Journal of Economics and Administrative Sciences
2024; 30 (139), pp. 149-160

P-ISSN 2518-5764
E-ISSN 2227-703X

Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq
DOI: https://doi.org/10.33095/rtzbhh20

Performance Classification for Lasso Weights with Penalized Logistic Regression for High-Dimensional Data

Afiaa Raheem khudhair*
Department of Economics,
College of Administration and Economics ,
University of Thi-Qar, Iraq
afya-rahim@utq.edu.iq
***Corresponding author**

Saja Mohammed. Hussein
Department of Statistics
College of Administration **and** Economics,
University of Baghdad, Iraq
saja@coadec.uobaghdad.edu.iq

Received:5/6/2023 **Accepted: 1/8/2023** **Published Online First: 29 /2/ 2024**

Abstract:

In high-dimensional data, classification performance is a crucial consideration. One method of interest is the penalized binary logistic regression. However, (Least Absolute Shrinkage and Selection Operator) Lasso method may face problems when the appropriate penalty for each coefficient is not determined. For this reason, different weights are used in weighted Lasso estimates to address this issue and improve classification performance. To overcome this limitation, we employ various Weighted Lasso Estimates, each with unique weight assignments, and compare their performance with our fifth proposed weight configuration. This application of Lasso weighting schemes aims to uncover the most effective approach for high-dimensional classification tasks while considering the optimal set of variables.

The evaluation criteria for these methods include the number of selected variables, classification accuracy, and mean squared error. We then apply these techniques to real-world data to identify the most effective classification mode and select the optimal set of variables. This rigorous and precise investigation aims to provide a robust and reliable classification approach for high-dimensional systems.

Paper type: Research paper.
Keywords: Classification, Penalized, Binary, Weighted, Lasso, High-dimensional, Weighted lasso

P-ISSN 2518-5764
E-ISSN 2227-703X

1.Introduction:

In recent years, the rapid advancements in modern science and technology have led to the prevalence of high-throughput and non-parametric complex data in various scientific fields such as gene-biology, chemometrics, and neuroscience. This has resulted in challenges like the "large p, small n" paradigm, where the number of covariates (p) exceeds the sample size (n), making it difficult classify data and select optimal explanatory variables effectively. Researchers have been exploring various regularization techniques to address these challenges.

This paper focuses on regression cases involving binary responses (or dichotomous responses). The responses $\{y_i\}$ can take only two values: "1, 0", "1, -1" or some other codes representing dichotomous responses such as: good and bad, big and small, win and lose, alive and dead, or healthy and sick. The challenge arises when classifying this type of data, as the number of explanatory variables (p) exceeds the sample size (n), leading to increased model complexity. This increased in complexity makes it difficult to effectively classify the data and select the optimal set of explanatory variables.

1.1 Literature review:

Many studies discussed the Least Absolute Shrinkage and Selection Operator (Lasso). Tibshirani (1996) used Lasso for variable selection and estimation in high-dimensional data. Subsequent advancements, such as the Adaptive Lasso proposed by Zou and Hastie (2005) improved the accuracy of variable by using data-driven weights. Sun and Wang (2012) developed a penalized logistic regression model specifically for high-dimensional DNA methylation data, outperforming existing regularization techniques.

El Anbari and Mkhadri (2014) introduced the (lasso-Correlation Based Penalty) L1CP method, which combined the L1 criteria and correlation-based penalty criteria to improve variable selection and estimation in partial regression models.

Algamal and Lee (2015) proposed the Adjusted Adaptive Elastic Net penalty for gene selection in high-dimensional cancer classification, demonstrating competitive results in classification accuracy and gene selection consistency. Saleh (2016) employed semi-parametric methods, such as (Least Absolute Shrinkage and Selection Operator -Minimum average variance estimation) LASSO-MAVE, to enhance estimation accuracy and flexibility in single-index models.

