Apply data mining to analyze the rainfall of landslide

Chou-Yuan Lee[1], Zne-Jung Lee[2,a] , Bin-Yu Peng[3,4], Chen-chen Lin[2], and Hsiang Huang[4]

[1]*Department of Information Technology, Fuzhou University of International Studies and Trade, Fujian, China*
[2]*Department of Information Management, Huafan University, Taiwan*
[3]*Department of Information Management, University of Kang Ning, Taiwan*
[4]*Department of Mechatronic Engineering, Huafan University, Taiwan*

Abstract. Taiwan is listed as extremely dangerous country which suffers from many disasters. The disasters from the landslide result in the loss of agricultural productions, life and property and so on. Many researchers concern about the disasters of landslide, but there are few discussions for the threshold of rainfall for landslide. In this paper, data mining is applied to establish rules and the threshold of rainfall for landslide in Huafan University, Taiwan. These used variables include rainfall, insolation, insolation rate, averaged humidity, averaged temperature, wind speed, and the tilt of inclinometer. The inclinometer is an important instrument for measuring tilt, elevation or depression of an object with respect to gravity. There are 26 inclinometers in Talun mountain area of Huafan University. In this research, the used data were collected from January 2008 to July 2014. In the proposed approach, the regression analysis is used to predict rainfall first. Then, decision tree is used to obtain decision rules and set the threshold of rainfall for landslide. The output of this approach can provide more information for understanding the change of rainfall. The threshold of rainfall could also provide useful information to maintain the security for Huafan University.

1 Introduction

The United Kingdom risk management consultant company, Verisk Maplecroft, announced that United States, Japan, China, and Taiwan were suffered extreme dangerous natural disasters in 2011 [1]. These above countries are still suffering severe damages due to the impact of climate in 2015. The flood is one of the severe damages that concentrate in Asia, especially in India, Bangladesh, China, and Taiwan. These floods cause 75% risk of death in the world, and about annual 2.2 million people are affected by the effect of mountain landslide. Additionally, there are about annual 78 million people expose in the dangerous of tropical cyclone [2]. Taiwan is listed as extremely dangerous country which suffers from many natural disasters. The rainfall of tropical cyclone is one of the most important issues in Taiwan, because it direct links to voluminous rainfalls [3].

Recently, there are many researches focus on rainfall. These researches include statistical approaches, ensemble, Bayesian model, and data mining [3-12]. Above techniques provide useful information for the rainfall. Unfortunately, there are no information of inclinometer and few discussions for the threshold of rainfall for landslide. The inclinometer is an important instrument for measuring tilt, elevation or depression of an object with respect to gravity. It provides the direct

[a] Corresponding author : johnlee@cc.hfu.edu.tw

response to the settlement and displacement of various stratigraphic depth distributions via the slopes of the tilt tube [13]. There are 26 inclinometers at Huafan University where is located on Talun mountain area of Shihding District in New Taipei City, Taiwan. Talun mountain area belongs to slope land. Because data mining techniques provide useful information for the rainfall of landslide [14], it is an important issue to analyze the rainfall of landslide and inclinometer data for Huafan University. In this paper, data mining is applied to analyze the rainfall of landslide for Huafan university.

The rest of this paper is organized as follows. Because regression analysis and decision tree play important roles in the proposed approach, section 2 provides the brief literature overview of multiple regression and decision tree. Section 3 describes the proposed approach. Section 4 outlines the simulation results. It also provides detailed comparisons. Finally, conclusions are drawn in the last section.

2 The brief description of multiple regression analysis and decision tree

In this paper, multiple regression analysis is used to provide the predict model of the rainfall of landslide. Furthermore, decision tree is used to find the threshold of rainfall. In this section, multiple regression analysis and decision tree are brief described first.

For multiple regression analysis, the data consist of m observations on a dependent variable Y and n independent variables, $x_1, x_2, \cdots, x_n$, The relationship between Y and $x_1, x_2, \cdots, x_n$ is expressed as Eq. (1).

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n + \varepsilon \tag{1}$$

Where $\alpha_0, \alpha_1, \cdots, \alpha_n$ are constants, α_0 is referred to as the intercept. $\alpha_1, \alpha_2, \cdots, \alpha_n$ represent the corresponding n regression coefficients, and ε is the random error. Using matrix in Eq. (1), the model equation can be expressed as [15]

$$Y = XB + E \tag{2}$$

The normal equation can be expressed as

$$(X'X)\tilde{B} = X'Y \tag{3}$$

Where $\tilde{B}$ is a vector of least squares estimates of B. The solution to matrix equation is written as [16]

$$\tilde{B} = (X'X)^{-1}X'Y \tag{4}$$

Decision tree is based on the greedy algorithm that utilizes a divide-and-conquer strategy to recursively construct decision rules [17]. It consists of the root node, internal nodes, branches, and leaves. Each decision tree represents a rule which categorizes data according to these attributes [18]. A node specifies an attribute (feature) in the dataset. A branch connects either two nodes or one node and a leaf. Each node has a number of branches which are labeled as the possible value of attribute in the parent node. Leaves are labeled as the decision value of classification. The tree-like structure is composed of a root node, a set of internal nodes, and a set of leaf nodes. When applied to the set of train patterns, *Info(S)* measures the average amount of needed information to identify the class of the pattern *S*.

