Catena 145 (2016) 164–179

Contents lists available at ScienceDirect

Catena

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

Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size

Paraskevas Tsangaratos [a,*], Ioanna Ilia [b]

[a] Mining and Metallurgical Engineering, National Technical University of Athens, School of Mining and Metallurgical Engineering, Department of Geological Studies, Zografou Campus: Heroon Polytechniou 9, 15780 Zografou, Greece
[b] National Technical University of Athens, School of Mining and Metallurgical Engineering, Department of Geological Studies, Zografou Campus: Heroon Polytechniou 9, 15780 Zografou, Greece

ARTICLE INFO

Article history:
Received 28 August 2015
Received in revised form 14 April 2016
Accepted 3 June 2016
Available online 10 June 2016

Keywords:
Landslide susceptibility
Logistic regression classifier
Naïve Bayes classifier
Geographical information system
Epirus
Greece

ABSTRACT

The main objective of the present study was to compare the performance of a classifier that implements the Logistic Regression and a classifier that employs a Naïve Bayes algorithm in landslide susceptibility assessments. The study provides an evaluation concerning the influence of model's complexity and the size of the training data, while it identifies the most accurate and reliable classifier.

The comparison of the two classifiers was based on the assessment of a database containing 116 sites located at the mountains of Epirus, Greece, where serious landslides events have been encountered. The sites are classified into two categories, non-landslide and landslide areas. The identification of those areas was established by analysing airborne imagery, extensive field investigation and the examination of previous research studies. The geo-environmental conditions in those locations where analyzed in regard with their susceptibility to slide. In particular, seven variables where analyzed: engineering geological units, slope angle, slope aspect, mean annual rainfall, distance from river network, distance from tectonic features and distance from road network.

Multicollinearity analysis and feature selection was implemented in order to estimate the conditional independence among the variables and to rank the variables according to their significance in estimating landslide susceptibility. By the above processes the construction of nine different datasets was accomplished. Further partition allowed creating subsets of training and validating data from the original 116 sites. Each dataset was characterized by the number of the variables used and the size of the training datasets.

The comparison and validation of the outcomes of each model was achieved using statistical evaluation measures, the receiving operating characteristic and the area under the success and predictive rate curves. The results indicated that model's complexity and the size of the training dataset influence the accuracy and the predictive power of the models concerning landslide susceptibility. In particular, the most accurate model with high predictive power was the eighth model (five variables and 92 training data), with the Naïve Bayes classifier having a slightly higher overall performance and accuracy than the Logistic Regression classifier, 87.50% and 82.61% on the validation datasets, respectively. The highest area under the curve was achieved by the Naïve Bayes classifier for both the training and validating datasets (0.875 and 0.806 respectively) while the Logistic Regression classifier achieved a lower AUC values for the training and validating datasets (0.844 and 0.711, respectively). When limited data are available it seems that more accurate and reliable results could be obtained by generative classifiers, like Naïve Bayes classifiers. Overall, landslide susceptibility assessments could serve as a useful tool for the local and national authorities, in order to evaluate strategies to prevent and mitigate the adverse impacts of landslide events.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Landslides are natural, geological phenomena that involve a wide range of soil, debris or rock mass movements that may occur in offshore, coastal and inland areas, driven by the force of gravity and the aid of water. These movements are identified as the result of the progressive

[*] Corresponding author.
E-mail addresses: ptsag@metal.ntua.gr (P. Tsangaratos), gilia@metal.ntua.gr (I. Ilia).

http://dx.doi.org/10.1016/j.catena.2016.06.004
0341-8162/© 2016 Elsevier B.V. All rights reserved.

or extreme evolution of natural events that are developed due to the action of geological, tectonic, geomorphological and climatic processes. In addition to these processes, it has been widely reported that many cases of landslides are developed as a result of the negative impact of human activities on the environment (Alexander, 1992; Hutchinson, 1995; Baioni, 2011; Alimohammadlou et al., 2013).

