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*

ABSTRACT

Keywords:

Earthquake prediction

Seismicity indicators

Feature selection

Time series

Supervised classification

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 infor-mation 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 charac-terized 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).

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 aftershocks occurring within the next two days after the main shock were predicted.

Finally, a general-purpose methodology, based on ANN, was introduced in [22] to calculate the probability of earthquake's inter-arrival time for a particular zone, given a magnitude interval or a magnitude greater than a preset threshold. To show its efficiency, the authors validated their methodology on a wide variety of datasets.

Note that despite the great effort done by the scientific community to develop effective methods to predict earthquakes, no successful method has yet been found. The main weakness of the methods is related with the lack of features analysis. Thus, works such as [1,2,26,35,40] applied a set of inputs intuitively selected but no correlation with the output in test sets was studied or shown. Furthermore, the spatial uncertainty is typically too large to produce accurate predictions for particular areas. In particular, [7,9,37] present such shortcoming. Temporal uncertainty is not usually considered in statistical-based methods [14,21,32,48], i.e. it is highly probable that, in active zones, an earthquake occurs within one year.

For all the aforementioned, it becomes essential to develop methods able to make predictions with an optimal set of inputs, for a reduced area and for a short-time prediction horizon. That is exactly what it is proposed in this work: analysis of several features proposed by some works, predictions for reduced areas of a maximum of 1° × 1°, and a temporal horizon of five to seven days.

3. Methodology

This section describes the tasks accomplished to improve the prediction strategies followed in [40,30,35]. Fig. 1 illustrates the full process. Every task is described below:

1. First of all, it is worth noting that a prediction problem has been turned into a binary classification one. To perform such a task, the labels assigned to every event or earthquake have information about the future. That is, every sample has been labeled with an *1* if an earthquake with a magnitude larger than a preset threshold is occurring within the next days; and with a *0* if not. The horizon of prediction has been set to five days for Chile (as in [40]) and to seven days for the Iberian Peninsula (as discussed in [30]). The triggering thresholds are those that ensure balanced training sets as proposed in [30,40].
2. Analysis of the quality of the features used in [40,30,35]. The information gain has been measured for each set of features separately. Table 1 lists the features considered in this work, where b_i are the ith Gutenberg-Richter's b-values calculated as in Eq. (9) from [40], OU stands for Omori-Utsu and GR for Gutenberg-Richter.
3. Selection of the best features, in terms of information gain (see Section 3.1), to be used as ANN's inputs. That is, given the sets $F_1 = \{f_1^1, f_2^1, \ldots, f_7^1\}$ – corresponding to the features proposed in [40,30] – and $F_2 = \{f_1^2, f_2^2, \ldots, f_9^2\}$ – corresponding to the features proposed in [35] –, every feature is evaluated and ranked according to their information gain. To have a more detailed description of the features, please refer to such papers. Then, the new set of features, F', with the seven features with higher gain of information is formed. That is, $F' = \{f'_1, f'_2, \ldots, f'_7\}$ where every f'_i can belong to either F_1 or F_2, and all these seven features exhibit higher information gain than the discarded nine ones ($\#(F_1 \cup F_2) = 16$). Note that F' is composed of seven features as the ANN architecture applied in both [40,30] is going to be used.
4. Evaluation of the new set of features by means of a wide variety of quality parameters (see Section 3.2), typically used for assessing classifiers performance.
5. Application of tests to show the statistical relevance of the results achieved (see Section 3.3). That is, to show that the results obtained with the new set of features outperform former sets, not only on average but also by means of statistical tests. Additionally, the analysis of others classifiers' performance is also provided to confirm that the ANN proposed is the one that fits better in this particular problem.

Table 1
Summary of the set of features evaluated, including formulas and description.

Feature	Notation	Description
f_1^1	x_1	$b_i - b_{i-4}$
f_2^1	x_2	$b_{i-4} - b_{i-8}$
f_3^1	x_3	$b_{i-8} - b_{i-12}$
f_4^1	x_4	$b_{i-12} - b_{i-16}$
f_5^1	x_5	$b_{i-16} - b_{i-20}$
f_6^1	x_6	OU's law
f_7^1	x_7	Dynamic GR's law
f_1^2	T	Elapsed time
f_2^2	M_{mean}	Mean magnitude
f_3^2	$dE^{1/2}$	Square root of seismic energy
f_4^2	β	Slope of magnitude-log plot
f_5^2	a	a-value from GR's law
f_6^2	ΔM	Magnitude deficit
f_7^2	μ	Mean time
f_8^2	σ	Coefficient of variation
f_9^2	η	Mean square deviation

