Istanbul Commerce University Journal of Science
İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 18(36), Güz 2019
http://dergipark.gov.tr/ticaretfbd

Research Article

TUNING PARAMATER SELECTION IN PENALIZED LOGISTIC REGRESSION WITH APPLICATION IN CANCER[*]

Sahar Fadhil AL-KHATEEB

Istanbul Commerce University, Graduate School of Science, Statistics, Kucukyalı, Istanbul, Turkey.
sahar711192@yahoo.com, Orcid.org/0000-0003-1539-8763

Abstract

Variable selection is an important subject in regression analysis intended to select the best subset of predictors. In cancer classification, gene selection plays an important issue. The Least Absolute Shrinkage and Selection Operator (LASSO) is one of most used penalized method. In logistic regression, Lasso right the traditional parameter estimation method, maximum log-likelihood, by adding the L1-norm of the parameters to the negative log-likelihood function. Lasso depends on the tuning parameter. Finding the optimal value for the tuning parameter is one of the most important topics. There are three popular methods to select the optimal value of the tuning parameter: Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and Cross-Validation (CV). The aim of this paper is to evaluate and compare these three methods for selecting the optimal value of tuning parameter in terms of coefficients estimation accuracy and variable selection through simulation studies and application in cancer classification.

Keywords: Cancer classification, gene selection, Lasso, penalized logistic regression

Araştırma Makalesi

KANSER SINIFLANDIRILMASINDA UYGULAMAYA SAHİP CEZALANDIRILMIŞ LOJİSTİK REGRESYONUNDA PARAMETRE SEÇİMİNİN AYARLANMASI

Öz

Değişken seçim, regresyon analizinde en iyi öngösterge alt kümesini seçmeyi amaçlayan önemli bir konudur. Kanser sınıflamasında gen seçimi önemli bir konudur. En az mutlak büzülme ve seçme operatörü (Lasso) en çok kullanılan ceza yöntemlerinden biridir. Lojistik regresyonda Lasso, parametrelerin L1-normunu negatif log-olabilirlik fonksiyonuna ekleyerek, geleneksel parametre tahmin yöntemini, maksimum log olasılığını değiştirir. Kement ayarlama parametresine bağlıdır. Ayar parametresi için en uygun değeri bulmak en önemli konulardan biridir. Ayar parametresinin en uygun değerini seçmek için üç popüler yöntem vardır: Bayesian bilgi kriteri (BIC), Akaike bilgi kriteri (AIC) ve çapraz doğrulama (CV). Bu çalışmanın amacı, simülasyon çalışmaları ve kanser sınıflandırma uygulamalarında katsayılar tahmin doğruluğu ve değişken seçimi açısından en uygun ayarlama parametresini seçmek için bu üç yöntemi değerlendirmek ve karşılaştırmaktır.

Anahtar kelimeler: Cezalandırılmış lojistik regresyon, gen seçimi, kanser sınıflandırması, Lasso.

[*] *Bu çalışma, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü'nde yapılan "Kanser Siniflandirilmasinda Uygulamaya Sahip Cezalandirilmiş Lojistik Regresyonunda Parametre Seçiminin Ayarlanmasi" başlıklı yüksek lisans tezinden hazırlanmıştır.*

Received / Geliş tarihi: 07/11/2019 Accepted / Kabul tarihi: 24.12.2019
Corresponding Author/ Sorumlu Yazar : sahar711192@yahoo.com

1. INTRODUCTION

In recent years, the framework of penalized methods has been gained popularity among the statisticians as the situation for performing variable selection and model estimation in high dimensional data simultaneously (Algamal, 2016).Accordingly, a family of penalized methods was proposed with a penalty term added to the likelihood function. The advantage behind the penalty term is to control the complexity of the model and provide criterion for variable selection by introducing some constraints on the parameters, which these constraints force some parameters to be exactly zero (Abdalteef, 2018). Therefore, a proper preference for the penalty expression will enhance the prediction accuracy and make an effortlessly interpretable model.

