AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Performance of Classification for Adjusted Adaptive Elastic Net Penalty and Adaptive LASSO for Breast Cancer Data

Afiaa R. Khudhair	Saja M. Hussein
afya-rahim@utq.edu.iq	saja@coadec.uobaghdad.edu.iq
Statistics Department -College of Administration and Economics -University of ThiQar, ThiQar, Iraq	Statistics Department -College of Administration and Economics -University of Baghdad, Baghdad, Iraq

Article Information

Article History:
Received: February, 7, 2024
Accepted: April, 12, 2024
Available Online: 31, December, 2024

Keywords:
Classification, Penalized, Binary, Lasso, High-dimensional

Correspondence:
Afiaa R. Khudhair
afya-rahim@utq.edu.iq

Abstract

In The current time witnesses a significant surge in data, fueled by technology's rapid advancement. This increase in data volume has led to the emergence of high-dimensional data (where the number of variables exceeds the sample size), creating challenges in precision and target identification. Consequently, binary response variable classification becomes intricate due to the multicollinearity in explanatory variables. To tackle this, response variable classification has prompted the utilization of penalization techniques, reduced variables and selecting best variables in the model. This aids in simplifying the model complexity to attain the specific binary outcome (0,1). In this paper, penalization methods were applied, including Adaptive Least Absolute Shrinkage and Selection Operator, With the Adjusted Adaptive Elastic Net Penalty with logistic regression model. The application involved a set of real data. A sample collected by the researcher (p=49, n=41), and it produced positive results for classification in the sample collected by the researcher, in resulted we found these methods have achieving high classification accuracy while efficiently selecting an optimal number of variables using a range of packages and functions in the R programming language.

DOI: https://doi.org/10.55562/jrucs.v56i1.27

1. Introduction

In recent years, technological advancements have led to an exponential increase in the amount of data generated across a wide range of fields, including medicine, economics, and social sciences. High dimensional data, where the number of explanatory variables(p) is larger than the size of sample (n) ($p > n$), has become increasingly common in these fields. The analysis of such data presents unique challenges due to the complex relationships between variables and the large number of potential predictors.

One example of high-dimensional data is cancergene expression data based on microarray analysis. Such data can contain thousands of genes, leading to an increase in complexity and the potential for multicollinearity, making traditional regression models ineffective. The challenge is to identify the best set of explanatory variables that can accurately classify patients based on their gene expression data.[11]

In this paper used the high dimensional data is 49 variables for 41 observation and applying two methods (Adjusted Adaptive Elastic Net Penalty AAEN) and (The Adaptive LASSO (Least

Absolute Shrinkage and Selection Operator)) for choose the best variables to classify the response variable.

In (2012) researchers (Sun and Wang) studying Penalized logistic regression model and used the high-dimensional DNA) methylation data. They specifically focused on (CpG) it refers to correlated DNA methylation sites within variables (genes) derived from high dimensional data. The Penalized procedure is developed is based on a union of the (Lasso) penalty and squared (Elastic net) penalty applied to coefficients of CpG sites within one gene. Their proposed method outperformed existing mainstream regularization techniques, particularly at what time data correlated with some.[14]

In (2014) (EL.Anbari and Mkhadri) proposed a new method for simultaneous identification of variables that favors the aggregation effect, applied to a partial regression model where highly correlated variables tend to be both inside and outside the model. The new method based on least squares with penalty function that added the (lasso (L1)) criteria and the Correlation Based Penalty criteria. It was called by the researchers the L1CP method. The results showed that it is more accurate, predictive, and more adaptive than Elastic net method work in tow case $p <= n$ (the number of variables is less than or equal to the sample size). If $p > n$, this method remains as good as and allows selection of many variables.[18]

In (2015) the researchers (Algamal , Hisyam Lee) Applying the Adjusted Adaptive Elastic Net penalty to select genes that are consistent and promote a simultaneous clustering effect in cancer classification for the high dimensional . They applied this method to three real microarray datasets and found that the Adjusted Adaptive Elastic Net was competitive and effective. It achieved useful results in Classification Accuracy, Sensitivity, Specificity, and convergence of in gene selection.[2]

In this paper, our focus is on binary regression and classification for (Yi), which deals with a binary response or dichotomous response variable. In This type of response variable in paper is (0,1) , Zero means infected and one is not infected, can only take two values, typically represented as "1, 0," "1, -1," or using other codes such as "good and bad," "big and small," "win and lose," "alive and dead," or "healthy and sick." Binary regression models have found widespread application in various fields, including business, computer science, education, and genetic or biomedical research.

