Knowledge-Based Systems 101 (2016) 15–30

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction

G. Asencio-Cortés [a], F. Martínez-Álvarez [a,*], A. Morales-Esteban [b], J. Reyes [c]

[a] Division of Computer Science, Universidad Pablo de Olavide, ES-41013 Seville, Spain
[b] Department of Building Structures and Geotechnical Engineering, University of Seville, Spain
[c] TGT-NT2 Labs, Santiago, Chile

ARTICLE INFO

Article history:
Received 22 June 2015
Revised 16 November 2015
Accepted 20 February 2016
Available online 16 March 2016

Keywords:
Sensitivity analysis
Earthquake prediction
Seismicity indicators
Supervised learning

ABSTRACT

The use of different seismicity indicators as input for systems to predict earthquakes is becoming increasingly popular. Nevertheless, the values of these indicators have not been systematically obtained so far. This is mainly due to the gap of knowledge existing between seismologists and data mining experts. In this work, the effect of using different parameterizations for inputs in supervised learning algorithms has been thoroughly analyzed by means of a new methodology. Five different analyses have been conducted, mainly related to the shape of training and test sets, to the calculation of the *b*-value, and to the adjustment of most collected indicators. Outputs sensitivity has been determined when any of these factors is not properly taken into consideration. The methodology has been applied to four Chilean zones. Given its general-purpose design, it can be extended to any location. Similar conclusions have been drawn for all the cases: a proper selection of the sets length and a careful parameterization of certain indicators leads to significantly better results, in terms of prediction accuracy.

1. Introduction

The problem of predicting earthquakes has fascinated the human being. Although this problem seems to be irresolvable, recent works have proposed new paradigms of prediction that should be taken into consideration [1]. In particular, the use of data mining techniques has emerged in this field as a powerful tool with undeniable benefits [2–5].

This work is focused on the analysis of the inputs used in several supervised machine learning classifiers in order to improve earthquake prediction accuracy. In particular, some studies recently conducted propose the use of various seismicity indicators (or attributes containing geophysical information associated with earthquake occurrence) for earthquake prediction [6–8].

The correlation of such indicators with the binary class (either an earthquake is coming or not) was analyzed in [9], showing that some of them exhibited information gain close to zero. This work goes one step ahead because all of these indicators have been used with a baseline configuration. This is, none of the works above referenced considered that most of the indicators are functions of certain variables. These works just used standard values omitting the fact that different configurations may lead to different results and, in some cases, to better results. And this is the main goal of this research: to conduct an exhaustive analysis on how an adequate adjustment of the seismicity indicators may improve the accuracy of the classifiers.

In particular, in [8] a new set of seismicity indicators was proposed as inputs for earthquake prediction. Later, in [9], such set of indicators were combined with those published in [7] and applied feature selection methods to discover that some of the indicators proposed in both [7] and [8] exhibit null information gain with the class. In this work, it should be highlighted that some indicators are highly dependent with their initial parameterization. A sensitivity study is performed to show that results can be highly improved if an adequate initialization is done.

A new methodology is thus proposed and the following issues have been explored. First, the size of the most adequate training sets and whether training and test sets must be contiguous or not. Second, how the *b*-value (a key predictive value [10]) must be calculated. Finally, how certain attributes introduced in [7,8] must be configured so that the best possible prediction is achieved. In other words, it provides some guidelines in order to properly parameterize the seismicity indicators proposed to date. Also, the best training set selection is performed.

Four zones of Chile, the country with the highest seismic activity [11], have been analyzed to validate the applicability of this

* Corresponding author. Tel.: +34 954 977 370.
E-mail addresses: guaasecor@upo.es (G. Asencio-Cortés), fmaralv@upo.es (F. Martínez-Álvarez), ame@us.es (A. Morales-Esteban), daneel@geofisica.cl (J. Reyes).

http://dx.doi.org/10.1016/j.knosys.2016.02.014
0950-7051/

methodology. Nevertheless, it has been defined so that it can also be applied to any zone in the world.

The rest of the paper is structured as follows. Section 2 provides a general overview on the state-of-the-art. Section 3 describes the new methodology proposed in order to find the set of seismicity indicators with the most adequate initialization (when used as input in supervised classifiers). All the results of applying the methodology to four cites in Chile have been presented in Section 4. Finally, the conclusions drawn from this study have been summarized in Section 5.

2. Related works

The possibility of predicting earthquakes has been questioned and answered in various ways, from denial to optimism, including the contribution of mathematical proofs and empirical support for each hypothesis [8,12–15].

To ensure that statements related to earthquake prediction are rigorous, the following information must be simultaneously provided according to [16]:

1. A specific location or zone.
2. A specific span of time.
3. A specific magnitude range.
4. A specific probability of occurrence.

