Available online at: https://ijact.in

Date of Submission	29/02/2020
Date of Acceptance	26/03/2020
Date of Publication	30/04/2020
Page numbers	3624-3628 (5 Pages)

Cite This Paper: Han Gao, Pei SF, Lea Tien Tay, Heng Chin Low. Logistic regression techniques based on different sample sizes in landslide susceptibility assessment: which performs better?, 9(4), COMPUSOFT, An International Journal of Advanced Computer Technology. PP. 3624-3628.

ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

LOGISTIC REGRESSION TECHNIQUES BASED ON DIFFERENT SAMPLE SIZES IN LANDSLIDE SUSCEPTIBILITY ASSESSMENT: WHICH PERFORMS BETTER?

Han Gao[1], Pei Shan Fam[1*], Lea Tien Tay[2], Heng Chin Low[3]

[1]School of Mathematical Sciences, Universiti Sains Malaysia, 11800, USM, Penang
[2]School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300, Nibong Tebal, Penang
[3]Research and Innovation Unit, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
alyssagaohan@gmail.com, fpeishan@usm.my, tay@usm.my, hclow@usm.my

Abstract: The main objective of this paper is to compare the landslide spatial prediction performance of logistic regression (LR) with different regularization methods, namely, Lasso LR and Ridge LR. Three types of training datasets with different sample sizes of 40,000, 4,000 and 400 are used to train and validate the models. ROC curves are used to evaluate the models' performance. The results show that Lasso and Ridge LR models have comparative performance compared to the ordinary LR models based on the AUC values, which indicates that there are no redundant input features to remove from the models for the available data in this work to some degree. The penalty terms play a negligible role in the LR models trained with the three types of datasets. Lasso LR has a better performance than ridge LR, which may be due to that the L_1 penalized parameter which can be exactly equal to zero. According to the AUC values, the group of models trained and validated using the dataset of 20,000 samples outperform the other two groups.

Keywords: landslide susceptibility; penalty term; machine learning; ridge logistic regression; lasso logistic regression; receiver operating characteristic;

I. INTRODUCTION

Landslide is a natural disaster which can cause severe damage to life and property. Penang Island is a popular landslide-prone area in Malaysia during the monsoon season. Landslide spatial prediction is of great importance to landslide mitigation and management. Landslide susceptibility analysis (LSA) is a commonly used way to visualize the landslide occurrence distribution using various methods, such as frequency ratio [1, 2], fuzzy logic [3-5], support vector machine [4, 6], decision tree [7, 8].Other review papers on the methodology of LSA is provided in [9]. Logistic regression (LR) is a simple but powerful binary classification tool, which is also a popular machine learning (ML) algorithm applied in various fields. In recent decades, LR techniques are widely used in landslide susceptibility research area [10-13]. However, few researchers applied lasso and ridge LR into this area, since they are usually unavailable in most of the statistical software [1]. Therefore, the objective of this paper is to compare the spatial prediction performance of ordinary LR, lasso LR as well as ridge LR based on the datasets with different sample sizes in Penang Island, Malaysia.

The rest of the paper is organized as follows. Section 2 provides a detailed description of the methodology used in this study, namely, the ordinary LR as well as the lasso and ridge LR. The experimental design is displayed in Section

3, including the sampling procedure for training and validation datasets as well as the testing datasets. The results and discussions of the paper are given in Section 4 followed by a brief conclusion.

II. Methodology

LR model, developed by [14] , is a widely used classification rather than a regression algorithm. The independent variables $\mathbf{X}=(x_1,x_2,...,x_n)^{\mathrm{T}}$ can be continuous or discrete and one or more, while the dependent variable y is dichotomous. The goal of LR is to find the best fitting model to describe the relationship between the dichotomous dependent variable and a set of independent variables **X**. Let

$$\sigma(z)=P(y=1\,|\,\mathbf{X})=1-P(y=0\,|\,\mathbf{X})\,. \qquad (1)$$

where $P(y=1\,|\,\mathbf{X})$ and $P(y=0\,|\,\mathbf{X})$ denote the event occurrence and non-occurrence probability given the input feature matrix **X**, which is composed by the input feature x_i, for i=1, 2,..., n. In this research, they denote the posterior probability of landslide occurrence and non-occurrence of a pixel data, respectively. The simplest form of the ordinary logistic function [2, 15] is given by:

$$\sigma(z)=1/(1+e^{-z})\,,\quad (2)$$

which is also called a sigmoid function because it is an S-shaped curve (see Figure 1).