When discussing binary regression, one of the most prevalent and widely used techniques is logistic regression, also known as logit regression or the logit model. Logistic regression is extensively employed for associating binary responses with covariates. It forms the basis for many binary classification tasks and has a significant impact on fields where predicting a binary outcome is essential.

To enhance the binary logistic regression model and introduce a penalty term, we move from binary logistic regression to binary logistic regression with penalization. The penalized technique adds a penalty to the model to prevent overfitting and improve generalization. This transformed model is often referred to as logistic regression with penalized binary logistic regression.

Binary logistic regression is a valuable tool for modeling binary responses in various applications, and by incorporating regularization techniques, we can enhance its performance and make it more robust in handling high dimensional datasets and preventing overfitting.[8]

These models add a penalty term to the Likelihood function to shrink the coefficient estimates towards zero, effectively performing variable selection and reducing the complexity of the model. The performance of these models is evaluated based on classification accuracy and error rate.

The classification similar forecasting processes but the main difference between forecasting and classification lies in their approach. Forecasting relies on specific inputs and outputs representing distinct values, while classification uses algorithms to generate the best fit curve from input data. The goal of automatic classification is to accurately predict the class of an object x based on observations.[10].

2. Penalized Logistic Regression

Penalized logistic regression applies a penalty to the logistic model when there are too many variables, causing the coefficients of less significant variables to shrink towards zero. The main goal of variable selection in high-dimensional data is to identify the most relevant and informative subset of explanatory variables to enhance classification accuracy and simplify model interpretation. [3][13]

Penalization is a common technique employed in high-dimensional data for variable selection. It involves introducing a penalty term denoted as Pλ(β) into the log-likelihood function. This addition serves the purpose of improving prediction accuracy while preventing over fitting.

Recently times, there has been a growing interest in using penalization methods within logistic regression models. These methods aim to identify the most important explanatory variables in classification tasks. Various forms of penalty terms have been proposed to tailor the approach to the application's specific requirements. [12]

Penalized logistic regression incorporates a nonnegative regularization term into the negative log-likelihood function, ℓ(β), allowing for better control over the size of variable coefficients in high-dimensional situations. Conventional logistic regression is not suitable for high-dimensional data, given the situation where there are significantly more variables than observations, this can result in challenges related to multi co-linearity and over fitting. To address these challenges, penalized logistic regression is employed.

Logistic regression analysis is used to classify the occurrence of an event as not occurring, such as determining whether a person is sick or healthy or whether a process has succeeded or failed. The logistic transformation generates a vector of probability estimates to inform these predictions. [4]

$\pi i = p(yi=1|xi)$ represents the probability of the binary event ($y_i = 1$) for the ith observation.

The logit transformation,

$$\text{Ln}\left[\left(\frac{\pi_i}{1-\pi_i}\right)\right] = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \qquad \text{for i = 1, 2, ..., n} \qquad (1)$$

β_0 : the intercept terms

β_j : p*1 vector of unknown coefficients.