Sur (2019) developed inferential tools for determining the correct number of principal components under a general noisy latent variable model, including the noisy independent component model as a special case. The problem is approached using hypothesis testing.

Araveeporn (2021) presented an interesting exploration of Lasso and elastic net methods, as well as their higher-order adaptive counterparts, in the context of high dimensional data classification using logistic regression models. The author conducts a series of simulations with varying numbers of independent variables and sample sizes smaller than the number of independent variables to study the performance of these methods.

The main problem in this research is the challenge of dealing with high-dimensional data, where an extensive number of variables are present, making it difficult to identify the most relevant variables for model building. And choose the best set of variables for the classification of the observation.

This research aims to reduce the high dimensions of the data and choose the optimal set of explanatory variables by using the latest penal methods to impose a different penalty on the transactions. In addition, the main objective is to classify the binary response variable (y) into two categories (0 or 1).

P-ISSN 2518-5764
E-ISSN 2227-703X

2.Material and Methods:

2.1 Data set:

The data set used in the study is a binary cancer classification data set which contains 100 samples, 53 of which are prostate tumor samples and 47 are non-tumor tissues (Ghaddar and Naoum-Sawaya, 2018).

The dataset was used to evaluate the effectiveness of penalty methods for the binary logistic model for classification purposes, where multicollinearity and overfitting were observed as major problems. Each sample in the data set contains information on 12600 genes. The prostate cancer data set is commonly used in research on cancer classification due to its large number of genes and its suitability for evaluating classification models.

The small sample consisted of 40 women with breast cancer at the Oncology Hospital. The researcher collected the sample at the Cancer Oncology Hospital in (Thi Qar) Governorate, and it was found that the sample included 27 females with breast cancer and 13 females who were not infected. The sample was subjected to a total of 49 medical examinations (variables).

2.2 Penalized logistic regression model:

Penalized logistic regression imposes a penalty on the logistic model for having too many variables. This results in shrinking the coefficients from the less contributive variables toward zero. We will select an optimal subset of explanatory variables in order to improve the classification accuracy and to make the model's interpretation easier is the main objective of the variable selection in high dimensional data (James, 2013).

Although logistic regression is one of the most popular classification methods, it does not choose variables (Huang, 2016).

A procedure called penalization, which is always used in variable selection in high dimensional data, attaches a penalty term $P_\lambda(\beta)$ to the log-likelihood function to get a better estimate of the prediction error by avoiding overfitting for parameters. Lately, there is growing interest in applying the penalization method in the logistic regression model (Sun and Wang, 2012).

In order to extract the most important explanatory variables in classification problems, a series of penalized logistic regression many methods have been proposed. and There are varieties of different forms of the penalty term, depending on the application requirement for the main target Penalized logistic regression adds a nonnegative regularization term to the negative log-likelihood function, $\ell(\beta)$, such that (Algamal, 2015).

The size of variables coefficients in high-dimension can be controlled. Because there are many more variables than observations, conventional logistic regression does not apply to high dimensions. Also, there Multicollinearity and overfitting are specific issues. Because of this, we have used penalized logistic regression. When attempting to forecast whether or not an event has a place, such as when determining whether a person was sick, healthy, or failed, logistic regression analysis is utilized. From the vector of probability estimates after logistic transformation (Algamal and Lee, 2015).

The general formula of logistic regression is written by:

$$y_i = \pi(x_i) + \varepsilon_i \ , \ i=1,2,\ldots,n \qquad (1)$$

Where y_i denotes the value of a dichotomous outcome variable, $p(x_i)$ denotes the probability of the Bernoulli distribution dependent or independent variable, X_i, and ε_i is called the error and follows a normal distribution with mean zero and variance equal to

$$p(x_i)\,[1- p(x_i)], \qquad (2)$$

P-ISSN 2518-5764
E-ISSN 2227-703X

the logistic regression model is considered as the probability by:

P (X_t; β_0, $\beta_1 \ldots \beta_j$)=p (y_i=1|X_{ij}; β_0 , $\beta_1 \ldots \beta_j$)=

$$p(x_i)= \frac{exp\ (\beta_0+X_i^t\ \beta_j)}{1+\ exp\ (\beta_0+X_i^t\ \beta_j)} \quad (3)$$

$$p(y_{\ i}=1|X_i) = \frac{exp\ (\beta_0+X_i^t\ \beta_j)}{1+\ exp\ (\beta_0+X_i^t\ \beta_j)} \quad (4)$$

$p(x_i)$= p(y_i=1|x_i) is modeled by a linear function, logit transformation:

$$\text{Ln}\ [\frac{p(x_i)}{(1-p(x_i))}] = \beta_0 + \sum_{j=1}^{p} \beta_j\ X_{ij}{}^{T} \qquad i=1,2\ldots..,n, \quad (5)$$

β_0: the intercept terms
β_j : p*1 vector of unknown coefficients.
The log-likelihood function:

$$L\ (\beta_0, \beta\ ) = \sum_{i=1}^{n}\{y_i \ln p(x_i)) + (1 - y_i)\ln(1 - p(x_i))\} \quad (6)$$

Where:

$$p(x_i)= p(y_{\ i}=1|X_i) \quad (7)$$

$$(1 - p(x_i)) = p(y_{\ i}=0|X_i) \quad (8)$$

The probability of classifying (i=1, 2,.n) for the sample in class 1 is estimated by

$$p(x_i) = exp(\beta_0 + \sum_{j=1}^{p} \beta_j\ X_{ij}^t)/1 +\ exp\ (\beta_0 + \sum_{j=1}^{p} \beta_j\ \ X_{ij}^t) \quad (9)$$

and the predicted class is then obtained by I ($p(x_i)$ > 0.5), where I(.) is an indicator function.

The penalized method for the logistic regression is obtained by adding the penalty term to the negative log-likelihood function:

$$PLR = -\sum_{i=1}^{n}\{y_i \ln(p(x_i)) + (1 - y_i)\ln(1 - p(x_i))\} + \lambda P(\beta) \quad (10)$$

$P(\beta)$ is the penalty term that penalizes the estimates. The penalty term depends on the positive tuning parameter, λ the tuning parameter should find the right balance between the bias and the variance to minimize the misclassification error (Sun and Wang, 2012).

The estimation of the vector β is obtained by minimizing:

$$\widehat{\beta}_{PLR}= \arg\min_{\beta} [\sum_{i=1}^{n}\{y_i \ln(p(x_i)) + (1 - y_i)\ln(1 - p(x_i))\} + \lambda P(\beta)\ ] \quad (11)$$

The tradeoff between fitting the data to the model and the penalty's effect is controlled by the positive tuning parameter.

P-ISSN 2518-5764
E-ISSN 2227-703X

2.3 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's role is to find the right balance between bias and variance to minimize misclassification errors. To determine the optimal value, cross-validation is employed. In this thesis, 10-fold cross-validation was conducted based on the training set to find the optimal value of λ (Algamal, 2015).

Cross-validation involves dividing the dataset into multiple smaller subsets or "folds." The model is then trained on the majority of these folds and tested on the remaining fold. This process is repeated, each fold is used as a test set once, resulting in a collection of performance metrics that can be averaged to estimate the model's performance.

By varying the value of the tuning parameter λ, different models can be compared and evaluated using cross-validation. The optimal value of λ is the one that yields the lowest average misclassification error or another appropriate performance metric. This optimal λ value will balance the trade-off between model complexity and prediction accuracy, resulting in a model that performs well on new (Sun and Wang, 2012).

2.4 Weighted Lasso Estimates:

The limitations of the Lasso (Least Absolute Shrinkage and Selection Operator) method for variable selection and regularization in linear regression models. Lasso can have difficulties when the penalties of different coefficients are the same and not related to the data. This can lead to suboptimal performance in certain cases, particularly with high-dimensional data.

