Geoderma 305 (2017) 314–327

Contents lists available at ScienceDirect

Geoderma

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

Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques

Wei Chen[a], Hamid Reza Pourghasemi[b,*], Aiding Kornejady[c], Ning Zhang[a]

[a] College of Geology & Environment, Xi'an University of Science and Technology, Xi'an 710054, China
[b] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran
[c] Department of Watershed Sciences and Engineering, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran

ARTICLE INFO

Keywords:
Spatial modeling
SVM
ANN
Maximum entropy
Ensemble models

ABSTRACT

"Spatial contraindication" is what exactly landslide susceptibility models have been seeking. They are designed for depicting perilous land activities, be it natural or anthropological. To find this pattern, three well-known machine learning models namely maximum entropy (MaxEnt), support vector machine (SVM), and Artificial Neural Network (ANN) were used accompanied by their ensembles (i.e. ANN-SVM, ANN-MaxEnt, ANN-MaxEnt-SVM, and SVM-MaxEnt) in Wanyuan area, China. The models were designed by eleven conditioning factors such as elevation, slope degree, slope aspect, profile and plan curvatures, topographic wetness index, distance to roads, distance to rivers, normalized difference vegetation index (NDVI), land use/land cover (LU/LC), and lithology along with two sets of training (213#) and testing (91#) landslide data. A statistical index (SI) model was implemented to examine the mutual relationship between classes of each factor and the landslide occurrences. Concerning the areal differentiation, the chi-square test was used where SVM and MaxEnt gained the highest and the lowest values, respectively. Afterward, the practicality — as an indicator of producing a focused susceptibility map and addressing highly susceptible classes (IV and V) in a compendious manner with a reduced spatial area — was calculated for models. Accordingly, SVM and MaxEnt were found to be the most and the least practical models having the highest and the lowest spatial area in highly susceptible classes, respectively. The receiver operating characteristic (ROC) curve was used to examine generalization and prediction accuracy of the models. As a result, in the case of validating models separately, ANN gained the highest area under the curve (AUC) with a value of 0.824, followed by SVM (0.819), and MaxEnt (0.75). In the case of validating ensemble models, the ANN-SVM had the highest AUC of all (0.826), followed by ANN-MaxEnt (0.803), SVM-MaxEnt (0.792), and ANN-MaxEnt-SVM (0.811). With regard to the premier model results, three factors namely distance from roads, elevation, and distance from rivers had the highest effect on landslide occurrence. The results of the SI values showed that the spatial combination of the main drivers namely farmlands, − 0.06–0.2 range in NDVI, rocks with inter-bedded limestone and other susceptible classes therein can make at least a prone area of about 30% to landsliding. Such spatial combination of environmental condition and human-made activities can be considered as a contraindication for the residents of the study area, especially at highly susceptible locations. This also addresses areas that further mitigation plans should be taken into account with urgency.

1. Introduction

Globally, landslides cause approximately 1000 deaths per year and a property damage of about US$ 4 billion (Pradhan and Youssef, 2010). According to Raja et al. (2016), landslides account for 17% of all deaths caused by natural hazards. Hence, one of the most tasks for landslide hazard and risk mitigation is to prepare landslide susceptibility maps (Aleotti and Chowdhury, 1999; Kelarestaghi and Ahmadi, 2009). Landslide susceptibility can be defined as the proneness of an area to generate landslides (Guzzetti et al., 2006; Raja et al., 2016). The reliability of landslide susceptibility maps mostly depends on the quantity and quality of available data, the working scale, and the selection of the appropriate methodology of analysis and modeling (Ayalew and Yamagishi, 2005).

