Relaxed Adaptive Lasso for Classification on High-Dimensional Sparse Data with Multicollinearity

Narumol Sudjai[1], Monthira Duangsaphon[2,*] and Chandhanarat Chandhanayingyong[1,*]

[1]*Department of Orthopaedic Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand*

[2]*Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand*

Abstract: High-dimensional sparse data with multicollinearity is frequently found in medical data. This problem can lead to poor predictive accuracy when applied to a new data set. The Least Absolute Shrinkage and Selection Operator (Lasso) is a popular machine-learning algorithm for variable selection and parameter estimation. Additionally, the adaptive Lasso method was developed using the adaptive weight on the l_1-norm penalty. This adaptive weight is related to the power order of the estimators. Thus, we focus on 1) the power of adaptive weight on the penalty function, and 2) the two-stage variable selection method. This study aimed to propose the relaxed adaptive Lasso sparse logistic regression. Moreover, we compared the performances of the different penalty functions by using the mean of the predicted mean squared error (MPMSE) for the simulation study and the accuracy of classification for a real-data application. The results showed that the proposed method performed best on high-dimensional sparse data with multicollinearity. Along with, for classifier with the support vector machine, this proposed method was also the best option for the variable selection process.

Keywords: High-dimensional sparse data, machine learning, multicollinearity, penalized logistic regression, variable selection method.

INTRODUCTION

Presently, advances in technology are growing rapidly, which have resulted in computers being able to store huge amounts of data effectively. With such enormous volumes of data, we require tools that can extract useful information. Particularly needed are predictive modeling techniques that can provide accurate results to help decision-making. Logistic regression is one of the techniques that is widely employed in data analysis and machine learning communities [1-4]. This predictive modeling technique describes the relationships between independent and outcome variables and predicts the outcome variables' future values [5, 6]. For a binary outcome variable, the classical method used to estimate coefficients in the logistic regression algorithm is maximum likelihood estimation (MLE). However, the MLE is only stable when the volume of data is large enough and there is no multicollinearity problem [5-7]. A critical problem that commonly arises in model building is high-dimensional data. High-dimensional data refers to a data set in which the number of independent variables (p) is large compared with the number of observations (n). This condition can lead to model overfitting [8]. Furthermore, it can result in the development of complex models that may be difficult to interpret. Another problem in model construction is the presence of multicollinearity. Multicollinearity refers to some of the independent variables are highly correlated. When n is substantially smaller than p, multicollinearity occurs [9]. This situation can inflate the variance of the maximum likelihood estimators in the logistic regression model [5, 6]. The hybrid of the two above problems can also lead to instabilities in a predictive model [7-9]. Therefore, the MLE used for coefficient estimation in logistic regression is inappropriate for constructing classification models [10].

To solve the problems of high-dimensional data with multicollinearity, the penalized method can be applied in the logistic regression model. This method proposes to reduce variance in parameter estimation and help mitigate model overfitting [11, 12]. Currently, popular penalty function methods are ridge regression, Lasso, and elastic net [13-15]. The choice of penalty function is part of the model constructing procedure, bearing in mind that the performance of each method is not the same for each data item. In previous studies, several researchers concentrated on developing an adaptive weight for the penalty function. For example, Zou [16] proposed adaptive Lasso in 2006, which enjoyed oracle properties and led to stable estimation. Next, Meinshausen [17] proposed a relaxed Lasso for linear

*Address correspondence to these authors at the Department of Orthopaedic Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand; E-mail: chandhanarat.c@gmail.com

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand;
E-mail: monthira@mathstat.sci.tu.ac.th

regression in 2007. In 2009, Zou and Zhang [18] proposed an adaptive elastic net; it has oracle properties and superior performance to the elastic net. However, no studies have compared the performances of penalized methods in logistic regression, focusing on 1) the power of adaptive weight on the penalty function under the scenario of sparse data with multicollinearity, and 2) the two-stage Lasso methods for classification. “Sparse” data indicate that the logistic regression model has several nonsignificant predictors whose coefficients are zero [19].

Hence, this study focused on 1) the power of adaptive weight on the penalty function, and 2) the two-stage variable selection method. The aim was to propose the relaxed adaptive Lasso sparse logistic regression. Additionally, the performances of six methods (i.e., ridge, Lasso, elastic net, adaptive Lasso, adaptive elastic net, and relaxed adaptive Lasso) were compared with the mean of the predicted mean squared errors (MPMSE) value obtained from Monte Carlo simulations. Along with this, in a real-data application, classification accuracy was used to assess the performances of each method.

MATERIALS AND METHODS

Logistic regression is a commonly used statistical method for classification. We let the dependent (or called outcome) variable be a dichotomous variable (i.e., 0 = negative class or 1 = positive class). Thus, $y_i \in \{0,1\}$, which is a $n \times 1$ vector where n is the sample of size. X is a $n \times (p+1)$ data matrix of p independent variables when $\underset{\sim}{x}_i$ denotes the independent variables for the i^{th} row of X. The dependent variable (Y_i) is a binary outcome that has a Bernoulli distribution ($Y_i \sim Bernoulli(\pi_i)$). The random error ($\varepsilon_i$) is assumed to follow a distribution with a mean of zero and a variance of $\pi_i(1-\pi_i)$. Therefore, the binary logistic regression model is the following [6]:

$$\pi_i = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right\}}{1+\exp\left\{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right\}}, \qquad i = 1,2,3,...,n \qquad \text{and} \qquad j = 1,2,3,...,p \tag{1}$$

where π_i denotes a probability that an observation is in a specified category of the binary outcome variable. Thus, y_i represents the value of a dichotomous outcome variable. For $y_i = 0$, the conditional probability that $y_i = 0$, given as $\underset{\sim}{x}_i$ can be written as $1-\pi_i = P(Y_i = 0 \mid \underset{\sim}{x}_i)$. In this case, $\pi_i = P(Y_i = 1 \mid \underset{\sim}{x}_i)$ is the conditional probability that $y_i = 1$, given as $\underset{\sim}{x}_i$.