The log-likelihood function:

$$L(\beta_0, \beta) = \sum_{i=1}^{n} \{y_i - ln(\pi_i) + (1 - y_i)\, ln(1 - \pi_i)\} \qquad (2)$$

Hint:

$$\pi_i = p(y_i = 1/X_i) \quad , \quad 1 - \pi_i = p(y_i = 0/X_i)$$

The Probability of classifying the (i) the sample in class 1 is estimated by:

$$\pi_i = exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}) / 1 + exp = (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}^{t}) \qquad (3)$$

and the predicted class is then obtained by I (π_i > 0.5) where I(.) is an indicator function. The penalized approach in logistic regression is achieved by incorporating a penalty term into the negative log-likelihood function. [1][7][14]

$$PLR = -\sum_{i=1}^{n} \{y_i - \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)\} + \lambda P(\beta) \qquad (4)$$

P(β) : is the penalty term .

λ : the tuning parameter.

The estimation of the vector $\hat{\beta}$ is obtained by minimizing:

$$\hat{\beta}PLR = \arg\min_{\beta} \left[\sum_{i=1}^{n} \{y_i \ln(\pi_i) + (1 - y_i)\ln(1 - \pi_i)\} + \lambda P(\beta)\right] \qquad (5)$$

The positive tuning parameter controls the tradeoff between fitting the data to the model and the penalty's effect. This parameter balance between the bias and the variance to minimize the misclassification error.[2] [3]

3. Tuning Parameter

The tuning parameter is a crucial component in selecting the best-fitting model. It is a non-negative parameter, and the penalty limit depends on the value of λ and a control quantity that

influences the degree of shrinkage of the parameters. When λ = 0, the tuning parameter reduces to the maximum likelihood estimation (MLE) estimator, while as λ approaches 1, the regularization term forces all variable coefficients to be zero.

In classification problems, the tuning parameter plays a crucial role in achieving an optimal balance between bias and variance, ultimately minimizing misclassification errors. Cross-validation is a widely used technique for determining the optimal the tuning parameter value (typically denoted as λ) in penalized models.[3][4]

Cross-validation involves dividing the dataset into smaller subsets or folds, often using a 10-fold scheme. The model is trained on the majority of these folds, leaving one-fold out for testing. This process is repeated, with each fold as a test set once. [1][5]

By systematically varying the value of the tuning parameter the optimal value of λ is determined based on the metric most relevant to the problem, such as the lowest average misclassification error or another appropriate performance measure.

The advantage of cross-validation is that it provides an unbiased estimate of the model's generalization performance and helps in avoiding over fitting. The tuning parameter can be fine-tuned Through this iterative process to strike the right balance between model complexity and predictive accuracy, resulting in improved classification performance. [16]

4. Adjusted Adaptive Elastic Net Penalty (AAEN)

The adjusted adaptive elastic net penalty is a regularization method for variable selection in logistic regression. It was introduced to address the limitations of LASSO and combines ridge and LASSO to deal with highly correlated variables while simultaneously selecting the best set of variables. Elastic net relies on two non-negative tuning parameters, lambda (1) and lambda (2), which control the regularization in penalized logistic regression solutions. It performs well when there are strong pairwise correlations among variables, but its reliability may decrease when the absolute correlation between variables falls below 0.95. Additionally, the elastic net does not explicitly account for the correlation structure among variables.

The penalized logistic regression model using elastic net (PLR with EN) is defined as follows:

$$\hat{\beta}^{*}_{\text{Elasticnet}} = \text{argmin}_{\beta}\,[\sum_{i=1}^{n}\{(y\text{i}\ln\left(\pi\left(x_{ij}\right)\right)+(1-y_i)\ln\left(1-\pi\left(x_{ij}\right)\right)\}+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2] \quad (6)$$

Elastic Net (EN) proposed by Zou and Hastie (2005) penalized method for logistic regression is defined by minimizing a function with two non-negative tuning parameters, (λ1) and (λ2). [17]

El Anbari and Mkhadri (2014) found that the reliability of the elastic net may decrease slightly when the absolute correlation between variables is less than 0.95 [7] . the elastic net does not consider the correlation structure among variables, which has been noted by Bühlmann, Rütimann, van de Geer. [6]

Although the elastic net shows better classification accuracy, it lacks the oracle property. Additionally, it does not solve the grouping effect issue of adaptive LASSO. [6][8]

It combines ridge regression regularization with the adaptive LASSO to address the variable selection and grouping effect problems. The improved adaptive elastic net method provides better variables selection consistency and grouping effect than adaptive LASSO. Hastie and Zou proposed an (Elastic Net) EN penalized method with a fixed λ2. [16]

In the case of high-dimensional classification problems, MLE may not be feasible, rendering the application of the Adaptive Elastic Net unworkable. [17].