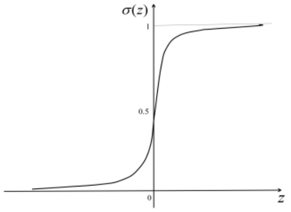


Figure 1. The LR model

In Equation (1), z is usually considered as a linear function which can be expressed as:

$$\begin{aligned} z &= \omega_0+\omega_1x_1+\omega_2x_2+...+\omega_nx_n \\ &= \boldsymbol{\omega}^{\mathrm{T}}\mathbf{X}+\omega_0 \end{aligned}\,,\quad (3)$$

where ω_0 is the intercept of the linear model and ω_i, $i=1,2,...,n$ denote the coefficients of the input variables and $\boldsymbol{\omega}=(\omega_0,\omega_1,\omega_2,...,\omega_n)^{\mathrm{T}}$.

Given the training dataset $D=\{(\mathbf{X}_1,y_1),(\mathbf{X}_2,y_2),...,(\mathbf{X}_m,y_m)\}$, the working process of an LR model in this work is displayed in Figure 2. A total of eleven landslide influencing factors, i.e., *aspect*, *curvature*, *geology*, *soil type*, *landuse*, *rainfall*, *height*, *distance to drainage*, *distance to fault*, *distance to road*, *slope*, are considered as the input features. The output feature is landslide occurrence or non-occurrence after being processed by the LR models.

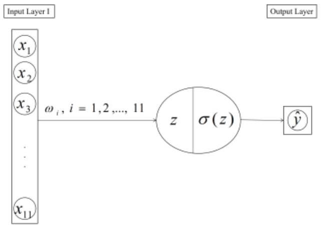


Figure 2. The LR algorithm

A. Ordinary Logistic Regression

The mathematical expression of LR algorithm for one example $\mathbf{x}^{(j)}$ is shown below:

$$z^{(j)}=\omega_0+\boldsymbol{\omega}^T\mathbf{x}^{(j)}\,,\qquad (4)$$

$$\sigma(z^{(j)})=1/(1+e^{-z^{(j)}})\,,\qquad (5)$$

where $\boldsymbol{\omega}^T$ denotes the transposition of the variables coefficient vector; $\hat{y}^{(j)}$, $j=1,2,...,m$ denotes the output value of the LR model and m denotes the number of samples. Based on Equation (5), each $\hat{y}^{(j)}$ can be regarded as the probability of landslide occurrence.

The least squared error (LSE) method is a widely accepted loss function for linear regression models. For LR models, however, LSE is unqualified to be regarded as a loss function. Since it will result in a non-convex curve with more than one local minimums [16]. Therefore, the cross-entropy loss function is considered for LR models, which can be expressed in Equation (6):

$$L(y^{(j)},\hat{y}^{(j)})=-[y^{(j)}\ln(\hat{y}^{(j)})+(1-y^{(j)})\ln(1-\hat{y}^{(j)})]\,,\quad (6)$$

where $y^{(j)}$ and $\hat{y}^{(j)}$ denote the true sample label and the predicted label, respectively. The cost function for all training examples is shown in Equation (7):

$$J(y^{(j)},\hat{y}^{(j)})=\frac{1}{m}\sum_{j=1}^{m}L(y^{(j)},\hat{y}^{(j)})\quad (7)$$

In order to avoid over-fitting when training models using datasets with a large number of input features but relatively small sample size, the regularization techniques are of great importance. Lasso and ridge LR are considered in this study.

B. Lasso Logistic Regression

Lasso stands for least absolute shrinkage and selection operator, which was originally proposed for linear regression models by [17]. It is a widely used variable selection and shrinkage technique. Lasso LR can be obtained by minimizing the cross entropy cost function with an L_1 penalized parameter applied to all variable coefficients except the intercept. The mathematical expression is shown in Equation (8):

$$J(y^{(j)},\hat{y}^{(j)})_{L_1}=\frac{1}{m}\sum_{j=1}^{m}L(y^{(j)},\hat{y}^{(j)})+C_1\sum_{i=1}^{n}|\,\omega_i\,|\,,\quad (8)$$

where C_1 denotes the regularization parameter of L_1.

C. Ridge Logistic Regression

Ridge regression was proposed by [18] and can be obtained by applying an L_2 penalized term to the cost function of a linear regression model. Compared to ridge regression, ridge LR can be obtained by maximizing the likelihood function with an L_2 penalized term applied to all variable coefficients except the intercept. The mathematical expression can be shown in Equation (9):

$$J(y^{(j)},\hat{y}^{(j)})_{L_2}=\frac{1}{m}\sum_{j=1}^{m}L(y^{(j)},\hat{y}^{(j)})+C_2\sum_{i=1}^{n}\omega_i^2\ , \quad (9)$$

where C_2 denotes the regularization parameter of L_2.