According to Varnes (1984), the risk against landslide phenomena, can be thought as the probability of occurrence of a potentially injurious event in a specified time period and in a given area. This definition contains two elements; space and time. The spatial variable specifies those areas that are susceptible to the development of failure at a given time, while the time variable specifies the time the event will occur in a given area.

In this context, the variables that determine the probability of occurrence of a landslide event can be divided into two categories (Dai et al., 2002): intrinsic variables that contribute to landslide susceptibility, such as geological and morphological characteristics of the earth's surface, geotechnical parameters, effects of vegetation cover, the hydrographic network, and external variables that are responsible for triggering landslides such as rainfall and seismic activity. If the external variables are not taken into account, the term susceptibility refers to the probability of the presence of a landslide event considering only the spatial dimension of the problem.

Korup and Stolle (2014) reports that the prediction through the use of various methods and techniques of the spatial distribution of landslides is by far the most investigated topic that aids land – use planning, decision making and overall landslide risk reduction strategies.

According to van Westen et al. (2006) a region could be considered to be prone to landslide phenomena when the geo – environmental conditions of the region share common features with a region where a failure has been manifested in the past. Thus, the susceptibility of the area could be defined by a set of geological, tectonic and hydrologic conditions, morphological characteristics, soil and vegetation features, land use and human practices.

In general, the analysis of landslide phenomenon is attempted through qualitative, semi – quantitative, quantitative methods and various modeling techniques (Fell et al., 2008; Rozos et al., 2008). The majority of the applied methods are based either on the experience and knowledge provided by experts or on statistical or probabilistic theories or even the use of deterministic models (Soeters and van Westen, 1996; Aleotti and Chowdhury, 1999; Castellanos Abella and van Westen, 2007; Fell et al., 2008).

Relatively recently, new techniques and methods where utilized as promising tools to evaluate the susceptibility and risk against landslides that come from the domain of Machine Learning and Data Mining (Flentje et al., 2007; Miner et al., 2010; Marjanovic et al., 2011; Pradhan, 2013; Tsangaratos et al., 2013; Korup and Stolle, 2014; Goetz et al., 2015). These methods are characterized by the ability of learning and discovering hidden and unknown patterns from large multi-thematic databases (Tsangaratos and Ilia, 2015).

Numerous papers can be found through the scientific literature that take advantage of their ability to sufficiently assess data, including: the logistic regression approach (Lee, 2004; Ayalew and Yamagishi, 2005; Lee and Sambath, 2006; Van Den Eeckhaut et al., 2006; Lee and Pradhan, 2007; Nefeslioglu et al., 2008; Das et al., 2010; Oh and Lee, 2010; Suzen and Kaya, 2011; Yalcin et al., 2011; Felicisimo et al., 2013; Pourghasemi et al., 2013a; Regmi et al., 2014; Hong et al., 2015), fuzzy logic method (Ercanoglu and Gokceoglu, 2002, 2004; Champatiray et al., 2007; Muthu et al., 2008; Pradhan et al., 2009, 2010a; Pradhan, 2010, 2011a, b; Akgun et al., 2012; Pourghasemi et al., 2012a; Tien Bui et al., 2012b; Feizizadeh et al., 2013, 2014; Zhu et al., 2014), artificial neural network method (Lee et al., 2003, 2004; Neaupane and Achet, 2004; Ermini et al., 2005; Ferentinou and Sakellariou, 2007; Caniani et al., 2008; Melchiorre et al., 2008; Nefeslioglu et al., 2008; Choi et al., 2010; Pradhan and Lee, 2010a, 2010b, 2010c; Pradhan et al., 2010a; Poudyal et al., 2010; Yilmaz, 2010; Tien Bui et al., 2012a; Zare et al., 2013; Alimohammadlou et al., 2014; Conforti et al., 2014; Tsangaratos and Benardos, 2014), Bayes theorem based on weights of evidence (Regmi et al., 2010a, 2010b; Kayastha et al., 2012; Pourghasemi et al., 2012b; Kouli et al., 2014; Ilia and Tsangaratos, 2015), neural–fuzzy method (Vahidnia et al., 2010; Pradhan et al., 2010b; Oh and Lee, 2011; Oh and Pradhan, 2011; Sezer et al., 2011; Sdao et al., 2013; Pradhan, 2013), support vector machines (Yao et al., 2008; Yilmaz, 2010; Xu et al., 2012; Ballabio and Sterlacchini, 2012; Tien Bui et al., 2012c; Pourghasemi et al., 2013b; Pradhan, 2013; Hong et al., 2015) and decision tree method (Saito et al., 2009; Yeon et al., 2010; Nefeslioglu et al., 2010; Tien Bui et al., 2012c; Pradhan, 2013; Tsangaratos and Ilia, 2015).

