|+\lambda(1-\alpha)\sum_{l\neq j}^{p}\sum_{k\neq j}^{p}\left(\frac{(\tilde{\beta}_l-\tilde{\beta}_k)^2}{1-\rho_{lk}}+\frac{(\tilde{\beta}_l+\tilde{\beta}_k)^2}{1+\rho_{lk}}\right).\end{aligned} \tag{27}$$

Assume that all values of $\tilde{\beta}_k$ for $k \neq j$ and β_0 are fixed. The objective function (26) can be minimized with respect to β_j. When $\beta_j \neq 0$, the Eqn. (26) is differentiated at β_j, which can be expressed as:

$$\begin{aligned}\frac{\partial f\left(\tilde{\boldsymbol{\beta}}\right)}{\partial \beta_j} &= -\sum_{i=1}^{n} t_i x_{ij}\left(z_i - \tilde{\beta}_0 - \sum_{k\neq j}^{p} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j\right) + \lambda\alpha w_j \operatorname{sign}\left(\beta_j\right) \\ &\quad +4\lambda(1-\alpha)\sum_{k\neq j}^{p}\frac{\beta_j - \rho_{jk}\tilde{\beta}_k}{1-\rho_{jk}^2}, \\ &= -\sum_{i=1}^{n} t_i x_{ij}\left(z_i - \tilde{z}_i^{(j)}\right) + \sum_{i=1}^{n} t_i x_{ij}^2\beta_j + \lambda\alpha w_j \operatorname{sign}\left(\beta_j\right) \\ &\quad +4\lambda(1-\alpha)\sum_{k\neq j}^{p}\frac{\beta_j - \rho_{jk}\tilde{\beta}_k}{1-\rho_{jk}^2},\end{aligned} \tag{28}$$

when $\beta_j = 0$, we have

$$\left|\sum_{i=1}^{n} t_i x_{ij}\left(z_i - \tilde{z}_i^{(j)}\right) + 4\lambda(1-\alpha)\sum_{k\neq j}^{p}\frac{\rho_{jk}\tilde{\beta}_k}{1-\rho_{jk}^2}\right| \leq \lambda\alpha w_j \tag{29}$$

The coordinate update of β_j in the ALCP model can be expressed as:

$$\tilde{\beta}_j \leftarrow \frac{S\left(\sum_{i=1}^{n} t_i x_{ij}\left(z_i - \tilde{z}_i^{(j)}\right) + 4\lambda(1-\alpha)\sum_{k\neq j}^{p}\frac{\rho_{jk}\tilde{\beta}_k}{1-\rho_{jk}^2}, \lambda\alpha w_j\right)}{\sum_{i=1}^{n} t_i x_{ij}^2 + 4\lambda(1-\alpha)\sum_{k\neq j}^{p}\frac{1}{1-\rho_{jk}^2}}, \tag{30}$$

where $\tilde{z}_i^{(j)} = \tilde{\beta}_0 + \sum_{k\neq j}^{p} x_{ik}\tilde{\beta}_k$ is the partial residual with respect to β_j. $S(\zeta, \gamma)$ is the soft-thresholding operator [41] as:

$$S(\zeta,\gamma) \equiv \operatorname{sign}(\zeta)(|\zeta| - \gamma)_+ = \begin{cases}\zeta - \gamma, & \text{if } \zeta > 0 \text{ and } \gamma < |\zeta| \\ \zeta + \gamma, & \text{if } \zeta < 0 \text{ and } \gamma < |\zeta| \\ 0, & \text{if } \gamma \geq |\zeta|.\end{cases} \tag{31}$$

The coordinate update in (30) is repeated for $j = 1, \cdots, p$ until convergence and the detailed algorithm for solving the ALCP is shown in Algorithm 1 . The algorithm for solving the ALCP model is based on the simple least squares coefficient on the

Algorithm 1: Algorithm for ALCP.

Input: Data set $(X, \boldsymbol{y})$, maxiteration number $\mathcal{Q}$, parameters τ, α and λ.
Output: $\hat{\boldsymbol{\beta}}$, $\mathcal{F}$.

1 $\mathcal{F} \leftarrow \emptyset$, $\hat{\beta}_j^{(0)} \leftarrow 0$ $(j = 1, \cdots, p)$;
2 **for** $k = 1$ **to** p **do**
3 Compute $H(\boldsymbol{x}_{(k)})$ by Eqn (10) in terms of the reference [35];
4 Compute $H(\boldsymbol{y})$ by Eqn (10) in terms of the reference [35];
5 Compute s_k by Eqn. (12) in terms of the reference [37];
6 $w_k \leftarrow 1/s_k^2$;
7 **for** $l = 1$ **to** $\mathcal{Q}$ **do**
8 **for** $j = 1$ **to** p **do**
9 Update the $\tilde{\beta}_j^{(l)}$ by Eqn. (30) based on the reference [39];
10 $\hat{\boldsymbol{\beta}}^{(l)} \leftarrow (\tilde{\beta}_1^{(l)}, \cdots, \tilde{\beta}_p^{(l)})^T$;
11 **if** $\|\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)}\|_2 \leq \varepsilon$ **then**
12 Break;
13 $\hat{\boldsymbol{\beta}} \leftarrow (\hat{\beta}_1^{(\mathcal{Q})}, \cdots, \hat{\beta}_p^{(\mathcal{Q})})^T$;
14 Select the non-zero coefficients of $\hat{\boldsymbol{\beta}}$ and form a feature set $\mathcal{F}$;
15 **return** $\hat{\boldsymbol{\beta}}$, $\mathcal{F}$.

partial residual $(z_i - \tilde{z}_i^{(j)})$. Since only the significant non-zero coefficients are updated and many parameters will be skipped in each iterative step for high-dimensional and sparse settings, coordinate descent is a very efficient method for dealing with high-dimensional sparse data. This implies that the coordinate descent algorithm is an efficient tool for regression modeling. The time complexity of lines 1–6 is $O(p)$. The time complexity in lines 7–14 is $O\left(\mathcal{Q} * p\right)$. Therefore, the overall computational complexity of the Algorithm 1 is $O\left(\mathcal{Q} * p\right)$.

