Applied Soft Computing 68 (2018) 147–161

Contents lists available at ScienceDirect

Applied Soft Computing

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

Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5′ and CART algorithms

S.M. Hamze-Ziabari[a], T. Bakhshpoori[b,*]

[a] Department of Civil Engineering, Iran University of Science and Technology, Tehran, 16, Iran
[b] Faculty of Technology and Engineering, Department of Civil Engineering, East of Guilan, University of Guilan, Rudsar, Vajargah, Iran

ARTICLE INFO

Article history:
Received 28 May 2017
Received in revised form 25 February 2018
Accepted 30 March 2018
Available online 5 April 2018

Keywords:
Earthquake prediction
Ground motion parameters
Ensemble modeling
M5′
CART

ABSTRACT

In the present study, an efficient bagging ensemble model based on two well-known decision tree algorithms, namely, M5′ and Classification and Regression Trees (CART) is utilized so as to estimate the peak time-domain strong ground motion parameters. Four different predictive models, namely, CART, Ensemble M5′, Ensemble CART, and Ensemble M5′ + CART are developed to evaluate Peak Ground Acceleration, Peak Ground Velocity, and Peak Ground Displacement. A big database from the Pacific Earthquake Engineering Research Center is employed so as to develop the proposed models. Earthquake magnitude, earthquake source to site distance, average shear-wave velocity, and faulting mechanisms are considered as the predictive parameters. The superior performances of the developed models are observed in the validation against the most recent soft computing based models available in the specialized literature. Parametric as well as sensitivity analyses are carried out to ensure the robustness of the predictive models in discovering the physical concept latent in the nature of the problem.

1. Introduction

The peak time-domain strong ground motion parameters, Peak Ground Acceleration (PGA), Peak Ground Velocity (PGV), and Peak Ground Displacement (PGD), are well known to characterize an earthquake, and helpful in risk assessment and seismic analysis of structures. Predicting the fore-mentioned parameters via methods other than the comprehensive approaches such as physical modeling and field investigation has been the subject of interest among many researchers, which known as ground motion prediction equations (GMPEs), or empirical ground motion models, or attenuation relations. Attenuation relations developed in the Pacific Earthquake Engineering Research Center (PEER) in two phases, NGA-West1 [1,2] and NGA-West2 [3], which are known as CB08 and CB14, respectively, are the most common predictive models among many others [4]. These models are developed based on the regression analysis approach using 1561 records from 64 earthquakes and 15,521 records from 322 earthquakes (in California and the worldwide known as strong ground motion databases), respectively. In the regression analysis, the peak time-domain strong ground motion parameters are frequently calculated as a function of the independent earthquake parameters as well as the physical parameters. The recent breakthroughs in Soft Computing (SC) based methods have enabled researchers to develop highly predictable SC-based attenuation relations which have merits such as the ability in learning and generalizing interactions among many variables, and the capability of assuming no equation form. Güllü and Erçelebi [5] firstly developed a neural network approach using strong ground motion data from Turkey. Neural networks are also used by Rees et al. [6] to predict earthquakes in Chile. Furthermore ANN besides feature selection techniques have been used by Martínez-Álvarez et al. [7] to improve prediction ability of ground motion parameters. A survey of earthquake magnitude prediction based on ANNs is made by Florido et al. [8]. Very recently Asencio-Cortés et al. [9] used ANN for Medium–large earthquake magnitude prediction in Tokyo. Alavi et al. [10] used a new variant of genetic programming, namely multi expression programming (MEP) for ground motion prediction problem. Genetic expression programming besides the regression models and likelihood estimation are also used by Güllü [11] for prediction of PGA using case records of Turkish strong ground motion data. Genetic programming [12] and gene expression programming [13] are also used in other studies to develop the GMPEs. Hybrid soft computing techniques [14–17], i.e. coupling the available methods with metaheuristics with the aim of improving the accuracy also attracted some attentions. Other methods such as lagramge [18], conic multivariate adaptive regression splines [19], support vector machine [20], M5′ algorithm [21], and randomized adaptive neuro-fuzzy inference system [22] are benefited to develop the GMPEs. Tree based ensemble learning [23] and

* Corresponding author.
E-mail addresses: ziabari@civileng.iust.ac.ir (S.M. Hamze-Ziabari), tbakhshpoori@Guilan.ac.ir (T. Bakhshpoori).

https://doi.org/10.1016/j.asoc.2018.03.052
1568-4946/© 2018 Elsevier B.V. All rights reserved.

imbalanced classification and ensemble learning [24] have been used successfully for short and large earthquake magnitude prediction terms in Hindukush region and Chile, respectively. Very recently four machine learning techniques including pattern recognition neural network, recurrent neural network, random forest and linear programming boost ensemble classifier have been used for earthquake magnitude prediction in Hindukush region [25].

