Selection of landslide affecting factors based on strong association analysis

Luyao Li¹, Rui Liu¹, ²*, Xin Yang¹, 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

Abstract: The performance of models in landslide susceptibility mapping largely depends on the selection and combination of affecting factors. The Apriori algorithm proposed in this paper is a factor selection method of strong association analysis, which can select the factors that are most likely to induce landslides from 15 affecting factors. Then combine the origin and optimized factors to build the prediction model of landslide susceptibility by support vector machine (SVM) in machine learning. Afterwards, we verifying the landslide points in the dataset to measure the accuracy of the model. Ultimately, ROC curve was adopted to evaluate the prediction results of the two models. The result reveals that the model based on the combination of optimized factors (AUC 1=0.930) is superior to that based on 15 affecting factors (AUC 2=0.898).

1. Introduction
Jiuzhaigou Valley Scenic and Historic Interest Area is a world natural heritage[1]. At 21:19:46, August 8, 2017, the MS7.0 earthquake occurred in Jiuzhaigou County, Sichuan Province, causing a host of landslides[2]. In order to avoid more hazards caused by landslides, it is indispensable to evaluate the prone areas of landslides[3]. When mapping landslide susceptibility, the random selection of factors may bring the noise[4]. In this paper, we propose to use Apriori algorithm for combination optimization of landslide affecting factors, this algorithm can fully consider the correlation between factors, so as to select the affecting factors objectively and combine factors with better prediction effects[5].

2. Study area and Materials

2.1 Study area
Jiuzhaigou Valley Scenic and Historic Interest Area, situated in Aba Prefecture, Sichuan Province, China, 103°46-104°4′E, 32°54′-33°N, covers an area of 651.34 m² (Figure 1). Since the Quaternary, the strong tectonic movement in this region has led to frequent regional seismicity[6].
2.2 Data Preparation
To build a landslide inventory map, we collected 136 landslide points by visual interpretation of remote sensing images and field surveys. Through statistical analysis, we found that the area of these landslides is mostly below 30m, so in figure 1 we use points to represent them.

Based on a large amount of data and related papers[7-9] we selected a total of 15 affecting factors (Figure 2). All factors were processed with ArcGIS 10.5 and resampled to the same resolution as the Digital Elevation Model (DEM):30 m. The digital elevation model data was derived from the Geospatial Data Cloud (http://www.gscloud.cn/).
3. Methods

The main technical flow chart is as follows: (i) Extract relevant affecting factors and reclassify them respectively. (ii) Coding the factors and using the Apriori algorithm to analyze the association rules. (iii) Select factors based on the strong association rules obtained. (iv) Using SVM model for landslide susceptibility mapping. (v) Model comparison and correlation factor analysis.

3.1 Landslide affecting factors selection using Apriori

The steps of the algorithm are as follows: (i) Generate a list of datasets for all individual items (i.e. categories of landslide affecting factors). (ii) Scan the dataset and then calculate the support. The set less than the min. support is removed. (iii) Combine the remaining collections, generate a set of items with two elements, rescan the dataset, and remove the item set less than the min. support, repeating the step until all item sets are removed. (iv) Generate association rules based on frequent itemsets. (v) Calculate the confidence and lift of the association rule, and the association rule less than the min. feasibility or the lift less than 1 is removed.

The A=>B support is expressed as follows:

\[ S_{A=>B} = \frac{|T(A \cup B)|}{|T|} \]  

(1)

And the A=>B confidence is expressed as:

\[ C_{A=>B} = \frac{|T(A \cup B)|}{|T|} \]  

(2)

Lift is defined as:

\[ L_{X=>Y} = \frac{|T(A \cup B)|}{|T(X)|} \cdot \frac{|T(Y)|}{T} \]  

(3)

where \(|T(A \cup B)|\) is the number of occurrences of the item set A=>B in the landslide record. \(|T(x)|\) is the number of occurrences of a certain landslide affecting factor in the landslide record. T is the total number of landslide records. We assumed that the factor with high predictive power appeared more frequently in the landslide records and harbored the idea that the factors appearing in obtained association rules were the optimal landslide affecting factors[10].

3.2 Landslide susceptibility models

The SVM for regression was proposed in 1996 by Vapnik. To solve a non-linear problem, three kernel functions have been introduced commonly[11]. The Radial Basis kernel Function (RBF) is defined as:

\[ K(x_i, x_j) = e^{-\gamma(x_i, x_j)^2} \]  

(4)

The probability of a landslide was estimated by the model output (between 0 and 1) in each grid cell.
3.3 Model assessment and comparison
The confusion matrix shown is a commonly used tool for evaluating model performance in machine learning, and simultaneously is widely used in the evaluation of landslide models. The sensitivity can be used to measure whether the model can classify positive categories correctly, which means there will be landslides, while the specificity measures whether the ability to correctly classify negative classes, which means there will be no landslides.

The ROC curve can represent changes in sensitivity and specificity as the threshold changes, the average performance value of the model can be measured by the area under the curve AUC. Consequently, in this paper the AUC value was used to objectively evaluate the model[12].