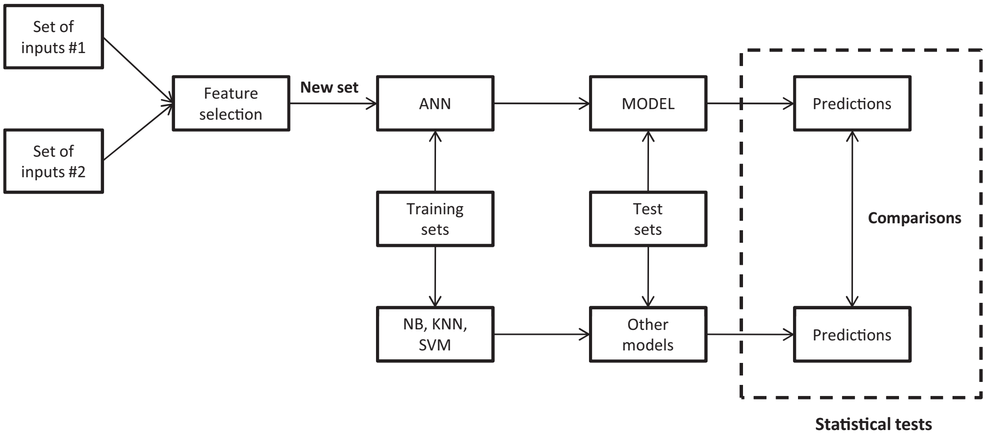

Fig. 1. Steps involved in the methodology.

Finally, it is important to remark that a prediction is made every time that an earthquake of magnitude larger than 3.0 occurs. In [40,30] it was shown that the cutoff magnitude for the earthquakes' database of Chile and the Iberian Peninsula is 3.0. Ought to the seismic activity of the areas under study, a prediction is made almost daily.

3.1. Feature selection based on information gain

The open source Weka software [34] has been used to measure the information gain associated with each feature with respect to the class. This is a widely used standard feature selection method that does not take into account feature interaction. This is the first time that it has been applied for earthquake prediction. Information gain has shown its usefulness in many fields such as image processing [13,43], text categorization [53], spam filtering [49] or brain-computer interface classification [24,45].

The information gain of a given feature f, regarding the class attribute C, is the reduction in uncertainty about the value of C, when the value of f is known, $I(C; f)$. The uncertainty about the value of C is measured by means of the entropy, denoted by $H(C)$. The uncertainty about the value of C, when f is known, is given by the conditional entropy of C given f, $H(C|f)$. In other words:

$$I(C;f) = H(C) - H(C|f) \tag{1}$$

where the entropy of C, for discrete variables is defined as:

$$H(C) = -\sum_{i=1}^{k} P(c_i) log_2(P(c_i)) \tag{2}$$

assuming that C can take values in $\{c_1, \ldots, c_k\}$, and $P(c_i)$ is the probability that $C = c_i$.

Then, the conditional entropy of C given f, assuming that f can take values in $\{f_1, \ldots, f_m\}$ is defined as:

$$H(C|f) = H(C) = -\sum_{j=1}^{m} P(f_j) H(C|f_j) \tag{3}$$

where $P(f_j)$ is the probability that $f = f_j$.

If the input feature f is continuous then, in order to compute its information gain with the class attribute C, all possible binary attributes, f_θ, are considered that arise from f when a cutoff threshold θ is chosen on f. θ may take values from all the range of f. In this case, the information gain formula is reduced to:

$$I(C;f) = argmax_{f_\theta} I(C, f_\theta) \tag{4}$$

In this particular context, Weka has to deal with continuous features, as all ANN's inputs are real values. To calculate the information gain associated with every feature Weka performs a previous discretization of all the variables, turning the continuous problem into a discrete one, which is typically easier to solve.

3.2. Quality parameters

To assess the performance of the ANN's designed, several parameters have been used. In particular:

1. True positives (TP). The number of times that an upcoming earthquake was properly predicted.
2. True negatives (TN). The number of times that neither the ANN triggered an alarm nor an earthquake occurred.
3. False positives (FP). The number of times that the ANN erroneously predicted the occurrence of an earthquake.
4. False negatives (FN). The number of times that the ANN did not trigger an alarm but an earthquake did occur.

The combination of these parameters leads to the calculation of:

$$P_0 = \frac{TN}{TN + FN} \tag{5}$$

$$P_1 = \frac{TP}{TP + FP} \tag{6}$$

where P_0 denotes the well-known negative predictive value, and P_1 the well-known positive predictive value.