Logistic regression is the logit transformation, which is represented as

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j . \tag{2}$$

The left term of equation (2) is called the logit function. A vector composed of logistic regression coefficients is $\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$. The log-likelihood function for equation (2) can be written as:

$$\ell(\underset{\sim}{\beta}) = \sum_{i=1}^{n}\left[y_i\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right) - \ln\left(1+\exp\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)\right)\right]. \tag{3}$$

The estimated parameters of equation (3) can be estimated by using the classical method, which is given below.

$$\hat{\underset{\sim}{\beta}}_{MLE} = \arg\max_{\underset{\sim}{\beta}}\left(\sum_{i=1}^{n}\left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right]\right) \tag{4}$$

where $\hat{\underset{\sim}{\beta}}_{MLE}$ is a $(p+1)\times 1$ vector of the maximum likelihood estimator. Although this method has good performance, it has some limitations [6, 7]. Thus, the penalized logistic regression model is applied as an alternative to the classical method (see Appendix). The principle of this technique is given below.

Penalized Logistic Regression Analysis

The purpose of penalized logistic regression analysis is to estimate logistic regression coefficients when the data are high dimensional and multicollinearity is present. Penalized logistic regression coefficients are determined as follows:

$$\hat{\underset{\sim}{\beta}}_{PLR} = \arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[\begin{matrix} y_i \ln(\pi_i) + \\ (1-y_i)\ln(1-\pi_i)\end{matrix}\right]\right\} + P_\lambda(\underset{\sim}{\beta})\right) \; ; \lambda \geq 0. \tag{5}$$

Equation (5) is similar to equation (4), but a different term is the penalty function ($P_\lambda(\underset{\sim}{\beta})$). λ is the tuning parameter, which is more than or equal to zero. If λ is large, then the effect of the penalty term on the coefficient estimation increases. Selecting the value of λ is important because it can affect the balance between variance and bias [20]. To select the optimal value of λ, cross-validation is commonly used, which depends on the real data [21].

The Proposed Method: Relaxed Adaptive Lasso

Definition

Relaxed Lasso was originally designed to solve the disadvantage of Lasso in a linear regression model [17]. In this study, we proposed the relaxed adaptive Lasso sparse logistic regression, which was developed based on the work of Meinshausen [17]. We defined the relaxed adaptive Lasso estimator ($\underset{\sim}{\hat{\beta}}^{\lambda,w}$) on the set $M^{\lambda,w} \subseteq \{1,2,3,...,p\}$ when p is the number of nonzero variables selected into the ultimate model. The novel procedure of variable selection and shrinkage of $\underset{\sim}{\beta}$ are controlled by two constraints (i.e., λ and ϕ) and the weight vector (w) to the penalty term. The relaxed adaptive Lasso estimator in logistic regression is defined as follows:

$$\underset{\sim}{\hat{\beta}}_{RAlasso} = \arg\min_{\underset{\sim}{\beta}} \left(\begin{array}{l} -\left\{ \sum_{i=1}^{n} \left[\begin{array}{l} y_i \left(\beta_0 + \sum_{j=1}^{p} x_{ij} \left\{ \beta_j \cdot 1_{M^{\lambda,w}} \right\} \right) \\ -\ln \left(1 + \exp \left(\beta_0 + \sum_{j=1}^{p} x_{ij} \left\{ \beta_j \cdot 1_{M^{\lambda,w}} \right\} \right) \right) \end{array} \right] \right\} \\ +\phi\lambda \sum_{j=1}^{p} w_j \left| \beta_j \right| \end{array} \right) \quad (6)$$

where $1_{M^{\lambda,w}}$ is an indicator function $\left\{1_{M^{\lambda,w}}\right\}_k = \begin{cases} 0, & k \notin M^{\lambda,w} \\ 1, & k \in M^{\lambda,w} \end{cases}$, for all $k \in \{1,2,3,...,p\}$; $\phi \in [0,1]$; define the weight vector $w_j = \left|\hat{\beta}_j\right|^{-\gamma}$, $\gamma > 0$. γ is the power of the adaptive weight. For λ and ϕ, the cross-validation is used to evaluate the optimal value of the tuning parameters [16, 17]. In the case of $\lambda = 0$ or $\phi = 0$, $\underset{\sim}{\hat{\beta}}_{RAlasso} = \underset{\sim}{\hat{\beta}}_{MLE}$. Additionally, the adaptive Lasso and relaxed adaptive Lasso are the same when $\phi = 1$ (see Appendix).

Algorithm

The algorithm for the relaxed adaptive Lasso is as follows:

Step 1. Let $\gamma > 0$, we use $\underset{\sim}{\hat{\beta}}_{MLE}$ to construct the weight in an adaptive Lasso based on the work of Zou [16], bearing in mind that this initial weight can be determined by using $\underset{\sim}{\hat{\beta}}_{MLE}$ unless collinearity is a concern, in which case we can use $\underset{\sim}{\hat{\beta}}_{ridge}$ from the best ridge logistic regression fit, because it is more stable than $\underset{\sim}{\hat{\beta}}_{MLE}$ [16].

Step 2. Define $X_j^* = X_j / \hat{w}_j, j = 1,2,3,...,p$, when $w_j = \left|\hat{\beta}_j^{ridge}\right|^{-\gamma}$.

Step 3. The procedure of relaxed adaptive lasso estimation is based on solving the relaxed lasso solutions in Meinshausen [17]. First, we compute $\underset{\sim}{\hat{\beta}}^{**}$, which is determined as

$$\underset{\sim}{\hat{\beta}}^{**} = \arg\min_{\underset{\sim}{\beta}} \left(\begin{array}{l} -\left\{ \sum_{i=1}^{n} \left[\begin{array}{l} y_i \left(\beta_0 + \sum_{j=1}^{p} x_{ij}^{*} \left\{ \beta_j \cdot 1_{M^{\lambda,w}} \right\} \right) \\ -\ln \left(1 + \exp \left(\beta_0 + \sum_{j=1}^{p} x_{ij}^{*} \left\{ \beta_j \cdot 1_{M^{\lambda,w}} \right\} \right) \right) \end{array} \right] \right\} \\ +\phi\lambda \sum_{j=1}^{p} \left| \beta_j \right| \end{array} \right). \quad (7)$$

Step 4. Then, the resulting estimates in step 3 to compute the relaxed adaptive lasso estimators. The solution of the relaxed adaptive Lasso can be defined as follows: $\hat{\beta}_j^{RAlasso} = \hat{\beta}_j^{**} / \hat{w}_j$, $j = 1,2,3,...,p$.