Owing to the complex nature of the peak time-domain strong ground motion parameters estimation problem, accurate SC-based prediction models development is indispensable. Although numerous SC-based estimation models have been proposed since the last decade, their accuracy level is not satisfying enough and is not meticulously considered. This survey is aimed at implementing the ensemble method to improve the prediction capability of soft computing based approaches so as to reach more accurate predictive models. The main idea behind the ensemble method is to rank a set of preceding candidate predictive models and incorporate them into a single ensemble model. This method often leads to a more precise model compared to single soft computing based model since it employs benefits of each base learner algorithms [26,27]. Very recently Roselli et al. [28] numerically evaluate and score a set of GMPEs and then an ensemble modelling is employed to merge the results of these GMPEs. In the current study, the bagging ensemble method known as the most popular ensemble algorithm [29] is employed to combine the predictions of M5′ and Classification and Regression Trees (CART) algorithms. The M5′ algorithm has been successfully used recently by the authors [21] so as to develop attenuation relations and has shown comparable performance against the existing models. The CART algorithm is also a well-known decision tree algorithms; however, the performance of the CART based model is often relatively weaker than other machine learning approaches [30]. The leading causes for such weak performances can be attributed to its high affection to the noise data as well as the redundant attributes of data [31]. In this study, the bagging ensemble algorithm is employed to enhance the performances of the CART and M5′ methods by combining their results.

In this regard, four different predictive models including CART, Ensemble CART, Ensemble M5′, and Ensemble M5′ + CART are developed. Almost all the SC-based models available in the literature are developed based on the NGA-West1 strong motion database. The same database is utilized herein to evaluate the predictive capability of the developed models against other available SC-based models. The contribution of four different influential parameters including earthquake magnitude, earthquake source to site distance, average shear-wave velocity, and faulting mechanisms are taken into account. Results show that the developed models can significantly improve the performance metrics. The parametric and sensitivity analyses are also performed to make sure that the developed models capture the physical concept of the problem.

The remaining sections of this article are organized as follows. Section 2 outlines the methodology used in this study. Section 3 describes the model development procedure. Results and discussions are made in three subsections within Section 4 containing performance analysis and comparative study, sensitivity analysis, and also parametric analysis. At the end, the article is concluded in Section 5.

2. Methodology

In the present study, Decision Tree (DT) is used to discover the relationship between peak ground motion parameters as outputs and the effective input parameters involved in the phenomenon. DTs are classified as data mining approaches, which can be utilized to solve both regression and classification problems. DTs consist of several nodes and leaves which are connected to each other as an inverted tree-shaped structure. The root node of the tree is located at the top and leaves are placed at the bottom of the nodes. Input instances are inserted into each node and different branches are generated based on a test criterion. The process of producing branches continues until all instances have been classified in a special branch.

In the problems related to regression (i.e. the output is continuous), Model Tree (MT) and Regression Tree (RT) are often applied as predictive algorithms. The main difference between two approaches is that the predictive model in RT is stated as constant values while MT presents its model as a linear combination of input parameters in each leave. The Classification and Regression Tree (CART) as a RT approach and the M5′ as a MT approach have been widely used to solve real-world problems. These methods are outlined in the following subsections.

2.1. CART

Classification and Regression Tree (CART) is one of the most common decision tree algorithms, in which the DT is developed based on a recursive partitioning method. CART algorithm is introduced by Breiman et al. [32] and can be used for either regression or classification problems. In case of classification problem, a classification tree is developed based on the CART algorithm while a regression tree is generated by the algorithm for a regression problem. In this study, the regression tree is used and therefore, its algorithm is described. CART algorithm divides the whole database into several subsets using different predictor variables. These subsets are generated repeatedly and begin with the entire database. The procedure of constructing a regression tree can be divided into three steps: (i) creating a set of questions based on predictor variables (X) (e.g. is $X \geq c$? where c is a constant, the answers to such questions can be yes or no); (ii) defining a splitting or goodness of fit criteria to choose the best split on variables; and (iii) providing a summary of statistics at the internal node.

The CART algorithm implements the Least Squared Deviation (LSD) impurity so as to determine the splitting rules and goodness of fit. The LSD measure (R(t)) can be simply calculated as follows:

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} \omega_i f_i \left(y_i - \bar{y}(t)\right)^2 \tag{1}$$

$$\bar{y}(t) = \frac{1}{N_w(t)} \sum_{i \in t} \omega_i f_i y_i \tag{2}$$

$$N_w(t) = \sum_{i \in t} \omega_i f_i \tag{3}$$

where $N_w(t)$ is the weighted number of samples in node t, ω_i is the value of the weighting response for record i (if any), f_i is the value of the record response (if any), y_i is the response value, and $\bar{y}_i$ is the mean value of response values. To split s at node t, the LSD uses the following criterion:

$$Q(s, t) = R(t) - R(t_L) - R(t_R) \tag{4}$$

where t_R and t_L are the left and right child nodes of node t, respectively. The split s is determined to maximize the $Q(s, t)$.

2.2. M5′

The M5′ algorithm is one of the most common decision tree algorithms so as to analyze complex systems with very high dimensionality up to hundreds of attributes. The M5 algorithm was firstly