Spatial prediction of landslide susceptibility based on the neighborhood rough set

Xin Yang¹, Rui Liu¹,²*, Luyao Li¹, Mei Yang¹ and Yuantao Yang¹
¹College of Geophysics, Chengdu University of Technology, Sichuan 610059, China
²Key Lab of Earth Exploration and Information Techniques of Ministry Education, Chengdu University of Technology, Sichuan 610059, China
*Corresponding author’s e-mail: lr@cdut.edu.cn

Abstract: This paper discusses the feasibility of reducing the landslide inducing factors by the neighborhood rough set algorithm in data processing section, which could improve the accuracy and timeliness of landslide susceptibility prediction models effectively. 15 predisposing factors for a continuous value that has not been graded were reduced by neighborhood rough set, a granularity calculation method, based on the importance degree of each factor. Then the combination of factors before and after optimization was put into random forest (RF) and support vector machine (SVM) for modelling. ROC curve and statistical indicators show that: the average performance of the reduced factors combination is superior to that before optimization. In addition, we used the RF which has a better performs in evaluation to map the landslides susceptibility in Jiuzhaigou area, discuss the timeliness of the assessment of landslides prediction and the weight of the predisposing factors.

1. Introduction
In China, the toll of death caused by landslides between 1949 and 2011 is conservatively estimated at more than 25,000[1], the frequent occurrence of landslides in China, for example, 4,800 landslides happened in Jiuzhaigou which triggered by the 7-magnitude earthquake in 2017, with a total area of 9.6 square kilometres, has brought immeasurable losses. Therefore, it is of great significance to study the landslide susceptibility.

At present, scholars mainly focus on the study of landslide vulnerability from two ways: one is the influence factor of landslides; the other is the improvement of prediction models[2]. But either way in previous studies, researchers are usually used to discretize the values of continuous attributes extracted near the landslide area, and the discrete classification standards are often based on previous experience. Actually, the change of data attributes inevitably leads to the loss of important information in these processes, which affects the prediction accuracy.

In this study, we use Jiuzhaigou earthquake co-seismic landslide point as the research data, select 16 landslide influence attributes. The neighborhood rough set algorithm is used to process the continuous data directly and reduce the redundancy factors at the same time, so as to avoid missing information and improve the prediction accuracy.

2. Study area
The study area is located in Jiuzhaigou County, near the northern boundary of Sichuan Province.
On 8 August, 2017, a Ms 7.0 earthquake occurred in Jiuzhaigou County, Aba Autonomous Prefecture, Sichuan Province, China with the location of (33.2° N, 103.8° E) and focal depth of 20 km (CENC, 2017: http://www.cenc.ac.cn/)[3-4]. Landslide are secondary disasters triggered by earthquakes[5], with many casualties and damage to infrastructure as well as environmental, historical and cultural heritage sites[6-7].

3. Data preparation

3.1 Landslides data sources
In this paper, we constructed a landslide inventory map with 270 landslides which are triggered by the quake locations in the study area. In order to balance the data, we randomly generate the same amount of non-landslide points in the study area. In addition, two groups of data were annotated with “1” and “0”, representing the landslide points the non-landslide points respectively.

3.2 Landslide predisposing factors
By summarized the data and related researches[8-9], this paper has selected the following static predisposing factors and shown in(Figure 1):
4. Method

4.1 Landslide affecting factors selection using Neighborhood Rough Set

Rough set is the initial form of neighborhood rough set, which was invented to possess with uncertainty by Pawlak[10], it is used to analyse and process incomplete and imprecise data directly, thus, obtain the concise and expressive form of knowledge expression from the inherent interconnection of data[11].

In the neighborhood rough set theory, let U be a certain set called the universe and A be the non-empty finite set of union of condition attribute C and decision attribute D, then \( S = \langle U, A \rangle \) is said to be an information system. In this study, U is the landslide inventory, C is the attribute of predisposing factors, D is the decision attribute presents the status of landslide (happened or not)[12]. Generally, the information system is also called decision system (as Table 1). For an information system S can be approximated by the lower approximation and upper approximation which are defined as \( N^-(X) = \{x_i | \text{IND}(A) \cap X \neq \emptyset, x_i \in U \} \) and \( N_-(X) = \{x_i | \text{IND}(A) \subseteq X, x_i \in U \} \), Where \( \text{IND}(-) \) represents the indiscernible relationship.

In addition, the boundary region and positive region (the lower approximation) are defined as \( \text{BNp}(X) = N^-(X) - N_-(X) \) and \( \text{Pos}_B(X) = N_-(X) \) respectively.