Using the standard error of the ridge estimator, $s_{j(\hat{\beta}_{\text{Ridge}})}$, provides an advantage in adjusting the regularized logistic regression when using ridge regression or elastic net estimates as an initial weight.

The Penalized logistic regression using the Adjusted Adaptive Elastic Net Penalty (AAEN) is defined as follows:

$$\hat{\beta}^{**}AAEN = \operatorname{argmin}\beta \left[-\sum_{i=1}^{n}\{(y_i^* \ln\left(\pi\left(x_{ij}\right)\right) + (1-y_i^*)\ln\left(1-\pi\left(x_{ij}\right)\right)\} + \lambda 1\sum_{j=1}^{p} w_{Ratio\,j}|\beta_j| + \lambda 2\sum_{j=1}^{p}\beta_j^2\right] \quad (7)$$

$w_{Ratio,j}\ j = (s_{j(\hat{\beta}_{Ridge})} / |\hat{\beta}_{j(Ridge)}|)^{-\gamma}$, j =1,2,….,p ,

γ *: constant positive*

$$\hat{\beta}AAElastic = (1+\lambda 2)\,\hat{\beta}^{**}AAElastic / w_{j\,Ratio} \quad (8)$$

The Adjusted Adaptive Elastic Net (AAEN) demonstrates the oracle property. [3][5]

5. The Adaptive LASSO (Least Absolute Shrinkage and Selection Operator)

The Lasso method is a popular technique for variable selection in high-dimensional data. However, it has some shortcomings, especially when the penalties of different coefficients are all the same and not related to the data. To compensate for these shortcomings, researchers have proposed various suggestions to make penal methods more accurate and effective in classifying high-dimensional data.[9]

Lasso is one of the most popular penalizations that has gained widespread popularity and served as the foundation for other penalized methods due to its unique capability to perform continuous shrinkage of descriptor coefficients and descriptor selection simultaneously. This approach effectively addresses limitations in traditional regression methods. And his idea is to multiply the penalty function by a certain weight. [15]

The value of this weight is the reciprocal of the absolute value of the parameters estimated in an elementary way (Lasso) (Tishirani 1996).I is a method for estimation parameters in the linear model by minimizing the residual sum of the square to the sum of the absolute values of the coefficients. The lasso estimate β is defined by [16][17]:

$$\hat{\beta}LASSO = \text{argBmin}\left[\sum_{i=1}^{n}(yi - \beta 0 - \sum_{j=1}^{k}\beta jxij)\,2 + \lambda\sum_{j=1}^{k}|\beta j|\right] \quad (9)$$

where $\lambda\sum_{j=1}^{k}|\beta j|$ *is the penalty function.*

For the binary dependent variable, the lasso estimate β is regularized from:

$$\hat{\beta}LASSO = \operatorname{argmin}\beta\left[-\sum_{i=1}^{n}\{(y_i \ln(\pi(x_i)) + (1-y_i)\ln(1-\pi(x_i))\} + \lambda\sum_{j=1}^{k}|\beta j|\right] \quad (10)$$

The Adaptive LASSO introduces weights that penalize different coefficients within the L1-penalty. This approach allows for a more flexible and adaptive regularization, particularly useful when dealing with complex data and models. [2][3]

The Adaptive LASSO is defined as (proposed by Zou, (2006))(18):

$$\hat{\beta}APLR = \operatorname{argmin}\beta\left[-\sum_{i=1}^{n}\{(y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i))\} + \lambda\sum_{j=1}^{k} w_j|\beta j|\right] \quad (11)$$