Additionally, two more parameters that correspond to common statistical measures of supervised classifiers performance have been used to evaluate the performance of the ANN's. These two parameters, sensitivity or rate of actual positives correctly identified as such (denoted by S_n) and specificity or rate of actual negatives correctly identified (denoted by S_p), are defined as:

$$S_n = \frac{TP}{TP + FN} \tag{7}$$

$$S_p = \frac{TN}{TN + FP} \tag{8}$$

3.3. Statistical tests

A statistical analysis is proposed to evaluate the significance of the new approach, following the non-parametric procedures discussed in García et al. [16].

When several classifiers need to be compared two different kind of tests can be applied, depending on the previous statistical knowledge of the data. Thus, when data are normally distributed and variances among populations are equal (homogeneity), it is usual to apply the ANOVA (ANalysis Of VAriance) test [12,41]. It can also be used when the number of hypothesis is small enough (typically less than 50), regardless data distribution.

However, the non-parametric version of ANOVA, the Friedman test [15], is used when any assumption about data can be made. This test is precisely the selected method to be applied in this work, since a priori no previous knowledge is known from data. This test is has been selected for this work since no previous knowledge from data is known a priori. This method would still be valid even if data followed a normal distribution. Indeed, it is a generalization of ANOVA suitable for any kind of data distributions.

For these reasons, the Friedman test has been selected in this work. The steps are now detailed. Given a matrix $n \times k$ of data $\{x_{ij}\}$, where n is the number of rows (experiments) and k the number of columns (the algorithms to be tested), the ranks within each row is calculated. In case of tied values, the average of the ranks that would have been assigned without ties is assigned to each tied value. The data are replaced with a new matrix $\{r_{ij}\}$ (also with $n \times k$ elements), where each element r_{ij} is the rank of x_{ij} within row i. Afterwards, the following values need to be found:

$$\bar{r}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} r_{ij} \tag{9}$$

$$\bar{r} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij} \tag{10}$$

$$SS_t = n\sum_{j=1}^{k}(\bar{r}_{\cdot j} - \bar{r})^2 \tag{11}$$

$$SS_e = \frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{r}_{ij} - \bar{r})^2 \tag{12}$$

Once (11) and (12) have been calculated, the result of the statistical test is given by $Q = SS_t/SS_e$. Last, when n or k are big enough (typically $n > 10$ or $k > 5$), which is the situation occurred in this work,

Q's probability distribution can be approximated by that of a χ^2 distribution. In such case, the p-value is given by $P(\chi^2_{k-1} \geqslant Q)$.

Finally, once that it has been verified that p has a significant value, a post hoc analysis is carried out in order to find differences between couples of algorithms. There are many strategies than can be followed, however, in this work, the Holm and Hochberg tests are applied to follow the methodology described in [29]. Let $H_1, \ldots, H_m$ be a family of hypotheses and $P_1, \ldots, P_m$ their associated p-values. First, the p-values are sorted $P_{(1)}, \ldots, P_{(m)}$, denoting their associated hypotheses as $H_{(1)}, \ldots, H_{(m)}$. Given a level of significance α, let be k the minimal index such that:

$$P_{(k)} \geqslant \frac{\alpha}{m - 1 + k} \tag{13}$$

The procedure then rejects $H_{(1)}, \ldots, H_{(k-1)}$ and does not reject $H_{(k)}, \ldots, H_{(m)}$. Note that k could be equal to 1. In this case, any hypothesis would be rejected. Also, there could not be any k satisfying (13); in that case, all hypotheses are directly rejected. Finally, the method is satisfactory if it ensures that $P(k \geqslant i) < \alpha$.

4. Case study #1: Chile

This section introduces the results of applying the methodology proposed to four main Chilean regions. In particular, the regions with cells varying from 0.5° × 0.5° to 1° × 1° around the cities of Talca, Santiago, Pichilemu and Valparaíso have been considered. These sets can be downloaded from the National Service of Seismology of the University of Chile upon request [33].

In order to make easier comparisons, for the four zones, the training and test sets chosen are those analyzed in [40]. Additionally, a comparison with Naive Bayes (NB) [18], K-nearest neighbors (KNN) [11] and support-vector machines (SVM) [10] has been calculated to show that ANN's have better performance than any other classifier. The default configuration provided by Weka 3.6 [34] has been used in all cases for the setup of NB, KNN and SVM.