In a decision system, the influence of factors attribute subset on decision attribute can be defined as importance. Implementation of attribute reduction based on neighborhood rough set is to eliminated attribute \( a_i \) by iteration. Obviously, the greater influence after elimination, the greater importance of \( a_i \) for decision system. Based on the dependence value, \( \text{Sig}(a_i, B, D) = \gamma_B(D) - \gamma_{B-a_i} \), the attribute without importance will be eliminated from B. The dependence is defined as Eq.1, which represents the ratio that fall into decision D correctly under attribute subset B[13-14].

\[
\gamma_B(D) = \frac{|\text{pos}_B(D)|}{|U|}
\]

Where \(|\cdot|\) is used to calculate the number of elements.

Table 1. The decision system

| \(\text{sn} \) | Roads  | Rivers  | DSF    | Elevation | …… | aspect | landslide |
|---|---------|---------|--------|-----------|-----|--------|-----------|
| 1 | 10311.36| 924.3909| 8376.461| 3482      |     | 228.742981| 1         |
| 2 | 10.86098| 73.77489| 6667.01 | 1581      |     | 104.036247| 1         |
| 3 | 72.01941| 73.84597| 7496.172| 1665      |     | 126.740562| 1         |
| … | …       | …       | …      | …         | …   | …      | …         |
| 540| 19059.02| 2310.402| 22578.36| 4456      |     | 333.434937| 0         |

Rough set show a good performance under discrete values, but it can’t be applied to the continuous
decision system. Hu (2008) proposed the concept of neighborhood rough set, which extended the domain to continuous universe after replacing the indiscernible relationship with neighbourhood.[13].

In the data pre-processing part of landslide prone problem, the subjectivity in the process of discretization of original data will inevitably affect the randomness problem. Therefore, this paper uses neighborhood rough set method to reduce the landslide predisposing factors and realize it with python.

4.2 Landslide susceptibility models

4.2.1 Random forest (RF). The Random Forest, developed by Breiman[15], was an ensemble-learning technique for classification and regression tasks. More formally, RF is integrated by stochastic subspace, decision tree (CART) and bagging learning strategy[16-18]. It operates by constructing different trees with bootstrap samples and outputting the class that is the best result of the individual trees after voting.

4.2.2 Support vector machine (SVM). Support vector machine, leverage the strategy of Structural Risk Minimization (SRM), is a supervised classifier for classification and regression[19]. SVM can applied for linear separable situation as well as more complex non-liner problem by constructing a hyperplane in a high or infinite dimension space and leveraging kernel method[20]. In this study, the radical basis function (RBF) was selected as kernel function of SVM.

4.3 Landslide conditioning factor analysis
Multicollinearity refers to the non-independence of landslide conditioning factors., the factors combination with strong correlation can leads to potential model analysis error. The tolerance and inflation factor (VIF) method are commonly used as one of the quantification of multicollinearity in statistical literature[21]. In this study, tolerance and VIF are used to diagnose the multicollinearity among the landslide predisposing factors.

4.4 Model performance assessment and comparison
In this study, the ROC curve, usually leveraged in landslide susceptibility assessment, is used as an objective evaluation index to qualitatively describe the accuracy changes after using the neighborhood rough set algorithm. In addition, statistical measures, root mean squared error (RMSE) and mean absolute error (MAE), have also been used as an indicator for the assessment of model.

5. Result

5.1 Predisposing factors reduction by neighborhood rough set
In neighborhood rough set method, the neighborhood value was set to 0.15 through parameter selection. Each factors’ dependency is shown in Table 2. Factors without dependency have been removed from the combination of predisposing factors, and the distance to roads, the distance to rivers, DSF, elevation, vertical slope curvature(VSC), TWI, SPI, PGA, SLOPE, NDVI, LT were selected for the future landslide susceptibility model.

| Factors                  | Dependency | Factors | Dependency |
|--------------------------|------------|---------|------------|
| The distance to roads    | 0.19       | VSC     | 0.05       |
| The distance to rivers   | 0.19       | TWI     | 0.05       |
| DSF                      | 0.05       | SPI     | 0.05       |
| Elevation                | 0.14       | PGA     | 0.19       |
| Profile curvature        | 0.00       | SLOPE   | 0.19       |
| LT                       | 0.14       | NDVI    | 0.29       |
5.2 Multicollinearity diagnose of landslide predisposing factors

The tolerance and VIF of the factor combination after neighborhood rough set optimization are shown in Figure 2(a). No obvious outliers appeared as the result and all these values were satisfied the standard (tolerance < 0.1 or VIF > 0), which indicated that no multicollinearity existed among the landslide predisposing factors.

![Figure 2](image)

Figure 2. (a)The VIF of factors; (b)The ROC of the RF and SVM: AUC 2 is using the optimised factors and the AUC 1 is using origin factors

5.3 Model validation and landslide susceptibility mapping

The ROC curve of two models constructed using a different combination of predisposing factors are shown in Figure 2(b). Other statistical measures are shown in Table 3. It can be observed that the accuracy of the model is improved after optimization, which indicated the feasibility of the neighborhood rough set method in landslide predisposing factors reduction. In addition, the result verified the two assumption, information loss with discretization and factor redundancy. Finally, random forests show the best average predictive performance through validation.