$P\lambda(|\beta_j|) = \lambda\sum_{j=1}^{p} wj\,|\beta_j|$

w_j: represents the weights dependent on the data and is calculated as follows:

$w_j = \frac{1}{|\hat{\beta}_{lasso}|^{\gamma}}$, $\gamma > 0$ positive constant, γ : shrinkage parameter

w_j=(w1,w2,….,wp)T is p*1

The parameter γ is a positive constant. [2][3]

6. Results

The real data collected by the researcher from breast cancer patients at the Cancer Center of the Dhi Qar Health Department was used. have 49 variables (p) and 41 (n) observations:

\Table 1: Symbols and Description of the Variables

Variable	Name	Full name of test
X1	Marital status	Marital status
X2	Nutrition	Nutrition
X3	Sport	Sport
X4	N.child	N.child
X5	Living location	Living location
X6	Genetic factor	Genetic factor
X7	Age	Age
X8	H	Height
X9	W	Weight
X10	BSA	Body Surface Area
X11	WBC	white blood cell
X12	RBC	Red blood cell
X13	HGB	*Hemoglobin*
X14	HCT	Hematocrit Test
X15	MCV	Mean Corpuscular Volume
X16	MCH	Mean Corpuscular Hemoglobin
X17	MCHC	Mean Corpuscular Hemoglobin Concentration
X18	PLT	Platelet Count Test
X19	LYM	Lymphocytes Percentage Test
X20	MXD	Mixed Cells Absolute Count.
X21	NEUT	Neutrophils
X22	RDW-SD	Red Cell Distribution Width - Standard Deviation
X23	RDW-CV	Red cell distribution width
X24	PDW	Platelet Distribution Width
X25	MPV	Mean platelet volume
X26	P-LCR	platelet large cell ratio
X27	PCT	Procalcitonin
X28	S.C	Serum creatinine
X29	PR	Progesterone
X30	ER	*Estrogens* and Oestrogens
X31	Her2\neu	Human Epidermal growth factor Receptor
X32	Ki-67	Antigen KI-67
X33	CA15-3	Cancer antigen 15-3
X34	S.calsuim	Serum Calcium
X35	CEA	Carcinoembryonic antigen
X36	blood urea	blood urea
X37	S.A.L.T(G.P.T)	Serum Glutamic-pyruvic Transaminase
X38	S.A.S.T(G.O.T	aspartate aminotransferase
X39	total bilirubin	total bilirubin
X40	R.B.Sugar	casual blood glucose
X41	s.sodium	Serum sodium
X42	S.Albumin	Serum albumin
X43	S.cholesterol	Serum cholesterol
X44	S.chloride	Serum chloride
X45	ALP	Alkal.phospphatsase
X46	v d3	Vitamin D3
X47	S.Triglyceride	serum Triglyceride
X48	S.Mg+	Magnesium Blood Test
X49	smoking	smoking

In Table (2) the provided Lasso (0.17) misclassification rate, and (AAEL, 0.09) The misclassification rate. A lower rate implies the method is more successful in accurately classifying instances. Conversely, a higher rate indicates a greater likelihood of misclassifying

instances. indicating its strong predictive capability and relatively accurate classification performance.

Table 2: Number Variables and Classification Accuracy and Misclassification rate

Methods	(No.Var.)	CA	misclassification rate
The Adaptive LASSO	5	0.83	0.17
AAEL	9	0.91	0.09

Table (3), Considering the results values in the confusion matrix, most of the methods achieved high TP, TN, and specificity values, indicating an excellent performance in correctly classifying both tumor and non-tumor instances. The sensitivity values are also high, indicating a good ability to detect positive cases. It is important to note that the evaluation was conducted using 70% of the data was used for model training and 30% for testing. They correctly classified 7 instances as class 1. The sensitivity (Sen.) values for methods is 1 , indicating that they correctly identified all instances of class 1. For class 0 (non-tumor), The methods achieved true negatives (TN) ranging from 3 to 4, indicating that they correctly classified 3 to 4 instances as class 0. The specificity (Spe.) values for all methods are either 0.6 or 0.8, indicating a relatively good ability to identify instances of class 0 correctly. Considering the distribution of classes in the test data (7 instances of class 1 and 5 instances of class 0), the methods performed reasonably well in correctly classifying instances of both classes.