Monte Carlo Simulation

The key factors affecting the accuracy of a predictive model are the number of predictors (p), the sample size (n), and the high correlation between the independent variables. In this study, we considered two scenarios in a simulation study:

1) High-dimensional sparse data [19]. Let $p > n$. Under the sparsity assumption on the true coefficients ($\underset{\sim}{\beta}$), we assumed that the number of significant predictors is equal to q when $q < p$. $\underset{\sim}{x}_i = \left(\underset{\sim}{x}_{iA}, \underset{\sim}{x}_{iB}\right)$ with $\underset{\sim}{x}_{iA} = \left(x_{i1}, x_{i2}, x_{i3}, ..., x_{iq}\right)^T \in \mathbb{R}^q$ and $\underset{\sim}{x}_{iB} = \left(x_{i(q+1)}, x_{i(q+2)}, x_{i(q+3)}, ..., x_{ip}\right)^T \in \mathbb{R}^{p-q}$. Hence, $\underset{\sim}{X} = \left(\underset{\sim}{x}_A, \underset{\sim}{x}_B\right)^T \in \mathbb{R}^{n \times p}$ is the matrix of all independent variables when $\underset{\sim}{x}_A = \left(\underset{\sim}{x}_{iA}, ..., \underset{\sim}{x}_{nA}\right)^T \in \mathbb{R}^{n \times q}$ and $\underset{\sim}{x}_B = \left(\underset{\sim}{x}_{iB}, ..., \underset{\sim}{x}_{nB}\right)^T \in \mathbb{R}^{n \times (p-q)}$.

2) The independent variables are correlated using the Toeplitz correlation structure, which is given below [22].

$$\sum\nolimits_k = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{k-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{k-3} \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{k-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \rho^{k-4} & \cdots & 1 \end{pmatrix}_{k\times k} \tag{8}$$

where k is a positive integer (the number of independent variables) and $0 \le \rho \le 1$.

The Monte Carlo simulations were performed using 25, 50, and 100 independent variables (p). The sample size (n) equaled 30 and 40. The independent variables were generated from the multivariate normal distribution with a mean of zero and covariance $\sum$ ($X \sim N(0,\sum)$). The dependent variables were generated from the Bernoulli distribution with parameter π_i ($Y_i \sim Bernoulli(\pi_i)$). The degree of correlation (ρ) was set to 0.75, 0.85, and 0.95. The number of significant predictors (q) equaled 15. The logistic regression coefficients were set the constant values as $\underset{\sim}{\beta}$. After generating the data set, we split the data into two subsets: 80% of the learning data set, and 20% of the testing data set. The simulation study compared the performances of the six penalized methods (ridge, Lasso, elastic net, adaptive Lasso, adaptive elastic net, and relaxed adaptive Lasso) in terms of the mean of the predicted mean square errors (PMSE). The estimated PMSE was determined from:

$$\text{PMSE} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \tag{9}$$

where y_i and $\hat{y}_i$ were the i^{th} actual and predicted values of the dependent variables, respectively. For the tuning parameter (λ), we found the optimal value of λ using a tenfold cross-validation strategy [12, 15, 23]. In this study, the experiment was repeated 1000 times to obtain a stationary result. Therefore, the MPMSE was calculated from the average of 1000 estimates of PMSE$_j$.

$$\text{MPMSE} = \frac{1}{1000} \sum_{j=1}^{1000} \text{PMSE}_j \ . \tag{10}$$

The methods providing the lowest MPMSE were considered the best option for high-dimensional sparse data with multicollinearity. The flowchart of the simulation process is illustrated in Figure **1**.

Regarding the real-data application, the workflow diagram of the machine-learning process with different penalized methods is presented in Figure **2**. The classification accuracy was used to assess the performance of each method. This accuracy value was evaluated from:

$$\text{Accuracy (\%)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100 \ . \tag{11}$$

When the true positive (TP) indicates that the prediction is correct, the predicted value is positive, and the actual value is positive. A true negative (TN) shows that the prediction is correct: the predicted value is negative, and the actual value is negative. False positive (FP) indicates that the prediction is wrong (also called a type I error): the predicted value is positive, but the actual value is negative. A false negative (FN) indicates that the prediction is wrong (also called a type II error): the predicted value is negative, but the actual value is positive.

Software

All simulations and analyses were carried out in R version 4.2.1 (R Foundation for Statistical Computing, Vienna, Austria). We used the package 'glmnet' to fit models using all the above penalized methods. The tuning parameters were selected tenfold cross-validation and the experiment was repeated 1000 times to obtain a stationary result. For support vector machine (SVM), the package 'e1071' was used to construct the models [24].

RESULTS

Simulation Study

Table **1** lists the MPMSE values for the six methods for different ρ when p = 25, 50, and 100 and n = 30 and 40. When ρ was increased while holding n and p fixed, the MPMSE values of the methods increased. With an increase in n, the MPMSE values of the methods decreased. In the cases of p = 50 and 100, we found that the MPMSE values of the relaxed adaptive Lasso method were the smallest compared with the other methods. However, for p = 25, the adaptive Lasso method was preferred.

Real-Data Applications

In this section, the application of the six penalized methods with a real-data set with high-dimensional sparse data and multicollinearity is presented to compare their classification performances.

The tumor data set was compiled from the medical records of 40 patients with soft-tissue tumors (20 intramuscular lipomas and 20 well-differentiated