Table 1
The ACA of the seven models on the test datasets across over 100 runs. The number in parenthesis is the standard deviation.

n	p	Lasso	EN	L1CP	WLR	AENCMI	ASGL	ALCP
60	1500	0.891 (0.041)	0.906 (0.044)	0.891 (0.067)	0.902 (0.032)	0.933 (0.027)	0.920 (0.041)	0.936 (0.034)
	5000	0.877 (0.051)	0.889 (0.044)	0.880 (0.076)	0.883 (0.065)	0.900 (0.044)	0.883 (0.050)	0.904 (0.044)
	10000	0.866 (0.057)	0.881 (0.051)	0.869 (0.080)	0.881 (0.058)	0.892 (0.045)	0.883 (0.051)	0.893 (0.050)
80	1500	0.919 (0.038)	0.935 (0.035)	0.924 (0.039)	0.930 (0.028)	0.946 (0.027)	0.948 (0.030)	0.958 (0.024)
	5000	0.910 (0.039)	0.924 (0.044)	0.912 (0.078)	0.908 (0.037)	0.938 (0.033)	0.926 (0.045)	0.948 (0.033)
	10000	0.897 (0.038)	0.913 (0.036)	0.898 (0.033)	0.900 (0.038)	0.927 (0.030)	0.923 (0.033)	0.937 (0.029)
100	1500	0.924 (0.030)	0.941 (0.034)	0.928 (0.040)	0.942 (0.028)	0.952 (0.022)	0.959 (0.027)	0.967 (0.021)
	5000	0.932 (0.026)	0.944 (0.026)	0.940 (0.030)	0.945 (0.034)	0.954 (0.024)	0.957 (0.024)	0.967 (0.020)
	10000	0.919 (0.037)	0.933 (0.034)	0.937 (0.051)	0.934 (0.035)	0.944 (0.028)	0.944 (0.031)	0.957 (0.027)

5. Experimental results

In this paper, we compare the proposed ALCP model with the other six competing sparse learning models: Lasso [12], Elastic Net (EN) [18], L1CP [20], Wilcoxon.penalized Logistic Regression (WLR) [30], Adaptive Elastic Net with Conditional Mutual Information (AENCMI) [22] and Adaptive Sparse Group lasso (ASGL) [32]. All the aforementioned seven sparse learning models are applied to deal with binary-classification tasks on both simulated datasets and benchmark datasets. For the sake of fairness, logistic regression is applied as the loss function for all compared methods. The main metrics for assessing the performance of the different models are the classification accuracy and the number of features selected (NFS). The classification accuracy can be defined as $\frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$ which is a general evaluation index of classification problem. The number of features selected (NFS) is the selected features corresponding to the non-zero coefficients, which can reflect the feature selection performance. In addition, the number of correct relevant features selected (NCFS) is also a key metric to reflect the feature selection performance.

5.1. Simulation results

To evaluate the performance of the ALCP on the simulated dataset. We generate the simulated datasets involving n observations from the model:

$$\log\left(\frac{Y_i}{1-Y_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j. \tag{32}$$

In every simulation, the dimension p of the predictor vector is 1500, and the observations (rows of X) are iid from a $N(0, \Sigma)$ distribution, where Σ is a $p \times p$ block diagonal matrix whose (i, j) entry is $\Sigma_{i,j} = 1$ when $i = j$, $\Sigma_{i,j} = 0.95$ when $1 \le i \ne j \le 30$ and $31 \le i \ne j \le 60$, and $\Sigma_{i,j} = 0$ otherwise. A threshold for binary value of Y is considered as 0.5. Moreover, $\beta_j = 1$ when $1 \le j \le 10$, $\beta_j = -1$ when $31 \le j \le 40$, and $\beta_j = 0$ otherwise. In other words, there are two sets of 30 highly correlated features, the first ten true coefficients in each set are nonzero, and many irrelevant features are in the simulated high-dimensional and low sample size data. Nine groups of datasets with training sample size $n = 60, 80, 100$ and feature size $p = 1500, 5000, 10000$ are considered for simulation studies. Each dataset includes a test dataset with testing sample size of 100 and $1500, 5000, 10000$ features to assess the seven regularized logistic regression models.

To avoid the one-time occasionality in experiments, we repeat the experiments for 100 times and the Average Classification Accuracy (ACA) are reported in Table 1. From Table 1, we can obtain that the predication performance of all the seven models are all improved as the number of training samples increases. For example, the ACA of the proposed ALCP model is 0.933, 0.958 and 0.967 with the training sample size $n = 60, 80, 100$, respectively when $p = 1500$. Similarly, the ACA of the proposed ALCP model is 0.904, 0.948 and 0.967 with the training sample size $n = 60, 80, 100$, respectively when $p = 5000$. The ACA of the proposed ALCP model is 0.893, 0.937 and 0.957 with the training sample size $n = 60, 80, 100$, respectively when $p = 10000$. Overall, the ACA and standard deviations of ALCP are superior to those of the six compared models on all the nine simulated datsets. In particular, as the number of features increases, the standard deviations of ALCP becomes more stable. This implies that ALCP achieves the best and stable predication performance among all the seven regularized regression models.

In order to evaluate the feature selection performance, we also show the average number of the features selected and the correct features selected in 100 runs for each model. The recovery rate (RR) is defined as the ratio of the average number of the correct relevant features selected (ANCFS) to the average number of features selected (ANFS), which can assess the comprehensive performance of feature selection. As shown in the Table 2, we can obtain that the ANCFS of ALCP is similar to that of ASGL and L1CP and much higher than that of the remaining four models. This implies that the proposed ALCP model has the ability to select more correct relevant features. In addition, the RR of ALCP is the highest among the seven models, which implies that the feature selection performance of ALCP is better than the other six models.