These high sensitivity and specificity values further highlight the efficiency of the method. The method "AAEL" also performed well, with high accuracy and sensitivity, and "AAEL" selected a smaller number of genes while maintaining high performance, suggesting their efficiency in gene selection for classification tasks.

Table 3: Confusion matrix for methods

Methods	TP	TN	FP	FN	Sen.	Spe.
The Adaptive LASSO	7	3	2	0	1	0.6
AAEL	7	4	1	0	1	0.8

In Table 4, the AAEL method has selected nine variables as the most important and relevant for determining whether a patient is infected with breast cancer or not. The method has selected the following nine variables: X7 (Age), X18 (Platelet Count Test), X19 (Lymphocytes Percentage Test),X21 (Neutrophils), X22 (Red Cell Distribution Width - Standard Deviation),X33 (Cancer Antigen 15-3), X35 (Carcinoembryonic Antigen),X40 (Casual Blood Glucose), X45 (Alkaline Phosphatase) .The explanation suggests that the variables selected by the AAEL method demonstrate strong associations and significant impact on breast cancer diagnosis. After consultation with oncology specialists, it has been clarified that these variables are among the most important ones for determining whether a patient has the disease or not. The AAEL method has identified a set of nine variables that, according to the analysis and expert consultation, are considered significant in predicting breast cancer diagnosis. The positive coefficients indicate the direction of the association between each variable and the predicted outcome. Fig.1 explains this method.

Table 4 :(method = AAEL)

No. Var.	Coefficient
Intercept	-10.357539081
X7	0.044688403
X18	0.005107768
X19	0.013485439
X21	0.022346436
X22	0.005996942
X33	0.021792440
X35	0.028998833
X40	0.030486953
X45	0.025652228

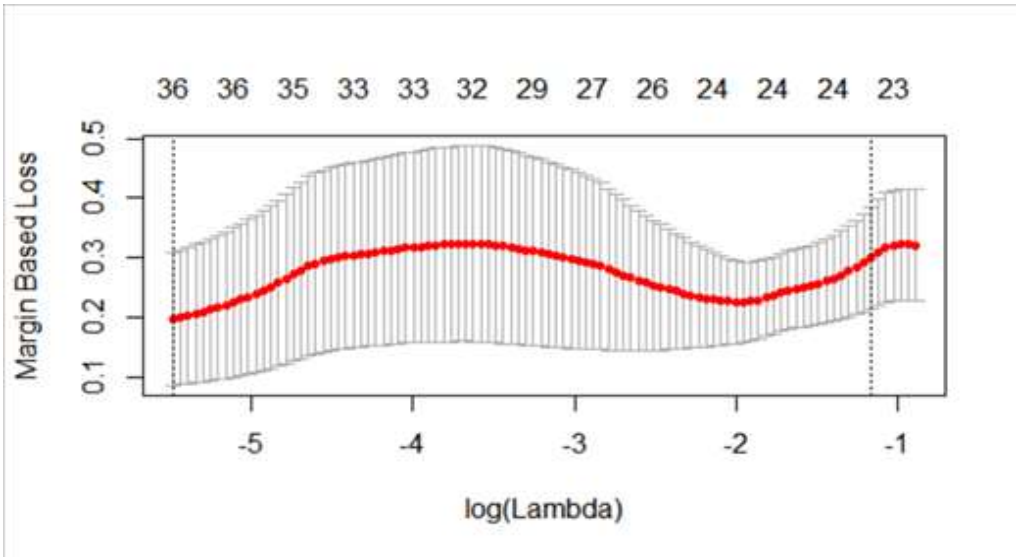

Figure 1: Display AAEL

In Table (5) (Adaptive lasso) selected five variables (X28: Serum creatinine, X29: Progesterone, X30: Estrogens and Oestrogens, X39: total bilirubin, X49: smoking). These variables represent the critical important variables included in the model according to the first weight of Lasso, but they are not the most essential variables present within the sample. Moreover Fig.2 explains this method.

