Science of the Total Environment 627 (2018) 744–755

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran

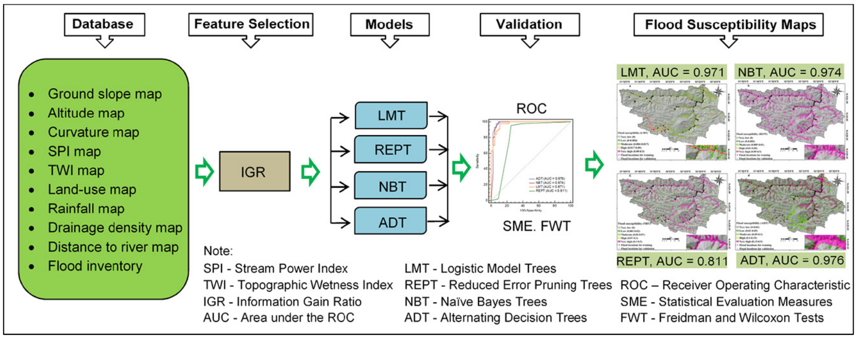
Khabat Khosravi [a], Binh Thai Pham [b,c,*], Kamran Chapi [d], Ataollah Shirzadi [d], Himan Shahabi [e], Inge Revhaug [f], Indra Prakash [g], Dieu Tien Bui [h]

[a] Department of Watershed Management Engineering, Faculty of Natural Resources, Sari Agricultural Science and Natural Resources University, Sari, Iran
[b] Geographic Information Science Research Group, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[c] Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[d] Departments of Watershed Sciences Engineering, Faculty of Natural Resources, University of Kurdistan, Sanandaj, Iran
[e] Departments of Geomorphology, Faculty of Natural Resources, University of Kurdistan, Sanandaj, Iran
[f] Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, PO-5003IMT, AAs, Norway
[g] Department of Science & Technology, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Government of Gujarat, Gandhinagar, India
[h] Geographic Information System Group, Department of Business and IT, University College of Southeast Norway, Gulbringvegen 36, N-3800 Bø i Telemark, Norway

HIGHLIGHTS

- Machine learning models namely LMT, REPT, NBT and ADT were used for flood assessment.
- Out of four models, the ADT has the highest performance for flood assessment.
- Advanced Decision Trees methods are promising for flood assessment in prone areas.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 September 2017
Received in revised form 12 January 2018
Accepted 26 January 2018
Available online xxxx

Editor: Ouyang Wei

Keywords:
Flood susceptibility assessment
Machine learning
Alternating Decision Trees
Logistic Model Trees
Reduced Error Pruning Trees
Naïve Bayes Trees

ABSTRACT

Floods are one of the most damaging natural hazards causing huge loss of property, infrastructure and lives. Prediction of occurrence of flash flood locations is very difficult due to sudden change in climatic condition and manmade factors. However, prior identification of flood susceptible areas can be done with the help of machine learning techniques for proper timely management of flood hazards. In this study, we tested four decision trees based machine learning models namely Logistic Model Trees (LMT), Reduced Error Pruning Trees (REPT), Naïve Bayes Trees (NBT), and Alternating Decision Trees (ADT) for flash flood susceptibility mapping at the Haraz Watershed in the northern part of Iran. For this, a spatial database was constructed with 201 present and past flood locations and eleven flood-influencing factors namely *ground slope, altitude, curvature, Stream Power Index (SPI), Topographic Wetness Index (TWI), land use, rainfall, river density, distance from river, lithology, and Normalized Difference Vegetation Index (NDVI)*. Statistical evaluation measures, the Receiver Operating Characteristic (ROC) curve, and Freidman and Wilcoxon signed-rank tests were used to validate and compare the prediction capability of the models. Results show that the ADT model has the highest prediction capability for flash flood susceptibility assessment, followed by the NBT, the LMT, and the REPT, respectively. These techniques have proven successful in quickly determining flood susceptible areas.

© 2018 Elsevier B.V. All rights reserved.

[*] Corresponding author at: Ton Duc Thang University, Ho Chi Minh City, Viet Nam.
E-mail address: phamthaibinh@tdt.edu.vn (B.T. Pham).

https://doi.org/10.1016/j.scitotenv.2018.01.266
0048-9697/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

All over the world, floods affect >20,000 lives per year (Sarhadi et al., 2012). In Asia, about 90% of all human losses are due to natural hazards mostly caused by floods (Dutta and Herath, 2004; Smith, 2013). Flooding occurs when a river's discharge exceeds its channel's capacity causing the river to overflow its floodplain. The most common cause of flooding is prolonged heavy rainfall (Casale and Margottini, 1999). A flash flood is a rapid flooding of geomorphic low-lying areas caused by extremely heavy rainfall in short time and also due to sudden dam or levee breaks, rock slide and/or mudslides (debris flow) (Elkhrachy, 2015). Iran has recently experienced many devastating flash floods in northern parts of the country at Noshahr (2012), Neka (2013), Behshahr (2013), and Sari City (2015) (Khosravi et al., 2016).

Main aim of the present flood modeling is to develop flood susceptibility maps in a frequently flood affected watershed. Due to complex, non-linear and dynamic structure of watersheds, floods cannot be modeled using simple non-linear hydrological models (Sahoo et al., 2009). Therefore, the problem of flood forecasting and mapping of some physically-based rainfall-runoff models still exist (Sahoo et al., 2009). One of the key solutions in future flash flood management and mitigation is the detection of flood-prone areas using appropriate methods with high precision (Youssef et al., 2011a).