With the increasing emphasis on the use of Geographic Information System (GIS) and Remote Sensing (RS), many researchers have applied heuristic, deterministic, statistic, and soft computing models to assess landslide susceptibility in different areas. Heuristic methods are commonly used to evaluate landslide susceptibility (Pourghasemi et al., 2012; Mandal and Maiti, 2015; Chen et al., 2016b), but these methods

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

http://dx.doi.org/10.1016/j.geoderma.2017.06.020
Received 10 March 2017; Received in revised form 30 May 2017; Accepted 21 June 2017
Available online 29 June 2017
0016-7061/

are subjective and need more sophisticated techniques to be integrated into the overall methodology (Raja et al., 2016). The deterministic approach is only feasible in small areas, where landslide types are simple and the geologic properties are fairly homogeneous (Dou et al., 2014). Therefore, in order to reduce bias of the above two kinds of methods, statistical and soft computing models are usually chosen for landslide susceptibility assessment in large areas (Pham et al., 2015).

Over the last decades, statistical models such as frequency ratio (Shahabi et al., 2015; Wang et al., 2016), weights of evidence (Regmi et al., 2014; Wang et al., 2016), certainty factors (Devkota et al., 2013; Tsangaratos and Ilia, 2016b; Youssef et al., 2016a), evidential belief function (Pourghasemi and Kerle, 2016), statistical index (Regmi et al., 2014; Nasiri Aghdam et al., 2016), index of entropy (Youssef et al., 2015; Hong et al., 2017), logistic regression (Tsangaratos and Ilia, 2016a; Raja et al., 2017; Chen et al., 2017a), and multivariate adaptive regression spline (Conoscenti et al., 2016; Pourghasemi and Rossi, 2016) have been used for landslide susceptibility mapping throughout the world, and these models have produced reliable results (Tien Bui et al., 2011).

However, in addition to the above mentioned methods, some machine learning algorithms have been applied for analysis of landslide susceptibility such as artificial neural network (Wang et al., 2016; Tien Bui et al., 2016b; Ngadisih et al., 2016; Chen et al., 2017a), support vector machines (Tien Bui et al., 2012; Dou et al., 2015; Chen et al., 2016c; Colkesen et al., 2016; Hong et al., 2016b), maximum entropy (Park, 2015; Moosavi and Niazi, 2016; Lombardo et al., 2016), kernel logistic regression (Hong et al., 2015; Tien Bui et al., 2016b), naive Bayes (Tien Bui et al., 2012; Pham et al., 2015), neuro-fuzzy (Lee et al., 2015), decision trees (Pradhan, 2013; Tsangaratos and Ilia, 2016b), random forests (Hong et al., 2016a; Chen et al., 2017b), classification and regression tree (Felicísimo et al., 2013; Chen et al., 2017b), alternating decision trees (Hong et al., 2015), and boosted regression trees (Lombardo et al., 2015; Youssef et al., 2016b).

Recently, several ensemble methods have also been developed such as bivariate statistical index and ANFIS (Nasiri Aghdam et al., 2016), CHAID and multivariate logistic regression (Althuwaynee et al., 2014), adaptive neuro-fuzzy inference system (ANFIS) (Pradhan, 2013; Dehnavi et al., 2015), functional trees with AdaBoost, Bagging, and MultiBoost ensemble frameworks (Tien Bui et al., 2016a), ANN-fuzzy logic (Kanungo et al., 2006), ANN-Bayes analysis (Lee et al., 2004), and stepwise weight assessment ratio analysis (SWARA) (Dehnavi et al., 2015).

The important capability of ensemble methods in landslide susceptibility studies is that this method is more precise in recognizing and has a higher prediction power in comparison with single machine learning models (Althuwaynee et al., 2016). This ability can lead to increase the popularity of this technique and assist the researchers for future landslide related studies. Therefore, the exploration of new ensemble approaches is really necessary in natural hazards cases. In the present study, we aimed to propose and verify ensembles of artificial neural network (ANN), maximum entropy (MaxEnt), and support vector machine (SVM) models (i.e. ANN-SVM, ANN-MaxEnt, ANN-MaxEnt-SVM, and SVM-MaxEnt) for spatial prediction of landslide susceptibility in Wanyuan area (China). The main difference between this study and the aforementioned studies — those who have only used models separately — is that hybrid integration approaches of ANN, MaxEnt, and SVM are new contribution that has been seldom used for spatial modeling of landslide susceptibility and other natural hazards.