$$Info(S) = -\sum_{j=1}^{k}\left\{[freq(C_j, \frac{S}{|S|})]\log_2[freq(C_j, \frac{S}{|S|})]\right\} \tag{5}$$

Where $|S|$ is the number of cases in the training set. C_j is a class for $j = 1,2,\cdots,k$ where k is the number of classes and $freq\left(C_j, \frac{S}{|S|}\right)$ is the number of cases used in C_j. To consider the expected information value $Info_x(S)$ for attribute X to the partition S, it can be stated as:

$$Info_x(S) = -\sum_{j=1}^{n}\left\{[(\frac{|S_j|}{|S|})Info(S_j)]\right\} \quad (6)$$

where n is the number of output for the attribute X, S_j is a subset of S corresponding to the j^{th} output and $|S_j|$ is the number of cases of the subset S_j. The information gain, *Gain(X)*, according to attribute X is shown as:

$$Gain(X) = Info(S) - Info_x(S) \quad (7)$$

For constructing the decision tree, the root node will branch to child node first. The recursive processes will be continued to branch the child node of training datasets until each leaf node in the dataset contains only one class of patterns, or until there is no improvement [18].

3 The proposed approach

The rainfall of Huafan University can be divided into: (a) From December to next April, this dry season rains more days than other periods. However, the cumulative rainfall leads to less rainfall intensity. (b) Plum rain season is from May to June, and the typhoon season is from July to September. From September to November, it is also typhoon season and northeast monsoon. Due to the rainfall levels in typhoon season increased significantly, the intensity of rainfall rises to the top from July to September in one year. Occasional typhoons bring rainfall and the northeast monsoon rainfall from October to November but accumulated rainfall and average rainfall intensity become less. In this paper, the multiple regression analysis and decision tree are applied to analyze the rainfall of landslide in Huafan University.

In this study, these used variables include rainfall, insolation, insolation rate, averaged humidity, averaged temperature, wind speed, and the tilt of inclinometer. The dependent (target) variable is rainfall and others are independent variables. The proposed approach is based on multiple regression analysis and decision tree (DT). The pseudo code of the proposed approach is listed as follow.

Procedure:

Pre-process data

Begin

Apply multiple regression analysis;

Apply decision tree;

Output the multiple regression results, threshold of rainfall, and decision rules;

End

In pre-process data, the missing values in the inclinometer data are imputed by 0. It means that there are no tilts for these used inclinometers. In the proposed approach, the multiple regression analysis is first used to predict rainfall. After multiple regression analysis, decision tree uses Eq. (7) to generate decision rules. These obtained threshold of rainfall and decision rules are provided for decision-making. Finally, the proposed approach outputs the multiple regression results, threshold of rainfall, and decision rules for decision-making.

4 Results

There are 79 collected datasets and 33 used variables in this study. These collected datasets are from Jan. 2008 to Oct. 2014 in Huafan University. These variables include rainfall, insolation, insolation rate, averaged humidity, averaged temperature, wind speed, and others are inclinometers (*C5~C32*). First, multiple regression analysis is applied to predict the rainfall of landslide in Huafan University. After the process of multiple regression analysis, the predict model of multiple regression analysis is shown in Eq. (8).

$$\begin{aligned} Rainfall = {} & (-240.2114) + (-1.88) \times Insolation + (-4.1553) \times InsolationRate \\ & +0.3586 \times AverageHumidity + 28.9926 \times AverageTemperature \\ & + 94.1142 \times WindSpeed + (-6.9486) \times C6 + (-5.2393) \times C7 \\ & +1.7772 \times C8 + 9.0128 \times C9 + 5.1209 \times C10 + (-1.2573) \times C11 \\ & +(-0.9317) \times C12 + 3.1387 \times C13 + 5.768 \times C14 + 9.2376 \times C15 \\ & +12.3311 \times C16 + 7.2773 \times C17 + 0.7927 \times C18 + 9.1294 \times C19 \\ & +16.2721 \times C20 + (-27.4859) \times C21 + (-4.3699) \times C22 \\ & +(-5.3602) \times C23 + (-3.1363) \times C24 + (-0.6239) \times C25 \\ & +(-9.1850) \times C26 + (-0.1621) \times C27 + (-3.8095) \times C28 \\ & +(-8.4103) \times C29 + (-6.1479\&) \times C30 + 4.5759 \times C31 \\ & +(-5.2223) \times C32 \end{aligned} \quad (8)$$

The residual standard error of multiple regression analysis is 132.1029. Thereafter, decision tree is used to generate decision rules. There are total 7 decision rules generated in this study, and these generated decision rules are shown in Table 1.