To address these shortcomings, researchers have proposed various improvements and extensions to the Lasso method. One such improvement is the weighted Lasso, which involves assigning different weights to the penalties of the coefficients. These weights can be data-dependent, and they typically consist of an unknown constant and a tuning parameter. The weighted Lasso aims to provide better convergence rates and more accurate variable selection compared to the ordinary Lasso (Algamal, 2017).

However, it's essential to note that the weighted Lasso is not a perfect solution either. Like any other method, it comes with its own set of assumptions and limitations. For instance, selecting appropriate weights can be challenging, and the method's performance can be sensitive to the choice of weights. Moreover, the weighted Lasso still may not be suitable for all types of data or problems, and researchers should consider alternative regularization methods or model selection techniques depending on the specific context.

In summary, the weighted Lasso improves the ordinary Lasso, aiming to provide better convergence rates and more accurate variable selection. However, it has its limitations, and researchers should consider the appropriateness of this method depending on the specific problem and data at hand (Huang, 2021).

In high-dimensional settings, where the number of variables (p) is much larger than the number of observations (n), the Lasso and its variants, including the weighted Lasso, can be quite useful. These methods help in variable selection, shrinkage, and regularization, leading to more interpretable and accurate models.in high-dimensional data, the ordinary Lasso may struggle to identify the correct set of variables due to the equal penalty assigned to all coefficients. This issue can be mitigated by using the weighted Lasso, as it allows for data-dependent weights on the penalties of the coefficients. This flexibility can lead to better performance in variable selection and prediction in high-dimensional settings.

P-ISSN 2518-5764
E-ISSN 2227-703X

However, it is crucial to remember that the performance of the weighted Lasso depends on the choice of weights, which can be challenging to determine in practice. Additionally, high-dimensional data can present other challenges, such as multicollinearity, sparsity, or noise, which may require alternative methods or additional preprocessing steps.

Weighted lasso:

2.4.1 The first Weighted :[Adaptive LASSO]

Lasso is one of the most popular penalization terms. where gained popularity and became a basis for other penalized methods because of its ability to simultaneously perform continuous shrinkages of the descriptor coefficient and descriptor selection. This method appeared to overcome the shortcomings and his idea is to multiply the penalty function by a certain weight.

As we observe, Zou and Zhang pointed out that the adaptive LASSO outperforms LASSO in terms of achieving the oracle property, even though the grouping effect problem for adaptive LASSO remains (Algamal, 2017).

The value of this weight is the reciprocal of the absolute value of the parameters estimated in an elementary way appeared Lasso from (Tibshirani, 1996) 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. (Lin, 2009)

The Lasso estimate β is defined by:

$$\hat{\beta}_{\text{LASSO}}=\arg_{\text{B}}\min\left[\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{k}\beta_j X_{ij})^2+\lambda\sum_{j=1}^{k}|\beta j|\right], \quad (12)$$

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

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

$$\hat{\beta}_{\text{LASSO}}=\operatorname{argmin}_{\beta}\left[-\sum_{i=1}^{n}\{(y_i\ln(p(x_i))+(1-y_i)\ln(1-p(x_i)))\}+\lambda\sum_{j=1}^{k}|\beta j|\right] \quad (13)$$

The Adaptive LASSO proposed weights are used for penalizing different coefficients in the L1-penalty. The main idea behind the Adaptive LASSO is that by assigning inga higher weight to the small coefficients and a lower weight to the large coefficients it is possible to reduce the bias.

