Knowledge-Based Systems 50 (2013) 198–210

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula

F. Martínez-Álvarez [a,*], J. Reyes [b], A. Morales-Esteban [c], C. Rubio-Escudero [d]

[a] Department of Computer Science, Pablo de Olavide University of Seville, Spain
[b] TGT-NT2 Labs, Santiago, Chile
[c] Department of Continuum Mechanics, University of Seville, Spain
[d] Department of Structures and Geotechnical Engineering, University of Seville, Spain

ARTICLE INFO

Article history:
Received 19 December 2012
Received in revised form 9 May 2013
Accepted 14 June 2013
Available online 27 June 2013

Keywords:
Earthquake prediction
Seismicity indicators
Feature selection
Time series
Supervised classification

ABSTRACT

This work explores the use of different seismicity indicators as inputs for artificial neural networks. The combination of multiple indicators that have already been successfully used in different seismic zones by the application of feature selection techniques is proposed. These techniques evaluate every input and propose the best combination of them in terms of information gain. Once these sets have been obtained, artificial neural networks are applied to four Chilean zones (the most seismic country in the world) and to two zones of the Iberian Peninsula (a moderate seismicity area). To make the comparison to other models possible, the prediction problem has been turned into one of classification, thus allowing the application of other machine learning classifiers. Comparisons with original sets of inputs and different classifiers are reported to support the degree of success achieved. Statistical tests have also been applied to confirm that the results are significantly different than those of other classifiers. The main novelty of this work stems from the use of feature selection techniques for improving earthquake prediction methods. So, the information gain of different seismic indicators has been determined. Low ranked or null contribution seismic indicators have been removed, optimizing the method. The optimized prediction method proposed has a high performance. Finally, four Chilean zones and two zones of the Iberian Peninsula have been characterized by means of an information gain analysis obtained from different seismic indicators. The results confirm the methodology proposed as the best features in terms of information gain are the same for both regions.

1. Introduction

The prediction of natural disasters has always been a challenging task for the human being. Currently, the prediction of tsunamis [38], volcanic eruptions [19], thunderstorms [5], hurricanes [52] or typhoons [46] has been addressed from many different points of view. Nevertheless, the prediction of earthquakes stands out due to the devastating effect they may cause in human activity, as thoroughly discussed by Panakkat and Adeli in 2008 [36] and, later in 2012, by Tiampo and Shcherbakov [47].

Despite the efforts made there is no system apparently capable of simultaneously fulfilling all the requirements demanded by the Seismological Society of America [3] to make an accurate prediction: to predict when, where, how big and how probable is an earthquake to occur.

This work is focused on the application of artificial neural networks (ANN) to improve earthquake prediction. In particular, based on three previous works [30,35,40], it aims to obtain an optimal set of seismicity indicators as ANN's inputs. These three works successfully applied completely different sets of inputs at Chile, the Iberian Peninsula and southern California, respectively, three regions with different geophysical properties. Moreover, Chile and southern California are two of the areas with larger seismic activity in the world, whereas the Iberian Peninsula is considered a moderate activity area.

However, none of them provided an analysis on the correlation exhibited between the inputs and the output. It is reasonable to think that not all the features have the same predictive ability and, even, that some of them could have decreased the prediction quality. And this is precisely the main goal of this work: to apply feature selection techniques to obtain a better set of features as ANN's inputs. It is expected, then, that the selection of the features with higher correlation will lead to more accurate predictions. In this sense, it is the first time that feature selection techniques have been applied for earthquake prediction.

* Corresponding author.
E-mail addresses: fmaralv@upo.es (F. Martínez-Álvarez), daneel@geofisica.cl (J. Reyes), ame@us.es (A. Morales-Esteban), crubioescudero@us.es (C. Rubio-Escudero).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.06.011

Feature selection (or variable selection or feature reduction) emerges as a crucial step to build robust models, especially when too many variables form the set of input features. Although many complex approaches have been proposed during the last decade [6,17,42], the analysis of the information gain that every seismicity indicator (or feature) presents is carried out to discover which ones show larger correlation with the output.

The Chilean zones described in [39] and studied in [40] – Talca, Santiago, Valparaíso and Pichilemu – have been subjected to analysis in order to assess the performance of such proposal. Also, the two most seismic areas of the Iberian Peninsula – the Alborán Sea and the Western Azores-Gibraltar fault –, described in [31], have been analyzed and compared with the Chilean zones.

More specifically, the features proposed in [35] have been used, for the first time, as inputs for both Chile and the Iberian Peninsula. Then, a set containing all the features proposed in [35,30,40] has been created for the six areas. Results reported after the application of feature selection show that the optimal set of features is the same for Chile and the Iberian Peninsula. Moreover, the use of this new set generates better results, for all the metrics studied, than those of sets in [35,30,40] individually applied. This fact confirms the need of assessing the adequacy of the seismicity indicators and suggests that several patterns can be found for active seismic areas regardless the physical properties of the area under study. Additionally, this work provides the reader with a ranking of all the features analyzed in terms of information gain, revealing that some of them have null contribution.

