MATEMATIKA: MJIAM, 2020, Volume 36, Number 1, 43–49

Logistic Regression Ensemble (LORENS) Applied to Drug Discovery

T. Dwi Ary Widhianingsih*, Heri Kuswanto and Dedy Dwi Prastyo
Department of Statistics, Faculty of Mathematics, Computing and Data Sciences
Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya 60111, Indonesia
*Corresponding author: t.dwiary@outlook.com

Article history
Received: 16 April 2019
Received in revised form: 5 August 2019
Accepted: 14 October 2019
Published online: 1 April 2020

Abstract Logistic regression is one of the commonly used classification methods. It has some advantages, specifically related to hypothesis testing and its objective function. However, it also has some disadvantages in the case of high-dimensional data, such as multicolinearity, over-fitting, and a high computational burden. Ensemble-based classification methods have been proposed to overcome these problems. The logistic regression ensemble (LORENS) method is expected to improve the classification performance of basic logistic regression. In this paper, we apply it to the case of drug discovery with the objective of obtaining candidate compounds to protect the normal non-cancerous cells, which is considered to be a problem with a data-set of high dimensionality. The experimental results show that it performs well, with an accuracy of 69.41% and Area Under Curve (AUC) of 0.7306.

Keywords Drug Discovery; Ensemble; Logistic Regression; Radio-protection.

Mathematics Subject Classification 62-07.

1 Introduction

Classification is a multivariate method dealing with the partition of the training samples and allocation of new observations into certain classes or categories. The purpose is to obtain the optimal discriminant function that can separate the observations coming from different classes, or obtain a rule which can be used to determine the category of each new observation [1]. Among the methods used for classification, logistic regression is very popular due to its interpretability and its nice output, e.g. the probability that can assign the object to the appropriate class. Nevertheless, it may lead to poor performance when it is applied to high-dimensional data, e.g. a data-set with more variables than observations.

In [2], it is explained that there are four problems to be faced when applying logistic regression to high-dimensional data. First of all, is no unique solution to the estimation process using a small data-set. Secondly, there will be multicolinearity because of the highly correlated predictors. Thirdly is over-fitting, i.e. a model that has very good performance on the training set becomes significantly worse on the testing set. Over-fitting will happen when the model is complex, as the result of the implementation in a data-set with a large number of variables [3]. The last is the computational burden. Since there is no unique solution, the parameters should be estimated by a numerical approach, which consumes more time and space than an analytical solution. All these problems lead to the restricted applicability of logistic regression for high-dimensional data.

Approaches to overcome these problems with logistic regression have been proposed. Some of them apply the ensemble technique to train the logistic regression. The classic concept of an ensemble is to obtain a better classifier from a set of weak classifiers [4]. This method can obtain better classification results than a single weak classifier [5]. A recent development of logistic regression using the ensemble technique is logistic regression ensemble (LORENS) technique [6]. Briefly, LORENS does the training by applying logistic regression to a data-set which is basically a partition of the whole of the original data-set in a way that satisfies the implicit assumption of logistic regression, namely, $p < n$, where p is the number of variables and n is the number of samples.

LORENS has been applied to some cases, such as the prediction of consumer defection [7], the classification of gene expression for Alzheimer's disease [8], and the classification analysis of enzymes [9]. The data has been used for analysis using some machine learning methods, such as random forest, SVM [10, 11], K-Nearest Neighbor (KNN), and Extreme Gradient Bossting (XGB) Kimura. Feature selection for this data-set has also been done in [11] using the importance of each variable.

In the present paper, an analysis has been carried out using LORENS of drug discovery data, which is considered to be high-dimensional data. The observations in this data-set are drug-forming compounds and the variables are the characteristics of the cells that have been injected with the compounds. The objective of the analysis is to get a good classifier that can classify the compounds into high radio-protection (positive category) or low radio-protection (negative category).

2 Classification Methods

LORENS is one of the classification methods using the ensemble technique. It was proposed by Lim [6] to overcome the disadvantages of logistic regression when it is implemented on high-dimensional data. The idea of this approach is to train the logistic regression using a data-set that has a lower dimension than the original one. Therefore, feature selection is not necessary, since the training has already been carried out using some independent sub-spaces, which basically is a partition of the whole of the data-set. Thus, its implementation can be done efficiently and it will be very effective, because it is as if one step, the feature selection or extraction step, has been eliminated while still yielding a model with a good performance. Basically, LORENS has a different idea than the regularization approach in logistic regression which estimates the parameters along with the selection of the variables by adding a penalty function to the objective function, e.g. the Least Absolute Shrinkage and Selection Operator (LASSO) method [12, 13].