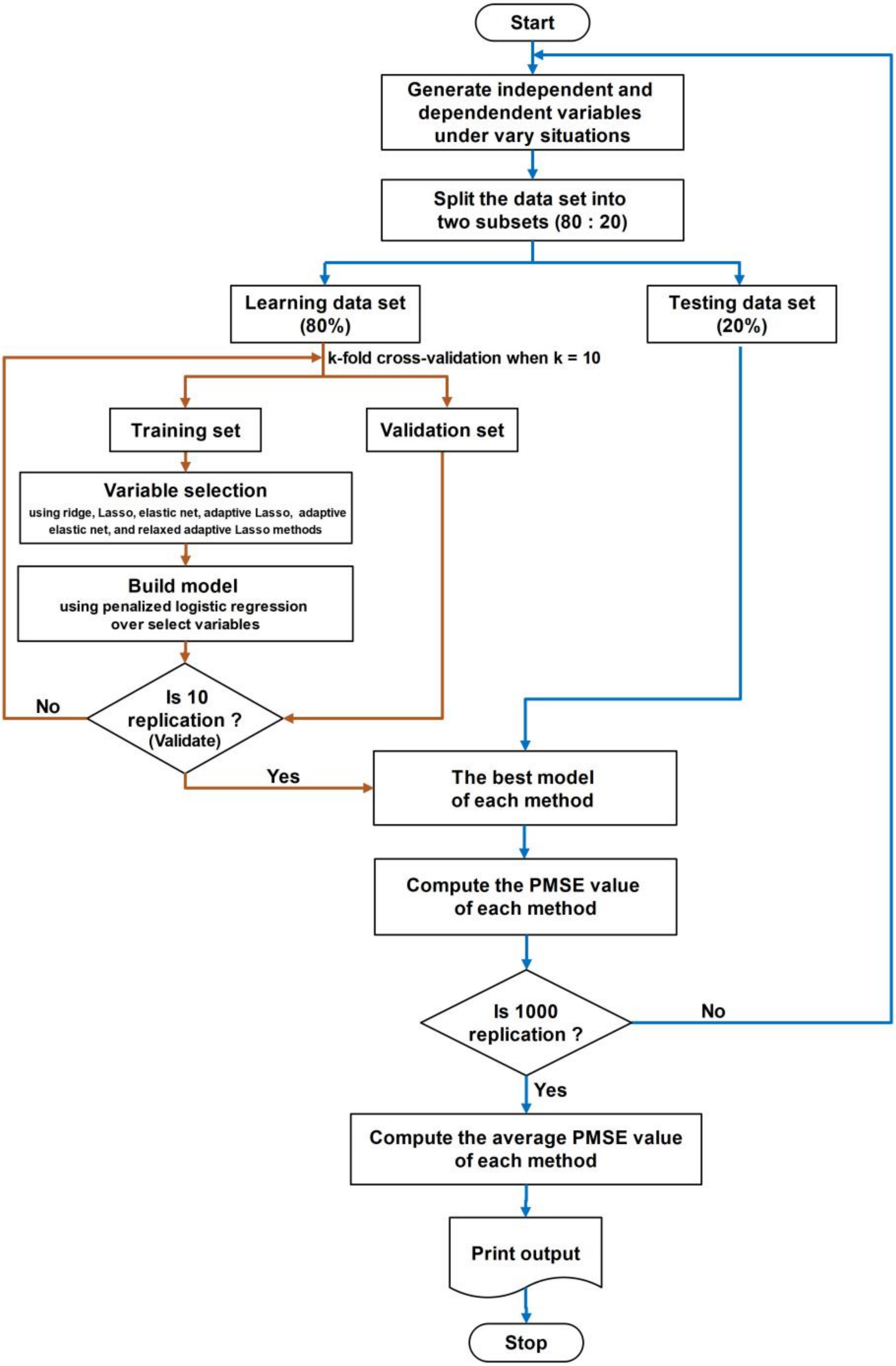
Start
Generate independent and dependent variables under vary situations
Split the data set into two subsets (80 : 20)
Learning data set (80%)
Testing data set (20%)
k-fold cross-validation when k = 10
Training set
Validation set
Variable selection
using ridge, Lasso, elastic net, adaptive Lasso, adaptive elastic net, and relaxed adaptive Lasso methods
Build model
using penalized logistic regression over select variables
Is 10 replication ? (Validate)
No
Yes
The best model of each method
Compute the PMSE value of each method
Is 1000 replication ?
No
Yes
Compute the average PMSE value of each method
Print output
Stop

Figure 1: Flowchart of the simulation process.

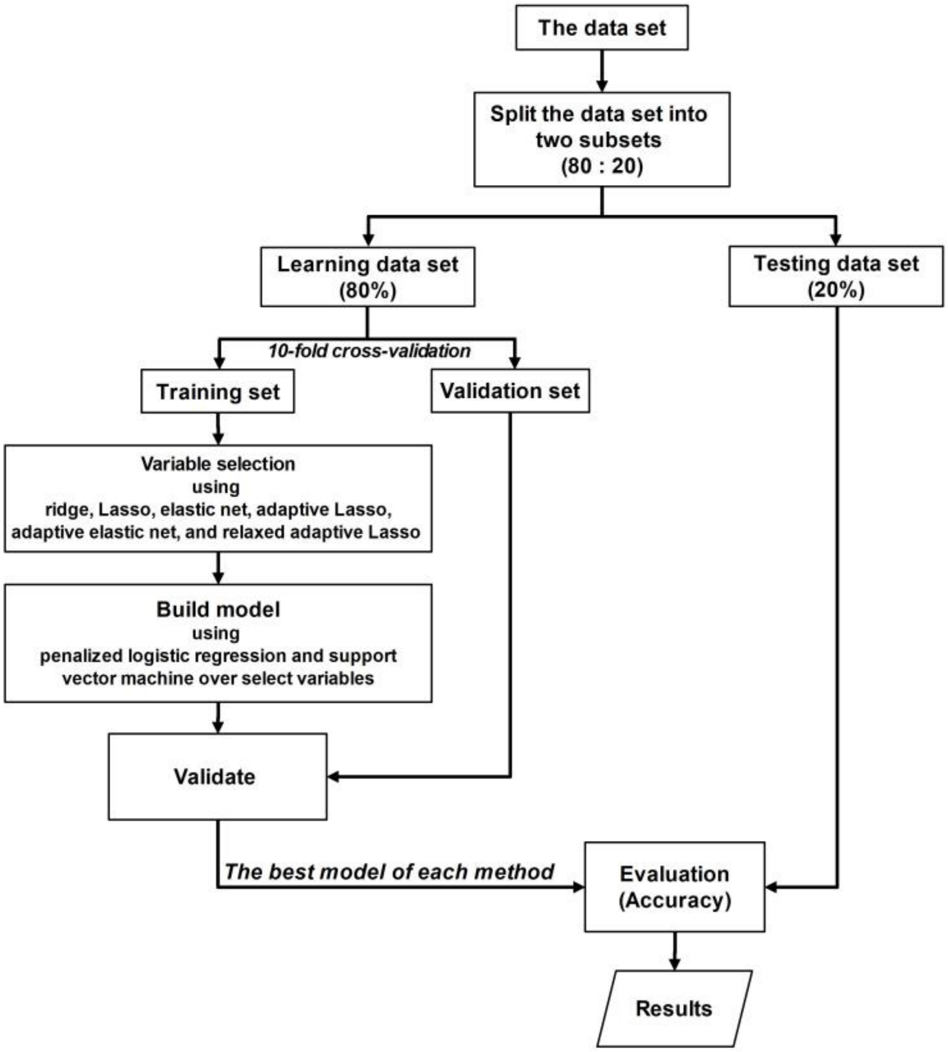
The data set
Split the data set into two subsets (80 : 20)
Learning data set (80%)
Testing data set (20%)
10-fold cross-validation
Training set
Validation set
Variable selection using ridge, Lasso, elastic net, adaptive Lasso, adaptive elastic net, and relaxed adaptive Lasso
Build model using penalized logistic regression and support vector machine over select variables
Validate
The best model of each method
Evaluation (Accuracy)
Results

