ICA
International Cartographic Association
Association Cartographique Internationale
ACI

Landslide susceptibility evaluation based on optimized support vector machine

Liu Jiping [a], Lin Rongfu [a,b,*], Xu Shenghua [a], Wang Yong [a], Che Xianghong [a], Chen Jie [a]

[a] *Chinese Academy of Surveying and Mapping, liujp@casm.ac.cn, xushh@casm.ac.cn, wangy@casm.ac.cn, chexh@casm.ac.cn, giserchenj@gmail.com*

[b] *School of Geomatics, Liaoning Technology University, giserlin1@gmail.com*

Abstract: Landslide is a natural disaster that has caused great property losses and human casualties in the world. To strengthen the target prevention and management level, ZhaShui county, Shaanxi province, is selected as the research area to evaluate the landslide susceptibility. First of all, under the premise of considering the correlation, 10 evaluation factors closely related to landslide disaster (i.e., elevation, rainfall, rock group, slope, slope aspect, vegetation index, landform, distance to residential area, distance to road, distance to river system) are taken together with non-landslide points, which are selected under multi-constraint conditions to form a sample data-set. Secondly, the sample dataset is substituted into the Support Vector Machine (SVM) model optimized by firefly algorithm for training and prediction. Finally, the result map was partitioned according to the natural discontinuous point method, and the landslide susceptibility map was obtained. The results show that the model optimized by the firefly algorithm has higher accuracy, and the landslide susceptibility results are more consistent with the actual distribution of disaster points.

Keywords: evaluation of landslide susceptibility, support vector machine

1. Introduction

China has a vast territory. Affected by plate movement, China's geological conditions and landforms are very complex, with mountainous areas accounting for nearly 70% of the territory area. The complex geological and topographic conditions in mountainous areas lead to a large number of geological disasters. According to the annual report on natural disasters issued by the Ministry of Natural Resources of China in 2020, a total of 7,210 geological disasters occurred in the country in 2020, an increase of 26.8% compared with 2019 and 14.6% compared with the average annual average during the 13th Five-Year Plan period. Among them, 4,810 landslide disasters occurred, accounting for 66.71% of the total number of geological disasters. Due to the vast territory of China, the different geological and topographic conditions in different regions, it is difficult to evaluate the vulnerability of landslides. At the same time, the evaluation of the susceptibility of landslides in China is not perfect at present, and there is still a large room for improvement in the related research work of the vulnerability assessment. In the face of the increasingly severe challenges of landslide disasters, to reduce the loss of life and property caused by landslides to the people as much as possible, it is necessary to effectively evaluate the areas where landslide disasters occur frequently and make scientific predictions about the areas where landslides are likely to occur. As a scientific prediction method for the possibility of landslides, the final prediction results can provide an important scientific reference for landslide disaster warning and urban development planning, and have important practical application value (Nguyen et al., 2019).

In recent years, with the rapid development of Artificial Intelligence (AI) field, machine learning has been used by more and more researchers in the field of landslide disaster prediction. (Bui et al., 2019) (Roy et al., 2019) (Chang et al., 2019) (Sahin et al., 2020). Tran Van Phonga et al. (Phong et al., 2019) selected Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression (LR) and Reduced Error Pruning Tree (REPT), and nine landslide condition factors were used to generate data sets for training and validation of the model. The results show that SVM is superior to all other methods, namely ANN, LR and REPT. The support vector machine model can effectively solve the problem of constructing high-dimensional data model under the condition of limited number of samples, and has good applicability in the field of landslide susceptibility evaluation. Therefore, this paper chooses this model as the vulnerability evaluation model. Aiming at the difficulty in selecting the hyperparameters of the support vector machine model, the firefly algorithm is introduced to optimize the model hyperparameters, and the optimized hyperparameter results are brought into the support vector machine. It is compared with the classic support vector machine model to verify the superiority of the optimization algorithm in this paper.

2. Model and Methods

SVM model is a machine learning algorithm proposed by Vapnik. SVM combines two learning techniques, i.e.,Vapnik-Cherbonenkis (VC) dimensional theory and statistical learning theory, and improves the generalization ability of learning machine by seeking structured risk minimization, so as to obtain good statistical rules in the case of relatively small statistical sample size. It is a widely used supervised learning model (Nhu et al., 2020). Therefore, SVM model was selected as the vulnerability evaluation model in this paper.

Proceedings of the International Cartographic Association, 4, 2021.
30th International Cartographic Conference (ICC 2021), 14–18 December 2021, Florence, Italy. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-4-67-2021 | © Author(s) 2021. CC BY 4.0 License.