In the following, we compare the percentage of features selected by proposed vs the other six models on the once of the 100 runs. The percentage of features captured by ALCP and the other six methods is 0.214, 0.416, 0.528, 0.450, 0.580, 0.586

Table 2
The average number of features selected by the six models on the test data sets across over 100 runs.

n	p	Model	ANCFS	ANFS	RR
60	1500	Lasso	2.28	14.09	0.162
		EN	11.63	45.05	0.258
		L1CP	17.37	75.27	0.231
		WLR	16.45	56.87	0.289
		AENCMI	16.17	58.36	0.277
		ASGL	16.20	59.71	0.271
		ALCP	16.83	51.31	0.328
	5000	Lasso	2.15	15.76	0.136
		EN	11.56	52.76	0.219
		L1CP	14.95	83.86	0.178
		WLR	16.27	83.06	0.195
		AENCMI	16.79	77.14	0.218
		ASGL	16.66	69.26	0.241
		ALCP	16.86	69.21	0.244
	10,000	Lasso	1.92	14.08	0.136
		EN	11.45	57.56	0.199
		L1CP	13.98	90.45	0.155
		WLR	16.94	91.05	0.186
		AENCMI	16.40	89.13	0.184
		ASGL	16.94	80.42	0.211
		ALCP	16.95	77.41	0.219
80	1500	Lasso	2.93	12.58	0.233
		EN	12.84	46.85	0.274
		L1CP	18.72	88.16	0.212
		WLR	15.36	79.59	0.193
		AENCMI	15.76	52.35	0.301
		ASGL	17.40	55.97	0.311
		ALCP	16.00	47.01	0.340
	5000	Lasso	2.52	14.12	0.178
		EN	12.62	48.19	0.262
		L1CP	18.07	96.57	0.187
		WLR	16.23	72.73	0.223
		AENCMI	16.35	57.24	0.286
		ASGL	16.30	57.17	0.285
		ALCP	16.30	50.84	0.321
	10,000	Lasso	2.35	14.70	0.160
		EN	12.40	53.04	0.234
		L1CP	17.88	82.99	0.215
		WLR	16.29	78.78	0.207
		AENCMI	15.41	65.73	0.234
		ASGL	16.70	64.47	0.259
		ALCP	16.90	54.70	0.309
100	1500	Lasso	2.88	11.29	0.255
		EN	13.65	44.90	0.304
		L1CP	18.96	69.78	0.272
		WLR	14.97	45.56	0.329
		AENCMI	16.14	43.81	0.368
		ASGL	17.26	48.16	0.354
		ALCP	18.48	48.15	0.384
	5000	Lasso	2.66	10.60	0.251
		EN	13.51	45.73	0.295
		L1CP	18.88	76.87	0.246
		WLR	15.45	70.44	0.219
		AENCMI	16.29	60.95	0.267
		ASGL	17.00	53.69	0.317
		ALCP	17.42	47.06	0.366
	10,000	Lasso	2.60	13.38	0.194
		EN	12.94	47.59	0.272
		L1CP	18.46	81.05	0.228
		WLR	15.90	73.34	0.217
		AENCMI	16.14	74.09	0.218
		ASGL	17.04	63.24	0.278
		ALCP	17.60	49.42	0.356

when $n = 60$ and $p = 1500$. The percentage of features captured by ALCP and the other six methods is 0.190, 0.441, 0.582, 0.443, 0.639, 0.601 when $n = 60$ and $p = 5000$. The percentage of features captured by ALCP and the other six methods is 0.196, 0.610, 0.528, 0.595, 0.622, 0.607 when $n = 60$ and $p = 10000$. The percentage of features captured by ALCP and the other six methods is 0.203, 0.506, 0.532, 0.427, 0.624, 0.650 when $n = 80$ and $p = 1500$. The percentage of features captured by ALCP and the other six methods is 0.144, 0.602, 0.511, 0.507, 0.716, 0.724 when $n = 80$ and $p = 5000$. The percentage of features captured by ALCP and the other six methods is 0.226, 0.548, 0.408, 0.496, 0.790, 0.887 when $n = 80$ and $p = 10000$. The percentage of features captured by ALCP and the other six methods is 0.195, 0.640, 0.780, 0.465, 0.691, 0.719 when $n = 100$ and $p = 1500$. The percentage of features captured by ALCP and the other six methods is 0.242, 0.716, 0.883, 0.549, 0.759, 0.955 when $n = 100$ and $p = 5000$. The percentage of features captured by ALCP and the other six methods is 0.229, 0.463, 0.770, 0.571, 0.786, 0.976 when $n = 100$ and $p = 10000$.

5.2. Application to benchmark dataset

In order to assess the performance of the proposed ALCP model in the field of cancer classification, text classification and digit recognition, we compare ALCP with the other six models on six public benchmark datasets: Colon, Prostate, Leukemia, PCMAC, RELATHE and Gisette. Colon cancer dataset [42] contains gene expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes achieved with an Affymetrix oligonucleotide array. This data is available on-lineat http://wws.weizmann.ac.il/mcb/UriAlon/download/downloadable-data. A subset of 2000 genes with the highest minimal intensity across the samples is adopted in our experiment. Prostate cancer dataset [43] consists of 102 samples of 52 prostate tumor samples and 50 non-tumor tissues, where each sample has 12,600 genes. The address of this data is http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi. Leukemia dataset consists of 72 samples of 47 ALL samples and 25 AML samples, where each sample has 7129 genes. The Leukemia data is available on-line http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43. PCMAC and RELATHE [44] are two text datasets, the former includes 1943 samples and 3289 features and the latter includes 1427 samples and 4322 features. Gisette [44] is a digit recognition dataset including 7000 samples and 5000 features. The last three data can be download by https://jundongl.github.io/scikit-feature/datasets.html. To reduce the computational overhead, we lower the numbers of samples from 7000 to 100 for the Gisette data.