Lasso a new penalized method, which used L1-norm alternatively of L2-norm.This technique can reduce the regression coefficients closer to zero and some coefficients are precisely set to zero. Therefore, Lasso can produce interpretable models. Because of its functionality in performing variable selection, Lasso receives many functions in a distinctive of types that belong to Generalized Linear Model (GLM) household such that logistic regression and Poisson regression (Park and Hastie, 2007) have given an extremely good survey of L1-norm in penalized regression. In genomics studies, for instance, where tens of heaps of genes can be acquired with only a few lots of patients (Adragni, 2014). In the medication and biology fields, the DNA microarray technology is a very essential and important technology that provides more realism on the gained results.

In cancer research, this technological know-how helps the determination of the expression values of thousands of genes simultaneously. In most purposes of the bioinformatics and computational biology using microarray technology, the wide variety of genes, p, is higher than the number of patients (tissues), n. Cancer classification, given gene expression data, has grown to be an active subject in biomedical research. procedure with the case p > n poses a challenging mission in the utility of the statistical classification methods due to the fact the classical classification techniques bear the damn of dimensional. Using all genes often outcomes in model overfitting, especially if there are inappropriate and genes is an essential goal when dealing with high-dimensional cancer classification in rule, gene determination targets to pick out a rather few collection of genes from a high-dimensional gene dataset, and consequently obtain excessive classification accuracy(Abdalteef, 2018). Furthermore, selecting essential genes can additionally assist in aiding the clinical specialist in previous diagnosis and medicine find for most cancers patients.

2. METHODOLOGY

2.1 The Logistic Regression Model

Based on the basic assumption, the dependent variable (y) is the response variable, which we are acting in our studying. The binary variable following the distribution of Bernoulli that he takes value (1) and (0) with (1-π) probability of the occurrence of the response and no longer occurring, as properly in linear regression, whose explanatory and variable take constant values, the model that links the variables is as follows: (Azhaar, 2014:13).

$$\log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x \qquad (1)$$

2.2 Penalized Logistic Regression Model

Logistic regression (LogiR) is regarded as a statistical technique to model a binary response variable, like a cancer classification issue in which the response variable only has two values: 1 for the tumour type and 0 for the regular class. In logistic regression, the regression equation has a nonlinear link with the linear collection of the explanatory variables. The response variable follows Bernoulli distribution with density function (Algamal and Hisyam, 2015:37).

$$f\ (y_i) = \pi^{y_i}\,(1-\pi)^{1-y_i}\ , \qquad (2)$$

2.3 Lasso

The Lasso penalty function has received extensive reputation and has grown to be one of the fundamental penal method in choosing variables. This is due to their capability to operate both the downsizing of parameters and the determination of variables simultaneously. It alleviate the regression coefficients to be zero (Abdalteef, 2018:28).

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{x\beta})^T\ (\boldsymbol{y} - \boldsymbol{x\beta}) + \lambda \sum_{j=1}^{p} \left|\boldsymbol{\beta}_j\right| \right\} \qquad (3)$$

2.4 Tuning Parameter Estimation

Accurate estimation of the tuning parameter(s) value λ is very essential due to the fact it can have an increased impact on the overall performance of the penalized likelihood methods (Androulakis et al., 2011). In the different words, it performs a consistent function in steady variable selection, the place its value will figure out how many chosen explanatory variables are as properly as the bias imposed on the estimated regression coefficients. (CV), Generalized Cross-Validation (GCV) and the records criteria, such as (AIC), and (BIC), are regarded the most broadly used techniques for discovering the estimation value of (λ).

2.5 Evaluation Criteria

The overall performance evaluation of the proposed penalized likelihood method and evaluating capability present penalized method is normally evaluated in phrases of variable determination and prediction accuracy assessment criteria. (Fan and Tang, 2013).