Table 5: (method =Adaptive lasso)

No. Var.	Coefficients
(Intercept)	2.244614e+01
X28	-1.877738e+01
X29	-1.585368e+00
X30	-3.325463e-15
X39	-8.040432e+00
X49	6.234680e-01

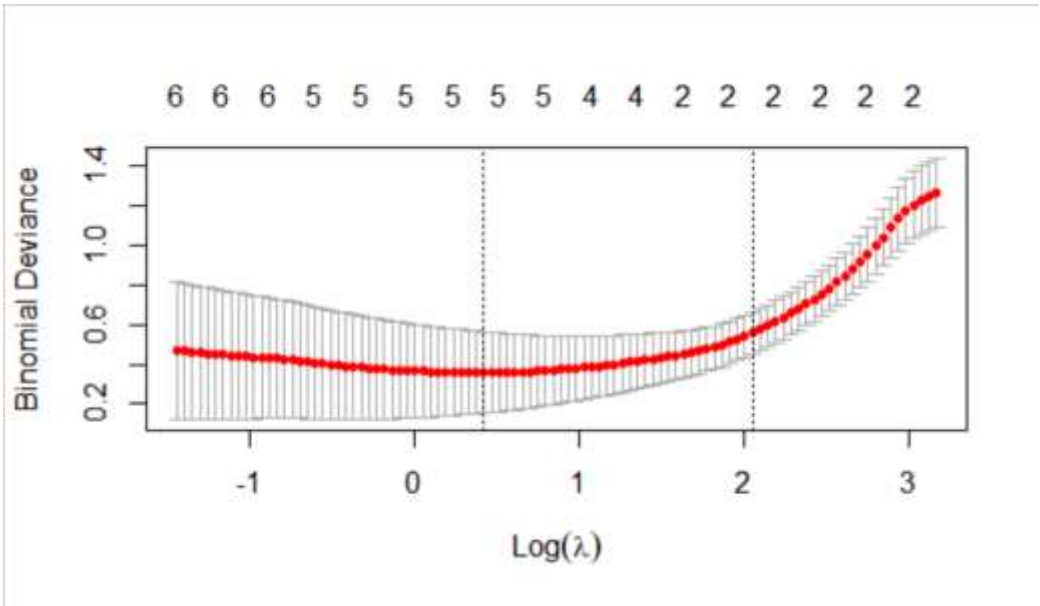

Figure 2: Display The Adaptive lasso

7. Conclusions:

We have concluded that the Adaptive lasso method chose the smallest number of variables, while the AAEL method chose a similar number of variables, but through our application of the sample collected by the researcher, the number of variables chosen through the AAEL method

is the best and actual variables that are included in the model to know the category of the dependent variable (0 or 1). Therefore, we have found a critical solution to our research question, which involves selecting a specific number of variables from a larger set of unimportant variables. This will allow us to classify the variable dependent as either affected or unaffected based solely on the selected variables.

References

[1] Ahmed, A. Y., Kahya, M. A., & Altami, S. A. (2022). Classification improvement of gene expression for bipolar disorder using weighted sparse logistic regression. Bulletin of Electrical Engineering and Informatics, 11(2).

[2] Algamal, Z. Y., & Lee, M. H. (2015). Applying penalized binary logistic regression with correlation-based elastic net for variables selection. Journal of Modern Applied Statistical Methods, 14(1), 168-179.

[3] Algamal, Z. Y., & Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Systems with Applications, 42(20), 9328-9332.

[4] Algamal, Z. Y., & Lee, M. H. (2017). A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives. SAR and QSAR in Environmental Research, 28(1), 75-90.