Suppose there are two types of samples, including landslides and non-landslides in this study:

$$(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\,/\,x\in R^n, y\in\{-1,+1\} \quad (1)$$

The general form of the classification function :

$$g(x)=\omega\cdot x+b \quad (2)$$

Linear classification surface can be described as:

$$\omega\cdot x+b=0 \quad (3)$$

In view of the requirement of the SVM model to correctly divide all samples, the restriction conditions are set as follows:

$$y_i\left[\left(\omega\cdot x_i\right)+b\right]-1\geq 0,\ i=1,2,\cdots,l \quad (4)$$

When $|g(x)|=1$, according to the Euclidean distance calculation formula, the classification interval d is equal to $\frac{2}{\|\omega\|}$. Regarding the maximum d required in the model, from another perspective, it is equivalent to the minimum $\frac{1}{2}\omega^2$. Combining the above two points to transform the optimal classification surface problem into a constrained optimization problem, can be written as:

$$\min\frac{1}{2}\|\omega\|^2 \quad (5)$$

$$\text{s.t. } y_i\left[\left(\omega\cdot x_i\right)+b-1\right]\geq 0 \quad (6)$$

Introduce the Lagrangian function and rewrite the formula (5-6) as:

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{l}\alpha_i\left[y_i\left(\omega\cdot x_i+b\right)-1\right] \quad (7)$$

In the formula, α_i is the Lagrangian multiplier vector. The partial differential equation is solved by KKT (Karush-Kuhn-Tucker) condition, The minimum value of formula (7) is obtained according to the following equation:

$$\begin{cases}\dfrac{\partial L}{\partial\omega}=0\Rightarrow\omega=\sum_{i=1}^{l}\alpha_i y_i x_i\\ \dfrac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{l}\alpha_i y_i=0\end{cases} \quad (8)$$

According to the duality theory, the dual problem of the original problem (5-6) can be obtained:

$$\min\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)-\sum_{i=1}^{l}\alpha_i \quad (9)$$

$$\text{s.t.}\,\alpha_i\geq 0,\ i=1,2,\cdots,l \quad (10)$$

$$\sum_{i=1}^{l}\alpha_i y_i=0 \quad (11)$$

Among them, α_i is the corresponding Lagrangian multiplier. Equation (9-11) is also the convex quadratic programming problem, and the unique solution α_i^* can be obtained by solving it. Substituting α_i^* into equation (8) and equation (4) can obtain the coefficient ω^* and the classification threshold b^* . So far, the optimal classification function expression can be obtained:

$$f(x)=\text{sgn}\{(\omega\cdot x)+b\}=\text{sgn}\left\{\sum_{i=1}^{l}\alpha_i^* y_i\left(x_i,x\right)+b^*\right\} \quad (12)$$

The evaluation factor low-dimensional feature space is mapped into the high-dimensional feature space, and the kernel function is used instead of the dot product operation, so the convex quadratic programming expression (9-11) is changed to:

$$\max\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K\left(x_i\cdot x_j\right) \quad (13)$$

$$\text{s.t.}\,\alpha_i\geq 0,\ i=1,2,\cdots,l \quad (14)$$

At the same time, the optimal classification surface function (12) is changed to:

$$f(x)=\text{sgn}\left\{\sum_{i=1}^{l}\alpha_i^* y_i K\left(x_i,x\right)+b^*\right\} \quad (15)$$

In this case, this paper considers the balance between the correct division and the classification interval. For this reason, slack variables can be introduced to allow the possibility of misclassification to solve the nonlinear classification problem.

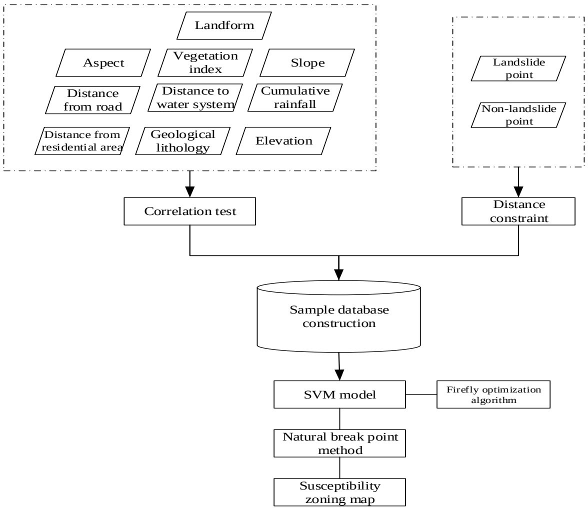


Fig 1. Technical route

When solving nonlinear problems in classification or regression model, kernel function is crucial. Therefore, kernel function plays a very important role in SVM model (Li & Chen, 2020). The kernel function is selected according to the characteristics of nonlinear problem. When solving the nonlinear problem, SVM models with different kernel functions have different effects. Therefore, the appropriate degree of kernel function is also very important when dealing with nonlinear problems. In this paper, based on previous research results, radial basis kernel function (Li & Chen, 2020) was selected. To solve the difficulty of kernel function parameter selection in support vector machine, firefly algorithm is introduced to optimize the selection of model hyper-parameters to ensure the applicability of model hyper-parameters.