4. Results and Discussion
4.1 Landslide affecting factors selection
We obtained many strong association rules by setting the support to 0.2, 0.3, 0.4 and the confidence to 0.7, 0.8, 0.9. We iterated through all the generated strong association rules and removed the landslide affecting factors that were absent in the rule set. Thus, 8 factors were obtained which have the strongest correlation with landslides, they are: Lithology, Elevation, SPI, Undulation, Drainage, LT, PGA, DF. As is shown in Table 1, when we consider the case that support is 0.4 and confidence is more than 0.9, these rules indicate that towards a geographic location, if the geographic features of the location meet the above rules, the probability of the landslide is high. Especially when elevation is in class 1, the lithology is in class 7, the SPI is in class 1, the drainage is in class 1 and the undulation is in class 1.

Table 1. The affecting of support and confidence concerning affecting selection

| Support | Confidence | Factor Class                  |
|---------|------------|-------------------------------|
| 0.4     | 1          | Lithology 6                   |
|         |            | Elevation(m) 1: (1894,2501)   |
| 0.4     | 1          | SPI 1: (0,11)                 |
|         |            | Undulation(m) 1: (1,31)       |
| 0.4     | 1          | Lithology 7                   |
|         |            | Drainage(m) 1: < 300          |
| 0.4     | 0.98       | LT 1: water                   |
|         |            | Drainage(m) 1: <300           |
| 0.4     | 0.95       | Undulation(m) 1: (1,31)       |
|         |            | SPI 1: (0,11)                 |
| 0.4     | 0.94       | Drainage(m) 1: <300           |
|         |            | Elevation(m) 1: (1894,2501)   |
| 0.4     | 0.92       | LT 1: water                   |
|         |            | Drainage(m) 1: (1894,2501)    |
| 0.4     | 0.91       | PGA(g) 5: >2.3                |
|         |            | Elevation(m) 1: (1894,2501)   |

4.2 Landslide susceptibility mapping
The origin and the screened factors were used to map the landslide susceptibility, we get the following figures:
4.3 Evaluation and analysis of landslide susceptibility mapping model
As it shows in figure 3, we get a more accurate map to predict the landslides in Jiuzhaigou area when we use the screened factors to build model. The streamlined factor combination is effectively reduced the noise in the result, which can more avoid misjudgement. We use ROC curve to evaluate two models and result shown in figure 4.

![Figure 4. The ROC of the SVM model: AUC 1 is using the optimised factors and the AUC 2 is using origin factors.](image)

5. Conclusions
In this paper, we optimized and reduced 8 landslide influence factors out of 15 based on the Apriori algorithm in the strong correlation analysis method, then use SVM model, adopted to the landslide susceptibility map of Jiuzhaigou Natural Scenic Area. Eventually, the ROC curve was used to evaluate the results of 15 factors combination and the combinations of the factors were optimized by Apriori algorithm. The results demonstrate that the model established by the latter (AUC 1=0.930) performs superior to the model established by the former (AUC 2=0.898). The results of this study can provide...
a reference for the establishment of landslides hazard prediction models in Jiuzhaigou and other places in western Sichuan, safeguard the security of human life, property and ecological environment effectively.

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] Li W, Ge X and Liu C (2005) Hiking trails and tourism affecting assessment in protected area: Jiuzhaigou biosphere reserve, china. Environmental monitoring and assessment 108: 279-293. doi: 10.1007/s10661-005-4327-0.
[2] Guo AN, Ren D, Zhao CC, Zhang XH, Wu JH (2017) Review of mid-long term prediction for jiuzhaigou ms7.0 earthquake in sichuan province. China Earthquake Engineering. doi: 10.3969/j.issn.1000-0844.2017.04.0639.
[3] Zhao B, Luo YH, Li J, Shen XZ (2018) Landslides and damage resulting from the Jiuzhaigou earthquake (2017), Sichuan, china. doi: 10.1098/rsos.17141810.6084/m9.
[4] Costanzo D, Rotigliano E, Irigaray C, Jiménez-Perálvarez JD and Chacón J (2012) Factors selection in landslide susceptibility modelling on large scale following the gis matrix method: Application to the river beiro basin (spain). Natural Hazards and Earth System Science 12: 327-340. doi: 10.5194/nhess-12-327-2012.
[5] Meinhardt M, Fink M and Tünschel H (2015) Landslide susceptibility analysis in central vietnam based on an incomplete landslide inventory: Comparison of a new method to calculate weighting factors by means of bivariate statistics. Geomorphology 234: 80-97. doi: 10.1016/j.geomorph.2014.12.042.
[6] Sun JB, Yue H, Shen ZK, Fang LH, YZ and Sun aX (2017) The 2017 jiuzhaigou earthquake: A complicated event occurred in a young fault system. Geophysical Research Letters. doi: 10.1002.
[7] Kavzoglu T, Kutlug Sahin E and Colkesen I (2015) Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. Engineering Geology 192: 101-112. doi: 10.1016/j.enggeo.2015.04.004
[8] Devkota KC, Regmi AD, Pourghasemi HR, Yoshida K, Pradhan B, Ryu IC, Dhital MR and Althuwaynee OF (2012) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in gis and their comparison at mugling–narayanghat road section in nepal himalaya. Natural Hazards 65: 135-165. doi: 10.1007/s11069-012-0347-6.
[9] Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z and Ma J (2017) A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 151: 147-160. doi: 10.1016/j.catena.2016.11.032.
[10] Qiu MY. Fire warming and countermeasures on glassland based on Aporio association rule. [J]. Science and technology economics guide, 2019,27(26):106-108.
[11] Vapnik V, Cortes C. (1995). Support-Vector Networks. Machine Learning. 20(3):273-297.
[12] 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.