We first randomly select 2/3 samples as the training samples and 1/3 samples as the testing samples. We first learn the different models on the training data applying the 5-fold cross-validation method, and then we test all the models on the testing data. Similar to Section 5.1, we also repeat the experiments for 100 times on the six public benchmark datasets to avoid the one-time occasionality in experiments. The results of ACA and Average Number of Feature Selected (ANFS) are shown in Table 3. Table 3 shows that ALCP achieves the maximum ACA as compared with the other methods and the ACA achieved by ALCP are monotone on Colon, Prostate and Leukemia datasets. Obviously, ALCP outperforms the L1CP in terms of ACA on both of the three datasets. The improvements are five point three percent, three point four percent and three point nine percent on Colon, Prostate and Leukemia datasets, respectively. In addition, the ACA obtained by the ALCP are also higher than those of the most recent three models ASGL, AENCMI and WLR on Colon, Prostate and Leukemia datasets. Since the network penalty of the ALCP can infer the correlation network of the genes, based on which ALCP can select the highly correlated genes in the complete network, which will improve the classification performance on the cancer data. As shown in Table 3, the ACA of ALCP is larger than that of the other six models on the PCMAC, RELATHE and Gisette datasets. The superior classification performance is due to the adaptive weights construction method in ALCP model. The ANFS on PCMAC and RELATHE datasets are suddenly much higher than the other models which seems like outliers. In addition, Table 3 shows that the standard deviations of WLR are significantly higher than the other six models on PCMAC and RELATHE datasets. The possible cause of those large errors may be because the two data network structures are not obvious, which cannot be fully identified by WLR. The main function running time in seconds by the six models on the six datasets are shown in Table 4. The running time of L1CP and WLR is close to the ALCP, which is much faster than AENCMI. Although Lasso and EN has the less running time than other compared models, the feature selection performance and classification performance of them are not as good as its running time, which has been verified in Table 3. In summary, the running time of the proposed ALCP is acceptable. In order to be useful to know how many genes were overlapping between the proposed ALCP and the existing models on the three biological datasets in Table 3, we compare the percentage of genes captured by proposed vs existing six models and show the detailed results in the following. The percentage of genes captured by the proposed ALCP vs the other six models on Colon dataset are 0.408, 0.533, 0.545, 0.556, 0.623, 0.733, respectively. The percentage of genes captured by the proposed ALCP vs the existing models on Prostate dataset are 0.319, 0.614, 0.340, 0.491, 0.593, 0.640, respectively. The percentage of genes captured by ALCP vs the existing models on Leukemia dataset are 0.372, 0.554, 0.583, 0.482, 0.619, 0.615, respectively.

As can be seen from Table 3, ALCP selects more ANFS than the Lasso, AENCMI and ASGL and fewer ANFS than WLR, EN and L1CP on the most of the datasets. In the Colon dataset, for example, ALCP selects 36.09 ANFS compared to 12.61, 54.51, 41.31, 36.39, 25.40 and 32.88 NGS for Lasso, EN, L1CP, WLR, AENCMI and ASGL respectively. In the Prostate dataset, for example, ALCP selects 75.98 ANFS compared to 18.20, 76.85, 76.89, 99.03, 69.15 and 56.87 ANFS for Lasso, EN, L1CP, WLR, AENCMI and ASGL respectively. In the Leukemia dataset, for example, ALCP selects 75.98 NGS compared to 17.26, 80.32, 76.42, 73.14, 44.77 and 60.06 ANFS for Lasso, EN, L1CP, WLR, AENCMI and ASGL respectively. ALCP has the potential to select more features than the Lasso, which indicates that most of these additionally selected features are probably highly

Table 3
The average number of features selected by the seven models on the test data sets across over 100 runs. The number in parenthesis is the standard deviation.

Dataset	Model	ACA	ANFS
Colon	Lasso	0.816 (0.058)	12.61 (6.13)
	EN	0.829 (0.060)	54.51 (20.29)
	L1CP	0.816 (0.093)	41.31 (58.75)
	WLR	0.851 (0.060)	36.39 (93.27)
	AENCMI	0.856 (0.056)	25.40 (18.50)
	ASGL	0.855 (0.061)	32.88 (42.19)
	ALCP	0.869 (0.055)	36.09 (13.19)
Prostate	Lasso	0.910 (0.051)	18.20 (7.52)
	EN	0.925 (0.042)	76.85 (37.23)
	L1CP	0.901 (0.058)	76.89 (68.58)
	WLR	0.926 (0.045)	99.03 (81.70)
	AENCMI	0.924 (0.042)	69.15 (67.39)
	ASGL	0.922 (0.044)	56.87 (47.84)
	ALCP	0.935 (0.036)	75.98 (34.66)
Leukemia	Lasso	0.915 (0.054)	17.26 (6.49)
	EN	0.939 (0.052)	80.32 (23.90)
	L1CP	0.917 (0.089)	76.42 (64.79)
	WLR	0.945 (0.044)	73.14 (54.12)
	AENCMI	0.946 (0.041)	44.77 (22.08)
	ASGL	0.940 (0.042)	60.06 (30.94)
	ALCP	0.956 (0.035)	68.18 (19.42)
PCMAC	Lasso	0.905 (0.015)	295.22 (77.91)
	EN	0.902 (0.015)	442.04 (92.79)
	L1CP	0.905 (0.015)	614.41 (140.98)
	WLR	0.898 (0.011)	934.08 (373.29)
	AENCMI	0.905 (0.013)	602.87 (135.17)
	ASGL	0.906 (0.012)	302.70 (141.68)
	ALCP	0.910 (0.011)	583.94 (124.01)
RELATHE	Lasso	0.887 (0.015)	304.24 (52.77)
	EN	0.895 (0.013)	485.52 (85.29)
	L1CP	0.898 (0.014)	728.65 (105.44)
	WLR	0.899 (0.013)	925.65 (323.45)
	AENCMI	0.898 (0.013)	733.34 (104.44)
	ASGL	0.899 (0.019)	228.04 (62.71)
	ALCP	0.900 (0.013)	179.74 (42.49)
Gisette	Lasso	0.859 (0.056)	16.46 (10.51)
	EN	0.856 (0.055)	75.12 (42.52)
	L1CP	0.857 (0.058)	83.27 (53.41)
	WLR	0.861 (0.067)	169.47 (98.34)
	AENCMI	0.870 (0.059)	85.88 (53.17)
	ASGL	0.862 (0.059)	44.29 (32.87)
	ALCP	0.881 (0.051)	57.40 (24.15)

Table 4
The average running time by the seven models on the six benchmark datsets across over 100 runs.

Dataset	Lasso	EN	L1CP	WLR	AENCMI	ASGL	ALCP
Colon	0.07	0.08	2.23	0.92	33.62	2.28	2.05
Prostate	0.42	2.32	32.33	22.57	985.47	67.04	64.73
Leukemia	0.2	0.22	12.57	6.95	319.08	15.89	16.60
PCMAC	3.66	3.58	281.27	252.58	433.32	287.69	307.04
RELATHE	4.41	4.11	112.44	107.91	516.42	96.06	92.80
Gisette	0.12	0.13	5.34	2.39	121.30	4.97	4.96

correlated. Although the ALCP selects slightly more features features than AENCMI, the higher accuracy and lower stability obtained by the ALCP as compared with AENCMI. Compared with EN and L1CP, ALCP has the potential to select fewer but important features, which implies that most of these features are beneficial to classification. In addition, the least standard deviations of both classification performance and feature selection performance for ALCP are shown to be more stable than the other six models in most cases. It proves that the proposed weighting strategy improves the classification performance and feature selection performance.