The Adaptive LASSO is defined as:

$$\widehat{\beta}_{\text{APLR}}=\operatorname{argmin}_{\beta}\left[-\sum_{i=1}^{n}\{(y_i\,ln(p(x_i))+(1-y_i)ln(1-p(x_i)))\}\right.$$
$$\left.+\lambda\sum_{j=1}^{k}w_j|\beta j|\right] \quad (14)$$
$$P_{\lambda}(|\beta_j|)=\lambda\sum_{j=1}^{p}w_j|\beta_j| \quad (15)$$

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

$w_j=(w_1, w_2,\ldots,w_p)^T$ is p*1

P-ISSN 2518-5764
E-ISSN 2227-703X

Since there is no presented information about model parameters, we cannot directly compare the selection and prediction accuracy. The comparison will be done by model size and prediction error, formerly lots of coefficients estimated by weighted Lasso methods, four are very small but not zero (Algamal, 2017; Huang, 2021).

2.4.2 The second Weighted:

$$w_j \propto \max_{i=1,\ldots,n}|x_{ij}| \sqrt{\frac{2}{n}(r\log p + \log 2)}, \quad r = 1 \qquad (16)$$

i=1,….n j=1,…..p

(Where r > 0 is a constant.)

2.4.3The third Weighted:

$$w_j \propto \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2} \cdot \sqrt{\frac{2}{n}(r\log p + \log 2)}, \quad r=1$$

(17)

(Where r > 0 is a constant.)

i=1,….n j=1,…..p

2.4.4 The fourth Weighted:

$$w_j = \{ \frac{1}{n}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_{.j}\right)^2 \}^{-1}$$

(18)

$w_j = [s\hat{d}_j]^{-\gamma}$ i=1,….n j=1,2, 3,..,p

where ŝdj is the standard deviation for each variable.

$\gamma > 0$ positive constant. (Huang, 2021)

2.4.5 The fifth weighted: based mean (suggestion method)

Despite the ongoing issue of the aggregation effect of weighted averages, we have yet to find a weight that relies on the arithmetic mean of each column in the data. Therefore, we propose this weight to assess its performance compared to other weights.

$$w_j = \frac{(\max(X_{ij}) - (X_{ij}))}{(max(X_{ij}) - \min(X_{ij}))} \qquad (19)$$

$\max(X_{ij})$: max value in col.

$\min(X_{ij})$: min value in col.

All of the above weights (w1, w2, w3, w4, w5) are substituted into the following equation (14):

$$\hat{\beta}_{APLR} = \text{argmin}_{\beta}[-\sum_{i=1}^{n}\{(y_i\, ln(p(x_i)) + (1-y_i) ln(1-p(x_i)))\} + \lambda\sum_{j=1}^{k} w_j|\beta j|]$$

2.5 Evaluation criteria of classification performance:

The classification performance of the model, classification accuracy (CA), sensitivity (Sen), and specificity (SP):

$$CA = \frac{TP+TN}{TP+FP+FN+TN} * 100\% \qquad (20)$$

$$Sen = \frac{TP}{TP+FN} \qquad (21)$$

$$SP = \frac{TN}{FP+TN} \qquad (22)$$

P-ISSN 2518-5764
E-ISSN 2227-703X

$$\text{FP rate} = \frac{\text{FP}}{\text{FP+FN}} \quad (23)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (24)$$

TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

3. Discussion of Results:

In this section, we apply weights estimates and we propose to analyze prostate data, our target is to select useful genes for specifying 0 and 1.

Table (1) presents the results of applying different Lasso weights (Type 1 to 5) to a dataset to select a subset of important genes (variables) and evaluate the important genes (variables) and evaluating the performance of each method. The performance is measured using classification accuracy (CA), sensitivity (Sen.), and specificity (Sep.). Let's go through each method and explain the results.

1. The Type First Weight method selects 23 genes and achieves the highest classification accuracy of 0.9. The sensitivity and specificity are also high at 0.93 and 0.85, respectively. This method strikes a balance between the number of selected genes and performance metrics.

2. The Type Two Weight method selects 142 genes and achieves a classification accuracy of 0.5 The sensitivity and specificity are 0.6 and 0.47, respectively. This method identifies many genes but has relatively low performance metrics compared to other methods.