4.1. Talca

For this zone, a training set comprising earthquakes occurred from June 19th 2003 to March 21st 2010 has been used. Analogously, the test set was composed of the earthquakes occurred from March 24th 2010 to January 4th 2011.

Tables 2 and 3 represent the information gain calculated for every seismicity indicator used in both Reyes et al. [40] and Panakkat and Adeli [35] works, respectively. In particular, the column *Average merit* is the average information gain and its standard deviation obtained from the calculation of the information gain in a 10-fold cross-validation process. It can be observed that there are three features with null contribution in Table 2 and four in Table 3. Therefore, the inclusion of such features in the original prediction may have influenced negatively the final results.

The next step consists in selecting the best features or, in other words, those that presented greater information gain. The number of features is limited to seven, to fulfill the architectural constraints presented in [40], as this particular architecture of ANN is desired to be used. Therefore, the new set of input parameters is:

$$Inputs_{Talca} = \{x_7, T, \mu, M_{mean}, \sigma, x_6, x_1\} \tag{14}$$

Given this new set of seismicity indicators, the ANN's were applied again yielding the results summarized in Table 4. It can be concluded that the new set of inputs led to significantly better results.

Finally, to check the adequacy of using ANN, different classifiers have been applied using as input parameters those in [40], those in [35], and the new set selected before. This information can be found in Table 5. Note that NB stands for Naive Bayes, KNN for K-Nearest Neighbors and SVM for Support-Vector Machine.

Table 2
Average information gain for every feature in [40] for Talca.

Average merit	Feature
0.277 ± 0.030	x_7
0.017 ± 0.052	x_6
0.010 ± 0.030	x_1
0.009 ± 0.028	x_3
0 ± 0	x_2
0 ± 0	x_4
0 ± 0	x_5

Table 3
Average information gain for every feature in [35] for Talca.

Average merit	Feature
0.203 ± 0.019	T
0.177 ± 0.019	μ
0.160 ± 0.018	M_{mean}
0.084 ± 0.055	σ
0.009 ± 0.027	η
0 ± 0	$dE^{1/2}$
0 ± 0	ΔM
0 ± 0	a
0 ± 0	β

Table 4
ANN performance for Talca with the new set of seismicity indicators.

Parameter	ANN [40]	ANN [35]	New ANN
TP	3	2	7
TN	23	22	30
FP	14	15	7
FN	5	7	1
Sensitivity (%)	37.50	22.22	87.50
Specificity (%)	62.16	59.46	81.08
P_0 (%)	82.14	75.86	96.77
P_1 (%)	17.65	11.76	50.00
Mean (%)	49.86	42.33	78.84

Two conclusions can be drawn from the analysis of Tables 4 and 5. First, the use of the new set has led to a general improvement comparing the results obtained in [40,35], that is, none of the eleven configurations outperformed on average the results obtained by the ANN when the new set of inputs was used. Second, the comparison between the four classifiers presented shows that the ANN clearly presents the best results. However, for the SVM, all input sets generated the same results, showing the inability of this classifier to deal with this kind of data.

4.2. Santiago

For this zone, the training set contains the earthquakes occurred from May 13th 2003 to June 2nd 2004. The test set was composed of the earthquakes occurred from June 23rd 2004 to January 16th 2006.

Tables 6 and 7 represent the information gain calculated for every seismicity indicator described in [40,35], respectively. Again, the column *Average merit* stands for the average information gain and its standard deviation obtained by means of a 10-fold cross-validation process in the training set. Two features with null contribution are reported in Table 6 and seven in Table 7; therefore, it is also desirable to conduct an analysis to obtain a better set of features.

Table 5
Several classifiers performance with new set of inputs in Talca.

Parameter	NB [40]	NB [35]	New NB
TP	0	1	3
TN	35	23	23
FP	2	14	14
FN	8	7	5
Sensitivity (%)	0.00	12.50	37.50
Specificity (%)	94.59	62.16	62.16
P_0 (%)	81.40	76.67	82.14
P_1 (%)	0.00	6.67	17.65
Mean (%)	44.00	39.50	49.86
Parameter	KNN [40]	KNN [35]	New KNN
TP	0	3	3
TN	37	26	24
FP	0	11	13
FN	8	5	5
Sensitivity (%)	0.00	37.50	37.50
Specificity (%)	100.00	70.27	64.86
P_0 (%)	82.22	83.87	82.76
P_1 (%)	0.00	21.43	18.75
Mean (%)	45.56	53.27	50.97
Parameter	SVM [40]	SVM [35]	New SVM
TP	0	0	0
TN	37	37	37
FP	0	0	0
FN	8	8	8
Sensitivity (%)	0.00	0.00	0.00
Specificity (%)	100.00	100.00	100.00
P_0 (%)	82.22	82.22	82.22
P_1 (%)	0.00	0.00	0.00
Mean (%)	45.56	45.56	45.56