In the process of obtaining the model, LORENS splits the data into sub-spaces in such a way as to fulfill the primary assumption, $p < n$, needed for training the model using logistic regression. Suppose that Y_i is a dichotomous response variable with two possible values {0,1} and x_i is the $p \times 1$ vector of predictors for $i = 1, 2, \ldots, n$. Then the probability model of logistic regression with coefficients β_0 and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is given below.

$$P(Y_i|\boldsymbol{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta})}} \tag{1}$$

The estimation of β_0 and $\boldsymbol{\beta}$ using maximum likelihood is not in closed form [2], so an iterative or numerical method is used for it, such as the Newton–Raphson method [14]. Using this technique, the parameters will be estimated using the following equation.

$$\{\hat{\beta}_0, \hat{\boldsymbol{\beta}}\}^{(k)} = \{\hat{\beta}_0, \hat{\boldsymbol{\beta}}\}^{(k-1)} - \boldsymbol{H}^{-1}(\beta_0^{(k)}, \boldsymbol{\beta}^{(k)}) \nabla_l(\beta_0^{(k)}, \boldsymbol{\beta}^{(k)}) \tag{2}$$

Here, k is the iteration index in the estimation process. The vector $\nabla_l(\beta_0^{(k)}, \boldsymbol{\beta}^{(k)})$ is the gradient vector, which is the first derivative of the log-likelihood function, while $\boldsymbol{H}^{-1}(\beta_0^{(k)}, \boldsymbol{\beta}^{(k)})$ is the Hessian matrix for $l(\beta)$.

LORENS is constructed by using the Classification by Ensembles from Random Partitions (CERP) algorithm [15]. The basic concept of this technique is to combine weak classifiers to get a better model, while each classifier is built up from the model using a different set of variables. Let $\boldsymbol{\Theta}$ be the space of predictors in the whole data and $\boldsymbol{\theta}_m$ be the m-th subspace, for $m = 1, 2, \ldots, M$. The first step in this algorithm is to split the data into M sub-spaces: $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)$. The number of members of each subspace is selected so that they will be balanced or nearly the same. The prediction values or the instance probabilities of each model are combined thereafter using the average value. After that, the prediction class can be decided based on a threshold. Generally, this threshold is 0.5. However, this will not be reliable if the ratio between two classes is unbalanced. Therefore, the threshold can be calculated by using $\frac{1}{2}(\bar{y} + 0.5)$, where $\bar{y}$ denotes the ratio of the particular class or category.

To improve the performance of the classifier, LORENS generates multiple ensembles and the result will be calculated using majority voting. To avoid draws, it will be better if we generate an odd number of replicates; generally we use 11 ensembles.

Algorithm 1 Logistic Regression Ensemble

Require: Sampled data $\boldsymbol{X} \in \mathbb{R}^p$ and target data $\boldsymbol{y} \in \{-1, 1\}^n$, so that $\boldsymbol{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_n, y_n)\}$
Require: The number of subspaces m, threshold, and replication q
1: Randomly split $\boldsymbol{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_Q\}$
2: **for** $q = 1$ to Q **do**
3: **for** $m = 1$ to M **do**
4: Build the logistic regression model
5: Calculate $\hat{P}_m(Y|\boldsymbol{x})$ for D^{test}
6: **end for**
7: Calculate $\bar{\hat{P}}(Y|\boldsymbol{x}) = \frac{1}{m}\sum_{m=1}^{M} \hat{P}_m(Y|\boldsymbol{x})$
8: Determine $\hat{Y}$ with respect to the threshold
9: **end for**
10: Assign $\hat{\boldsymbol{Y}}$ using majority voting

3 Methodology

An analysis is conducted of the drug discovery data obtained from the experiments of [16]. The objective of this experiment is to get the compounds that can cover the normal non-cancerous cells, preventing them from being exposed to radiation. The expected result or positive category in this experiment is a high radio-protection effect of the compound. The number of samples (compounds) used in this analysis is 84 and the number of variables (characteristics of the cells) is 217. The experiment was done by injecting p53 inhibitor, then irradiating the normal cell with a dose of 10 Gy. After that, the compound is injected into the cell. The dose will be increased from 0 μM to 300 μM, then the death rate of these cells is calculated. The label of the data is determined according to the cell death rate. From all the experiments, the number of elements of the positive category in this data-set is 39, and there are 45 negatives, so it can still be considered as balanced data. The variables in the data are basically the characteristics of the cells that have been treated by some compounds and radiation. The features are related to some categories:

- hydrophobicity: Alogp98, AlogP_MR, logD, Apol,...
- structure: Br_Count, C_Count, Cl_Count,...
- size and molecular: Molecular_Mass, Molecular_Solubility,...
- others: ES_Count_aaaC, QED, QED_Alerts,...

The classification analysis in this paper was done by using the R programming language. The objective of the classification analysis here is to get the settings for the the best performance of LORENS. This has been sought by considering the following possible numbers of partitions: $5, 10, 15, \ldots, 70$. The evaluation criteria that have been used to calculate the performance of the model are accuracy and Area Under Curve (AUC). The accuracy can be simply calculated by counting the number of correct classifications, then dividing it by the total number of elements of the data-set. The AUC is calculated based on the Receiver Operating Characteristic (ROC) curve; its result can be interpreted as the average power of the test on default or non-default corresponding to all possible cut-off values [13]. In each setting, the training was replicated 10 times to overcome any uncertainty in the results because of the random step in the algorithm. Then, the average value of the performance was calculated as the result.

4 Experimental Results

Before discussing the experimental results, some statistics about the data will be shown. First of all is the index of the balancing ratio. Using the formula in [17], the imbalance ratio of the data used in this paper is 1.0103. In [17], it is explained that if the data have an index of 1, this means that it constitutes a completely balanced data-set, i.e. each class has the same number of instances as all the other classes. Thus, because the calculation of the imbalance ratio of the data used in this paper has no significant difference from the criterion for being balanced, we can say that the drug discovery data constitutes a balanced data-set.

Table 1: The Predictions of LORENS for the Testing Set

Id of the compounds	Actual class	Prediction class	Probability
3	-1	1	0.9435
11	-1	-1	0.0727
38	-1	-1	0.3086
46	-1	1	0.4908
47	-1	-1	0.1636
49	-1	1	0.5720
52	-1	-1	0.2006
66	-1	1	0.8162
78	-1	-1	0.4731
7	1	1	0.9034
20	1	-1	0.4545
30	1	-1	0.3065
31	1	1	0.5281
43	1	-1	0.4349
54	1	-1	0.4545
61	1	1	0.6364
84	1	1	0.6182

Secondly, the threshold that will be used for assigning the class prediction in the analysis will be presented. In Section 2, it was already mentioned that instead of using the balance threshold 0.5, we prefer to use the adjusted threshold using $\frac{1}{2}(\bar{y}+0.5)$. Suppose that $y = 1$, then $\bar{y}$ is the number of samples with $y = 1$ divided by the total number of elements in the whole data-set. So, we will get $\bar{y} = 0.4643$. Then, the adjusted threshold for this data-set is $\frac{1}{2}(0.4643 + 0.5) = 0.4822$. Using this threshold, we can later assign the prediction class as follows: if $\hat{P}(Y|\boldsymbol{X}) \geq 0.4822$, the sample will be classified as positive, and in the opposite case, as a member of the negative class.

As an illustration of an analysis using LORENS, the experimental procedure has been applied to 5 partitions and 11 ensembles; the number of variables for each partition or subspace is 44, 43, 43, 43 and 44. The decision to use at least 5 partitions is because we want to get the minimal number of the sub-spaces such that the requirements for applying logistic regression can be fulfilled. Then, using an 80:20 split between the training set and the testing set, the number of samples for the training set is 67, with 17 compounds for the testing set. Now for each sub-space, there are more samples than the variables, so we can directly apply logistic regression.

Table 2: Confusion Matrix for LORENS

		Actual	
		Positive	Negative
Prediction	Positive	4	4
	Negative	4	5

Using these settings, we get the following results. Table 1 shows the results of the testing prediction. The prediction probability denotes the tendency of the test data to be classified into the positive category. If this probability is greater than the optimal threshold, which is 0.48134, the prediction category is assigned to be 1, but if it is less than the optimal threshold, the prediction category is -1.