Figure 2: Workflow diagram of the machine-learning process used to appraise the classification performance of the penalized logistic regression models with six penalized methods with application on real data. The penalized methods were ridge, Least Absolute Shrinkage and Selection Operator (Lasso), elastic net, adaptive Lasso, adaptive elastic net, and relaxed adaptive Lasso. The initial data set was split into two subsets: one set with 80% of the learning data, and the other set with 20% of the testing data. Next, the learning data set underwent a tenfold cross-validation strategy as follows: (1) the training sets were used to select variables (the “variable selection step,” using the six penalized methods) and (2) the validation sets were used to evaluate the classification performance of each model (the “model building step,” in which we applied the penalized logistic regression and support vector machine over selected variables). The best model for each method was selected. These models were subsequently appraised on the testing data set.

liposarcomas). The patients had been treated at our institution between 2010 and 2020. Their data were retrieved after receiving approval from the Ethics Committee of our institute. The patients were diagnosed using their final pathological examinations, and all underwent magnetic resonance imaging examinations and total excision surgery. For our case study, the outcome of interest was an intramuscular lipoma or a well-differentiated liposarcoma as a dichotomous variable. The predictors of interest were 50 radiomic features as continuous variables; these were extracted from preoperative T1-weighted magnetic resonance images. The correlations between all predictors of the tumor data set are depicted in Figure **3**. The correlation matrix shows different shades. The dark shade signifies that the predictors have a high correlation, whereas the light shade represents a low correlation between the predictors. It is apparent that the multicollinearity problem was present in this sample data set.

In Table **2**, we evaluated the classification performances of the six penalized methods in differentiating between intramuscular lipomas and well-differentiated liposarcomas. When looking at the accuracy of their classifications, the highest accuracy values were obtained with the relaxed adaptive Lasso while the lowest accuracy values were obtained with

Table 1: MPMSE Values for Different Methods

p	n	ρ	Ridge	Lasso	Elastic net	Adaptive Lasso	Adaptive elastic net	Relaxed adaptive Lasso
25	30	0.75	0.230	0.190	0.192	0.185*	0.191	0.190
		0.85	0.233	0.195	0.198	0.190*	0.198	0.200
		0.95	0.233	0.198	0.201	0.194*	0.201	0.204
	40	0.75	0.203	0.189	0.190	0.185*	0.191	0.190
		0.85	0.208	0.191	0.196	0.188*	0.194	0.195
		0.95	0.213	0.196	0.198	0.194*	0.197	0.199
50	30	0.75	0.225	0.191	0.192	0.190	0.192	0.184*
		0.85	0.227	0.194	0.197	0.196	0.198	0.189*
		0.95	0.228	0.195	0.199	0.204	0.199	0.191*
	40	0.75	0.229	0.187	0.192	0.183	0.191	0.179*
		0.85	0.230	0.193	0.196	0.188	0.194	0.183*
		0.95	0.232	0.193	0.198	0.198	0.197	0.188*
100	30	0.75	0.220	0.188	0.192	0.180	0.188	0.175*
		0.85	0.221	0.190	0.194	0.185	0.193	0.178*
		0.95	0.227	0.196	0.199	0.194	0.196	0.187*
	40	0.75	0.217	0.184	0.187	0.178	0.181	0.168*
		0.85	0.219	0.188	0.189	0.182	0.187	0.176*
		0.95	0.225	0.194	0.198	0.193	0.195	0.183*

MPMSE, mean of the predicted mean square errors; Lasso, Least Absolute Shrinkage and Selection Operator; *The penalized methods providing the smallest MPMSE.

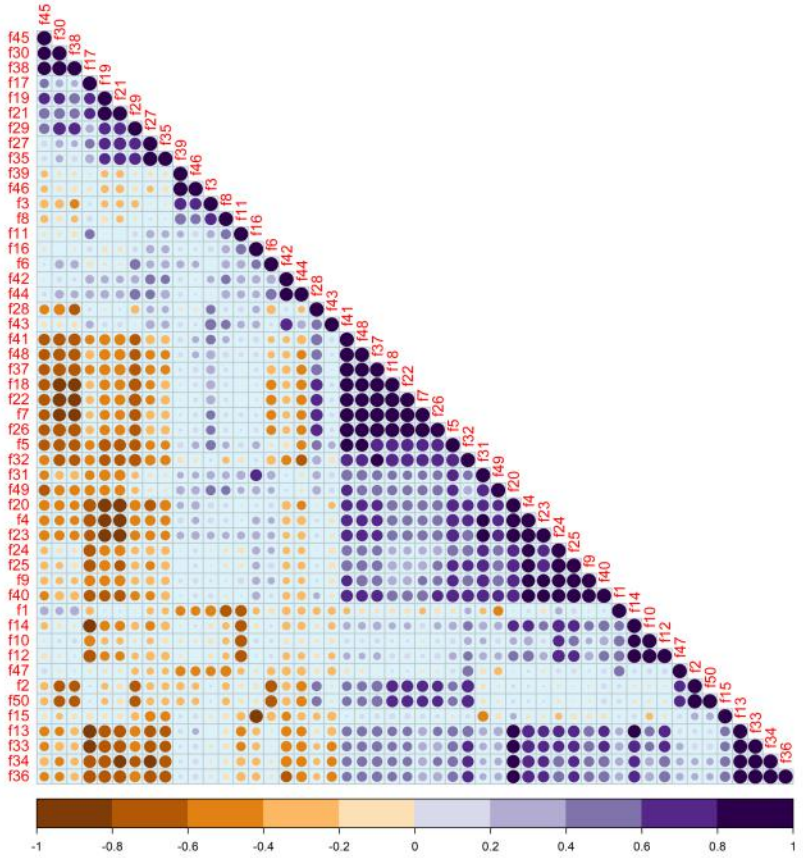

Figure 3: Correlation matrix of 50 radiomic features in 40 patients, showing the different shades. The dark shade indicates that the features have a high correlation. In contrast, the light shade represents a low correlation.

