Catena xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Catena

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

Prediction of the landslide susceptibility: Which algorithm, which precision?

Hamid Reza Pourghasemi[a,*], Omid Rahmati[b]

[a] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran
[b] Department of Watershed Management, Faculty of Agriculture and Natural Resources, Lorestan University, Lorestan, Iran

ARTICLE INFO

Keywords:
Machine learning
Landslide spatial modeling
Variables importance
GIS
Iran

ABSTRACT

Coupling machine learning algorithms with spatial analytical techniques for landslide susceptibility modeling is a worth considering issue. So, the current research intend to present the first comprehensive comparison among the performances of ten advanced machine learning techniques (MLTs) including artificial neural networks (ANNs), boosted regression tree (BRT), classification and regression trees (CART), generalized linear model (GLM), generalized additive model (GAM), multivariate adaptive regression splines (MARS), naïve Bayes (NB), quadratic discriminant analysis (QDA), random forest (RF), and support vector machines (SVM) for modeling landslide susceptibility and evaluating the importance of variables in GIS and R open source software. This study was carried out in the Ghaemshahr Region, Iran. The performance of MLTs has been evaluated using the area under ROC curve (AUC-ROC) approach. The results showed that AUC values for ten MLTs vary from 62.4 to 83.7%. It has been found that the RF (AUC = 83.7%) and BRT (AUC = 80.7%) have the best performances comparison to other MLTs.

1. Introduction

In general, landslides as one of the natural earth surface processes and an example of land degradation; it change features of landscape, reduces physical extent of the soil of ecosystems (http://www.ccma.vic.gov.au/soilhealth/resource/definitions.htm), causes to erosion and sediment yield and loss of soil resources (Montanarella, 2003; Keesstra et al., 2015), and subsequently damage to houses and basic infrastructures, agricultural lands, and economics and human welfare (Corominas et al., 2014; Papathoma-Köhle et al., 2015; Schilirò et al., 2016). Through destructive impacts of landslides and their consequences, research institutions and governments have long attempted to delineate landslide susceptible areas for improving disaster preparedness and damage prevention (An et al., 2016; Betts et al., 2017). Hence, the identification of landslide prone areas not only provides an insight into control of land degradation; but, can also form a basis for safer strategic planning of future developmental activities in the region (Atkinson and Massari, 1998; Mertens et al., 2016). This process aims to highlight the spatial distribution of potentially unstable slopes based on link the past landslide events according to landslide causing variables that are responsible for the occurrence of landslides in a region (Hong et al., 2015).

In landslide susceptibility modeling, quantitative approaches are grouped into physically based models, statistics based correlation analysis, and soft computing techniques. Physically based models (Thanh and De Smedt, 2014) require detailed data of geotechnical engineering and geological aspect of the slope failure at site specific in regional scale (Tien Bui et al., 2016). These models are quite expensive and not practical for large scale areas (van Westen and Terlien, 1996). Traditional statistical models, which assume an appropriate structural model and then focus on parameterizing it, are widely used for analyzing of natural hazards such as landslides (Yesilnacar and Topal, 2005; Pourghasemi and Kerle, 2016). Classification of each landslide conditioning factors in traditional statistical models is a key point that affects the quality of landslide susceptibility map (Costanzo et al., 2012) and has been deeply discussed in Chacon et al. (2006). In contrast, machine learning techniques, a powerful group of data driven tools, use algorithms to learn the relationship between a landslide occurrence and landslide related predictors, and avoids starting with an assumed structural model (Elith et al., 2008; Dickson and Perry, 2016). Romer and Ferentinou (2016) stated that to obtain more reliable results through the statistical methods, large amounts of data are required, whereas ML-based models can effectively overcome the limitation of data dependent bivariate and multivariate statistical methods (Pham et al., 2016). Furthermore, other advantages of these models are that no statistical assumptions are made, and the nonlinear character of landslides is also considered (Ferentinou and Chalkias, 2013; Zare et al., 2013; Kornejady et al., 2017a). These MLTs allow handle data from various measurement scales, any type of independent variable (i.e. ratio, interval, nominal, or ordinal), and without needing to define

* Corresponding author.
E-mail address: hr.pourghasemi@shirazu.ac.ir (H.R. Pourghasemi).

https://doi.org/10.1016/j.catena.2017.11.022
Received 30 May 2017; Received in revised form 31 October 2017; Accepted 27 November 2017

Table 1
MLTs adopted in the several studies on the landslide susceptibility mapping.

Technique	Examples of previous studies
Decision tree (DT)	Saito et al. (2009), Nefeslioglu et al. (2010), Pradhan (2013), Tien Bui et al. (2016), Kavzoglu et al. (2014a, 2014b), Wu et al. (2014)
AdaBoost (AB)	Micheletti et al. (2014)
Generalized additive model (GAM)	Brenning (2008), Goetz et al. (2011), Vorpah et al. (2012), Goetz et al. (2015)
Artificial neuronal networks (ANN)	Lee et al. (2004), Yesilnacar and Topal (2005), Pradhan and Lee (2010), Yilmaz (2010), Zare et al. (2013), Thai Pham et al. (2015), Arnone et al. (2016)
Adaptive Neuro-Fuzzy Inference System (ANFIS)	Pradhan et al. (2010), Oh and Pradhan (2011), Sezer et al. (2011), Nasiri Aghdam et al. (2016)
Random forest (RF)	Trigila et al. (2015), Pourghasemi and Kerle (2016), Hong et al. (2016b), Youssef et al. (2016),
Classification and regression trees (CART)	Felicísimo et al. (2013), Vorpah et al. (2012)
Multiple adaptive regression splines (MARS)	Vorpah et al. (2012), Felicísimo et al. (2013), Conoscenti et al. (2015)
Boosted regression tree (BRT)	Dickson and Perry (2016), Youssef et al. (2016)
Maximum entropy (MaxEnt)	Felicísimo et al. (2013), Park (2015), Hong et al. (2016a), Kornejady et al. (2017)
Naive Bayes (NB)	Tien Bui et al. (2012), Thai Pham et al. (2015), Tsangaratos and Ilia (2016)
Support vector machine (SVM)	Yao et al. (2008), Yilmaz (2010), Marjanović et al. (2011), Micheletti (2011), Tien Bui et al. (2012), Pourghasemi et al. (2013), Hong et al. (2015)
General linear model (GLM)	Ayalew et al. (2004), Brenning (2005), Vorpah et al. (2012), Youssef et al. (2016)
Quadratic discriminant analysis (QDA)	Rossi et al. (2010)