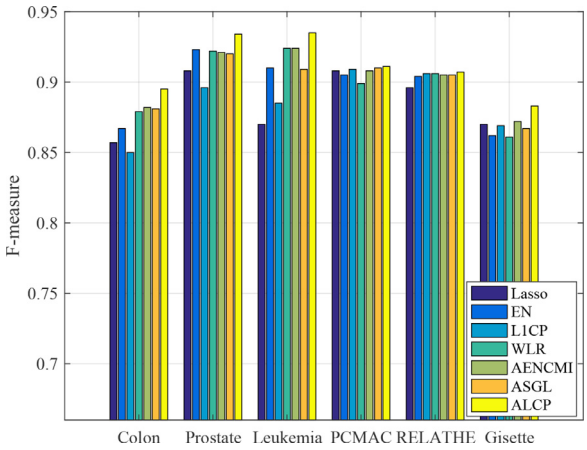

Fig. 3. F-measure of each model on the six benchmark datasets.

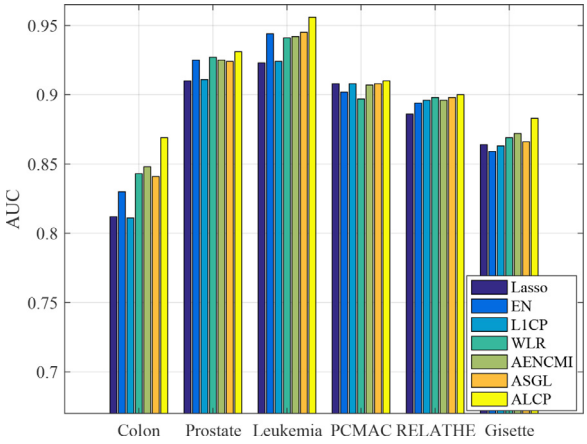

Fig. 4. AUC of each model on the six benchmark datasets.

Note that there exists an imbalance between the positive and negative samples in most classification problems, thus accuracy measurement of each class is important to provide further insight for the performance of the seven models. Therefore, we adopt the overall classification performance evaluation metrics: F-measure and Area Under Curve (AUC). F-measure is defined as:

$$F-measure=\frac{2PR}{P+R}=\frac{2TP}{2TP+FP+FN}, \tag{33}$$

where $P=\frac{TP}{TP+FP}$ is precision and $R=\frac{TP}{TP+FN}$ is recall, 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.

Suppose that there are n_1 positive class samples and n_2 negative class samples. $\{s_1,\ldots,s_{n_1}\}$ are the scores of the positive points and $\{t_1,\ldots,t_{n_2}\}$ are the scores of the negative points. The AUC is defined as:

$$\mathrm{AUC}=\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(I(s_i>t_j)+\frac{1}{2}I(s_i=t_j)\right), \tag{34}$$

where $I(s_i>t_j)=1$ when $s_i>t_j$, and 0 otherwise.

Figure 3 and Fig. 4 display the overall classification results of the seven models on all the six benchmark datasets. The F-measure of the seven models on the six datasets is shown in Fig. 3. Fig. 3 shows that the F-measure of ALCP is higher than that of the other six models on the six datasets. Fig. 4 reports the AUC of the seven compared models on all six datasets. It can be observed from Fig. 4 that ALCP achieves the highest AUC on all the six datasets as compared with the other six models. Therefore, the proposed ALCP model obtains the best overall classification performance on each benchmark dataset among all seven models.

In order to obtain the correlation structure of the six datasets, we calculate the similarity matrix of the datasets and remove the obvious weak correlation features, then draw the network structure diagram by Gephi software. We set different