From the results in Table 1, we can make the confusion matrix, as shown in Table 2. Then we can calculate the accuracy from this table to be $\frac{4+5}{17} = 0.5294$. The AUC of this result is obtained by calculating the area under the ROC curve. Using this criterion, the performance of the model is 0.5972. The performance of the

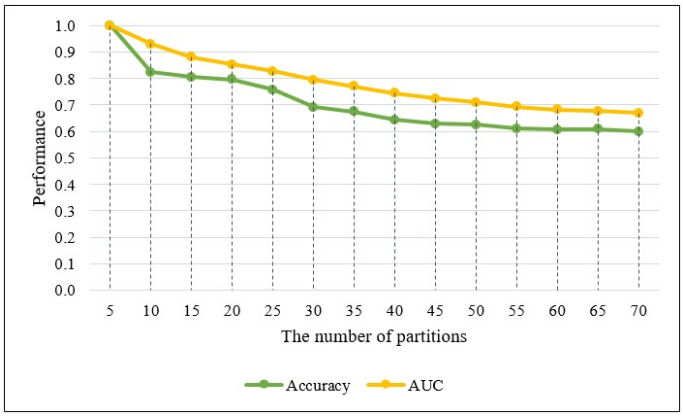

Figure 1: The Effect of Increasing the Number of Partitions

model is really low, because the developed model used only part of the whole data-set. So we need to do the classification for the rest of the sub-spaces and apply the LORENS procedure to get a better performance. And because of the random step in the splitting of the variables into partitions, this result will be variable, so we iterate 10 times to obtain a better result.

Table 3: ROC of the Result

The number of partitions	Accuracy	AUC
5	0.5882	0.6292
10	0.5882	0.6986
15	0.5882	0.7181
20	0.6176	0.7264
25	0.6941	**0.7306**
30	0.7059	0.7194
35	0.7059	0.7097
40	0.7059	0.7208
45	0.7059	0.7292
50	0.7059	0.7250
55	0.7059	0.7208
60	**0.7118**	0.7111
65	0.7059	0.7069
70	0.7000	0.7014

The overall performance of this implementation of LORENS for all possible partition settings for the training set is shown in Figure 1. The graph reveals that the number of partitions can influence the performance of the model. At first, using 5 partitions, the accuracy and the AUC reach 100%. Then the performance decreases as the number of partitions increases. The decreasing accuracy is more obvious than the AUC when the number of partitions used is increased to 10. Then the decrease is relatively stable until reaching 60% when using 70 partitions. However, the optimal LORENS model is not assigned based on this graph. It will be chosen using the performance on the testing set, as shown in Table 3.

To choose the best number of partitions, the results of LORENS for some scenarios are given in Table 3. It shows that the accuracy and AUC are not too different. The best performance is apparently obtained from 25 and 60 partitions. The AUC has the best value when using 25 partitions, but the best accuracy is when using 60 partitions. In detail, we can see from Table 3 that if we calculate the difference between those criteria in both scenarios (both numbers of partitions), then AUC has the greatest difference. So, in this analysis we assign

the best partition based on the AUC. Consequently, we choose 25 partitions as the best number of sub-spaces. Getting an AUC of about 73% means that the classifier is good at dividing the data-set into two classes, high and low radio-protection.

Table 4: Performance Comparison

Method	Feature Selection	AUC
Random Forest [10]	-	0.5810
SVM [10]	-	0.4110
LORENS	-	**0.7306**
Random Forest [11]	10% variables	0.6190
Random Forest [18]	15% variables	0.6880
SVM [18]	15% variables	0.6280
SVM [11]	5% variables	0.6460
KNN [18]	5% variables	**0.7570**
XGB [18]	5% variables	0.6350

A comparison between LORENS and other methods that have been applied to the same data is shown in Table 4. LORENS is compared to random forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Extreme Gradient Boosting (XGB). Among these methods, three of them were implemented without feature selection, viz. random forest, SVM and LORENS. In the feature selection column, 10% of the variables for random forest means that the model was implemented using the first 10% of the variables, in order of decreasing importance. In the original papers, the experiments are done using varied configurations, but in the present paper only the results from the best configuration are shown. From Table 4, we see that among all the methods, KNN can give us the best performance. However, this method uses feature selection. Compared to KNN, LORENS has a lower performance, but it is still comparable to the other methods that do not use feature selection.

