Landslide susceptibility mapping using machine learning for Wenchuan County, Sichuan province, China

Xin Yang^{1,2}, Rui Liu^{1,2*}, Luyao Li^{1,2}, Mei Yang^{1,2} and Yuantao Yang^{1,2}

¹ College of Geophysics, Chengdu University of Technology, Sichuan 610059, China

² Key Lab of Earth Exploration and Information technique of Ministry Education, Chengdu University of Technology, Sichuan 610059, China

Abstract: Landslide susceptibility mapping is a method used to assess the probability and spatial distribution of landslide occurrences. Machine learning methods have been widely used in landslide susceptibility in recent years. In this paper, six popular machine learning algorithms namely logistic regression, multi-layer perceptron, random forests, support vector machine, Adaboost, and gradient boosted decision tree were leveraged to construct landslide susceptibility models with a total of 1365 landslide points and 14 predisposing factors. Subsequently, the landslide susceptibility maps (LSM) were generated by the trained models. LSM shows the main landslide zone is concentrated in the southeastern area of Wenchuan County. The result of ROC curve analysis shows that all models fitted the training datasets and achieved satisfactory results on validation datasets. The results of this paper reveal that machine learning methods are feasible to build robust landslide susceptibility models.

1 Introduction

Landslides, as one of the natural disasters, thread human society around the world. Assessment of landslide susceptibility is one of practical methods to manage landslide, which mainly includes qualitative and quantitative methods[1]. Among them, machine learning methods have received a lot of attention from researchers and performed satisfactory results[2]. Different from the past, this paper used six popular machine learning algorithms to map the landslide susceptibility. The datasets contain the landslides triggered by the Great Wenchuan Earthquake were leveraged for model construction. Final, we used the trained models to predict

each pixel and map the susceptibility of Wenchuan County, a landslide-prone area.

2 Materials

2.1 Study area

Wenchuan County is located in the eastern margin of Qinghai Tibet Plateau, between longitude 102°51'E and 103°44'E, and latitude 30°45'N and 31°43'N. The area is approximately 173.184 km². A Ms 8.0 earthquake occurred here and triggered many landslides.

Fig 1. The study area, Wenchuan County

*Corresponding author's e-mail: lr@cdut.edu.cn

2.2 Landslide inventory and predisposing factors

For landslide susceptibility mapping, it is essentially a binary-classification approach[3]. So, a dataset with “positive sample” and “negative sample” is the basis for modeling. In this study, a total of 1365 landslide points were recorded by visual interpretation as well as filed investigation, after the Great Wenchuan Earthquake. Meanwhile, the 1365 non-landslide points were generated randomly from the landslide-free area. Subsequently, the all data were split into training (70%) and validation (30%) datasets.

Having the landslide inventory, the next step is to choose a proper combination of landslide predisposing factors according to the environment and the historical research in the study area[4]. In this study, 14 conditioning factors including Normalized difference vegetation index (NDIV), Landuse, Relief, Curvature, Plan curvature, Aspect, Slope, Stream power index (SPI), Distance to rivers, Distance to faults, Distance to roads, Topographic wetness index (TWI), PGA, and elevation.

3 Methods

3.1 Landslide susceptibility models

In this study, six machine learning algorithms was used to construct landslide susceptibility models. The outputs of models are considered as the probability of landslide occurrence of each pixel within the study area. The values of outputs were set to 0 (landslide) or 1 (non-landslide).

3.1.1 Logistic regression (LR)

Logistic regression, one of the mostly used regression and classification algorithms, leverages a combination of independent conditioning variables to describe the relationship between input and target. It shows a satisfaction performance with relatively little computer cost.

3.1.2 Multi-layer perceptron (MLP)

Multi-layer perceptron is a neural network that includes input, hidden, and output layers. It is different from logistic regression, the hidden layer can contain one or more non-linear activation function which can improve the capability of solving the complex problem[5]. With the Backpropagation algorithm, the weight between different layers can alter until yielding a model that describes the relationship between the input and output layers. In this study, the nodes in input layer corresponds to the number of landslide predisposing factors.

3.1.3 Random forests (RF)

Random forests method, proposed by Breiman, has been used in a wide range of landslide susceptibility mapping. As an ensemble-learning technique, RF combines independent decision trees to learn the pattern of datasets and then vote the best result. The bootstrap sampling allows the RF more resistant to overfitting compared to other tree-based algorithms.