Additionally, the U.S. Geological Survey (USGS) founded the Collaboratory for the Study of Earthquake Predictability (CSEP) in 2007 [17]. The goal of this organization is to develop a virtual and distributed laboratory that can support a wide range of scientific prediction experiments in multiple regional or global natural laboratories. This earthquake system science approach seeks to provide answers to the questions:

1. How should scientific prediction experiments be conducted and evaluated?
2. What is the intrinsic predictability of the earthquake rupture process?

In this context, several methods have been proposed to predict any of the features detailed by Allen [16]. According to the Accelerating Moment Release (AMR) method, the rate of seismic moment release for magnitude is rapidly increased before a large event occurs [18,19].

Variations of *b*-value have also been analyzed. For a large magnitude earthquake to occur, it is necessary a prior elastic potential energy accumulation. This fact causes a deficit of small and moderate earthquakes. This leads to an abnormal alteration of the Gutenberg–Richter law's *b*-value [10,20].

M8 algorithms study the occurrence of earthquakes of magnitude larger than 8.0. They are based on the evolution of several time series composed of earthquakes of moderate magnitude. The goal is to decide if a time of increased probability (TIP) exists for an event of larger magnitude [21,22].

Region–Time–Length (RTL) is an algorithm that analyzes temporal sequences of earthquakes. It only takes into consideration location, time, magnitude, and detects anomalies in seismicity prior to large events [23,24].

It is thought that for a large earthquake to occur, it is necessary that more energy is released during the loading period than during the unloading one. Based on this assumption, Load-Unload Response Ratio (LURR) uses the ratio of energy released as a potential precursor to make predictions [25,26].

Another widely used method is Every Earthquake is a Precursor According to Scale (EEPAS). This method is based on the observation of an increment of small earthquakes, as this is considered a precursory phenomenon for larger earthquakes [27,28].

Epidemic-Type Aftershock Sequence (ETAS) considers that every earthquake is a simultaneously potential aftershock, main shock or foreshock, with its own aftershock sequence. This way, anomalous configurations for temporal and spatial seismicity can be found [29,30].

The Simple Smoothed Seismicity model, or simply TripleS, provides space-rate-magnitude forecasts based on a spatial clustering of seismicity. To get this done, a Gaussian smoothed is applied to the seismic catalogue, which estimates the amount of foreseen earthquakes in particular zones for particular periods of time [31].

Increased attention is being payed to algorithms based on machine learning nowadays. These algorithms include a vast variety of solutions ranging from unsupervised learning [10,32] to supervised one [4,9]. It must be noted that in [10] clustering techniques were used to obtain patterns that model the behavior of seismic temporal data and can help to predict medium-large earthquakes. It is true that in [9] the seismicity indicators used were based on [10] and others. But, in [10], the results were patterns and in [9] the results were a probability of an earthquake to happen after the hit of every earthquake of magnitude larger than 3.0. Moreover, the results in [9] were improved thanks to feature selection techniques. In supervised learning, every earthquake is modeled by means of certain attributes that Panakkat and Adeli [7] named seismicity indicators. From its initial application, several works have proposed new indicators. Such is the case of [8] or [6], where the authors also added Bath and Omori–Utsu laws, as well as variations of *b*-value, to the set of proposed seismicity indicators. The model was assessed by artificial neural networks, a method also used in [2,5,33].

Recently, Ikram and Qamar [34] introduced an expert system for earthquake prediction, which extended [35]. They considered the historic record of earthquakes and divided the Earth into four zones. Then, association rules were applied to predict earthquakes in each of the four zones with a horizon of prediction equals to one day.

Nonetheless, some of the seismicity indicators proposed exhibit parametrical dependence, this is, there is a need of an initial setup so that they can properly work with supervised classifiers. Moreover, the original studies do not explicitly propose a specific tuning for them. In this context this work has been carried out: to determine the influence of either an adequate or wrong adjustment for all the existing seismicity indicators reported in the literature.

3. Methodology

This section introduces a methodology to systematically identify those values for certain parameters, somehow hidden in a set of seismicity indicators, that generate better results in terms of average accuracy when predicting earthquakes. In this sense, a set of parameters that may deeply influence the accuracy for predictions has been first identified. Later, a sensitivity analysis over such parameters has been conducted in order to determine how a wrong setup may lead to the occurrence of a major loss of accuracy in predictions.

Note that the proposed methodology must be applied to every geographical zone. In this work, the four Chilean zones studied in [8] and [9] have been considered.

Section 3.1 explains how supervised learning is applied to predict earthquakes. Once this strategy is defined, Section 3.2 details the five proposed studies to analyze the sensitivity of the parameters involved in the prediction.

3.1. Procedure for earthquake prediction

Generally speaking, the prediction of earthquakes is carried out in the context of supervised learning by means of well-known