Table 2: Accuracy of Machine-Learning Algorithms for Differentiating between Intramuscular Lipomas and Well-Differentiated Liposarcomas under 50 Radiomic Features in 40 Patients

Variable selection method	Ridge	Lasso	Elastic net	Adaptive Lasso	Adaptive elastic net	Relaxed adaptive Lasso
Classifier with penalized logistic regression						
Accuracy (%)	81.0	88.0	87.0	91.0	92.0	93.5*
Classifier with support vector machine						
Accuracy (%)	86.3	89.8	90.0	93.0	91.0	94.0*

Lasso, Least Absolute Shrinkage and Selection Operator; *The penalized methods providing the highest accuracy.

the ridge method. Regarding the classifier with support vector machine, the relaxed adaptive Lasso was also preferred.

DISCUSSION

From the simulated results in Table **1**, it can be seen that the important factors influencing the MPMSE values were the correlated independent variables (i.e., correlation coefficient level (ρ)), the penalty function, and the sample size (n). An increase in the correlation coefficient level lends to an increase in the MPMSE values for all methods when holding p and n fixed. The worst case was obtained when the correlation coefficient level was very high (ρ = 0.95). In the case of the power of the adaptive weight on the l_1-norm penalty, choosing the relaxed adaptive Lasso is preferred for $p > n$ while the adaptive Lasso for $p < n$. We can see that an increase in the sample size (n) affects a decrease in the MPMSE values for all methods while holding ρ and p fixed.

Regarding the results of the real-data applications in Table **2**, it is obvious that the relaxed adaptive Lasso method showed a better performance than the other methods for the classification of the high-dimensional sparse data with multicollinearity. This finding corresponds to the results of the simulation study. Along with, for classifier with the support vector machine, the relaxed adaptive Lasso method was also the best option for the variable selection process.

The limitations of the classical method can be overcome with penalized logistic regression. However, the performance of penalized logistic regression depends on the penalty function method used. Therefore, the performance of each method is not the same for each data item. Table **3** compares the advantages, disadvantages, limitations, and appropriateness of each penalty function in the penalized logistic regression model.

Regarding the Lasso method, the shrinkage of the Lasso causes the estimation of non-zero coefficients to be biased towards zero. To remedy this disadvantage, two-stage Lasso methods have been proposed (i.e., the adaptive Lasso [16] and the relaxed Lasso [17]). For the relaxed Lasso (known as the Lasso twice), the concept in the first stage of this method is to use cross-validation to estimate the initial penalty parameter for the Lasso. Subsequently, in the second stage, we perform Lasso again on the selected set of predictors obtained from the first stage. Since this second stage has less competition from noise variables, cross-validation may tend to pick a smaller value for the tuning parameter λ [17, 23]. Consequently, their coefficients will shrink less than those in the initial estimate, which can solve the above disadvantage and can help mitigate the slow convergence of the Lasso in the case of high-dimensional data. However, the relaxed adaptive Lasso methods provides better performance, compared with fitting Lasso twice [23] because these methods will be used in the second stage with weights imposed on the important measures (i.e., assigning a smaller weight to large coefficients and a higher weight to small coefficients). Thus, these methods can reduce the bias problems of the Lasso.

An alternative approach for reducing those biases, the Lasso can be considered using for the variable selection step followed by an advanced machine learning technique. This approach was used in clinical research, which has provided good classification performance results in classifying tumors into lipomas and atypical lipomatous tumors [4]. For our comparisons, the variable selection procedures with our proposed method can be used with advanced machine learning techniques as a classifier, which also showed a good performance.

CONCLUSION

The relaxed adaptive Lasso methods can be used to (1) reduce overfitting, (2) enable machine-learning algorithms to train faster, (3) improve the accuracy of the predictive model, and (4) alleviate the complexity of the model.

Table 3: Comparison of the Advantages, Disadvantages/Limitations, and Appropriateness of each Method

Method	Advantages	Disadvantages/limitations	Appropriateness of application
Ridge	- Able to solve the multicollinearity problem.	- Lacks selection of variables	- The data are low- or high-dimensional.
	- Able to deal with low-dimensional data ($p < n$) and high-dimensional data ($p > n$)	- When the number of independent variables increases, it may be difficult to interpret the obtained model.	- All independent variables relate to the dependent variable.
	- The estimated parameters $\underset{\sim}{\hat{\beta}}$ are stable.		- Multicollinearity is present.
Lasso	- Able to select the independent variables coupled with their computation.	- When $p > n$, Lasso selects at most n variables before it saturates.	- The data are high-dimensional.
		- If there is a high pairwise correlation between independent variables in the data set, Lasso selects only one variable or a few of them among a group of correlated variables, and it does not care which one is selected.	- The independent variables have low/medium collinearity.
		- When $n > p$ and the independent variables have high collinearity, Lasso is dominated by ridge regression.	
		- Lacks oracle properties	
Elastic net	- Able to enforce sparsity.	- Lacks oracle properties	- The data are high-dimensional.
	- No limitation on the number of selected variables.		- Multicollinearity is present.
	- Able to deal with multicollinearity.		
Adaptive Lasso	- The estimated parameters $\underset{\sim}{\hat{\beta}}$ using the adaptive Lasso are stable and have superior performance to Lasso.	-	- The data are high-dimensional.
	- The estimators have oracle properties.		- The independent variables are highly correlated.
Adaptive elastic net	- The estimators have oracle properties.		- The data are high-dimensional.
	- The adaptive elastic net has superior performance to the elastic net.	-	- The independent variables are highly correlated.
Relaxed adaptive Lasso	- The estimated parameters $\underset{\sim}{\hat{\beta}}$ using the relaxed adaptive Lasso are stable and have superior performance to the other methods.	-	- The data are high-dimensional.
	- Able to alleviate the slow convergence of the Lasso in the case of high-dimensional data		- The independent variables are highly correlated.
	- The estimators have oracle properties.		

Lasso, Least Absolute Shrinkage and Selection Operator.

AUTHOR CONTRIBUTIONS

Conceptualization, N.S., M.D., C.C.; methodology, N.S., M.D., C.C.; software, N.S., M.D.; validation, N.S., M.D., C.C.; formal analysis, N.S., M.D., C.C.; investigation, N.S., C.C.; resources, C.C.; data curation, N.S., C.C.; writing—original draft preparation, N.S.; writing—review and editing, N.S., M.D., C.C.; project administration, N.S.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Siriraj Research Fund, Faculty of Medicine Siriraj Hospital, Mahidol

University, grant number [IO] R016332025, and the Siriraj Foundation Fund for advanced sarcoma research, grant number D004146.

