Engineering Applications of Artificial Intelligence 26 (2013) 1339–1347

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

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

An equation-discovery approach to earthquake-ground-motion prediction

Štefan Markič, Vlado Stankovski *

Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 10 June 2012
Received in revised form
8 December 2012
Accepted 10 December 2012
Available online 4 January 2013

Keywords:
Equation discovery
Ground motion prediction equations
Peak ground acceleration
Lagramge

ABSTRACT

In active seismic regions an earthquake's peak ground acceleration (PGA) is required information when designing a building. In this study we employ the state-of-the-art, Lagramge, equation-discovery system to induce an equation that is suitable for modeling the PGA and investigate its applicability. In contrast to traditional modeling techniques the Lagramge system does not presume the structure of the equation and then identify the parameter values; instead, it finds the equation's structure as well. From the large amount of background knowledge on earthquake engineering we formalize a context-free grammar, which is then used as a guideline for the equation-building procedure. The PF-L data set used for the experiments is taken from the study of Peruš and Fajfar (2010), which is based on the data sets of Chiou et al. (2008) in the project Next Generation Attenuation of Ground Motion and the study of Akkar and Bommer (2010). The best model derived from the grammar is then quantitatively and qualitatively evaluated and compared. The presented results support the proposal to use an equation-discovery tool as an aid to the PGA modeling work and to potentially contribute new knowledge to the field of earthquake engineering.

1. Introduction

An earthquake is a natural phenomenon that manifests itself as a violent, rapid, earth tremor and happens unexpectedly, without prior notice. Strong earthquakes usually cause a lot of difficulties for people and communities; hence, the engineer's task is to properly design a structure, bearing in mind that a devastating earthquake could occur during its lifetime. In the earthquake engineering domain, the correspondence with physical reality must be taken as the strongest criterion for the acceptability of the developed models along with the estimated prediction accuracy. The ground-motion prediction equations (GMPEs) or attenuation relations, the common name that was used for them (Douglas, 2003), are some of the key elements used by engineers to estimate a possible earthquake load at the site of a structure.

One of the ground-motion parameters is the peak ground acceleration (PGA), the prediction of which is the focus of the present study. More than 250 papers concerning PGA modeling have been published over the past 50 years, which means the area has been well investigated (see Douglas, 2011). Traditionally, the PGA is modeled as a single mathematical formula based on an author's knowledge about the problem. The parameters included in such a formula are then fitted to the data by using a regression analysis for the prediction accuracy. Consequently, the resulting models are based on various assumptions and data sets and differ significantly in qualitative terms as well as quantitatively. Eq. (1) from the study of Akkar and Bommer (2010) is presented here for illustrative purposes and can be described as a typical example of a GMPE. The variables used in Eq. (1) are: (i) the PGA in [cm/s^2]; (ii) the moment magnitude M_w; (iii) the Joyner–Boore distance R_{jb} in [km]; (iv) the average soil shear-wave velocity in the upper 30 m of soil underneath the observation spot $V_{s,30}$ in [m/s]; and (v) the faulting mechanism F (Akkar and Bommer, 2010).

Recently, researchers involved in earthquake engineering have experimented with new approaches when predicting the PGA that do not assume an equation form and have drawn different conclusions. Peruš and Fajfar (2010) used a conditional average estimator (CAE) method, which in contrast to conventional approaches does not make any *a priori* assumption, and found this method to be a simple but powerful tool,

$$\begin{aligned}\log_{10}(PGA) = {} & 1.04159+0.91333\cdot M_w-0.08140\cdot M_w^2 \\ & +(-2.92728+0.28120\cdot M_w)\cdot\log_{10}\sqrt{R_{jb}^2+7.86638^2} \\ & +\begin{cases}0.08753 & \text{if } V_{s,30}<360\ \text{m/s}\\ 0.01527 & \text{if } 360\ \text{m/s}\le V_{s,30}<800\ \text{m/s}\\ 0 & \text{if } 800\ \text{m/s}\le V_{s,30}\end{cases} \\ & +\begin{cases}-0.04189 & \text{if } F=\text{normal}\\ 0 & \text{if } F=\text{strike-slip}\\ 0.08015 & \text{if } F=\text{reverse}\end{cases}\end{aligned} \quad (1)$$

* Corresponding author. Tel.: +386 41200565.
E-mail addresses: vlado.stankovski@fgg.uni-lj.si, vlado@stankovski.net (V. Stankovski).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.12.005

especially in the research environment. Kuehn et al. (2011) used Bayesian networks and concluded that the model they obtained is the maximum *a posteriori* model; i.e., the most probable model given the data. Alavi et al. (2011) used multi expression programming (MEP), a machine-learning technique, and found that the generated models predict better than, or comparable with, the previously published regression-based models and, in their opinion, provide relatively simple equations, as opposed to the more complicated models from the Next Generation Attenuation (NGA) project. In summary, the use of non-conventional methods has so far concentrated on improving the prediction results.

With the development of computers a new scientific area was founded, where authors propose machine algorithms that try to imitate learning as an important human property. In equation discovery (ED), a sub-area of machine learning, the algorithms try to find a proper equation formulation that best fits a given data set. All ED systems use some kind of language bias that limits the hypothesis space, which is the space of all the possible equations constructed from a given set of operators, functions and variables. Such a space is usually infinite, and is therefore restricted by the means of the algorithm. The state-of-the-art, Lagramge, ED system used in this study employs a declarative bias in the form of a context-free grammar (CFG) to limit the hypothesis space, which is given as input information to the system (Todorovski and Džeroski, 1997). With such a formalism, domain knowledge can be easily provided to the ED system and so guide it toward the expected equation formulations.