normally distributed transformed variables (Ferentinou and Chalkias, 2013; Zare et al., 2013; Hong et al., 2016b). Table 1 gives a summary of prior literature on the most used different machine learning techniques for landslide susceptibility modeling. However, a comparative study on landslide susceptibility maps produced by ANN, BRT, CART, GLM, MARS, NB, QDA, RF, SVM, and GAM has not been commonly encountered in the literature. For this reason, a comparison among these relatively new approaches is needed to estimate spatial landslide susceptibility to select the best model for regional analyses.

Landslide as a land degradation driving force is one of the major geo-environmental issues in humid areas, also involving the Mazandaran Province in northern Iran (Zare et al., 2013). Significant landslide hazard exists in Mazandaran Province with the highest landslide activity—due to their steep topography, wet climate, and high weathering rates (Choobbasti et al., 2009; Vahidnia et al., 2010; Pourghasemi and Kerle, 2016). Here, this study gives a comprehensive comparison of the performance of ten state-of-the-art machine learning models by focusing on their main distinctive characteristics. For this purpose, Ghaemshahr region (in the Mazandaran Province) was selected as study area. According to literature, the machine learning models are new yet in the area of landslide susceptibility assessment compared to other methods. The literature review showed that a further comparative study among different MLTs is needed to better understand the issue, develop innovative technologies, assessment of the effectiveness of models, and other assistance for improving the prediction of landslide susceptibility and land degradation mitigation (Rossi et al., 2010; Goetz et al., 2015; Youssef et al., 2016). Our main objectives are to: (1) investigating the predictive performance of ten machine learning (ML)-based techniques, including ANN, BRT, CART, GLM, GAM, MARS, NB, QDA, RF, and SVM, (2) comparing the accuracy of models through the receiver operating characteristic (ROC) curve method for selection of the best technique for regional landslide susceptibility modeling in the study area, and (3) analyzing the relative importance of landslide conditioning factors using LVQ method. This understanding is fundamental for producing accurate and reliable landslide susceptibility maps and allows drawing more robust conclusions about capability of different MLTs. In addition, the results of this study is a valuable help to local authorities in sustainable land management and combat land degradation and also add some knowledge to understand the relation between the geo-environmental variables and distribution of landslides.

2. Material and methods

2.1. Study area

The Mazandaran Province is situated in southern coast of the Caspian Sea (i.e. the largest lake in the world). The study area (Ghaemshahr Region) is located in the eastern part of the Mazandaran Province, Iran, between latitudes of 35° 54′ to 36° 26′ N, and longitudes of 52° 36′ to 53° 15′ E (Fig. 1). The climate of the study area is humid, which characterized by warm summer and mild winter (Rodionov, 1994), with rainfall varying between 600 and 940 mm, and a mean annual rainfall of 729 mm (Water Resources Company of Mazandaran (WRCM), 2015). It covers an area of about 2241 km^2 and its altitude ranges from 30 to 3810 m above sea level (a.s.l.), with an average of 1218.7 m a.s.l. In winter, the temperature ranges from −1.4 °C to 15.5 °C, while in summer it varies from 20.9 °C to 32.6 °C (Water Resources Company of Mazandaran (WRCM), 2015).

The population of the study area is about 300,000 inhabitants based on population census data from the Iranian Statistical Institute (ISI) (2016). In 2016, the total area of cultivated land in Ghaemshahr Region was around 264.5 km^2 (11.87%). The period of major deforestation was between 2006 and 2011, which was due to illegal wood cutting, cultivation of crops, and residential uses of fuel wood. These forests are highly susceptible to fire (i.e. especially during exceptionally dry years) which potentially is impressive on runoff, erosion, and landslide processes (Shafiei et al., 2010).

From a morphological viewpoint, the southern Coasts of the Caspian Sea have been classified into five zones including Western Gilan, Central Gilan, Central Mazandaran, Western Mazandaran, and Golestan (Khoshravan, 1998, 2007). The study area has been identified as "Central Mazandaran", which the most of its stratigraphic units are related to Miocene and Quaternary including marl, swamp and marsh, and dark shale that are susceptible to landslide occurrence (Choobbasti et al., 2009). The heavy rainfalls often trigger landslides in the region which are undoubtedly one of the mightiest and most devastating forces of nature. Therefore, landslides are a major geo-hazard in the study area, where numerous slope failures have occurred in the past decade and highly likely to occur in the future due to natural or anthropogenic causes. Most of them are shallow rotational landslides occurring in the last 15 years and along river incisions and/or failure surfaces. They have caused damages to infrastructure, buildings, and sources of livelihood earnings. Meanwhile, it should be noted that the population, economical, and ecological pressures of Ghaemshahr Region brings with it informal settlements and land degradation—due to over grazing and deforestation—onto potentially dangerous slopes in the area. In order to obtain a better visual understanding of landslides in the study area, some field photographs are presented in Fig. 2.

2.2. Methodology

The methodological approach carried out in this study is consisted