2.5.1 Cross-Validation Method

The CV is to partition the data matrix into several folds and use one fold of the data matrix to predict the rest of the data matrix, and then to find the tuning parameter λ which gives the smallest prediction error. When the original data matrix is partitioned into k folds, we call the cross-validation by $k -$ fold cross-validation ($k -$ CV). In the penalized likelihood methods, $k -$ CV will be randomly split the dataset into k mutually exclusive folds of approximately equal size. Among the k folds, one fold is retained as validation dataset for testing the penalized likelihood model, and the remaining $k - 1$ folds are used as training dataset to fit the penalized likelihood model. The CV process is repeated k times, and each of the folds is used exactly once as validation dataset. Different values of λ could result in different fitted penalized model using the same training dataset. Different values of λ could result in different fitted penalized model using the same training dataset. The optimal penalized model is the one that has the minimum CV prediction error, and the corresponding value of the λ for the optimal penalized model is preferred (Arlot and Celisse, 2010). Typically, the value of k is often chosen between 3 and 10. When k is equal to the sample size, then $k -$ CV is called leave-one-out cross-validation (LOOCV) .

2.5.2 Variable Selection Evaluation Criteria

i. The Model Size (MS), which represents the number of the selected explanatory variables.

$$MS = \#\{j : \hat{\beta}_j \neq 0, j = 1, ..., p\} \tag{4}$$

ii. True Positive (TP), determine as the numbers of non-zero variables for a given vector that represents the true variables that estimated as zero variables.

$$TP = \#\{j : \hat{\beta}_j \neq 0 \,\&\, \beta_j \neq 0, j = 1, ..., p\} \tag{5}$$

The high value of TP shows a better-penalized method. The greater range of TP is equal to the variety of the nonzero variables in the given true vector.

iii. False positive (FP) explained as a number of the zero variables of a given true vector that estimated as non-zero variables.

$$FP = \#\{j : \hat{\beta}_j \neq 0 \,\&\, \beta_j = 0,\ j = 1,\ldots,p\} \tag{6}$$

For penalized method, the good behavior was indicated when the value of FP approaching zero. In general, a penalized method is wanted when it has the highest TP and the lowest FP.

2.5.3 Prediction Evaluation Criteria

In the classification studies, the usual performance measure of the prediction accuracy is classification Accuracy (CA) or misclassification error (misclassification rate) (ME).

3. APPLICATION

3.1 Simulation Studies

Simulation studies are conducted based on the high dimensional linear regression model as well as the high dimensional logistic regression model. Two simulation models for each regression model are considered in order to investigate to practical cases: the existence of correlation among explanatory variables and the existence of correlation between a group of explanatory variables.

One simulation model is consider for logistic regression model, because the sample size has a direct influence on the prediction accuracy (Mkhadri and Ouhourane, 2015), three kind of dataset represent the sample size of each training dataset and the testing dataset (n_{train}, n_{test}) that considered with $(50,50)$, $(100,100)$ and $(150,150)$ respectively. Besides, the consideration of the quantity of variables described as, P equal 1000, 5000 and 10000 in order to mobilized the fact that the magnitude of p has an impact on the variable selection with mainly effect on the value of FP. The data records generated using the logistic regression model as:

$$\mathbf{y} : B\left(\frac{\exp(\mathbf{X}\beta_{true})}{1+\exp(\mathbf{X}\beta_{true})}\right) \tag{7}$$

For both training and testing processes for the used datasets and for explanatory variables, matrix $\mathbf{X}$ is promoted from multivariate normal distribution $N(\mathbf{0}, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is the covariance matrix with $\Omega_{i,j} = 0.5^{|i-j|}\,(i, j = 1, 2, \ldots, p)$.

Case 1 (Small effect): In this case, we set the true vector:
$\beta_{true} = (1.5,1,0.8,0.7,-0.6,9,-3,2,\underbrace{0,...,0}_{p-q})^T$, with nonzero variables $q = 8$.

Case 2 (Large effect): In this case, we set the true vector:
$\beta_{true} = (5,-5,10,-10,15,-15,20,-20,\underbrace{0,...,0}_{p-q})^T$, with nonzero variables $q = 8$.

The generated data in each simulation model is repeat 100 times. Counting on the records of used training dataset, the $k-CV$ method was adopted, with $k = 10$, to find the appropriate values of the tuning parameters.