3.1.4 Support vector machine (SVM)

Support vector machine, a supervised algorithm for regression and classification, was proposed in 1996 by Vapnik. SVM leverages the kernel functions to expand the linear space into a non-linear space. The commonly used kernel functions are linear kernel, polynomial kernel, sigmoid kernel, and radial basis function (RBF). In this study, the kernel function of SVM was set to RBF.

3.1.5 AdaBoost and Gradient Boosted Decision Trees (GBDT)

AdaBoost, short for “Adaptive Boosting”, is an ensemble classification technique proposed by Freund and Schapire. As a boosting algorithm, AdaBoost constructs different weak classifiers with the datasets, and then assembles these to form a stronger classifier. Besides, Gradient Boosted Decision Trees, another boosting algorithm, is also used for landslide susceptibility modeling.

3.2 Model performance evaluation

Receiver Operating Characteristic (ROC) curve is an important indicator for evaluating the performance the of the classification model, especially the binary-model. The axis-X and axis-Y are 1-specificity and sensitivity, respectively[6]. To quantify the ROC and compare the performance of each model we used, Area under the ROC Curve (AUC) is introduced.

4 Results and discussions

4.1 Landslide susceptibility mapping

Mapping the landslide susceptibility is aimed to calculate the probability of landslide susceptibility in the study area. The trained models were used to predict each pixel and obtained the outputs that represent the probability. Then, the outputs were categorized into five group namely very low, low, moderate, high, and very high. As shown in Fig. 2, the risk area is concentrated in the southeastern part of Wenchuan County, especially near the epicenter of the Great Wenchuan Earthquake.

Fig 2. Landslide susceptibility maps generated by (a) LR, (b) MLP, (c) RF, (d) SVM, (e) Adaboost, and (f) GBDT

4.2 Model performance evaluation

Figure 3(a) shows the roc curve of all models we used on the training datasets, which is also called success rate curve (SRC). SRC can evaluate how well the models fit on training datasets. RF has the highest SRC-AUC (0.922), followed by MLP (0.916), GBDT (0.885), Adaboost (0.883), SVM (0.860), and LR models (0.856).

Figure 3. ROC curve of landslide susceptibility models on (a) training and (b) validation datasets

5 Conclusion

In this paper, we leverage six popular machine learning methods to perform landslide susceptibility mapping in Wenchuan County. A total of 1365 landslide points and 14 predisposing factors were used to train the models. The landslide susceptibility maps showed that the main landslide zone was concentrated near the epicenter of Great Wenchuan Earthquake in the southeastern area of Wenchuan County. The result of ROC analysis showed that all models perform well on training and validation datasets, with SRC-AUC and PRC-AUC greater than 85%. The result of this paper demonstrate machine learning method can rapidly build robust models for landslide susceptibility mapping.

References

1. F. Wang, "An extreme rainfall-induced landslide susceptibility assessment using autoencoder combined with random forest in Shimane Prefecture, Japan," *Geoenvironmental Disasters*, vol. 7, 2020.
2. L. Li, R. Liu, S. Pirasteh, X. Chen, L. He, and J. Li, "A novel genetic algorithm for optimization of conditioning factors in shallow translational landslides and susceptibility mapping," *Arab. J. Geosci.*, vol. 10, no. 9, 2017, doi: 10.1007/s12517-017-3002-4.
3. G. L. Bennett, S. R. Miller, J. J. Roering, and D. A. Schmidt, "Landslides, threshold slopes, and the survival of relict terrain in the wake of the Mendocino Triple Junction," *Geology*, vol. 44, no. 5, pp. 363–366, 2016, doi: 10.1130/G37530.1.
4. A. Ozdemir and T. Altural, "A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan mountains, SW Turkey," *J. Asian Earth Sci.*, vol. 64, no. Mar.5, pp. 180–197, 2013, doi: 10.1016/j.jseae.2012.12.014.
5. M. Zare, H. R. Pourghasemi, M. Vafakhah, and B. Pradhan, "Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: A comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms," *Arab. J. Geosci.*, vol. 6, no. 8, pp. 2873–2888, 2013, doi: 10.1007/s12517-012-0610-x.
6. D. Tien Bui, T. A. Tuan, H. Klempe, B. Pradhan, and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides*, vol. 13, no. 2, pp. 361–378, 2016, doi: 10.1007/s10346-015-0557-6.

All models fit the training datasets well. Figure 3(b) is the prediction rate curve (PRC) used to evaluate the predictability on the unknown data. RF has the highest PRC-AUC (0.865), followed by GBDT (0.853), SVM (0.852), LR (0.850), MLP (0.850), and Adaboost models (0.850). The results of the PRC-AUC indicate that the trained models show high predictability when dealing with unknown data.