The present paper focuses on the quantitative methods that utilize statistical or probabilistic models and also Machine Learning and Data Mining methods, to assess the role of landslide – causative variables. Particular, the study addresses two methods, Logistic Regression (LR) and Naïve Bayes (NB) to develop appropriate classifiers in order to classify the research area into landslide or non-landslide zones. LR is a widely used statistical direct probability model, while NB is considered as a simple probabilistic model that is based on the Bayes' theorem.

In more details, LR has been utilized in numerous landslide susceptibility assessments, providing accurate and reliable results in a rather simple manner. Based on its learning mechanism it is characterized as a discriminative model which estimates the probability for a given feature (x) and the label (y) directly from the training data by minimizing error (Ng and Jordan, 2001). On the other hand NB has been employed in rather fewer studies presenting respectively high accuracy. Based on its learning mechanism it is referred to as a generative model since for the given features (x) and the label (y) it estimates a joint probability from the training data (Ng and Jordan, 2001). The two techniques further differ in the adopted assumptions and also limitations of the models performance. The NB model assumes that all the features are conditionally independent, while LR splits feature space linearly, thus it works even if some of the variables are correlated (John and Langley, 1995; Montgomery et al., 2001). As for models limitations, the NB has been reported to work well even with less training data, as the estimates are based on the joint density function, while LR produces results that over fit the data, a condition when a model begins to memorize training data rather than learning to generalize from trend (Melchiorre et al., 2008; Tsangaratos and Benardos, 2014). This was our initial intension, to find for each model its limitations. Specifically, through the implementation of the developed classifiers, two objectives were achieved; the construction of a landslide susceptibility map for each approach and the comparison of their performance in regard with the complexity of the developed models and the size of the training data used.

Concerning the second objective of the study, several studies have compared LR and NB with other qualitative, semi – quantitative and quantitative methods to determine the optimum mathematical method to assess landslide susceptibility (Yesilnacar and Topal, 2005; Lee and Sambath, 2006; Miner et al., 2010; Pradhan and Lee, 2010a, 2010b; Yilmaz, 2010; Akgun, 2012; Ballabio and Sterlacchini, 2012; Tien Bui et al., 2012c; Bijukchhen et al., 2013; Felicisimo et al., 2013; Pourghasemi et al., 2013a; Shahabi et al., 2014). However, there are relative few studies that evaluate the performance of the applied classifiers in regard with the complexity of the models and the size of the training data sets (Brenning, 2005; Nefeslioglu et al., 2008; Pradhan and Lee, 2010b; Wang et al., 2013; Heckmann et al., 2014).

The study area covers the mountains of Central Tzoumerka, which are located at the administrative unit of Epirus Greece, where serious landslides events have been encountered. The computation process was carried out using Microsoft Visual Studio 2010 Professional (Halvorson, 2010) for implementing the NB algorithm, Weka 3.7.6 for feature selection process (Hall et al., 2009) and SPSS 16.0 (SPSS, 2007) for implementing multicollinearity analysis and LR, while ArcGIS 10.1

Download English Version:

https://daneshyari.com/en/article/4570892

Download Persian Version:

https://daneshyari.com/article/4570892

Daneshyari.com