Again, the next step is to select the seven features that presented greater information gain. In this case, the new set of seismicity indicators to be used as ANN's inputs is:

$$Inputs_{Santiago} = \{T, \sigma, x_1, x_3, x_5, x_4, x_2\} \quad (15)$$

The ANN was applied with these new inputs yielding the results summarized in Table 8. It can be observed that the new set of seismicity parameters obtained better results (71.89% versus 65.68% and 53.33% accuracy).

Finally, to check the adequacy of using ANN, different classifiers have been applied using as input parameters those in [40], those in [35], and the combined set selected above. This information can be found in Table 9.

The followings conclusions can be drawn from the observation of Table 9. First, none of the eleven configurations outperformed on average the results obtained by the ANN when the new set of inputs was used. For SVM, all input sets generated the same results, showing the inability of SVM to deal with this kind of data. For NB, the new results were similar to those obtained by the parameters in [40] and significantly better than those of [35]. Only in the application of KNN the new set obtained worse results than those from [40] (62.40% versus 58.78%).

Table 6
Average information gain for every feature in [40] for Santiago.

Average merit	Feature
0.012 ± 0.002	x_1
0.009 ± 0.001	x_3
0.009 ± 0.001	x_5
0.009 ± 0.002	x_4
0.003 ± 0.002	x_2
0 ± 0	x_7
0 ± 0	x_6

Table 7
Average information gain for every feature in [35] for Santiago.

Average merit	Feature
0.113 ± 0.016	T
0.029 ± 0.061	σ
0 ± 0	M_{mean}
0 ± 0	$dE^{1/2}$
0 ± 0	μ
0 ± 0	β
0 ± 0	ΔM
0 ± 0	η
0 ± 0	a

4.3. Valparaíso

For this zone, a training set comprising earthquakes occurred from January 31st 2006 to December 19th 2008 has been used. Analogously, the test set was composed of the earthquakes occurred from December 20th 2008 to February 10th 2011.

Tables 10 and 11 represent the information gain and its standard deviation (column *Average merit*) that every seismicity indicator exhibited in [40,35], respectively, when a 10-fold cross-validation process was applied to the training sets. As for Talca and Santiago, several features had null contribution: five for the set in [40] and four for the set of [35]. Again, it becomes necessary to assess the quality of the features in order to obtain a better set of them.

Then, the features with greater information gain are selected. As commented before, the number of features is limited to seven to serve as input of the ANN presented in [40]. The new set of inputs is:

$$Inputs_{Valp.} = \{T, x_7, \mu, \sigma, x_6, dE^{1/2}, M_{mean}\} \quad (16)$$

Given this new set of seismicity indicators, the ANN are again applied generating the results reported in Table 12. The same conclusion is reached: the new set of inputs obtained significantly better results.

Finally, to assess the ANN performance, different classifiers have been applied using as input parameters those in [40], those in [35], and the new set defined in Eq. (16). This information is summarized in Table 13.

Again, there was no classifier with the set of seismicity indicators outperforming the new set applied to the ANN. The comparison between classifiers shows that the new sets proposed also performs better results, on average, than the results obtained with the sets proposed in [40,35], except for SVM.

Table 8
ANN performance for Santiago with the new set of seismicity indicators.

Parameter	ANN [40]	ANN [35]	New ANN
TP	5	2	5
TN	101	99	105
FP	7	9	3
FN	9	12	9
Sensitivity (%)	35.71	14.29	35.71
Specificity (%)	93.52	91.67	97.22
P_0 (%)	91.82	89.19	92.11
P_1 (%)	41.67	18.18	62.50
Mean (%)	65.68	53.33	71.89

Table 9
Several classifiers performance with new set of inputs in Santiago.

