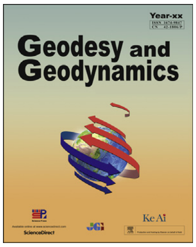
Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/geog;
http://www.jgg09.com/jweb_ddcl_en/EN/volumn/home.shtml



Groundwater level prediction of landslide based on classification and regression tree

Yannan Zhao[a,b,*], Yuan Li[a,b], Lifen Zhang[a,b], Qiuliang Wang[a,b]

[a] Key Laboratory of Earthquake Geodesy, Institute of Seismology, China Earthquake Administration, Wuhan 430071, China

[b] Wuhan Base of Institute of Crustal Dynamics, China Earthquake Administration, Wuhan 430071, China

ARTICLE INFO

Article history:
Received 23 June 2016
Accepted 5 July 2016
Available online 23 August 2016

Keywords:
Landslide
Groundwater level
Prediction
Classification and regression tree
Three Gorges Reservoir area

ABSTRACT

According to groundwater level monitoring data of Shuping landslide in the Three Gorges Reservoir area, based on the response relationship between influential factors such as rainfall and reservoir level and the change of groundwater level, the influential factors of groundwater level were selected. Then the classification and regression tree (CART) model was constructed by the subset and used to predict the groundwater level. Through the verification, the predictive results of the test sample were consistent with the actually measured values, and the mean absolute error and relative error is 0.28 m and 1.15% respectively. To compare the support vector machine (SVM) model constructed using the same set of factors, the mean absolute error and relative error of predicted results is 1.53 m and 6.11% respectively. It is indicated that CART model has not only better fitting and generalization ability, but also strong advantages in the analysis of landslide groundwater dynamic characteristics and the screening of important variables. It is an effective method for prediction of ground water level in landslides.

How to cite this article: Zhao Y, et al., Groundwater level prediction of landslide based on classification and regression tree, Geodesy and Geodynamics (2016), 7, 348–355, http://dx.doi.org/10.1016/j.geog.2016.07.005.

1. Introduction

Groundwater is a key factor in the formation, occurrence and development of landslides in reservoir bank. Long-term prediction of the underground water level is not only the prerequisite for long-term prediction of slope stability in reservoir bank, but also the key for ensuring the safe operation of the reservoir [1]. According to statistics, more than 90% of the rocky slopes damage and groundwater are related, and 30%–40% of

[*] *Corresponding author*. Key Laboratory of Earthquake Geodesy, Institute of Seismology, China Earthquake Administration, Wuhan 430071, China.

E-mail address: yannan-1985@163.com (Y. Zhao).

Peer review under responsibility of Institute of Seismology, China Earthquake Administration.

Production and Hosting by Elsevier on behalf of KeAi

http://dx.doi.org/10.1016/j.geog.2016.07.005

the dam accident damage caused by the groundwater flow [2]. The Three Gorges Dam begins to store high water level in October each year and to flood the low water level at next flood season, and the reservoir level since 2009 started to keep 145 m–175 m fluctuations. The significant and cyclical fluctuations of water level coupled with the impact of heavy rainfall produce dramatic changes in the groundwater power field of landslide area, thereby threatening a large number of landslides stability located in both sides of reservoir [3]. Therefore, the dynamic prediction of groundwater level plays a key role in the evaluation of landslide stability.

The groundwater level in the landslide has a complex nonlinear relationship between the natural and anthropogenic factors with uncertain characteristics of randomness and fuzziness, so it is difficult to express a deterministic model [4]. Jiang et al. [5] built radial basis function-neural network model, which was to predict the groundwater level of Huanglashi landslide. Liu [6] built a dynamic prediction model of underground water level by the genetic algorithm (GA) and the back propagation neural network (BPNN) which was used in a specific water resource spot. Peng et al. [7] analyzed the response relationship between influential factors and ground water level, and used a nonlinear genetic algorithm and support vector regression (GA-SVR) model to predict the values of ground water level in landslides. Sangrey [8] using deterministic methods and numerical simulation studied the groundwater level change with external factors (rainfall, reservoir storage, etc.). Such studies have achieved good results.

However, neural network method has the problems of local minima and selection initial weights. Although some methods such as genetic algorithms, support vector machines can avoid local minima, they generally require a huge amount of computation [9,10]. In contrast, the decision tree method can easily handle changing data and filter out scientifically important variables for classification and regression. At the same time, the method can be achieved simply with short training time and generated easily understandable rules.

Decision tree classification and regression tree algorithm (CART) is being explored in terms of statistical analysis and data mining, which can use the form of regression equation to predict continuous variables and effectively deal with non-modeling and solving linear problems [11–13]. Based on this, the groundwater level monitoring data of Shuping landslide in Three Gorges Reservoir area was used to analyze the response relationship between landslide groundwater level change and reservoir water, rainfall and other factors, and then the CART prediction model was established to dynamically predict the landslide groundwater level through the influence factors.

2. Model theory

2.1. CART building

CART, a recursive partitioning method, builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The classic CART algorithm was popularized by Breiman et al. [14,15].

In the case of a categorical variable, the number of possible splits increases quickly with the number of levels of the categorical variable. Thus, it is useful to tell the software the maximum number of levels for each categorical variable. In choosing the best splitter, the program seeks to maximize the average "purity" of the two child nodes. In CART, Gini coefficient is used to measure the "purity", and its mathematical definition is:

$$G(t) = 1 - \sum_{j=1}^{k} p^2(j|t) \tag{1}$$

where t, k and p are the node, the number of categories in output variables, and the probability when the sample output variables take the probability of j for the node t. It reaches its minimum (zero) when all cases in the node fall into a single target category.

CART uses Gini coefficient to measure reduction of the heterogeneity, and its mathematical definition is:

$$\Delta G(t) = G(t) - \frac{N_r}{N} G(t_r) - \frac{N_l}{N} G(t_l) \tag{2}$$

where $G(t)$ and N are the Gini coefficient of output variables and sample size before grouping, $G(t_r)$, N_r, $G(t_l)$ and N_l are respectively the Gini coefficient and sample size of right subtree, the Gini coefficient and the sample size of the left subtree after grouping. We can get point of division whose heterogeneity decreasing the fastest by repeating the method.

In the regression tree (continuous dependent variable), strategy of determining the optimal grouping variable is the same as the classification tree, and the main difference is variance as the measure indicator for output variable heterogeneity. Its mathematical definition is:

$$R(t) = \frac{1}{N-1} \sum_{i=1}^{N} (y_i(t) - \overline{y}(t))^2 \tag{3}$$

where t, N, $y_i(t)$ and $\overline{y}(t)$ are the node, the sample size for the node t, the output in a variable's value for t, and the average of the output variables for the node t. Therefore, the measure of heterogeneity decreasing is variance reduction, its mathematical definition is:

$$\Delta R(t) = R(t) - \frac{N_r}{N} R(t_r) - \frac{N_l}{N} R(t_l) \tag{4}$$

where R (t) and N are the variance of output variable and sample size before grouping, $R(t_r)$, N_r, $R(t_l)$ and N_l are respectively the variance and sample size of right subtree, the variances and the sample size of the left subtree after grouping. To achieve maximum of $\Delta R(t)$ variable should be the best grouping variable. The method for determining the best point of division is the same as the classification tree.

2.2. CART pruning

Due to the complete decision tree on the feature of training samples is described too accurate, it loses general representation and cannot be used in the classification or prediction of new data. This phenomenon is called over-fitting. Pruning is a