[5] Araveeporn, A. (2021). The Higher-Order of Adaptive Lasso and Elastic Net Methods for Classification on High Dimensional Data. Mathematics, 9(10), 1091. https://doi.org/10.3390/math9101091

[6] Bühlmann, P., & Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

[7] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417.

[8] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[9] Huang, H., Gao, Y., Zhang, H., & Li, B. (2021). Weighted Lasso estimates for sparse logistic regression: Non-asymptotic properties with measurement errors. Acta Mathematica Scientia, 41(1), 207-230.

[10] Hussein, S M. (2019). Comparison of Some Suggested Estimators Based on Differencing Technique in the Partial Linear Model Using Simulation. Baghdad Science Journal, 16(4), 0918-0918.

[11] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.

[12] Jiang, H., Hu, X., & Jiang, H. (2021). Penalized Logistic Regressions with Technical Indicators Predict Up and Down Trends. https://doi.org/10.21203/rs.3.rs-1098354/v1

[13] Khudhair, A.F., & Hussein, S M. (2023). Performance classification for Lasso weights with penalized logistic regression for high-dimensional data. Journal of Economics and Administrative Sciences JEAS.

[14] Sun, H., & Wang, S. (2012). Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics, 28(10), 1368-1375.

[15] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58, 267–288.

[16] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

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

[18] El Anbari, M. E., & Mkhadri, A. (2014). Penalized regression combining the L1 norm and a correlation-based penalty. Sankhya B, 76(1), 82-102.

مجلة كلية الرافدين الجامعة للعلوم (2024)؛ العدد 56؛ 297- 306

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

اداء التصنيف لطريقة الشبكة المرنة التكيفية المعدلة الجزائية وطريقة لاسو المعدلة لبيانات سرطان الثدي

م.م افياء رحيم خضير	أ.د سجى محمد حسين
afya-rahim@utq.edu.iq	saja@coadec.uobaghdad.edu.iq
قسم الاحصاء، كلية الادارة والاقتصاد، جامعة ذي قار، ذي قار، العراق	قسم الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد، بغداد، العراق

المستخلص

يشهد الوقت الحالي طفرة كبيرة في البيانات، يغديها التقدم التكنولوجي السريع. وقد أدت هذه الزيادة في حجم البيانات إلى ظهور بيانات عالية الأبعاد (حيث يتجاوز عدد المتغيرات حجم العينة)، مما خلق تحديات في الدقة وتحديد الأهداف. وبالتالي، يصبح تصنيف متغير الاستجابة الثنائية معقدًا بسبب التعددية الخطية في المتغيرات التوضيحية. ولمعالجة ذلك، أدى تصنيف متغير الاستجابة إلى استخدام تقنيات الجزاء وتقليل المتغيرات واختيار أفضل المتغيرات في النموذج. وهذا يساعد في تبسيط تعقيد النموذج لتحقيق النتيجة الثنائية المحددة (0،1). في هذا البحث، تم تطبيق أساليب الجزاء، بما في ذلك Adaptive Lasso، مع الشبكة المرنة التكيفية المعدلة الجزائية مع نموذج الانحدار اللوجستي الجزائي. ويتضمن التطبيق مجموعة من البيانات الحقيقية. العينة تم جمعها من قبل الباحثة (p = 49، n= 41)، وظهرت نتائج إيجابية للتصنيف في العينة التي جمعتها الباحثة، ونتيجة لذلك وجدنا أن هذه الأساليب قد حققت دقة تصنيف عالية مع اختيار العدد الأمثل للمتغيرات بكفاءة باستخدام مجموعة من الحزم والوظائف في لغة البرمجة R.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث:7/2/2024
تاريخ قبول البحث:12/4/2024
تاريخ رفع البحث على الموقع:
31/12/2024

الكلمات المفتاحية:

لاسو المعدلة، الجزاء، انموذج الانحدار اللوجستي الجزائي، التصنيف، عالية الابعاد

للمراسلة:

افياء رحيم خضير
afya-rahim@utq.edu.iq

DOI: https://doi.org/10.55562/jrucs.v56i1.27