The logistic regression simulation model is illustrated in table 1. BIC produced a very sparse model due to the fact it gave less MS values. For instance, when n=100 and p=5000, BIC selected 19 variables compared with 27 and 37 of AIC and CV, respectively. Regarding the TP criterion, the simulation results give that BIC carried out properly compared with CV and AIC. It yielded the absolute best TP of choosing the real nonzero explanatory variables as nonzero explanatory variables, which potential that BIC selected higher real nonzero explanatory variables than the different presentation methods.

Table 1: Variable Selection Evaluation Criteria Results of the Logistic Regression Model Based on 100 Replications or Case 1

n	p	Methods	MS	TP	FP
50	1000	BIC	22	6	16
		CV	26	4	22
		AIC	24	4	20
100	5000	BIC	19	6	13
		CV	37	4	23
		AIC	27	5	32
150	10000	BIC	30	8	22
		CV	46	5	41
		AIC	42	5	37

For instance, when n=150 and p=10000, BIC chose 8 compared variables out of 8 in selected with 5 and 5 selected real variables of AIC and CV, respectively. In terms of FP criterion, give value that BIC properly in contrast with CV and AIC. It yielded the smallest FP of selecting the authentic zero explanatory variables as nonzero explanatory variables, which means that BIC selected fewer true zero explanatory variables than the other current methods., when n=50 and p=1000, BIC selected 16 variables out of p-q compared with 20 and 22 selected real variables of AIC and CV, respectively.

It can be viewed from table 2, that BIC produced the highest classification accuracy in train dataset process and the lowest misclassification error in test dataset process for the logistic regression model. When n=150 and p=1000, BIC performed greater classification accuracy at 0.95 evaluating with AIC and CV of 0.84, 0.78, respectively.

Table 2: Logistic Regression Model Prediction Accuracy Criteria Results Based on 100 Replications for Case 1

n	p	Methods	Train data	Test data
			CA	ME
150	1000	BIC	0.95 (0.05)	0.06 (0.09)
		CV	0.78 (0.09)	0.22 (0.16)
		AIC	0.84 (0.09)	0.21 (0.14)

In a similar way, for case 2, it can be observe from table 3, that BIC significantly performs the best among the other competitor methods. In terms of variable selection, the average of the MS, TP, and FP for the logistic regression simulation model are record in table 3. It can be observed from these tables that the BIC produced a very sparse model because it gave less MS values. For instance, from table 3, when n=50 and p=1000, BIC selected 22 variables compared with 24 and 26 of AIC and CV, respectively. In a similar way, for case 2, it can be seen from tables 3 and 4 that BIC significantly performs the best among the other competitor methods. In terms of variable selection, the average of the MS, TP, and FP for the logistic regression simulation model are recorded in table 3. The BIC produced a very sparse model because it gave less MS values. For instance, from table 3, when n=50 and p=1000, BIC selected 22 variables compared with 24 and 26 of AIC and CV, respectively. Regarding the TP criterion, the simulation results suggested that BIC performed well compared with CV and AIC. It yielded the highest TP of selecting the true nonzero explanatory variables as nonzero explanatory variables, which means that BIC selected higher true nonzero explanatory variables than the other existing methods. For instance, from table 3, when n=100 and p=1000, BIC selected 6 true variables out of 8 compared with 4 and 4 selected true variables of AIC and CV, respectively. In terms of FP criterion, on the other hand, the simulation results suggested that BIC performed well compared with CV and AIC. It yielded the smallest FP of selecting the true zero explanatory variables as nonzero explanatory variables, which means that BIC selected fewer true zero explanatory variables than the other existing methods. For instance, from Table 3, when n=150 and p=10000, BIC selected 23 variables out of p-q compared with 37 and 40 selected true variables of AIC and CV, respectively. Regarding the prediction performance, it can be seen from table 4, that BIC produced the highest possible classification accuracy in the train dataset records and the lowest misclassification error in test dataset records for the logistic regression model, from table 4, when n=100 and p=1000, BIC performed greater classification accuracy at 0.95 evaluating with AIC and CV of 0.88, 0.79, respectively.