Compared to L_2, the advantage for L_1 is that L_1 penalized parameter is prone to get more sparse resolutions, which can reduce the difficulty of prediction for the LR models [16]. In this work, learning curve is used to determine the optimal C_1 and C_2 values in lasso and ridge LR, respectively. The range of the values is set from 0.0 to 2.0 and the interval is set to 0.05. Table 1 displays the optimal values of C_1 and C_2.

Table 1. The optimal *C* values for Lasso and Ridge LR

Models	Penalized parameter	Sample size of lasso and ridge LR		
		20000-20000	2000-2000	200-200
Lasso LR	C_1	0.05	0.55	0.25
Ridge LR	C_2	0.10	0.90	0.05

D. Performance Measure

The accuracy values for training, validation and testing datasets are considered to evaluate the model's performance. Since LR model's output is a probability value from 0 to 1, a threshold value is needed when transforming the probability into a binary label value. The selection of the threshold highly affects the results of performance measure. Therefore, a more robust evaluation measure is needed. Receiver operating characteristic (ROC) curve is a proper substitute to accuracy for model evaluation.

ROC curves are two-dimensional graphs depicting the performance of the classifiers [19]. The *x*-axis and *y*-axis denote the false positive rate and true positive rate, respectively. The area under the curve (AUC) is a commonly used index to compare the model's performance quantitatively.

III. Experimental Design

The total number of pixels are 3,004,631, including 20,245 landslide pixels and 2,984,386 non-landslide pixels. The imbalance ratio (IR) of the whole dataset is around 1:150. By using random sampling method, three types of datasets are selected from landslide (positive) and non-landslide (negative) pixels, independently. The total number of samples for the three types of datasets with balanced sample ratio are 40,000, 4,000 and 400, respectively. Each of the three datasets are randomly split into 90% and 10% for model training and validation, respectively. In order to test the prediction power of the LR models trained using dataset with relatively small sample size, the 2,000 samples are randomly selected from the 20,000 samples, and the 200 samples are randomly selected from the 2,000 samples. In other words, the models trained using a relatively large dataset consider more information than the models trained using a relatively small dataset. A testing dataset with 200 samples, including 100 positive samples and 100 negative samples, is selected from the rest of the whole dataset using random sampling method. Figure 3 displays the sampling procedure. The magnitude of the ellipse is unrelated to the number of samples.

In order to be clearer about the names of datasets and models, let us define the ordinary LR models trained and validated using dataset with 20,000 negative and 20,000 negative samples as Ordinary_20000. Similarly, LassoLR_2000 and RidgeLR_200 denote the lasso LR model trained using 2,000 negative and 2,000 positive samples and the ridge LR models trained using 200 negative and 200 positive samples, respectively.

During the model testing process, the testing dataset is divided into two sub-testing datasets each with 100 positive samples and 100 negative samples, respectively. The whole dataset can be considered as a special testing dataset to some degree, since most of the samples in the dataset are unknown to the LR models.

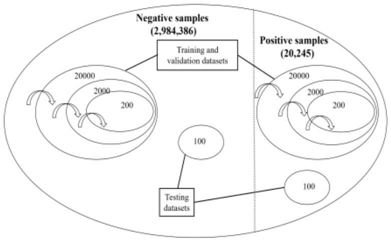


Figure 3. Random sampling procedure

IV. Results and Discussions

All the experiments are conducted in Windows 10 serverwith an Intel Core i5 2.40GHz processor. All the models are trained in Python 3.7.0 based mainly on sklearn, numpy and pandas libraries. SPSS 20.0 is used for some data preparation. ArcGIS is used to produce landslide susceptibility maps.Table 2 displays the training, validation and testing accuracy values for three types of models trained using three different datasets.

From the results, all the three types of LR models trained and validated using the datasets with 40,000 and 4,000 samples are not over-fitted or under-fitted based on the training and validation accuracy values. However, the models trained using the dataset with only 400 samples show severe under-fitting according to the big difference between training and validation accuracy. A probable explanation is that the variance of the training dataset (90%) is bigger than that of the validation dataset (10%) due to the small sample size and uneven split. Although all the models are trained and validated using balanced datasets, the prediction accuracy values for the dataset with only positive samples are always higher than the values for the datasets with only negative samples. The reason may be for landslide hazard research, the landslides may occur in the places without landslide occurrence previously.