The remainder of the work is structured as follows. Section 2 explores the works related with the application of ANN to earthquake prediction. Section 3 describes the methodology used, as well as the mathematical fundamentals underlying the approach. Sections 4 and 5 presents the results stemmed from the application of the ANN to Chile and the Iberian Peninsula, respectively. In this section a comparative analysis with other well-known classifiers is also provided. A statistical analysis has been carried out in Section 6 to verify that the results obtained by means of the new methodology are statistically different to all others. A discussion on the features selected is presented in Section 7. Finally, the conclusions drawn are summarized in Section 8.

2. Related works

This section is to provide the reader with a general overview of the latest published works related with earthquake prediction and all those that used ANN's.

Firstly, it should be noticed that earthquake forecasting has come in recent years to be synonymous with probabilistic statements about seismicity distributions, whereas predictions emphasize individual earthquakes. In this sense, the use of artificial intelligence techniques has recently emerged as a powerful tool for earthquake prediction. For instance, the use of a method called Pattern Informatics that identifies correlated regions of seismicity in recorded data that precede the main shock, was introduced in [32], as well as its extended version for 3D zones [48]. Also, the use of quantitative association rules and decision trees was applied to predict shocks in the Iberian Peninsula [28]. The prediction of medium-large earthquakes by means of the K-means algorithm was presented in [31], where the authors discovered some patterns preceding medium-large earthquakes. Also, hidden Markov models were applied to predict earthquakes in California [14] or cellular automata simulations in Turkey and Western Canada [21].

However, the application of ANN's for earthquake prediction highlights among other techniques, since it was first proposed for evaluating the seismicity of Azores in 2006 [4]. The author used techniques developed for forecasting another chaotic time series: the financial markets. A neural network with three basic inputs: time, intensity and location was used. Two earthquakes were correctly predicted using major groups of 1° longitude with a range of ±5–6 months. The author pointed out the necessity to integrate physical precursors in order to narrow the predicting window. Panakkat and Adeli [35] proposed three different ANN's to predict earthquakes magnitude for southern California and San Francisco bay. Especially remarkable is the novel set of seismicity indicators they used. This method yielded good results for earthquakes of magnitude between 6.0 and 7.5. Later, the same authors, predicted earthquake time and location in southern California, using a recurrent neural network [37]. To achieve such a task, they computed eight seismicity indicators of earthquakes taking into consideration the latitude and the longitude of the epicentral location as well as the time of occurrence of the following earthquake. Another kind of ANN, a probabilistic neural network was evaluated in [1] and also applied to southern California. The main novelty was the use of this kind of neural network for classification purposes, in particular, using the earthquake magnitude as a target label to classify. This model was accurate for magnitudes between 4.5 and 6.0, complementing the range of magnitude prediction of the method proposed in [35]. Recently, Zamani et al. [51] have studied the spatial–temporal variations in seismicity parameters before the Qeshm earthquake in South Iran. For that purpose, they used artificial neural networks and adaptive neural fuzzy inference system. The authors, also, point out the necessity to choose more appropriate seismicity parameters.

The seismicity of four zones of Chile, one of the countries with higher seismic activity, was explored by means of neural techniques in [40]. The authors proposed a particular architecture and used a novel set of inputs, mainly based on the variations of the b-value of the Gutenberg-Richter law, Bath's law and Omori-Utsu's law. Especially remarkable is the small spatial and temporal uncertainty their ANN's presented (cells varying from $0.5^\circ \times 0.5^\circ$ to $1^\circ \times 1^\circ$ and 5 days, respectively).

ANN's have also been applied to predict earthquake's magnitude in Greece [26]. In this work, the authors only used the magnitude of the previous earthquakes as inputs and obtained a high accuracy rate for medium earthquakes. However, the rate considerably decreased when major seismic events were considered.

The suitability of applying ANN's to the northern Red Sea area has also been analyzed in [2]. This time, the authors proposed a number of different architectures varying the number of hidden layers, the transfer functions and the number of nodes. Then, they compared their performance to several Box–Jenkins models [8].

Another hazardous area, India, has been subjected to study by means of ANN's [9]. After evaluating several architectures, the authors concluded that the best one must include two hidden layers and the sigmoid transfer function. Also the tectonic regions of Northeast India have been explored [44]. The authors retrieved earthquake data from NOAA and USGS catalogues and proposed two non-linear forecasting models. Both approaches are stable and suggest the existence of certain seasonality in earthquake occurrence in this area.

The East Anatolian fault system is known for causing many earthquakes. A multi-layer Levenberg–Marquardt ANN was applied to predict earthquakes in that area in [25]. The main novelty of this work lied on the use of variations of radon as ANN's inputs. Also in Turkey, an earthquake early warning system was developed in [7]. To achieve this goal, an ANN making use of the information provided by a seismic sensor network, that records ground motions, was proposed by the authors.

The unsupervised ANN's version – Kohonen's self-organized maps [23] – was applied to study the concentration and the trend of aftershocks occurred after the Sichuan (China) earthquake in 2008 [27]. The longitude, the latitude and the magnitude of the