INSTITUTIONAL REVIEW BOARD STATEMENT

The tumor data set in this study was retrieved after approval from the Ethics Committee of the Faculty of Medicine Siriraj Hospital, Mahidol University (approval number Si 874/2020).

ACKNOWLEDGEMENTS

We thank Supani Duangkaew for her research coordination and Pakorn Yodprom for physics consultation and magnetic resonance image retrieval. We are also indebted to David Park for the English-language editing of this paper.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

APPENDIX

Binary logistic regression can be written as [25]:

$$Y_i = \pi_i + \varepsilon_i,\ i = 1,2,3,...,n \tag{A.1}$$

where the dependent variable (Y_i) represents a binary outcome that has a Bernoulli distribution with the parameter $\pi_i = \frac{e^{\underset{\sim}{x_i}\underset{\sim}{\beta}}}{1+e^{\underset{\sim}{x_i}\underset{\sim}{\beta}}}$. ε_i is the random error that has a distribution with zero mean and a variance equal to $\pi_i(1-\pi_i)$ [6]. $\underset{\sim}{x_i}$ is the independent variables for the i^{th} row of X. X is a $n\times(p+1)$ data matrix with p independent variables and sample size n. $\underset{\sim}{\beta}$ represents an $(p+1)\times 1$ unknown coefficient vector.

The transformation of π_i is a central of logistic regression model that is called the logit function, which is as follows:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\ ,\ i = 1,2,3,...,n \text{ and } j = 1,2,3,...,p\,. \tag{A.2}$$

The classical method used for coefficient estimation in the binary logistic regression is the MLE. Y_i be coded as 0 or 1. The conditional probability that $Y_i = 0$, given as $\underset{\sim}{x_i}$ can be written as $1-\pi_i = P(Y_i = 0\,|\,\underset{\sim}{x_i})$. On the other hand, $\pi_i = P(Y_i = 1\,|\,\underset{\sim}{x_i})$ is the conditional probability that $Y_i = 1$, given as $\underset{\sim}{x_i}$. For a set of observations $(y_i, \underset{\sim}{x_i})$, if y_i = 0, then the contribution to the likelihood function is $1-\pi_i$ as well, as if y_i = 1, then the contribution to the likelihood function is π_i. Hence, the contribution to the likelihood function for the set of observations $(y_i, \underset{\sim}{x_i})$ can be written as:

$$P(Y_i = y_i) = \begin{cases} \pi_i^{y_i}(1-\pi_i)^{1-y_i} & , y_i = 0,1 \\ 0 & , \text{otherwise.} \end{cases} \tag{A.3}$$

The likelihood function can be obtained from the terms of (A.3) as follows:

$$\mathrm{L}(\underset{\sim}{\beta}) = \prod_{\mathrm{i}=1}^{n} \pi_i^{\,y_i}(1-\pi_i)^{1-y_i}\ . \tag{A.4}$$

From equation (A.4), we can be expressed by taking the log as:

$$\ln\left[\mathrm{L}(\underset{\sim}{\beta})\right] = \ln\left[\prod_{\mathrm{i}=1}^{n} \pi_i^{\,y_i}(1-\pi_i)^{1-y_i}\right]$$

$$\ell(\underset{\sim}{\beta}) = \sum_{i=1}^{n}\left[y_i\ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right]. \tag{A.5}$$

Thus, the log-likelihood function of the logit transformation of equation (A.2) can be written as:

$$\ell(\underset{\sim}{\beta}) = \sum_{i=1}^{n}\left[y_i\ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right]$$

$$= \sum_{i=1}^{n}\left[y_i\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right) - \ln\left(1 + \exp\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)\right)\right]. \tag{A.6}$$

The estimated parameters of equation (A.6) can be estimated by using the MLE, which is given below.

$$\hat{\underset{\sim}{\beta}}_{MLE} = \underset{\underset{\sim}{\beta}}{\arg\max}\left(\sum_{i=1}^{n}\left[y_i\ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right]\right) \tag{A.7}$$

where $\hat{\underset{\sim}{\beta}}_{MLE}$ is a $(p+1)\times 1$ vector of the maximum likelihood estimator. However, this method has some limitations. Therefore, the penalized logistic regression is employed as an alternative to the MLE.

From equation (A.6), we can be written in the form of penalized function as follows [26]:

$$\ell^{*}(\underset{\sim}{\beta}) = -\,\ell(\underset{\sim}{\beta}) + P_{\lambda}(\underset{\sim}{\beta}) \tag{A.8}$$

where λ is the tuning parameter and $P_{\lambda}(\underset{\sim}{\beta})$ is the penalty function.

Regarding the penalized logistic regression coefficients, the estimated parameters $\hat{\underset{\sim}{\beta}}_{PLR}$ is obtained by minimizing equation (A.8), which can be determined as follows:

$$\hat{\underset{\sim}{\beta}}_{PLR} = \arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[y_i\ln(\pi_i)+(1-y_i)\ln(1-\pi_i)\right]\right\}+P_\lambda(\underset{\sim}{\beta})\right)$$
$$=\arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[\begin{array}{l} y_i\left(\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j\right)\\ -\ln\left(1+\exp\left(\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j\right)\right)\end{array}\right]\right\}+P_\lambda(\underset{\sim}{\beta})\right). \qquad \text{(A.9)}$$

For the penalty function, adaptive Lasso [16] is one of the techniques employed in data analysis. The concept of this technique is a different weight for each parameter in the l_1-norm penalty. The adaptive Lasso penalty is defined as follows:

$$P_\lambda^{Alasso}(\underset{\sim}{\beta}) = \lambda\sum_{j=1}^{p}w_j\left|\beta_j\right|. \qquad \text{(A.10)}$$

Hence, the estimation of $\underset{\sim}{\beta}$ using the adaptive Lasso penalty is as follows:

$$\hat{\underset{\sim}{\beta}}_{Alasso} = \arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[y_i\ln(\pi_i)+(1-y_i)\ln(1-\pi_i)\right]\right\}+P_\lambda^{Alasso}(\underset{\sim}{\beta})\right)$$
$$=\arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[\begin{array}{l} y_i\left(\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j\right)\\ -\ln\left(1+\exp\left(\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j\right)\right)\end{array}\right]\right\}+\lambda\sum_{j=1}^{p}w_j\left|\beta_j\right|\right) \qquad \text{(A.11)}$$

where a vector composed of w is $w=(w_1,w_2,w_3,...,w_p)^T$ and $w_j=\left|\hat{\beta}_j\right|^{-\gamma}$; $\gamma>0$. γ is the power of the adaptive weight.