5 Conclusion

This paper has applied successfully the logistic regression ensemble (LORENS) technique to drug discovery data, which is considered to a high-dimensional data-set. Basically, LORENS is an extension of logistic regression using an ensemble technique, so it can be applied to high-dimensional data. The ensemble technique employed in LORENS is the classic ensemble algorithm that learns the model using a partition of the data, then combines them to get the ensemble result. Overall, LORENS obtains a good result. The best model is produced by using 25 partitions. The performance is about 0.7306 in terms of the AUC, with an accuracy of 69.41%. Compared to the previous methods of analysis, LORENS has a comparable performance to the other methods that do not use feature selection.

References

[1] Johnson, R. A. and Wichern, D. W. *Applied Multivariate Statistical Analysis*. 6th. Edition. New Jersey: Prentice-Hall. 2007.

[2] Bielza, C., Robles, V., and Larranaga, P. Regularized logistic regression without a penalty term: an application to cancer classification with micro-array data. *Expert Systems with Applications*. 2011. 38(5): 5110–5118.

[3] Romero, C., Ventura, S., Pechenizky, M. and Baker, R. *Handbook of Educational Data Mining*. Boca Raton: CRC Press. 2011.

[4] Dietterich, T. G. Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*. 2000. 1–15.

[5] Rokach, L. Ensemble-based classifier. *Artificial INtelligence Review*. 2010. 33(1–2): 1–39.

[6] Lim, N. *Classification by Ensemble from Random Partitions using Logistic Regression Models*. Stony Brook University: Ph.D. Thesis. 2007.

[7] Kuswanto, H., Asfihani, A., Sarumaha, Y. and Ohwada, H. Logistic regression ensemble for predicting customer defection with very large sample size. *Procedia Computer Science*. 2015. 72: 86–93.

[8] Kuswanto, H. and Werdhana, R. W. Logistic regression ensemble to classify alzheimer gene expression. *International Conference on Smart Cities, Automation, and Intelligent Computing Systems (ICON-SONICS)* 2017. 36–41.

[9] Kuswanto, H., Melasasi, J. and Ohwada, H. Enzyme classification on DUD-E database using logistic regression ensemble (Lorens). *Innovative Computing, Optimization, and Its Applicaitons*. 2018. 93–109.

[10] Matsumoto, A., Ito, T., Nishi, Y., Teraoka, T., Aoki, S. and Ohwada, H. Prediction of radio protectors targeting p53 for bioinformatics and biomedicine (BIBM). *IEEE International Conference*. 2015. 1725–1727.

[11] Matsumoto, A., Aoki, S. and Ohwada, H. Comparison of random forest and SVM for raw data in drug discovery: prediction of radiation protection and toxicity case study. *International Journal of Machine Learning and Computing*. 2016. 6(2): 145.

[12] Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society*. 2005. 67(2):301–320.

[13] Haerdle, W. and Prastyo, D. D. Embedded predictor selection for default risk calculation: a southeast asian industry study. In D. L. K. Chuen and G. N. Gregorious (Eds). *Handbook of Asian Finance, Vol.1, Financial Market and Sovereign Wealth Fund*. San Diego: Academic Press. 2014. 131–148.

[14] Lee, S. I., Lee, H., Abbeel, P. and Ng, A. Y. Efficient L_1 regualrized logistic regression. *American Association for Artificial Intelligence*. 2006. 6:401–408.

[15] Ahn, H., Moon, H., Fazzari, M. J., Lim, N., Chen, J. J. and Kodell, R. L. Classification by ensembles from random partitions of high-dimensional data. *Computational Statistics and Data Analysis*. 2007. 51(12): 6166–6179.

[16] Ariyasu, S., Sawa, A., Morita, A., Hanaya, K., Hoshi, M., Takahashi, I., Wang, B. and Aoki, S. Design and synthesis of 8-hydroxyquinoline-based radioprotective agents. *Bioorganic and Medical Chemistry*. 2014. 22(15): 3891–3905.

[17] Tanwani, A. K. and Farooq, M. The role of biomedical dataset in classification. *Conference on Artificial Intelligence in Medicine*. 2009. 370–374.

[18] Kimura, M., Aoki, S. and Ohwada, H. Predicting radiation protection and toxicity pf p53 targetting radioprotectors using machine learning. *Computational INtelligence in Bioinformatics and Computational Biology (CIBCB)* 2016. 1–6.