3. Experiment and analysis

3.1 Sample dataset optimization

The landslide hazard points data in the study area were ob tained from the "Spatial Distribution Data of Geological Hazards" of the Resource and Environmental Science Da ta Center of the Chinese Academy of Sciences(http://ww w.resdc.cn/data.aspx?DATAID=290). The data format is excel and vector shape file. The geological and lithologic al data of the study area comes from the "Spatial Distribut ion Data of Geological Lithology in China" of the Resour ce and Environmental Science Data Center of the Chinese Academy of Sciences(http://www.resdc.cn/data.aspx?DA TAID=307). The digital elevation model (DEM) data and residential point datas come from the results of the first n ational census of geographic condition s of China. Road d ata and water system data are public data-sets provided by OpenStreetMap.

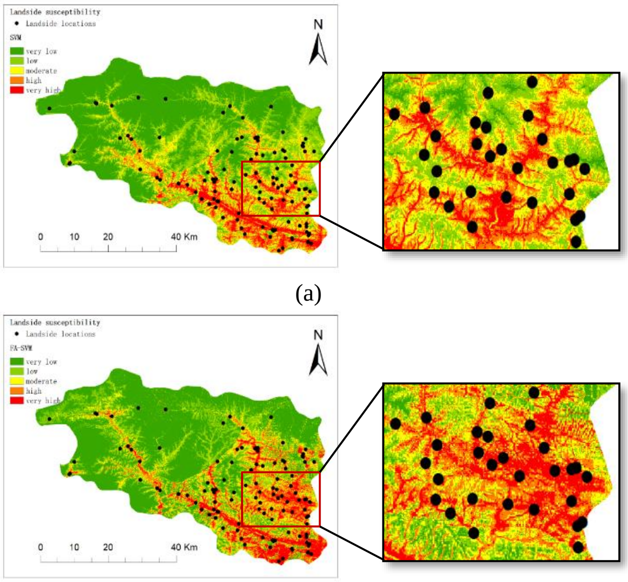
Based on the comprehensive collection of geological information and field investigation in the study area (Zhashui County, Shaanxi Province), this paper analysesthe geological environment and natural environment conditions in the study area, as well as the development characteristics and formation conditions of landslide disasters, and the distribution rule of each evaluation factor. First of all, according to the distribution of landslide disasters in the study area, 10 evaluation factors including elevation, rainfall, rock group, slope, slope aspect, vegetation index, landform, distance to residential area, distance to road and distance to river system were selected as input factors of the model. Pearson's correlation coefficient was used to test the correlation of the evaluation factors,We removed redundant evaluation factors, and used the evaluation factors that pass the correlation test as the final input factors of the model.

In previous studies, the selection of“non-landslide points” samples are often subjectively inferred by researchers or randomly selected on susceptibility evaluation. Because newly developed landslide hazards usually occur in areas where landslide hazards have not occurred temporarily. If non-landslide point samples are directly selected in such areas, it is very likely that grid cells with landslide possibility will be mistakenly selected as non-landslide samples, and it cannot ensure that the selected "non-landslide points" are truly "non-landslide points". "In the selection of non-landslide point samples, this study comprehensively considered the distance between the landslide point and the non-landslide point, and the distance between the generated non-landslide points. We constructed a negative sample selection method under multiple constraints to ensure that the landslide disaster in the area selected by the negative sample is not easy to occur, which ensures the accuracy of the negative sample. Finally, the same number of negative samples as the positive samples were selected to form the sample point data set.

3.2 Vulnerability model construction

The experiment was carried out in the MATLAB language environment, and the support vector machine model was selected as the vulnerability evaluation model. 70% of the sample points were randomly selected as the training data and 30% of the sample points were taken as the test data. When training and predicting the susceptibility evaluation model, due to the dimensional differences between different evaluation factors, there is an inner product problem in the kernel function, and the larger attribute value of the evaluation factor will be used in the calculation. The occurrence of anomalies will eventually lead to anomalies, affect the structure of the vulnerability evaluation model, and adversely affect the prediction results of the model. Therefore, before inputting the evaluation factor into the landslide susceptibility evaluation model, the dimension of the attribute value of the evaluation factor needs to be normalized.