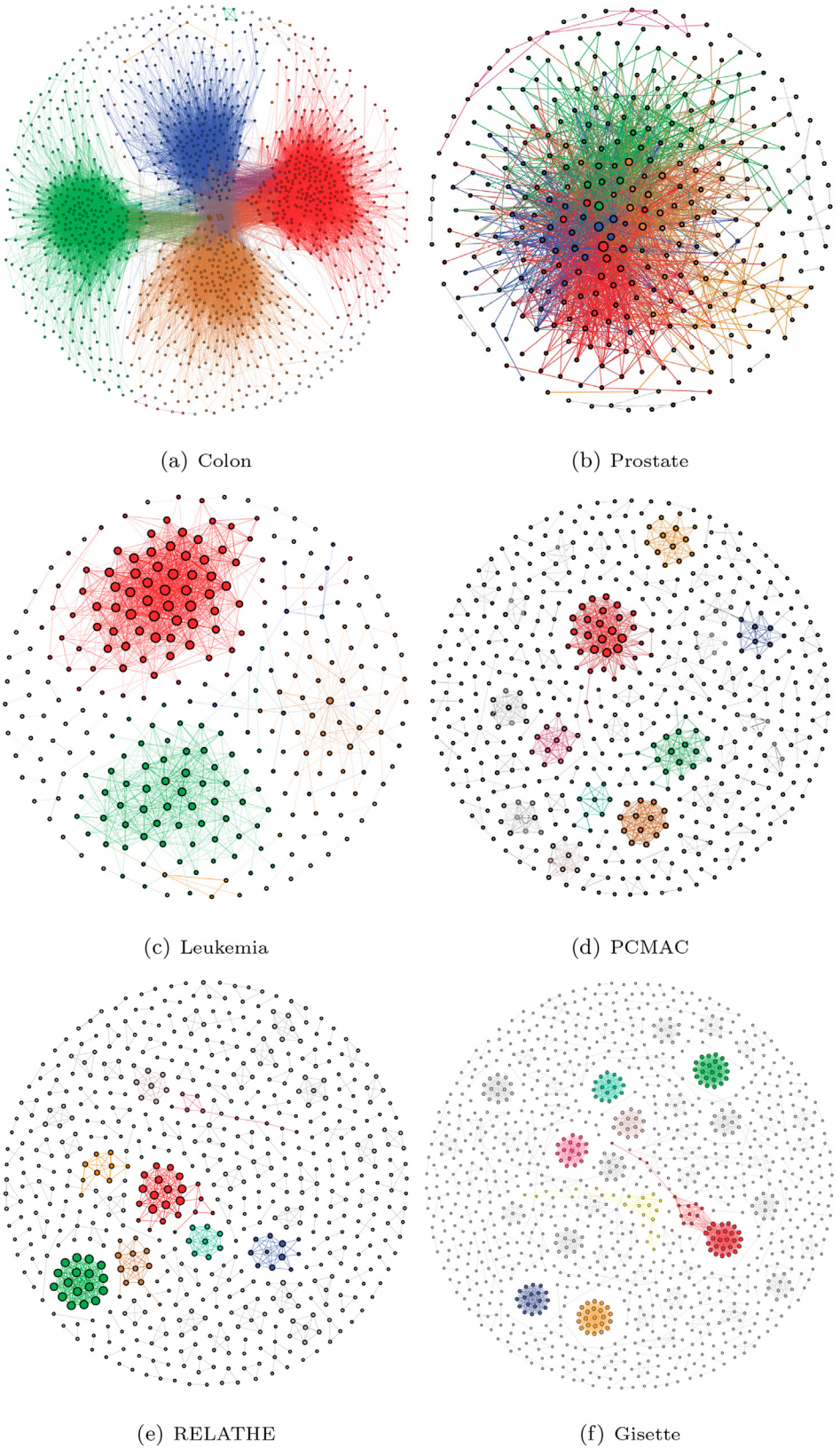

Fig. 5. The correlation structure of six benchmark datasets.

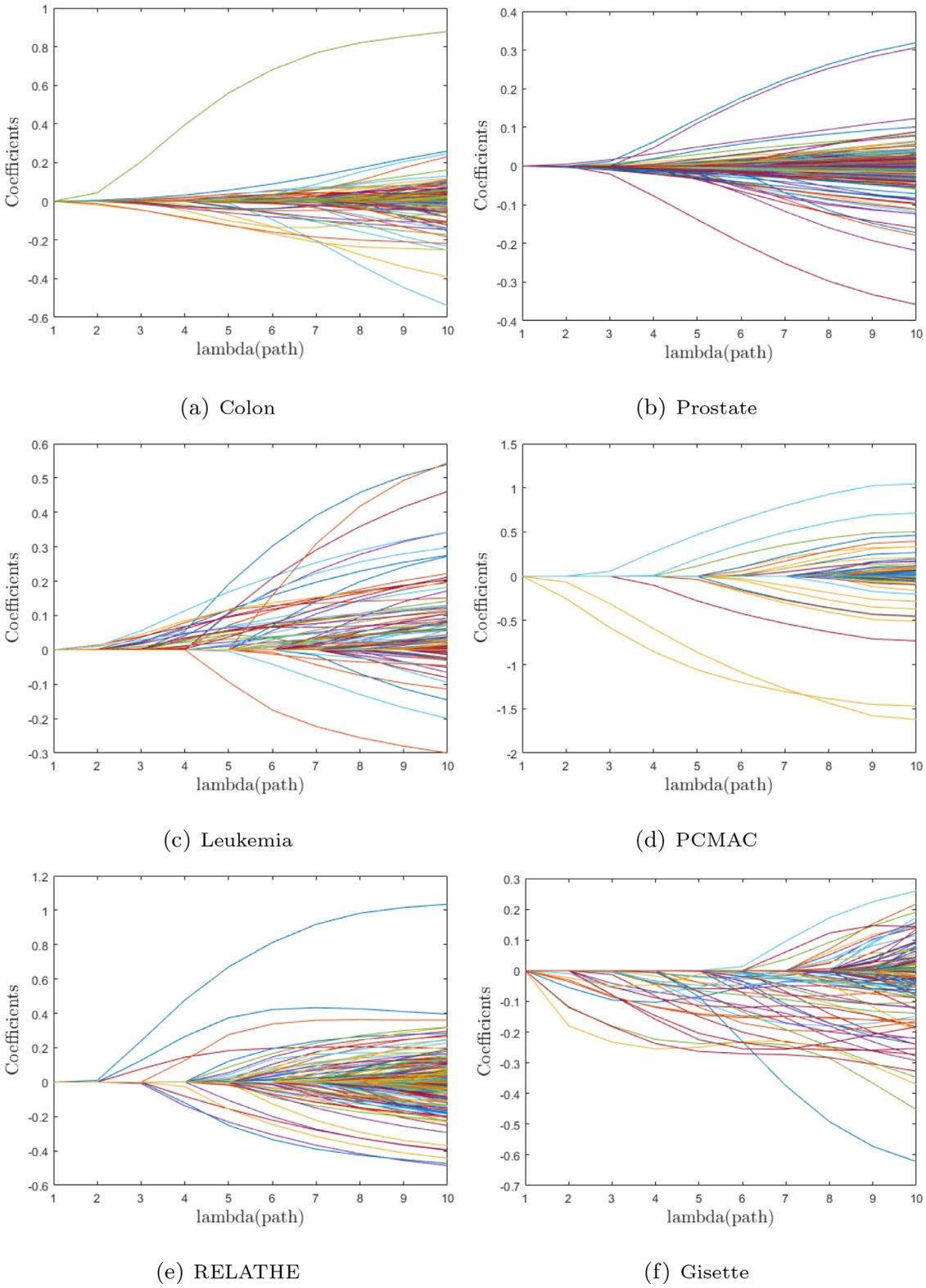

Fig. 6. The coefficient paths of the ALCP based parameter path.