Parameter	NB [40]	NB [35]	New NB
TP	4	6	6
TN	98	76	86
FP	10	32	22
FN	10	8	8
Sensitivity (%)	28.57	42.86	42.86
Specificity (%)	90.74	70.37	79.63
P_0 (%)	90.74	90.48	91.49
P_1 (%)	28.57	15.79	21.43
Mean (%)	59.66	54.87	58.85

Parameter	KNN [40]	KNN [35]	New KNN
TP	6	0	0
TN	93	82	101
FP	15	26	7
FN	8	11	11
Sensitivity (%)	42.86	21.43	21.43
Specificity (%)	86.11	75.93	93.52
P_0 (%)	92.08	88.17	90.18
P_1 (%)	28.57	10.34	30.00
Mean (%)	62.40	48.97	58.78

Parameter	SVM [40]	SVM [35]	New SVM
TP	0	0	0
TN	108	108	108
FP	14	14	14
FN	0	0	0
Sensitivity (%)	0.00	0.00	0.00
Specificity (%)	88.52	88.52	88.52
P_0 (%)	100	100	100
P_1 (%)	0.00	0.00	0.00
Mean (%)	47.13	47.13	47.13

Table 10
Average information gain for every feature in [40] for Valparaíso.

Average merit	Feature
0.244 ± 0.202	x_7
0.175 ± 0.016	x_6
0 ± 0	x_1
0 ± 0	x_2
0 ± 0	x_3
0 ± 0	x_4
0 ± 0	x_5

Table 11
Average information gain for every feature in [35] for Valparaíso.

Average merit	Feature
0.270 ± 0.025	T
0.193 ± 0.017	μ
0.188 ± 0.022	σ
0.045 ± 0.055	$dE^{1/2}$
0.009 ± 0.028	M_{mean}
0 ± 0	η
0 ± 0	ΔM
0 ± 0	a
0 ± 0	β

4.4. Pichilemu

For this zone, a training set comprising earthquakes occurred from August 10th 2005 to March 31st 2010 has been used. Analogously, the test set was composed of the earthquakes occurred from April 1st 2010 to October 8th 2011.

Table 12
ANN performance for Valparaíso with the new set of seismicity indicators.

Parameter	ANN [40]	ANN [35]	New ANN
TP	20	34	29
TN	59	47	60
FP	3	15	2
FN	24	10	15
Sensitivity (%)	45.45	77.27	65.91
Specificity (%)	95.16	75.81	96.77
P_0 (%)	71.08	82.46	80.00
P_1 (%)	86.96	69.39	93.55
Mean (%)	74.66	76.23	84.06

Tables 14 and 15 report the average information gain and its associated standard deviation calculated for every seismicity indicator introduced in [40,35], respectively. This time there were no features with null average information gain in Table 14 but four features had null contribution in Table 15. Again, the necessity to conduct an analysis to obtain a better set of features is highlighted.

The next step is to select the best features or, in other words, those that presented greater information gain. The number of features is limited to seven, to use the ANN in [40]. Therefore, the new set of input parameters is:

$$Inputs_{Pichilemu} = \{x_7, x_6, M_{mean}, dE^{1/2}, \sigma, x_4, x_5\} \quad (17)$$

Given this new set of seismicity indicators, the ANN's are applied yielding the results summarized in Table 16. Similarly to the other three zones, the new set of inputs generated significantly better results.

Finally, to evaluate the performance of using ANN's, different classifiers have been applied using as input parameters those in

Table 13
Several classifiers performance with the new set of inputs in Valparaíso.

Parameter	NB [40]	NB [35]	New NB
TP	18	21	19
TN	58	53	61
FP	4	9	1
FN	26	23	25
Sensitivity (%)	40.91	47.73	43.18
Specificity (%)	93.55	85.48	98.39
P_0 (%)	69.05	69.74	70.93
P_1 (%)	81.82	70.00	95.00
Mean (%)	71.33	68.24	76.87

Parameter	KNN [40]	KNN [35]	New KNN
TP	30	21	33
TN	52	46	53
FP	10	16	9
FN	14	23	11
Sensitivity (%)	68.18	47.73	75.00
Specificity (%)	83.87	74.19	86.48
P_0 (%)	78.79	66.67	82.81
P_1 (%)	75.00	56.76	78.57
Mean (%)	76.46	61.34	80.47

Parameter	SVM [40]	SVM [35]	New SVM
TP	35	12	38
TN	46	57	31
FP	16	5	31
FN	9	32	6
Sensitivity (%)	79.55	27.27	86.36
Specificity (%)	74.19	91.94	50.00
P_0 (%)	83.64	64.04	83.78
P_1 (%)	68.63	70.59	55.07
Mean (%)	76.50	63.46	68.80

Table 14
Average information gain for every feature in [40] for Pichilemu.

Average merit	Feature
0.512 ± 0.104	x_7
0.499 ± 0.086	x_6
0.113 ± 0.012	x_4
0.103 ± 0.010	x_5
0.101 ± 0.009	x_3
0.010 ± 0.030	x_2
0.010 ± 0.031	x_1

Table 15
Average information gain for every feature in [35] for Pichilemu.