Table 3: Variable Selection Evaluation Criteria Results of the Logistic Regression Model Based on 100 Replications for Case 2.

n	p	Methods	MS	TP	FP
50	1000	BIC	22	7	15
		CV	26	4	21
		AIC	24	5	20
100	1000	BIC	23	6	17
		CV	27	4	23
		AIC	25	4	21
150	10000	BIC	31	8	23
		CV	45	5	40
		AIC	41	6	37

Table 4: Prediction Accuracy Criteria Results of the Logistic Regression Model Based on 100 Replications for Case 2.

N	P	Methods	Train data	Test data
			CA	ME
100	1000	BIC	0.95 (0.05)	0.09 (0.08)
		CV	0.79 (0.09)	0.17 (0.16)
		AIC	0.88 (0.09)	0.21 (0.14)
	5000	BIC	0.94 (0.08)	0.11 (0.03)
		CV	0.84 (0.11)	0.16 (0.15)
		AIC	0.86 (0.09)	0.14 (0.13)
	10000	BIC	0.92 (0.09)	0.12 (0.09)
		CV	0.79 (0.09)	0.22 (0.16)
		AIC	0.83 (0.09)	0.21 (0.14)

3.2 Real Data Application

To evaluate the performance behavior of the BIC and compare it with AIC and CV in a real practical application, three real binary dataset records belong to three kind of cancer were used in this study. Diffuse large B-cell lymphoma (DLBCL), prostate cancer , and colon cancer are the three used dataset as illustrated in Table 5 that show some details for these dataset.

Table 5: The Detail Information for the Used Datasets

Dataset	**# samples**	**# genes**	**Classes**
DLBCL	77	7,129	DLBCL / FL
Prostate	102	5,966	Tumour / Non-tumour
Colon	62	2,000	Tumour / Normal

Seventy seven value with 7,129 gene expression in each value belong to DLBCL data represent the gene expression data. These sample were measured using high-density oligonucleotide microarrays which consist of 58 sample as diffuse large B-cell lymphomas and 19 samples of follicular lymphoma (FL). For prostate dataset 12,600 genes for each 52 prostate tumour samples and 50 non-tumour tissues was used in this study. A subset of 5,966 genes was adapted in the classification. The colon cancer dataset, contained gene expression of 40 tumour and 22 normal colon tissues for 6,500 human genes estimated by Affymetrix oligonucleotide array. A subset of 2,000 genes with the highest minimal intensity across the samples was used. To accurately assess the comparison, two datasets records were generated randomly from each test. From the original size of the used dataset 70% of samples were used in training process 30% used in testing process. In order to get better values for tuning parameters, the 10-fold CV was proposed using training dataset. The median of MS and CA calculated from the training dataset while the ME calculated from the testing dataset. Table 6 illustrate the values of each used methods.

Table 6: Classification Evaluation Performance Results of the Used Methods Over 50 Partitions

Datasets	Methods	Evaluation criteria		
		MS	CA	ME
Prostate	BIC	22	0.932 (0.381)	0.128 (0.241)
	CV	44	0.913 (0.472)	0.224 (0.428)
	AIC	27	0.901 (0.482)	0.235 (0.312)
DLBCL	BIC	18	0.951 (0.301)	0.133 (0.218)
	CV	55	0.937 (0.398)	0.277 (0.305)
	AIC	22	0.919 (0.401)	0.281 (0.289)
Colon	BIC	10	0.964 (0.642)	0.121 (0.207)
	CV	24	0.942 (0.901)	0.257 (0.364)
	AIC	14	0.917 (0.661)	0.249 (0.373)

From table 6, the value of the MS for BIC detect fewer genes compared with two other methods. In DLBCL, where only 18 gene detected for BIC while 55 and 22 genes for CV and AIC respectively. Maximum accuracy of 0.932, 0.951 and 0.964 for prostate, DLBCL, and colon datasets achieved in terms of classification accuracy respectively. Furthermore, the results show that the BIC outperformed the AIC in terms of classification accuracy for all datasets. Moreover, BIC improved the classification accuracy compared to CV. The improvements were 2.03%, 1.47%, and 2.28% for the prostate, DLBCL, and colon datasets. And in terms of ME, it can also be seen from table 6 that BIC has the lowest misclassification error of 0.128, 0.133, and 0.121 for the prostate, DLBCL, and colon datasets, respectively. As a result, BIC can correctly classify the outcome variable in the test datasets.