Because of the fact that almost all GMPEs take the form of equations, the use of an ED system as an aid in earthquake-engineering design studies may come as a natural choice. Our investigation revealed that ED systems have not been used in earthquake engineering, to the best of our knowledge. Therefore, a specific goal of the present study is to propose a method for using the Lagramge system when modeling the PGA, which is used as a case study because of a particularly large domain knowledge. Bearing in mind the extensive expert requirements when modeling GMPEs, a careful investigation of the ED system is necessary before its usage for modeling the PGA is proposed. Moreover, it is necessary to appropriately incorporate the existing domain knowledge into the ED process, because the experimental set-up itself, if correctly designed, has the potential to yield high-quality results. With the system's heuristic or exhaustive exploration of defined hypothesis space it is possible to investigate thousands of equation formulations and based on quantitative criteria, such as the mean squared error (MSE) and qualitative criteria like physicality, select the best equation. This procedure is crucial in order that the proposed ED method gains acceptance within the earthquake-engineering community. Fortunately, as we had access to powerful distributed-computing infrastructures, in our experiments all the calculations were pushed to their limits. The goal of this study was also to compare the results obtained with already existing GMPEs.

The rest of the paper is organized as follows. In Section 2 the Lagramge ED system and its input parameters along with the CFG and the data-set requirements are thoroughly explained. We describe the whole process of the application of the Lagramge ED system to the problem of predicting the PGA in Section 3. Descriptive tables and figures showing the results and the best equation found, together with their explanations, are presented in Section 4. We conclude this study with Section 5, where we discuss and evaluate the presented results and provide some ideas for future research.

2. Lagramge

Equation discovery (ED) is an emerging machine-learning discipline that is closely related to system identification, inductive logic programming and genetic programming. About a dozen ED systems have been described in the literature, among which Bacon of Langley et al. (1987), Lagrange of Todorovski and Džeroski (1995) and Lagramge of Todorovski and Džeroski (1997)[1] have received particular attention in the machine-learning community. The Lagramge system seems to be the most suitable for the PGA modeling task at hand, particularly because it uses CFG to specify prior knowledge. For this reason it was selected and used in the present study.

The Lagramge ED system has already been applied to several scientific fields of interest. The first experiments with the Lagramge system were made in the area of ecological modeling, e.g., the prediction of phytoplankton growth in the studies of Todorovski et al. (1998) and Kompare et al. (2001). Todorovski and Džeroski (2001) also applied it to population dynamics, predicting the behavior of prey–predator dependence and found that the integration of specific domain knowledge in the CFG significantly improved the prediction results. Some of the latest applications of the Lagramge system include discovering mathematical models of a mechanically ventilated lung by Ganzert et al. (2010) and the financial forecasting of commodity prices from the London Metal Exchange by Alzaidi and Kazakov (2011).

The problem given to the Lagramge system is denoted with two input files: a data set D and a CFG (Todorovski and Džeroski, 1997). The input data $D=\{M,v_d,W\}$ consists of one or more tables of measurements or records M of variables $W=\{v_1,v_2,\ldots,v_n\}$. Among the variables, one must be selected as a dependent variable $v_d \in W$. So as to make it easier to understand the grammar building described in the following paragraphs, let us assume that we want to design a CFG that will be able to generate the first three terms of Eq. (1).

A tuple $CFG=\{N,T,P,S\}$ prescribes the syntax of the right-hand side of an equation. It contains finite disjunctive sets of non-terminals (N) and terminals (T). The Lagramge system uses a special non-terminal symbol $V \in N$, which denotes any of the independent variables from the input data set $W \backslash v_d$; otherwise, any symbol can be used to denote a non-terminal. The set T consists of all the independent variables $v_i \in W \backslash v_d$ and a special symbol *const*, whose syntax in the Lagramge system is as follows:

$$const[name : lowest\ value : starting\ value : highest\ value] \qquad (2)$$

In the case of our example, the set of non-terminals is $N=\{Linear,Term,V\}$ and the set of terminals is $T=\{M_w,const[\ldots]\}$.

The productions $P=\{P_1,P_2,\ldots,P_n\}$ denote the grammatical rules that relate the non-terminals among themselves (recursion is possible) and to the terminals. The standard form of a production P is $A \rightarrow \alpha$, where $A \in N$, $\alpha \in N \cup T$ and the operators or functions used are (already or user-) defined in the programming language C. If we want to reference to an explicit variable in a grammar, we must use *variable_* in front of its name. However, the productions for V are added to the grammar automatically during the run-time, as the Lagramge system reads the variables' names from the input data file, i.e., $\forall v_i \in W \backslash v_d : V \rightarrow variable\_v_i \in P$. We use the annotation with the logical *or* operator $A \rightarrow \alpha_1 |\alpha_2| \ldots |\alpha_n$ for productions $A \rightarrow \alpha_1$, $A \rightarrow \alpha_2, \ldots, A \rightarrow \alpha_n$. In order to derive the first three terms of Eq. (1), only addition and multiplication are needed, which are both already predefined in C.

Finally, $S \in N$ is a special, non-terminal symbol, from which the derivation of the expressions starts. In the case of our example, it is denoted by the symbol *Linear*.

The definition of the developed example polynomial grammar is provided in Table 1. Its first four productions provide enough

[1] The Lagramge system release 2.2 used in this study is available as open-source software at URL: http://www.ai.ijs.si/~ljupco/ed/lagrange.html.