Average merit	Feature
0.415 ± 0.026	M_{mean}
0.298 ± 0.027	$dE^{1/2}$
0.152 ± 0.037	σ
0.102 ± 0.038	μ
0.100 ± 0.076	T
0 ± 0	η
0 ± 0	ΔM
0 ± 0	a
0 ± 0	β

[40], those in [35], and the new set above selected. This information is reported in Table 17.

From the observation of Tables 16 and 17 two conclusions can be drawn. In concordance with Talca, Santiago and Pichilemu, the best results have been obtained with the new set of seismicity indicators using the ANN as classifier. Second, the use of such a set led to a general improvement in all methods, except for NB, where only the use of the indicators proposed in [40,35] reached better results on average.

5. Case study #2: Iberian Peninsula

This section presents the results of applying the methodology proposed to the two most seismic zones of the Iberian Peninsula (the Alborán Sea and West Azores-Gibraltar Fault), with cells around 1° × 1°. Although rough data can be downloaded from [20], they have been preprocessed with the help of the Spanish's National Geographical Institute.

For both zones, to make easier comparisons, the training and test sets chosen are those analyzed in [30]. Additionally, a comparison with Naive Bayes (NB) [18], M5P [50] and support-vector machines (SVM) [10] has been performed to show that the proposed ANN's have better performance than any other classifier. The default configuration provided by Weka 3.6 [34] has been used for the setup of NB, M5P, and SVM.

Table 16
ANN performance for Pichilemu with the new set of seismicity indicators.

Parameter	ANN [40]	ANN [35]	New ANN
TP	13	14	21
TN	91	76	88
FP	2	17	5
FN	16	15	8
Sensitivity (%)	44.83	48.28	72.41
Specificity (%)	97.85	81.72	94.62
P_0 (%)	85.05	83.52	91.67
P_1 (%)	86.67	45.16	80.77
Mean (%)	78.60	64.67	84.87

Table 17
Several classifiers performance with new set of inputs in Pichilemu.

Parameter	NB [40]	NB [35]	New NB
TP	12	3	3
TN	91	93	90
FP	2	0	3
FN	17	26	26
Sensitivity (%)	41.38	10.34	10.34
Specificity (%)	97.85	100	96.77
P_0 (%)	84.26	78.15	77.59
P_1 (%)	85.71	100	50.00
Mean (%)	77.30	72.12	58.68
Parameter	KNN [40]	KNN [35]	New KNN
TP	20	13	21
TN	38	60	62
FP	55	30	28
FN	9	16	8
Sensitivity (%)	68.97	44.83	72.41
Specificity (%)	40.86	66.67	68.89
P_0 (%)	80.85	78.95	88.57
P_1 (%)	26.67	30.23	42.86
Mean (%)	54.34	55.17	68.18
Parameter	SVM [40]	SVM [35]	New SVM
TP	0	14	13
TN	93	82	87
FP	0	11	6
FN	29	15	16
Sensitivity (%)	0.00	48.28	44.83
Specificity (%)	100	88.17	93.55
P_0 (%)	76.23	84.54	84.47
P_1 (%)	0.00	56.00	68.42
Mean (%)	44.06	69.25	72.82

5.1. The Alborán Sea

For this area, the training set was composed of 122 linearly independent vectors occurred from December 5th 2004 to May 7th 2005. Analogously, the test set included 79 vectors generated from May 7th 2005 to August 10th 2005. A thorough discussion on the election of both sets can be found in [30].

Tables 18 and 19 report the average information gain and its associated standard deviation calculated for every seismicity indicator used in [30,35], respectively. Again, four features in Table 18 and four more in Table 19 had null contribution, which supports the necessity of selecting a better set of features.

Once the information gain has been obtained for all the features, the seven best ones are selected to be used as inputs of the ANN's architecture introduced in [30]. This new set of inputs is:

$$Inputs_{Alb.Sea} = \{x_6, x_7, M_{mean}, T, \sigma, \eta, dE^{1/2}\} \qquad (18)$$

Then, the ANN is applied with the new inputs and the results are summarized in Table 20. Note that as noticed in the four Chilean zones analyzed in the previous section, the new inputs significantly improved the results obtained by using separately the inputs proposed in [30,35].

Different classifiers have been applied to assess their performance with the new set of seismicity indicators. The results are included in Table 21.