3. The Type third Weight method selects 53 genes and achieves a classification accuracy of 0.7The sensitivity is quite high at 0.63, but the specificity is 0.90. This method identifies fewer genes and performs better than Type two Weight.

4. The Type Fourth Weight method selects only 8 genes and achieves a classification accuracy of 0.7 The sensitivity and specificity are 0.71 and 0.77, respectively. This method identifies the fewest genes and has a balanced performance in terms of sensitivity and specificity.

5. Type Fifth Weight method selects 24 genes and achieves the highest classification accuracy of 0.9 this method is outstanding for classification. The sensitivity and specificity are also high at 0.93 and 0.86,

In summary, the tables show that different Lasso weight types result in different numbers of selected genes and performance metrics. Type IV Weight appears to be the best-performing method, with a balanced number of selected genes and the highest classification accuracy, sensitivity, and specificity. These results highlight the importance of selecting appropriate weights in the weighted Lasso method to achieve the best performance in a given application.

Table 1: The number of variables and accuracy for all weights for big sample

Methods	Statistics			
	Selected genes	C.A	sensitivity	specificity
Adaptive Lasso1	23	0.9	0.93	0.85
Weighted Lasso 2	142	0.5	0.6	0.47
Weighted Lasso 3	53	0.7	0.63	0.90
Weighted Lasso 4	8	0.7	0.71	0.77
Weighted Lasso 5	24	0.9	0.93	0.86

Table 2 provides performance metrics for different methods using Lasso weights. The table includes the precision, false positive (FP) rate, false negative (FN), true negative (TN), false positive (FP), and true positive (TP) values for each method.

P-ISSN 2518-5764
E-ISSN 2227-703X

In summary, the table shows the performance of several variations of weighted Lasso methods, along with an adaptive Lasso method. The precision values range from 0. **5 to** 0.9, indicating the proportion of correctly identified positive cases. The FP rates vary from 0.**6** to 1, representing the proportion of falsely identified negative cases. The FN values range from 0 to 7, indicating the number of incorrectly identified positive cases. The TN values range from 7 to 13, representing the number of correctly identified negative cases. The FP values range from 1 to 3, only Lasso 2 has 9 false negatives representing the number of falsely identified negative cases. The TP values range from 12 to 15, indicating the number of correctly identified positive cases.

In general, the methods achieve relatively high precision values, ranging from 0.**8** to 0.9, indicating a high proportion of correctly identified positive cases. However, the FP rates vary, suggesting differences in the proportion of falsely identified negative cases among the methods. The FN values also differ, indicating variations in the number of incorrectly identified positive cases. The TN, FP, and TP values show variations in the number of correctly and falsely identified negative and positive cases among the methods.

Table 2: Performance Metrics for Lasso Weights

Methods	Statistics					
	TP	FP	TN	FN	FP rate	Precision
Adaptive Lasso1	15	2	12	1	0.6	0.8
Weighted Lasso 2	9	9	8	4	0.6	0.5
Weighted Lasso 3	12	1	10	7	0.12	0.9
Weighted Lasso 4	15	2	7	6	0.25	0.8
Weighted Lasso 5	15	2	13	0	1	0.8

* Tables and results from the researcher's work on the R program.

The following Tables 2 and 1 represent the number of variables and accuracy and the confusion matrix for classification, which is used as evaluation metrics for the model calculated from 30% of the data. The matrix elements were calculated for 30 samples out of a total of 100, where it included 17 within Class 1 and 13 within Class 0, where the actual model was built using 70% of the data. This is indicated by all the weights that were chosen, including our suggested weight, which proves the efficiency, quality and accuracy of the proposed weight (Liu and Wong, 2019).