For further proving for the stability of the obtained results for the BIC, in classifying the high dimensional cancer datasets with a high degree of accuracy compared to the other methods. A two-way analysis of variance (ANOVA) used in this study as a statistical test to check the relationship and the differences of classification accuracy statistically significant differences where the *p*-value was obtained as < 0.05as shown in table 7. The BIC and the two other used methods in terms of classification accuracy. In addition, it was obvious that the DLBCL, prostate, and the colon datasets had different classification accuracy values.

Table 7: ANOVA test result for CA

Source	df	SS	MS	F	*p*-value
Methods	2	0.4087	0.20435	232.691	0.0000
Datasets	2	0.1233	0.06165	70.200	0.0000
Error	445	0.3908	0.000878		
Total	449	0.9228			

Moreover, Duncan’s multiple range test proposed to gain more detailed results for the differences between the BIC and the other adopted methods in this study. Table 8 illustrate the p-value of each compared pair of methods. It was noted from Table 8 that the BIC showed statistical differences compared to the AIC and CV. Overall, the results of the real data application are encouraging and indicating that BIC yields the best classification accuracy results with a lower misclassification error and fewer selected genes compared with AIC and CV.

Table 8: The $p-$Value of Duncan’s Multiple Range test for CA between the used methods

	BIC	AIC	CV
BIC		0.0224	0.0000
AIC			0.0012
CV			

4. CONCULUSION

1- According to the three used methods, all are evaluate through extensive simulation studies applied for and veritable data analysis. The results of the simulation studies and the actual data applications explain that the performance of the BIC method yields very satisfactory results in expression of variable chosen and prediction accuracy. Comparing to AIC and CV, BIC efficiently outperformed them.

2- Comparing to AIC, BIC showed slight differences in terms of TP and prediction accuracy, but they are still preferable than the CV.

3- The simulation and practical results showed that the LASSO method is the penal method So the LASSO function is one of the most common methods by adding it to the sum of the remaining squares.

4- In the future, additional extensions can be made to transact for the issue of selecting the associated variables when both the response variable and the demonstration variables have outliers.

REFERENCES

Abdalteef A. M., (2018), “Variable selection in Poisson regression model using penalizedlikelihood methods”, University of Mosul ,Faculty of Mathematics and Statisti,Master Thesis in Statistics, Mosul.

Adragni, K. P., (2014), Independent screening in high-dimensional exponential family predictors space, Journal of Applied Statistics, 42(2), 347–359.

Algamal, Z., Hisyam M., (2015), “Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification”,Expert Systems with Applications;42(23):9326-9332.

Algamal, Z. Y., (2016) ,“Adaptive Penalized Likelihood Methods In High Dimension algeneralized Linear Models", Unpublished, phD Thesis, UniversitiTeknologi Malaysia.

Arlot, S., Celisse, A., (2010), A survey of cross-validation procedures for model selection. Statistics Surveys. 4, 40–79.

Androulakis, E., Koukouvinos, C., Mylona, K., (2011), Tuning parameter estimation in penalized least squares methodology. Communications in Statistics - Simulation and Computation. 40(9), 1444–1457.

Azhaar, J., (2014), “Multivariate Data Analysis for Diagnosis of phthalmic Diseases Using the Distributive Function and Logistic Regression Comparative Study ",Mustansiriya University, Faculty of Management and Economics, Master Thesis in Statistics, Baghdad.

Fan, Y., Tang, C.Y., (2013), Tuning parameter selection in high dimensional penalized likelihood, Journal of the Royal Statistical Society. Series B (Methodological). 75(3), 531–552.

Park, B.U., Hastie, T., (2007), L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society. Series B (Methodological). 69, 659677.

Mkhadri, A., Ouhourane, M., (2015), A group VISA algorithm for variable selection. Statistical Methods & Applications. 24, 41–60.