Regarding the proposed method, we defined the relaxed adaptive Lasso estimator on the set $M^{\lambda,w}\subseteq\left\{1,2,3,...,p\right\}$, where p is the number of nonzero variables selected into the ultimate model. The procedure of variable selection and shrinkage of $\hat{\underset{\sim}{\beta}}$ are controlled by two constraints (λ and ϕ) and the weight vector (w) to the penalty term. Thus, the relaxed adaptive Lasso estimator is defined as follows:

$$\hat{\underset{\sim}{\beta}}_{RAlasso} = \arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[\begin{array}{l} y_i\left(\beta_0+\sum_{j=1}^{p}x_{ij}\left\{\beta_j\cdot 1_{M^{\lambda,w}}\right\}\right)\\ -\ln\left(1+\exp\left(\beta_0+\sum_{j=1}^{p}x_{ij}\left\{\beta_j\cdot 1_{M^{\lambda,w}}\right\}\right)\right)\end{array}\right]\right\}+\phi\lambda\sum_{j=1}^{p}w_j\left|\beta_j\right|\right)$$
$$\text{(A.12)}$$

where $1_{M^{\lambda,w}}$ is an indicator function $\left\{1_{M^{\lambda,w}}\right\}_k=\begin{cases}0, & k\notin M^{\lambda,w}\\ 1, & k\in M^{\lambda,w}\end{cases}$, for all $k\in\left\{1,2,3,...,p\right\}$; $\phi\in\left[0,1\right]$. $w_j=\left|\hat{\beta}_j\right|^{-\gamma}$, $\gamma>0$.

REFERENCES

[1] Makalic E, Schmidt DF. Review of modern logistic regression methods with application to small and medium sample size problems. In: Li J, editor. AI 2010: Advances in artificial intelligence. Lecture notes in computer science. 1st ed. Berlin, Heidelberg: Springer 2010; p. 213-222. https://doi.org/10.1007/978-3-642-17432-2_22

[2] Sudjai N, Siriwanarangsun P, Lektrakul N, *et al*. Tumor-to-bone distance and radiomic features on MRI distinguish intramuscular lipomas from well-differentiated liposarcomas. J Orthop Surg Res 2023; 18: 255. https://doi.org/10.1186/s13018-023-03718-4

[3] Sudjai N, Siriwanarangsun P, Lektrakul N, *et al*. Robustness of radiomic features: two-dimensional versus three-dimensional MRI-based feature reproducibility in lipomatous soft-tissue tumors. Diagnostics 2023; 13: 258. https://doi.org/10.3390/diagnostics13020258

[4] Tang Y, Cui J, Zhu J, Fan G. Differentiation between lipomas and atypical lipomatous tumors of the extremities using radiomics. J Magn Reson Imaging 2022; 56: 1746-54. https://doi.org/10.1002/jmri.28167

[5] Kleinbaum DG, Klein M. Logistic regression: a self-learning text. 3rd ed. New York: Springer; 2010. https://doi.org/10.1007/978-1-4419-1742-3

[6] Hosmer DW, Lemeshow SJ. Applied logistic regression. 3rd ed. New Jersey: Wiley; 2013. https://doi.org/10.1002/9781118548387

[7] Senaviratna NAMR, Cooray TMJA. Multicollinearity in binary logistic regression model. In: Thapa N, editor. Theory and practice of mathematics and computer science. 1st ed. West Bengal: BP International 2021; p. 11-9. https://doi.org/10.9734/bpi/tpmcs/v6/2417E

[8] Brimacombe M. High-dimensional data and linear models: a review. Open Access Med Stat 2014; 4: 17-27. https://doi.org/10.2147/OAMS.S56499

[9] Belsley DA, Kuh E, Welsch RE. Regression diagnostics: identifying influential data and sources of collinearity. New York: John Wiley & Sons; 1980. https://doi.org/10.1002/0471725153

[10] Kastrin A, Peterlin B. Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. Expert Syst Appl 2010; 37: 5178-85. https://doi.org/10.1016/j.eswa.2009.12.074

[11] Hosseinnataj A, Bahrampour A, Baneshi M, *et al*. Penalized Lasso methods in health data: application to trauma and influenza data of Kerman. Journal of Kerman University of Medical Sciences 2019; 26: 440-9. https://doi.org/10.22062/jkmu.2019.89573

[12] Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Stat Med 2016; 35: 1159-77. https://doi.org/10.1002/sim.6782

[13] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970; 12: 55-67. https://doi.org/10.1080/00401706.1970.10488634

[14] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol 1996; 58: 267-88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[15] Zou H, Hastie T. Regularization and variable selection via the elastic Net. J R Stat Soc Series B Stat Methodol 2005; 67: 301-20. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[16] Zou H. The adaptive Lasso and Its oracle properties. J Am Stat Assoc. 2006; 101: 1418-29. https://doi.org/10.1198/016214506000000735

[17] Meinshausen N. Relaxed Lasso. Comput Stat Data Anal 2007; 52: 374-93. https://doi.org/10.1016/j.csda.2006.12.019

[18] 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

[19] Cherkassky V, Mulier F. Learning from data: concepts, theory, and methods. 2nd ed. New Jersey: John Wiley and Sons; 2006. https://doi.org/10.1002/9780470140529

[20] 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

[21] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. New York: Springer; 2013. https://doi.org/10.1007/978-1-4614-7138-7

[22] Hardin J, Garcia SR, Golan D. A method for generating realistic correlation matrices. Ann Appl Stat 2013; 7: 1733-62. https://doi.org/10.1214/13-AOAS638

[23] Hastie T, Tibshirani T, Friedman JB. The elements of statistical learning: data mining inference and prediction. 2nd ed. Berlin, Heidelberg: Springer; 2009. https://doi.org/10.1007/978-0-387-84858-7

[24] Kassambara A. Machine learning essentials: practical guide in R. 1st ed: STHDA; 2017.

[25] Sudjai N, Duangsaphon M. Liu-type logistic regression coefficient estimation with multicollinearity using the bootstrapping method. Science, Engineering and Health Studies 2020; 14: 203-14. https://li01.tci-thaijo.org/index.php/sehs/article/view/222465

[26] 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

Received on 05-08-2023 Accepted on 01-09-2023 Published on 17-09-2023

https://doi.org/10.6000/1929-6029.2023.12.13