Applying all methods with breast cancer (small sample n=40, p=49):

Table 3: The number of variables and accuracy for all weights for a small sample

Methods	Statistics			
	Selected genes	C.A	sensitivity	Specificity
Adaptive Lasso1	4	83.3	100	77
Weighted Lasso 2	5	83.3	100	77
Weighted Lasso 3	6	83.3	100	77
Weighted Lasso 4	2	83.3	100	60
Weighted Lasso 5	5	91	100	87

P-ISSN 2518-5764
E-ISSN 2227-703X

Table 4: Performance Metrics for Lasso Weights for small sample

Methods	Statistics					
	TP	FP	TN	FN	FP rate	Precision
Adaptive Lasso1	7	1	4	0	1	0.8
Weighted Lasso 2	7	2	3	0	0.7	0.7
Weighted Lasso 3	7	2	3	0	1	0.7
Weighted Lasso 4	7	2	3	0	1	0.7
Weighted Lasso 5	7	0	4	1	1	1

Tables and results from the researcher's work on the R program**.** The results of a proposed method in a small sample application. In The tables (3) and (4), we notice that proposed method had a high classification accuracy in the third table.
And the proposed method also gave results similar to the previous weights in the table (3). These observations indicate the quality and strength of the method in classification.

The following tables **3** and 4 represent the number of variables and accuracy and the confusion matrix for classification, which is used as evaluation metrics for the model calculated from 30% of the data. The matrix elements were calculated for 12 samples out of a total of 40, where it included 7 within Class 1 and 5 within Class 0, where the actual model was built using 70% of the data. This is indicated by all the weights that were chosen, including our suggested weight, which proves the efficiency, quality and accuracy of the proposed weight.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

In this context, "Positive" and "Negative" refer to the predicted classification of the model, while "True" and "False" indicate the accuracy of the predictions compared to the actual values. The confusion matrix assess the model's performance by measuring quantities such as true positives, false negatives, false positives, and true negatives.

3.Conclusion:

1.The performance of various Lasso weight types on this particular task demonstrates the importance of selecting the appropriate weight type based on classification with penalized logistic regression.
2.Both Type first and Type fifth weights exhibit strong performance in terms of classification accuracy.
3.The choice of the ideal Lasso weight type should be determined by considering the trade-offs between classification accuracy and number variables, and any other relevant factors or metrics that are crucial to the particular problem.
4.The fifth weight method, proposed by us, exhibits remarkable performance when dealing with high-dimensional data in both big data and small data scenarios. with penalized logistic regression model for classification tasks.

The effectiveness and success of our method are clearly evident in the results obtained. With a good classification matrix, high classification accuracy, and the fulfillment of criteria such as sensitivity, specificity, and other classification metrics, our method establishes its reliability and demonstrates immense potential for practical applications.

4.Further Work:

Applying this weighted with other models and using our proposal in multi-response model.

P-ISSN 2518-5764
E-ISSN 2227-703X

Authors Declaration:

Conflicts of Interest: None

-We Hereby Confirm That All The Figures and Tables In The Manuscript Are Mine and Ours. Besides, The Figures and Images, Which are Not Mine, Have Been Permitted Republication and Attached to The Manuscript.

- Ethical Clearance: The Research Was Approved By The Local Ethical Committee in The University.

References:

1. Lin, Z., Xiang, Y. and Zhang, C. 2009. "Adaptive Lasso in high-dimensional settings." *Journal of Nonparametric Statistics*, 21 (6),pp. 683-696.

2. Huang, H. H., Liu, X. Y. and Liang, Y. 2016." Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid L1/2 +2 Regularization". *PLOS ONE*, 11 (5). Retrieved from https://doi.org/10.1371/journal.pone.0149675.

3. Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (*Statistical Methodology*), 67(2), pp. 301-320. doi:10.1111/j.1467-9868.2005. 00503.x.

4. Sur, P., Candes, E. J., and Witten, D. 2019. "Estimating the number of components in a large matrix". *The Annals of Statistics*, 47 (6), pp.3152-3182.

5. Tibshirani, R. 1996."Regression shrinkage and selection via the Lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp. 267-288.