Table 2. The accuracy values for training, validation and testing datasets

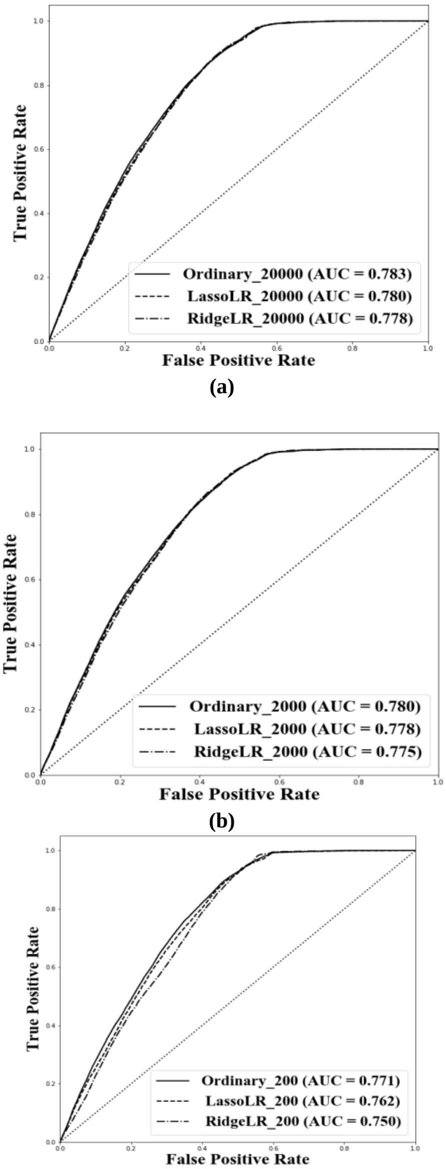
Sample size	Model type	Accuracy (%)						
		Train	Validate	Test	Test	Test (+)	Test	Overall
20000-20000	Ordinary	72.15	71.98	69.50	55.00	62.00	58.50	64.22
	Lasso	72.04	72.08	70.00	53.00	67.00	60.00	62.98
	Ridge	72.18	72.23	68.00	52.00	73.00	62.50	62.62
2000-2000	Ordinary	72.72	74.25	69.00	52.00	65.00	58.50	64.71
	Lasso	72.61	75.50	68.00	51.00	67.00	58.00	63.55
	Ridge	73.36	74.75	68.50	50.00	73.00	61.50	63.11
200-200	Ordinary	72.78	90.00	65.50	56.00	68.00	62.00	66.76
	Lasso	72.78	92.50	64.50	56.00	73.00	64.50	61.70
	Ridge	71.67	92.50	66.00	52.00	74.00	63.00	61.58

*Test(-) and Test(+) denote the test dataset with the number of 100 negative and positive samples, respectively.

For the places where landslides occurred, furthermore, it may become a landslide-free area in the near future. The prediction accuracy values for the testing dataset with 200 samples should be the arithmetic mean of the two accuracy values of the two testing datasets only with one type of samples. For example, the value of 58.50% is obtained by averaging the two values of 55.00% and 62.00%, as shown in the box of Table 2. But the results are not like that. The reason for such a phenomenon is that the models trained and validated using the training data with balanced sample ratio cannot predict the testing data with only one type of samples. In ML area, a fundamental assumption is that the training and testing samples come from the same independent identical distribution (i.i.d.). The big difference of samples ratio in training and testing datasets leads to the results. Figure 4 displays the ROC curves for all the three types of LR models based on three datasets. Table 3 displays the AUC values for the three types of LR models. Based on the AUC values, the group of models trained and validated using the dataset with 20,000 samples show the best performance than the other two groups. The model Ordinary_20000 with the AUC value of 0.783 outperforms LassoLR_20000 and RidgeLR_20000. The landslide susceptibility map produced by Ordinary_20000 is shown in Figure 5. The degree of hazardous for landslide occurrence is decreasing from the colour Red to Green to Blue until White. The predicted hazardous area mainly locates in the middle mountainous area, which is fitting to the real landslide occurrence situation to a high degree.