The prediction model trained with feature matrix composed of the selected predisposing factors of the landslide inventory can extended to every pixel of the whole study area after validation. Landslide susceptibility index (LSI) value was in the range of 0 to 0.99 in study area. Using LSI, we can get the possibility of landslide around the study area and landslide susceptibility map (LSM) shown in Figure 5. The LSI were reclassified into tree as very low (0-009), low (0.09-0.21), moderate (0.21-0.37), high (0.37-0.65), very high (0.65-0.99) and the produced landslide susceptibility maps were shown in figure 5, which indicated the distribution of hazardous areas. In addition, the proportion of pixel of the corresponding dangerous level in study area is shown in figure 3, which gives us a summary of the hazard level.
5.4 Importance of landslide predisposing factors

The importance of each predisposing factor is obtained after predictive model training, represents the correlation between predisposing factor and landslide event. Obviously, the greater importance of predictive factors, the greater its evoked effect in landslide event. Figure 4 shows the average importance of each factor under predictive models and the top three is elevation, distance to rivers, PGA, respectively. The result reveals the main predisposing factors in the study area by analyzing weight ranking.

| statistical measures | before | after | before | after |
|----------------------|--------|-------|--------|-------|
| SVM                  | 0.8699 | 0.87585 | 0.93934 | 0.94461 |
| MAE                  | 0.1358 | 0.129636 | 0.0617 | 0.05556 |
| RMSE                 | 0.3685 | 0.36004 | 0.24845 | 0.2357 |
6. Discussion

6.1 The timeliness about evaluation of landslide susceptibility prediction

The extraction of predisposing factors often takes a certain amount of time, but timeliness is very significant in the evaluation of hazards. In this study, after some factors were eliminated by neighborhood rough set, we selected the combination of the best factors for the correlation between the study area and landslide event. And the result of model validation demonstrates that the models established by less predisposing factors perform better than the one established by more factors. So neighborhood rough set method can not only reduce the training time of prediction model, but also speed up the time of data acquisition.

In this study, the number of reduced factors is not large due to the limitation of the number of factors we extracted, but the advantages of neighborhood rough set method when the number increases. It also inspires us to pay attention to the timeliness in the assessment of landslide susceptibility. This paper attempted to compare the training time of the prediction model, but because of the limitation of the amount of dataset, it cannot reflect the difference in time. Of course, the neighborhood rough set is one of the attribute reduction algorithms, and how to find a suitable algorithm for the landslide problem is the future research we will carry out.

6.2 The discussion about importance of landslide predisposing factors

Landslide predisposing factors in the top-three ranking based on the calculation of weight are elevation, distance to rivers, PGA, respectively. It is worth pointing out that the landslide points researched in this paper are triggered by earthquake and researchers have studied the significance of the PGA in co-seismic landslides assessment before. Through models analysis, this paper also concludes that PGA accounts for a large weight, which is consistent with previous studies[22].

Result reveals that the weight analysis is necessary for the regional landslide susceptibility assessment. For the Jiuzhaigou area, it is significant to strengthen the prevention of predisposing factors with high weight (elevation, distance to rivers and PGA). Therefore, the weight analysis method about landslide susceptibility can afford relevant department indicators that need to be focused on prevention, as to reduce the degree of landslide damage.

7. Conclusion

In this study, neighborhood rough set, a mathematical tool, is used to optimize the combination of inducing factors about the analysis of landslide susceptibility. First, in order to avoid losing information, 15 predisposing factors with continuous values are selected to be construct the dataset, directly. Then, optimized factor combination containing the distance to roads, the distance to rivers, DSF, elevation, vertical slope curvature, TWI, SPI, PGA, SLOPE, NDVI, LT were selected based on neighborhood rough set. The Random Forests and SVM model were adopted to generate the landslide susceptibility model in study area. After validation and comparison, the results demonstrate that the models established by the optimized combination containing less factor based on neighborhood rough set perform more preferable than the models established by the non-optimized combination. In addition, it also discussed the timeliness of evaluation and predisposing factors weight. The result of this study can prove the feasibility of reduction of landslide predisposing factors based on neighborhood rough set and also provide a reference for evaluation of landslide susceptibility in Jiuzhaigou and other places with similar geographical environment.