2. Study area

The study area (Wanyuan) belongs to the Sichuan Province, is located in the south-west part of China. The study area lies between longitudes of 107°28′53.5″ to 108°30′34.4″E and latitudes of 31°38′35.7″ to 32°20′21.8″, covering an area of about 4065 km^2 (Fig. 1). This area is located in the hinterland of Daba Mountain and prevail subtropical monsoon climate, with the characteristic of moderate climate, distinctive seasons, and abundant rainfall (http://www.wanyuan.gov.cn). The annual average temperature and rainfall is 14.70 °C and 1232.7 mm, respectively.

Topographically, the elevation ranges from 343 to 2404 m a.s.l. according to the DEM of 20 m regular grid, with the average elevation of 1017 m and the standard deviation of 320 m. The slope angles vary from 0 to 80.84°. Areas with slope angles of 10°–20° cover 27.35% of the total area, followed by 20°–30° (26.13%), 30°–40° (18.15%), 0°–10° (13.32%), 40°–50° (10.25%), 50°–60° (4.03%), 60°–70° (0.75%), and > 70° (0.03%), respectively.

The geological units of the study area comprise geological formations from Sinian to Quaternary, and were reclassified into ten groups according to litho-facies and geological ages. Limestone, dolomite, sandstone, mudstone, siltstone, and shale are the most out-cropped lithological formations.

3. Methodology

The diagram in Fig. 2 shows the steps involved in the generation of landslide susceptibility maps in the present study. There are four main phases: (i) dataset preparation for ensemble spatial modeling; (ii) data correlation analysis using bivariate statistical method namely SI; (iii) landslide spatial modeling using ANN, MaxEnt, SVM models and their ensembles; (iv) validation and selection of the optimal models.

3.1. Dataset preparation for ensemble spatial modeling

In the study area, the landslide inventory map was prepared using interpretation of the Landsat 7 and Google Earth satellite images, historical information of landslide events from earlier reports, and extensive field surveys with a handheld GPS. Finally, a total number of 304 landslides were mapped (Fig. 1) comprised of translational (80%) and rotational (20%) slides based on Varnes classification (Varnes, 1978; Hungr et al., 2014). Although the separate treating, mapping, and modeling of the landslide types have been encouraged by many scholars, the redundancy (frequency) and spatial dominance (larger expansion) of the translational slides over rotational slides led to production of one landslide inventory map for the whole study area so that type separation in this case would threaten the data enrichment for model training and validation. On the other hand, these two are usually classified as a subset of a bigger class known as "*slides*" based on movement types and they are completely discernible from falls, topples, lateral spreading, flows, and complex types. Therefore, the present study was carried out under the hypothesis that the relationship between the predictors and the two landslide types does not have a meaningful effect on the final susceptibility maps. The smallest size of landslides identified is about 200 m^2 (correspond to a rotational slide), whereas the largest is about 5×10^4 m^2 (related to a translational slide). Fig. 3 shows two landslides identified from Google Earth satellite images. Subsequently, 213 landslide locations were used for training the ensemble models and 91 landslides were used for validating the built models (Fig. 1). According to Tsangaratos and Benardos (2014), Lombardo et al. (2014), Cama et al. (2016), and Kornejady et al. (2017), a balanced (1:1) positive (landslide) and negative (non-landslide) dataset has been proposed. Therefore, in this research, the same number of positives (304) was applied to negatives, randomly generated from the landslide-free areas using Hawth's Tools. The positive and negative datasets were also randomly split into two parts with a ratio of 70/30% and the resulted positive-negative sets combined together to obtain the training (70%) and validation (30%) datasets for running different machine learning ensemble models.

There are a variety of inter-related factors that affect landslides. Based on the previous researches (Constantin et al., 2011; Dou et al., 2014; Pham et al., 2015; Akgun and Erkan, 2016; Chen et al., 2016a; Hong et al., 2016a; Pourghasemi and Kerle, 2016), the general features