The evaluation result of the susceptibility evaluation model is a continuous value with a value range of 0 to 1. This value represents the probability of landslide disaster occurring in each grid cell in the study area, with a range from 0% to 100%. The evaluation value is the landslide susceptibility evaluation index (Abedini et al., 2019) (Lee et al., 2017). To obtain the susceptibility zoning map, it is necessary to discretize the susceptibility index. Finally, the probability map output by the model is divided into five categories according to the Natural breaks, i.e. very low susceptibility area, low susceptibility area, medium susceptibility area, high susceptibility area and very high susceptibility area. Finally, the landslide susceptibility evaluation zoning map of the study area is obtained.



(a)

(b)

Fig 2. Landslide susceptibility zoning map

4. Conclusion

The distribution areas of the high-incidence areas in the two results are similar. The high-incidence areas of landslide disasters in this region are mainly concentrated in the southeast area and spread to the surrounding areas along the valley. It can be seen from the results that the development of landslide disasters in this region is mainly affected by the complex terrain, landform and geological structure of the mountainous area. Specifically, the optimized model is more robust. The results demonstrate that more landslide points in the result figure optimized by algorithm in the southeastern part of the study area fall into the high-incidence area, while the result figure obtained by the unoptimized classic support vector machine model fails to accurately identify the region. It can be seen that the optimized model prediction results are more accurate and more consistent with the actual distribution of landslide disasters. The prediction results have good practical value and will provide guidance for the subsequent disaster prevention and control.

5. References

Abedini, M., Ghasemian, B., Shirzadi, A., & Bui, D. T. (2019, Sep). A comparative study of support vector machine and logistic model tree classifiers for shallow landslide susceptibility modeling. Environmental Earth Sciences, 78(18). https://doi.org/ARTN 560

10.1007/s12665-019-8562-z

Bui, D. T., Shirzadi, A., Shahabi, H., Geertsema, M., & Lee, S. (2019). New Ensemble Models for Shallow Landslide Susceptibility Modeling in a Semi-Arid Watershed. Forests, 10(9), 743.

Chang, K. T., Merghadi, A., Yunus, A. P., Pham, B. T., & Dou, J. (2019). Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. Scientific Reports, 9(12296), 1603-1604.

Lee, S., Hong, S. M., & Jung, H. S. (2017, Jan). A Support Vector Machine for Landslide Susceptibility Mapping in Gangwon Province, Korea. Sustainability, 9(1). https://doi.org/ARTN 48

10.3390/su9010048

Li, Y., & Chen, W. (2020, Jan). Landslide Susceptibility Evaluation Using Hybrid Integration of Evidential Belief Function and Machine Learning Techniques. Water, 12(1). https://doi.org/ARTN 113

10.3390/w12010113

Nguyen, V. V., Pham, B. T., Ba, T. V., Prakash, I., & Bui, D. T. (2019). Hybrid Machine Learning Approaches for Landslide Susceptibility Modeling. Forests, 10(157), 1-27.

Nhu, V. H., Zandi, D., Shahabi, H., Chapi, K., Shirzadi, A., Al-Ansari, N., Singh, S. K., Dou, J., & Nguyen, H. (2020, Aug). Comparison of Support Vector Machine, Bayesian Logistic Regression, and Alternating Decision Tree Algorithms for Shallow Landslide Susceptibility Mapping along a Mountainous Road in the West of Iran. Applied Sciences-Basel, 10(15). https://doi.org/ARTN 5047

10.3390/app10155047

Phong, T. V., Phan, T. T., Prakash, I., Singh, S. K., Shirza di, A., Chapi, K., Ly, H. B., Ho, L. S., Quoc, N. K., & P ham, B. T. (2019, Sep 16). Landslide susceptibility mod eling using different artificial intelligence methods: a ca se study at Muong Lay district, Vietnam. Geocarto Inter national. https://doi.org/10.1080/10106049.2019.166571 5

Roy, J., Saha, S., Arabameri, A., Blaschke, T., & Bui, D. T. (2019). A Novel Ensemble Approach for Landslide Susceptibility Mapping (LSM) in Darjeeling and Kalimpong Districts, West Bengal, India. Remote Sensing, 11(23).

Sahin, E. K., Colkesen, I., Acmali, S. S., Akgun, A., & Aydinoglu, A. C. (2020). Developing comprehensive geocomputation tools for landslide susceptibility mapping: LSM tool pack. Computers & Geosciences, 104592.