Table 1. The decision rule of rainfall.

No.	Rule	Rainfall
1	C7>=7.025	616mm
2	C7<7.025, Insolation Rate<12.75	465mm
3	C7<7.025, Insolation Rate>=12.75, Temperature>=19.21, and C32>=2.28	432mm
4	C7<7.025, Insolation Rate>=12.75, Temperature>=19.21, C32<2.28 and C18<-1.325	351mm
5	C7<7.025, Insolation Rate>=12.75, Temperature>=19.21, C32<2.28, and C18>=-1.325	210mm
6	C7<7.025, Insolation Rate>=12.75, Temperature<19.21, and Insolation Rate<19.87	257mm
7	C7<7.025, Insolation Rate>=12.75, Temperature<19.21, and Insolation Rate>=19.87	121mm

In Table 1, these rainfalls for decision rules of No. 3, 4, 5, 6, and 7 are not so large enough, so the decision rules of No. 1 and 2 are set as the threshold. For decision rule of No. 2 (*C*7<7.025, Insolation Rate<12.75), its rainfall is 465 mm. Finally, the decision rule of No. 1 is set as the important threshold. These threshold values must take more attention to beware of disaster emergency and dangerous. However, it is safe when the rainfall is below 465mm. It needs to prepare all necessary plans for disaster emergency, when the rainfall is between 466 mm and 616 mm. Moreover, it is dangerous as the rainfall is greater than 617 mm.

5 Conclusions

In this paper, multiple regression analysis is applied to provide the model for the predict rainfall of landslide in Huafan University. Decision tree is used to obtain decision rules and set the threshold of rainfall for landslide. There are seven decision rules for the rainfall of landslide in Huafan University. From decision rules, the threshold of rainfall is set as 465 mm and 616 mm. It could provide useful information to maintain the security for Huafan University.

Acknowledge

The authors would like to thank the Ministry of Science and Technology (MOST) of Taiwan for financially supporting this research under Contract No. MOST 106-2221-E-211-002, MOST 106-2632-M-211-001, MOST 106-2622-E-211-002-CC3, and MOST 105-2632-M-211-001.

References

1. Maplecroft, Natural Hazards Risk Atlas 2011, https://maplecroft.com/about/news/natural_hazards_2011.html (2011)
2. M. Dilley, S Chen Robert, D. Uwe, A. L. Lerner-Lam and A. Margaret, Natural Disaster Hotspots: A Global Risk Analysis, World Bank, Washington DC, http://earth.columbia.edu/news/2005/story03-29-05.html (2005)
3. C.C. Wei, Soft computing techniques in ensemble precipitation nowcast, Applied Soft Computing, **13** (2013)
4. S.T. Wang, C.L. Yen, G.T. Chen and S.L. Shieh, The characteristics of typhoon precipitation and the prediction methods in Taiwan area (III) (in Chinese), Hazards Mitigation Program Technical Report 74-51, National Science Council, Taiwan (1986)
5. T. Hall, H.E. Brooks and C.A. Doswell, Weather and Forecasting, **14**(3), 338-345 (1999)
6. M. Nasseri, K. Asghari and M.J. Abedini, Expert Systems with Applications, **35**(3), 1415-1421 (2008)
7. G.A. Fallah-Ghalhary, M. Mousavi-Baygi and M.H. Nokhandan, Research Journal of Environmental Sciences, **3**(4), 400-413 (2009)
8. G. Srivastava, S.N. Panda, P. Mondal and J. Liu, Journal of Hydrology, **395**(3), 190-198 (2010)
9. A. Talei, L.H.C. Chua and T.S.W. Wong, Journal of Hydrology, **391**(3-4), 248-262 (2010)
10. S.A. Asklany, K. Elhelow, I.K. Youssef and M.A. El-Wahab, Atmospheric Research, **101**(1), 228-236 (2011)
11. C.C. Wei, Weather and Forecasting, **27**(2), 438-450 (2012)
12. C.C. Wei, Journal of Hydrometeorology, **13**(2), 722-734 (2012)
13. P.E. Mikkelsen, Advances in inclinometer data analysis, In Symposium on Field Measurements in Geomechanics (2003)
14. M. Kannan, S. Prabhakaran and P. Ramachandran, Rainfall forecasting using data mining technique (2010)
15. H.W. Altland, Regression Analysis: Statistical Modeling of a Response Variable, Technometrics, **41**(4) (1999)
16. Chatterjee, Samprit and Ali S. Hadi. Regression analysis by example, John Wiley & Sons (2015)
17. J.R. Quinlan, Induction of Decision Trees, Machine Learning (1986)
18. J.R. Quinlan, C4.5: Programs For Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA (1993)