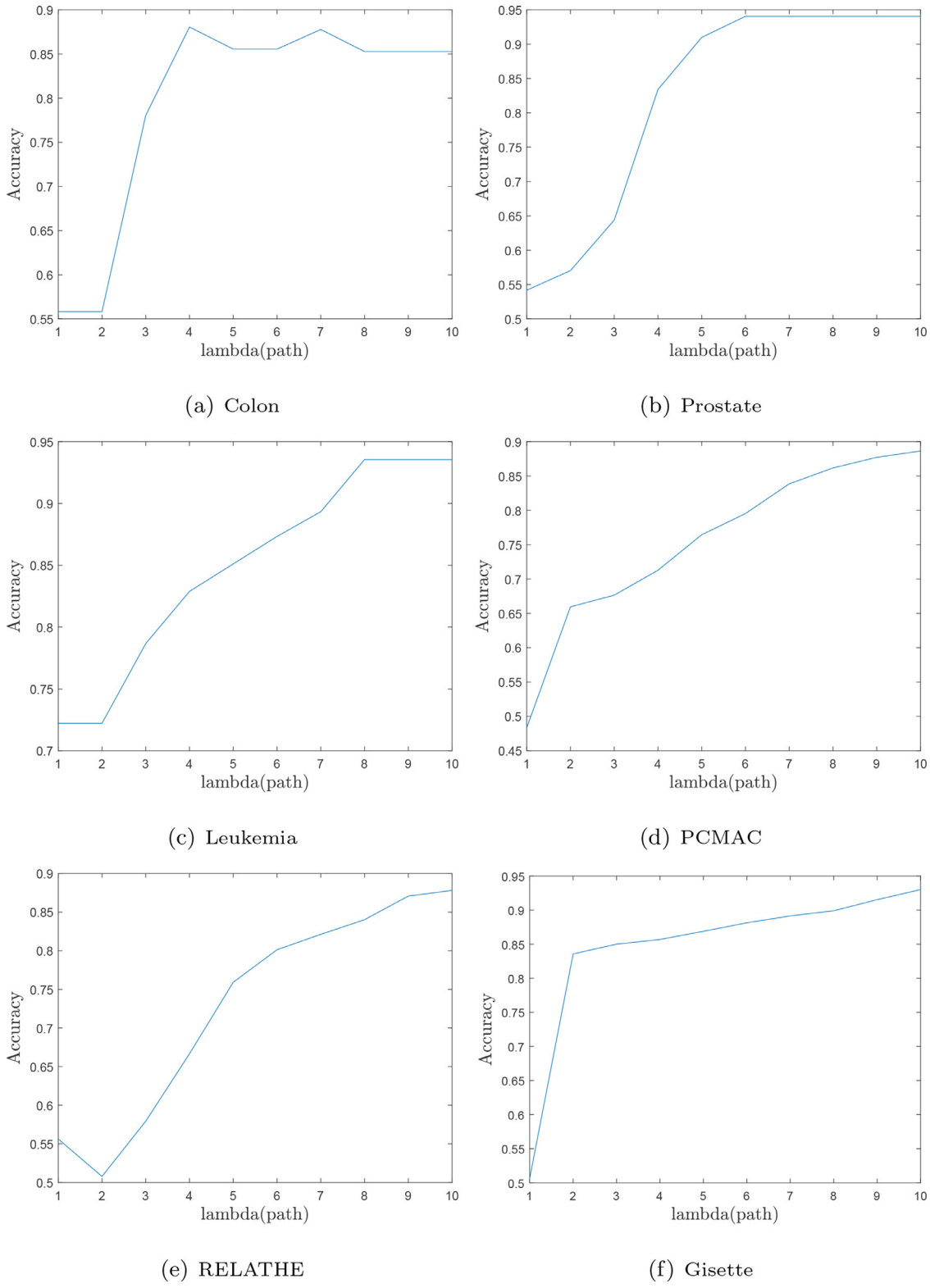

Fig. 7. The performance of the ALCP based parameter path.

colors for nodes (features) of different network modules, and adjust the size of nodes according to the degree of nodes. Finally, the network structure diagrams of six datasets are shown in Fig. 5.

5.3. Sensitivity analysis of parameters

In this section, we analyze the sensitivity of the learning performance of the proposed ALCP model to different parameter values by applying the six bechmark datasets. For a fixed α, we aim to calculate a solution path with a range of λ values. Suppose the regularization parameter α of the ALCP is determined as 0.5 in advance. The solution paths for the proposed ALCP model is shown in Fig. 6. The horizontal axis represents the natural path of the parameter lambda(path), the vertical axis represents the values of coefficient and each line corresponds to a coefficient path for a particular gene. It should be noted that any line segment between two inflection points is linear. Hence, every coefficient path of ALCP is piecewise linear in regarding to the lambda(path). The classification accuracy of the ALCP is tends to be stable with the increase of the lambda(path) that is shown in Fig. 7, which indicates that the learning performance of the ALCP is not sensitive to the value of the regularization parameter λ.

6. Conclusion

In this paper, we propose a novel regularized logistic regression with adaptive Lasso and correlation based penalty (ALCP), which is used for feature selection and encourages grouping effects simultaneously. In order to improve the performance of classification and feature selection, we introduce an adaptive weight construction method based on mutual information into Lasso type norm and form an adaptive Lasso penalty in penalized logistic regression. Then we impose these discriminative penalties on each feature, which leads to outstanding performance of classification and feature selection. According to the coordinate descent algorithm, a learning algorithm is further developed for the proposed ALCP model. The extensive experiments on simulated and six public benchmark datasets are conducted to demonstrate that the classification performance and feature selection performance of the proposed ALCP model are superior to those of the other six competing regularized logistic regression models. It is noticed from the simulated results that when the correlations between each pair of features are very high, the proposed method works efficiently. In particular, the results obtained by the proposed ALCP model are obviously better than the other models in the three cancer gene expression data with correlation structure. Hence, the ALCP model is helpful to select the highly correlated genes on the gene expression data, the potential interaction of which may affect the complex disease outcomes and have some implications and insights on the field of medical engineering. Since the proposed model is design for the binary classification, the proposed model may not be well suited to the multi-classification, which is the limitations of the study. Our future research will focus on the more detailed study of outliers and the possible cause of large errors.

Acknowledgment

This work is supported by grants from the National Natural Science Foundation of China (Nos. 62106066, 62006065), National Basic Research Program of China (No. 2019YFE0126600), the Major Project of Science and Technology of Henan Province (No. 201400210300) and Key Research and Promotion Projects of Henan Province (Nos. 212102210393, 202102110121, 202102210368), the Key Research Projects of Henan Higher Education Institutions (No. 22A520019) and Kaifeng Science and Technology Development Plan (No. 2002001), and the Research Foundation of HZNU (No. 4115C50220204003).