Acknowledgments

Foundation of SiChuan Educational Committee (No.14ZB0071), Science and Technology Department of SiChuan Province Technology Support Program (No.2012FZ0018) and Research Center for Petroleum Development of Sichuan Province (No.SKB10-11) has supported the research work. We thank all reviewers for their comments towards improving the technical quality of the manuscript.
References

[1] Sang K. (2013) Statistics and Analysis of Landslide Disaster of China in Recent 60 Years [J]. S&T Communication, (10):154+159. (In Chinese)

[2] Reichenbach, P., Rossi, M., Malamud, B., Mihir, M., & Guzzetti, F. (2018) A review of statistically-based landslide susceptibility models. Earth-Science Reviews, S0012825217305652.

[3] Li L, Liu R, Pirasteh S, Chen X, He L and Li J. (2017) A novel genetic algorithm for optimization of conditioning factors in shallow translational landslides and susceptibility mapping. Arabian Journal of Geosciences 10. Doi: 10.1007/s12517-017-3002-4.

[4] Zhang Y, Zhang G, Hetland EA et al. (2018) Source fault and slip distribution of the 2017 Mw 6.5 Jiuzhaigou, China, earthquake and its tectonic implications. Seismological Research Letters 89(4): 1345-1353.

[5] Juan Cao, Zhao Zhang, Chenzhi Wang, Jifu Liu, Liangliang Zhang. (2019) Susceptibility assessment of landslides triggered by earthquakes in the Western Sichuan Plateau[J]. Catena, 175(175):63-76.

[6] Juliev, M., Mergili, M., Mondal, I., Nurtaev, B., Pulatov, A., Hübl, J. (2019) Comparative analysis of statistical methods for landslide susceptibility mapping in the Bostanlik District, Uzbekistan. Sci. Total Environ. 653, 801–814.

[7] Sun J, Yue H, Shen Z, et al. (2018)2017 Jiuzhaigou earthquake: a complicated event occurred in a young fault system[J]. Geophysical Research Letters.

[8] Jiao YM, Zhao DM, Ding YP, Liu Y, Xu Q, Qiu YM, Liu CJ, Liu ZL, Zha ZQ, Li R. (2019). Performance evaluation for four GIS-based models purposed to predict and map landslide susceptibility: A case study at a World Heritage site in Southwest China[J]. Catena, 183 (2019) 104221.

[9] Zhou Hongjian. (2015)Risk Assessment of Disaster Chain: Experience from Wenchuan Earthquake-induced Landslides in China. Journal of Mountain Science. 12(5):1169-1180.

[10] Z. Pawlak. (1982)Rough sets, Int. J. Comput. Inf. Sci., vol. 11, no. 5, pp. 341–356.

[11] Q. Zhu, M. Huang, and G. Zhao. (2011)Apple mealiness detection based on neighborhood rough set and hyperspectral scattering image, Nongye Jixie Xuebao/Transactions Chinese Soc. Agric. Mach., vol. 42, no. 10.

[12] S. S. Pal and S. Kar. (2019) Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory, Math. Comput. Simul., vol. 162, pp. 18–30.

[13] Q. H. Hu, D. R. Yu, and Z. X. Xie. (2008) Numerical attribute reduction based on neighborhood granulation and rough approximation, Ruan Jian Xue Bao/Journal Softw., vol. 19, no. 3, pp. 640–649.

[14] B. Barman and S. Patra. (2019)A novel technique to detect a suboptimal threshold of neighborhood rough sets for hyperspectral band selection, Soft Comput., vol. 23, no. 24, pp. 13709–13719.

[15] Breiman, L. (2001) Random forests. Machine learning, 45(1), 5-32.

[16] A. M. Youssef, H. R. Pourghasemi, Z. S. Pourtaghi, and M. M. Al-Katheeri. (2016)Erratum to: Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia (Landslides, 10.10, Landslides, vol. 13, no. 5, pp. 1315–1318.

[17] K. Zhang, X. Wu, R. Niu, K. Yang, and L. Zhao. (2017) The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China, Environ. Earth Sci., vol. 76, no. 11, pp. 1–20.

[18] H. R. Pourghasemi and O. Rahmati. (2018) Prediction of the landslide susceptibility: Which algorithm, which precision?, Catena, vol. 162, no. October 2017, pp. 177–192.
[19] V. N. Vapnik and V. N. Vapnik. (2000) Introduction: Four Periods in the Research of the Learning Problem, Nat. Stat. Learn. Theory, pp. 1–15.

[20] Julian Broséus, Vallat, M., & Esseiva, P. (2011) Multi-class differentiation of cannabis seedlings in a forensic context. Chemometrics & Intelligent Laboratory Systems, 107(2), 343-350

[21] C. F. Dormann et al. (2012) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, no. February, pp. 1–20.

[22] Huang, H. P., Yang, K. C., & Lin, B. W. (2013) Statistical evaluation of the effect of earthquake with other related factors on landslide susceptibility: using the watershed area of shihmen reservoir in taiwan as a case study. Environmental Earth Sciences, 69(7), 2151-2166.