There are many statistical and machine leaning methods available for flood susceptibility modeling. Statistical models for the flood prediction which include frequency ratio (Lee et al., 2012; Youssef et al., 2016), weights-of-evidence (Tehrany et al., 2014; Youssef et al., 2015b), and multiple criteria decision methods (Papaioannou et al., 2015; Stefanidis and Stathis, 2013; Youssef et al., 2011b). In recent years, machine leaning methods such as artificial neural networks (Radmehr and Araghinejad, 2014), logistic regression (Youssef et al., 2015a), support vector machines (Tehrany et al., 2014) and decision trees (Tehrany et al., 2013) were investigated for flood modeling with promising results. Among these methods, Decision Trees (DT) is a good method for flood susceptibility mapping and it has shown high prediction performance (Tehrany et al., 2013) However, the use of DT models for flash flood assessment is still limited.

The DT provides a transparent tree-like structure with easily interpreted rules (Tien Bui et al., 2016a). Other advantages of the DT method are: (1) it is a type of statistical analysis with no statistical distribution assumption, (2) it can handle data from various scales, (3) it permits identification of homogeneous groups with various susceptibility levels, and (4) it facilitates the construction of rules for prediction of complex relationships (Tehrany et al., 2013). The DT can also be used for the real time flood forecasting with respect to water level rise and water flow (Han et al., 2002).

Logistic Model Trees (LMT), Reduced Error Pruning Trees (REPT), Naïve Bayes Trees (NBT) and Alternating Decision Trees (ADT) are advanced DT methods. Therefore, the main objective of this study is to apply these models (LMT, REPT, NBT, and ADT) in the study area and compare results for the selection of best flash flood susceptibility assessment model. The Haraz Watershed (Mazandarn Province) which is a flash flood prone area of northern Iran was selected as the study area. Statistical evaluation measures, the Receiver Operating Characteristic (ROC) curve, and Freidman and Wilcoxon signed-rank tests were used to validate and compare the predictive capability of the models. Data processing and modeling were done using Arc map 10.2 and Weka 3.7.12 software.

2. Background of the methods used

2.1. Decision trees algorithms

2.1.1. Logistic Model Trees (LMT)

The LMT is a classification method, which combines decision trees (C4.5 algorithm) and logistic regression machine learning methods. These methods are based on an earlier idea of a model that composes of a tree structure with a set of inner nodes and leaves or terminal nodes (Quinlan, 1993). The C4.5 algorithm is used at the nodes and logistic regression function which is used at the leaves (Quinlan, 1993). Linear logistic regression is used to find the posterior probability in a leaf node with the following equation:

$$P(N|x) = \exp(L_i(x))/\sum_{i=1}^{n} \exp(L_i(x)) \tag{1}$$

where $P(N|x)$ is the posterior probability in a leaf node of N number of classes in the input vector x and L_i (x) is the least-square fits given by the following equation:

$$L_i(x) = \sum_{i=1}^{n} \alpha_i x_i + \alpha_o = 0 \tag{2}$$

where n is the number of influencing factors, α_o, α_i are the coefficients of the component of vector $x = x_i$ which represents the influencing factors.

2.1.2. Reduced Error Pruning Trees (REPT)

The REPT is a hybrid approach of the Reduced Error Pruning (REP) method and the DT method, which builds a decision or regression tree based on the information gain/variance reduction (Quinlan, 1987). The DT builds the classification tree by looking for the input variable with the highest gain ratio calculated as below (Tien Bui et al., 2012):

$$Gain\ Ratio(x,Z) = \frac{Entropy(Z) - \sum_{i=1}^{n} \frac{|Z_i|}{|Z|} Entropy(Z_i)}{-\sum_{i=1}^{n} \frac{|Z_i|}{|Z|} \log_2 \frac{|Z_i|}{|Z|}} \tag{3}$$

where the attribute x belongs to a training dataset Z with subsets Z_i, $i = 1, 2, \ldots, n$.

The pruning method REP is used to decrease the complicity of the tree structure of the DT method (Mohamed et al., 2012) as it can remove some leaves and branches of the tree which provide little power for classification (Galathiya et al., 2012). The REP is one of the simplest and most popular pruning methods (Quinlan, 1987) in comparison with feature selection and cross validation (Galathiya et al., 2012). Advantage of this method is that it is not only able to reduce the complexity of the tree structure but also prevent the over-fitting problem during learning process without a significant accuracy loss (Polo et al., 2008).

2.1.3. Naive Bayes Trees (NBT)

The NBT is an ensemble of Naïve Bayes (NB) and the DT models based on Bayes' theorem (Kohavi, 1996). It is one of the popular classification methods due to its simplicity, efficiency, excellent performance, and interpretability. This method requires little computer memory and is very quick to learn from a training set (Wang et al., 2015a). During the construction of the tree, pre-pruning is done in either of the two ways: (1) the data at the node is split or (2) a leaf is created that contains a local NB model trained on the data at that specific node (Landwehr et al., 2005). Kohavi (1996) stated that the NBT model gives improved performance in comparison to individual DT and NB models.

The NBT model according to entropy concept follows the attribute selection measure to growing trees. If D be a set of cases and $|D|$ be a total number of cases, then these cases can be classified into m classes as Di $(i = 1,2, \ldots m)$ where $|Di|$ is the number of the cases that belongs to the class $|Di|$. Values of entropy for classifying the set D can be estimated as follows:

$$Entropy(D) = -\sum_{i=1}^{m} (|D_i|/|D|) \log_2[|D_i|/|D|] \tag{4}$$

The NB algorithm considers that predictive attributes $a_1, a_2, \ldots, a_m$ are independent. Given class attribute C_j, a class attribute of class set C,

Download English Version:

https://daneshyari.com/en/article/8860878

Download Persian Version:

https://daneshyari.com/article/8860878

Daneshyari.com