References

[1] S. Piramuthu, Evaluating feature selection methods for learning in data mining applications, Eur J Oper Res 156 (2) (2004) 483–494.
[2] S.C. Satapathy, S. Chittineni, S.M. Krishna, J.V.R. Murthy, P.P. Reddy, Kalman particle swarm optimized polynomials for data classification, Appl Math Model 36 (1) (2012) 115–126.
[3] J. Hao, W.K. Ching, W. Hou, On orthogonal feature extraction model with applications in medical prognosis, Appl Math Model 40 (19–20) (2016) 8766–8776.
[4] A.R. Patil, S. Kim, Combination of ensembles of regularized regression models with resampling-based lasso feature selection in high dimensional data, Mathematics 8 (1) (2020) 1–23.
[5] S. Das, Filters, wrappers and a boosting-based hybrid for feature selection, in: International Conference on Machine Learning, 2001, pp. 74–81.
[6] J. Wang, J. Wei, Z. Yang, S. Wang, Feature selection by maximizing independent classification information, IEEE Trans Knowl Data Eng 29 (4) (2017) 828–841.
[7] R. Ibrahim, M. Elaziz, A. Ewees, M. El-Abd, S. Lu, New feature selection paradigm based on hyper-heuristic technique, Appl Math Model 98 (2021) 14–37.
[8] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1–4) (1997) 131–156.
[9] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif Intell 97 (1–2) (1997) 273–324.
[10] Y. Yang, H. Yang, Adaptive and reversed penalty for analysis of high-dimensional correlated data, Appl Math Model 92 (2021) 63–77.
[11] X. Li, Y. Wang, R. Ruiz, A survey on sparse learning models for feature selection, IEEE Trans Cybern (2020), doi:10.1109/TCYB.2020.2982445.
[12] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B (Methodological) 58 (1) (1996) 267–288.
[13] J. Gui, Z. Sun, S. Ji, D. Tao, T. Tan, Feature selection based on structured sparsity: a comprehensive study, IEEE Trans Neural Netw Learn Syst 28 (7) (2016) 1490–1507.
[14] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society, Series B 68 (1) (2006) 49–67.

[15] L. Meier, S. van de Geer, P. Bühlmann, The group lasso for logistic regression, Journal of the Royal Statistical Society Series B 70 (2008) 53–71.
[16] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse-group lasso, Journal of Computational and Graphical Statistics 22 (2) (2013) 231–245.
[17] Y. Wang, X. Li, R. Ruiz, Weighted general group lasso for gene selection in cancer classification, IEEE Trans Cybern 49 (8) (2019) 2860–2873.
[18] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B 67 (2) (2005) 301–320.
[19] G. Tutz, J. Ulbricht, Penalized regression with correlation-based penalty, Stat Comput 19 (3) (2009) 239–253.
[20] M.E. Anbari, A. Mkhadri, Penalized regression combining the l_1 norm and a correlation based penalty, Sankhya B: The Indian Journal of Statistics 76 (1) (2014) 82–102.
[21] B. Chen, W. Zhai, Z. Huang, Low-rank elastic-net regularized multivariate huber regression model, Appl Math Model 87 (2020) 571–583.
[22] Y. Wang, X.G. Yang, Y. Lu, Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information, Appl Math Model 71 (2019) 286–297. JUL
[23] H. Zou, The adaptive lasso and its oracle properties, J Am Stat Assoc 101 (476) (2006) 1418–1429.
[24] H. Zou, H. Zhang, On the adaptive elastic-net with a diverging number of parameters, The Annals of Statistics 37 (4) (2009) 1733–1751.
[25] J. Li, Y. Jia, Z. Zhao, Partly adaptive elastic net and its application to microarray classification, Neural Computing and Applications 22 (2013) 1193–1200.
[26] H. Wang, C. Leng, A note on adaptive group lasso, Comput. Stat. Data Anal. 52 (12) (2008) 5277–5286.
[27] K. Fang, X. Wang, S. Zhang, J. Zhu, S. Ma, Bi-level variable selection via adaptive sparse group lasso, J Stat Comput Simul 85 (13) (2015) 2750–2760.
[28] A. Patil, B. Park, S. Kim, Adaptive lasso with weights based on normalized filtering scores in molecular big data, Journal of Theoretical and Computational Chemistry 19 (04) (2020) 2040010.
[29] 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 (23) (2015) 9326–9332.
[30] H. Park, Y. Shiraishi, S. Imoto, S. Miyano, A novel adaptive penalized logistic regression for uncovering biomarker associated with anti-cancer drug sensitivity, IEEE/ACM Trans. Comput. Biol. Bioinf. 14 (4) (2017) 771–782.
[31] C. Liu, H. Wong, Structured penalized logistic regression for gene selection in gene expression data analysis, IEEE/ACM Trans. Comput. Biol. Bioinf. 16 (1) (2017) 312–321.
[32] J. Li, Y. Wang, Y. Cao, C. Xu, Gene selection of rat hepatocyte proliferation using adaptive sparse group lasso with weighted gene co-expression network analysis, Comput Biol Chem 80 (2019) 364–373. June
[33] G. Brown, A. Pocock, M.J. Zhao, M. Lujan, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, Journal of Machine Learning Research 13 (1) (2012) 27–66.
[34] T. Naghibi, S. Hoffmann, B. Pfister, A semidefinite programming based search strategy for feature selection with mutual information measure, IEEE Trans Pattern Anal Mach Intell 37 (8) (2015) 1529–1541.
[35] T.M. Cover, J.A. Thomas, Elements of information theory, 2003,
[36] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (4) (1994) 537–550.
[37] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, Cambridge University Press, Cambridge, 1996.
[38] P. Maji, Mutual information-based supervised attribute clustering for microarray sample classification, IEEE Trans Knowl Data Eng 24 (1) (2012) 127–140.
[39] N. Simon, J. Friedman, T. Hastie, A blockwise descent algorithm for group-penalized multiresponse and multinomial regression, J Stat Softw 10 (2) (2013) 1–11.
[40] J. Fiedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J Stat Softw 33 (1) (2010) 1–22.
[41] J. Fiedman, T. Hastie, H. Hofling, R. Tibshirani, Path wise coordinate optimization, Ann Appl Stat 1 (2) (2007) 302–332.
[42] U. Alon, N. Barkai, D.A. Notterman, et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences 96 (12) (1999) 6745–6750.
[43] D. Singh, et al., Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2) (2002) 203–209.
[44] J. Li, K. Cheng, S. Wang, et al., Feature selection: a data perspective, ACM Computing Surveys (CSUR) 50 (6) (2016) 94.