6. James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013, "An introduction to statistical learning", Springer, New York.

7. Algamal, Z.Y. and 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), pp. 168-179.

8. Algamal, Z. Y., and Lee, M. H. 2015."High dimensional logistic regression model using adjusted elastic net penalty". *Pakistan Journal of Statistics and Operation Research*, 11 (4),pp. 667-676.

9. Sun, H., and Wang, S. 2012. "Penalized logistic regression for high-dimensional DNA methylation data with case-control studies". *Bioinformatics*, 28. (10), pp.1368-1375.

10. Huang, H., Gao, Y., Zhang, H., and Li, B. 2021. "Weighted Lasso estimates for sparse logistic regression: Non-asymptotic properties with measurement errors". *Acta Mathematica Scientia*, 41 (1), pp.207-230.

11. Algamal, Z. Y., and 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),pp. 75-90.

12. ELAnbari, M. E., and Mkhadri, A. 2014. "Penalized regression combining the L 1 norm and a correlation-based penalty". *Sankhya B* ,76 (1), pp.82-102.

13. Saleh, T.A. 2016." Some of the semi-parametric methods to estimate and variable selection for single index model". Ph.D. thesis, College of Administration and Economics, University of Baghdad.

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

15. Ghaddar, B., and Naoum-Sawaya, J. 2018. "High dimensional data classification and feature selection using support vector machines". *European Journal of Operational Research*, 265 (3),pp.993-1004.Retrievedfrom https://doi.org/10.1016/j.ejor.2017.08.040

16. Liu, C., and Wong, H. S. 2019. "Structured Penalized Logistic Regression for Gene Selection in Gene Expression Data Analysis". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16 (1), pp.312-321. doi:10.1109/TCBB.2017.2767589.

Journal of Economics and Administrative Sciences
2024; 30 (139), pp. 149-160

P-ISSN 2518-5764
E-ISSN 2227-703X

اداء التصنيف لأوزان لاسو مع انموذج الانحدار اللوجستي الجزائي للبيانات عالية الابعاد

افياء رحيم خضير
جامعة بغداد / كلية الادارة والاقتصاد
قسم الإحصاء
بغداد ، العراق
afya-rahim@utq.edu.iq

سجى محمد حسين
جامعة ذي قار/ كلية الإدارة والاقتصاد
قسم الاقتصاد
العراق
saja@coadec.uobaghdad.edu.iq

Received:5/6/2023 **Accepted: 1/8/2023** **Published Online First: 29 /2/ 2024**

مستخلص البحث:
في البيانات ذات الأبعاد العالية هناك مشكلة عدم معرفة اختيار المتغيرات ذات الاهمية لذلك يعد أداء التصنيف معياراً مهماً لمعرفة اهم المتغيرات الداخلة في النموذج حيث يلخص هذا البحث اداء تصنيف متغير الاستجابة للبيانات عالية الابعاد من خلال تطبيق اوزان مختلفة للاسو مع الوزن المقترح من قبل الباحث مع انموذج الانحدار اللوجستي الجزائي وتم تطبيق هذه الازوان على بيانات حقيقية تضمنت 12600 جين لعينة مؤلفة من 100 مشاهدة وعينة صغيرة تم جمعها من قبل الباحثة حيث تضمنت 40 انثى مصنفة (27 مصابة و13 غير مصابة) وتم استخدام برنامج R للحصول على النتائج حيث تم التوصل الى ان الاوزان تعمل بدقة عالية وجيدة لغرض التصنيف وحققت الطريقة المقترحة نتائج جيدة وعملية في اختيار افضل المتغيرات التوضيحية لتصنيف متغير الاستجابة .

نوع البحث: مستل من اطروحة دكتوراه .

المصطلحات الرئيسة للبحث: الجزاء ، التصنيف ، انموذج الانحدار اللوجستي الثنائي ، البيانات عالية الابعاد ، اوزان لاسو ، معلمة الضبط.