Table 3. The AUC values for the types of LR models

Model type	AUC values		
	20000-20000	2000-2000	200-200
Ordinary	**0.783**	0.780	0.771
Lasso	0.780	0.778	0.762
Ridge	0.778	0.775	0.750



(a)

(b)

(c)

Figure 4. The ROC curves for different types of LR models trained using **(a)** 20000 **(b)** 2000 and **(c)** 200 samples

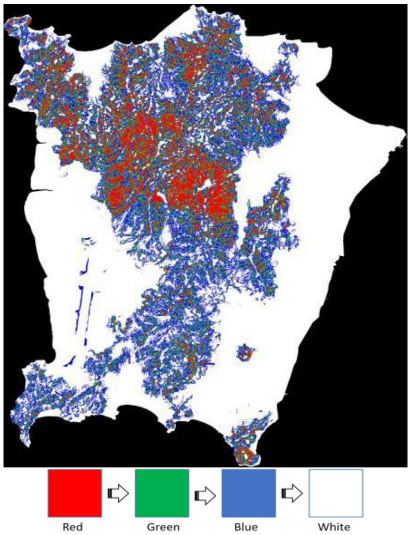


Figure 5. Landslide susceptibility map producedby Ordinary_20000

V. Conclusions

In this work, three types of LR algorithms, namely, ordinary, lasso and ridge LR, are applied in landslide spatial prediction research in Penang Island, Malaysia. Three types of datasets are generated to train and validate the LR models. An original testing dataset with 200 samples is used to test the models. Two sub-testing datasets are derived from the original testing dataset, each with 100 negative and 100 positive samples, respectively. The whole dataset is also considered as a testing dataset, since most of the samples are new to the trained models. The testing accuracy values for sub-testing dataset with 100 positive samples are all higher than the dataset with 100 negative samples.

Overall, the lasso and ridge LR models have comparative performance compared to the ordinary LR models based on the AUC values, which indicates that all the eleven input features contribute to the models for the available data in this work to some degree. The penalty terms of L_1 and L_2 play a negligible role in the LR models. Furthermore, lasso LR has a better performance than ridge LR, which may be due to the L_1 penalized parameter which can be exactly equal to zero.

Acknowledgement

This research work was supported by RUI Grant (1001/PMATHS/8011093) and from Universiti Sains Malaysia.

References

[1] J. M. Pereira, M. Basto, A. F. Silva, The logistic lasso and ridge regression in predicting corporate failure, Procedia Economics and Finance, Vol. 39, pp. 634-641, 2016.

[2] B. Pradhan, S. Lee, Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models, Environmental Earth Sciences, Vol. 60, No. 5, pp. 1037-1054, 2010.

[3] B. Pradhan, Landslide susceptibility mapping of acatchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches, Journal of the Indian Society of Remote Sensing, Vol. 38, No. 2, pp. 301-320, 2010.

[4] B. Pradhan, A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS, Computers & Geosciences, Vol. 5, pp. 350-365, 2013.

[5] H. R. Pourghasemi, B. Pradhan, C. Gokceoglu, Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran, Natural Hazards, Vol. 63, No. 2, pp. 965-996, 2012.

[6] X. Yao, L. Tham, F. Dai, Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China, Geomorphology, Vol. 101, No. 4, pp. 572-582, 2008.

[7] P. Tsangaratos, I. Ilia, Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece, Landslides, Vol. 13, No. 2, pp. 305-320, 2016.

[8] Y. K. Yeon, J. G. Han, K. H. Ryu, Landslide susceptibility mapping in Injae, Korea, using a decision tree, Engineering Geology, Vol. 116, No. 3-4, pp. 274-283, 2010.

[9] H. Gao, P. S. Fam, L. T. Tay, H. C. Low, An overview and comparison on recent landslide susceptibility mapping methods, Disaster Advances, Vol. 12, No. 12, pp. 46-64, 2019.

[10] L. Ayalew, H. Yamagishi, The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan, Geomorphology, Vol. 65, No. 12, pp. 15-31, 2005.

[11] S. Lee, Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data, International Journal of Remote Sensing, Vol. 26, No. 7, pp. 1477-1491, 2005.

[12] S. Lee, B. Pradhan, Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models, Landslides, Vol. 4, No. 1, pp. 33-41, 2007.

[13] G. C. Ohlmacher, J. C. Davis, Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA, Engineering Geology, Vol. 69, No. 3-4, pp. 331-343, 2003.

[14] D. R. Cox, The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), Vol. 20, No. 2, pp. 215-232, 1958.

[15] M. L. Süzen, V. Doyuran, A comparison of the GIS based landslide susceptibility assessment methods: multivariate versus bivariate, Environmental Geology, Vol. 45, No. 5, pp. 665-679, 2004.

[16] Z. H. Zhou, Machine Learning, Tinghua University Press, 2016 (Chinese).

[17] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological), Vol. 58, No. 1, pp. 267-288, 1996.

[18] A. E. Hoerl, R. W. Kennard, Ridge Regression: Applications to Nonorthogonal Problems, Technometrics, Vol. 12, No. 1, pp. 69-82, 1970.

[19] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Vol. 27, No. 8, pp. 861-874, 2006.