Similarly to the Chile's zones, the ANN with the new set of seismicity indicators reported the best results. Also, except for NB, the new use of the inputs described in [30,35] generated better results

Table 18
Average information gain for every feature in [30] for the Alborán Sea.

Average merit	Feature
0.237 ± 0.050	x_6
0.217 ± 0.042	x_7
0.009 ± 0.027	x_3
0 ± 0	x_1
0 ± 0	x_2
0 ± 0	x_4
0 ± 0	x_5

Table 19
Average information gain for every feature in [35] for the Alborán Sea.

Average merit	Feature
0.133 ± 0.021	M_{mean}
0.103 ± 0.010	T
0.094 ± 0.006	σ
0.060 ± 0.049	η
0.012 ± 0.035	$dE^{1/2}$
0 ± 0	μ
0 ± 0	ΔM
0 ± 0	a
0 ± 0	β

in terms of mean accuracy. This confirms the need of a study a priori of the information used as ANN's inputs.

5.2. Western Azores-Gibraltar Fault

For this area, the training set contained 122 linearly independent vectors occurred from July 28th 2003 to June 5th 2005. Analogously, the test set included 79 vectors generated June 8th 2005 to June 26th 2006. Again, the discussion on the election of both sets is described in [30].

The average information gain and its standard deviation calculated for all the indicators in [30,35] is reported in Tables 22 and 23, respectively. This time there were no features with null average information gain in Table 14 but four features had null contribution in Table 15. Again, it is confirmed the necessity to conduct an analysis to obtain a better set of features.

The best seven features to be used as inputs of the ANN's architecture introduced in [30] are then selected. The new set is:

$$Inputs_{W.Azores-Gib.F.} = \{x_7, T, \mu, M_{mean}, x_6, \sigma, \eta\} \tag{19}$$

Then, the ANN is applied with such inputs and the results are summarized in Table 24. Note that as observed in all the previous zones, the use of the new set obtained better results than those of [30,35].

The next step is to apply different classifiers to compare their performance when using the new set of inputs. The results are listed in Table 25.

Table 20
ANN performance for the Alborán Sea with the new set of seismicity indicators.

Parameter	ANN [30]	ANN [35]	New ANN
TP	10	34	21
TN	34	6	34
FP	5	33	6
FN	30	6	19
Sensitivity (%)	25.00	85.00	52.50
Specificity (%)	87.18	15.38	85.00
P_0 (%)	53.13	50.00	64.15
P_1 (%)	66.67	50.75	77.78
Mean (%)	57.99	50.28	69.86

Table 21
Several classifiers performance with new set of inputs in the Alborán Sea.

Parameter	NB [30]	NB [35]	New NB
TP	35	33	34
TN	7	5	4
FP	32	34	35
FN	5	7	6
Sensitivity (%)	87.50	82.50	85.00
Specificity (%)	17.95	12.82	10.26
P_0 (%)	58.33	41.67	40.00
P_1 (%)	52.24	49.25	49.28
Mean (%)	54.01	46.56	46.13
Parameter	M5P [30]	M5P [35]	New M5P
TP	2	38	11
TN	27	0	8
FP	12	39	23
FN	38	2	3
Sensitivity (%)	5.00	95.00	78.57
Specificity (%)	69.23	0.00	25.81
P_0 (%)	41.54	0.00	72.73
P_1 (%)	14.29	49.35	32.35
Mean (%)	32.51	36.09	52.36
Parameter	SVM [30]	SVM [35]	New SVM
TP	8	39	40
TN	17	0	0
FP	21	39	39
FN	32	1	0
Sensitivity (%)	20.00	97.50	100
Specificity (%)	44.74	0.00	0.00
P_0 (%)	34.69	0.00	0.00
P_1 (%)	27.59	50.00	50.63
Mean (%)	31.75	36.88	37.66

Table 22
Average information gain for every feature in [30] for West Azores-Gibraltar Fault.

Average merit	Feature
0.232 ± 0.020	x_7
0.140 ± 0.018	x_6
0 ± 0	x_1
0 ± 0	x_2
0 ± 0	x_3
0 ± 0	x_4
0 ± 0	x_5

Table 23
Average information gain for every feature in [35] for West Azores-Gibraltar Fault.

Average merit	Feature
0.203 ± 0.019	T
0.177 ± 0.019	μ
0.160 ± 0.018	M_{mean}
0.084 ± 0.055	σ
0.009 ± 0.027	η
0 ± 0	$dE^{1/2}$
0 ± 0	ΔM
0 ± 0	a
0 ± 0	β

From the observation of Table 25 several relevant conclusions can be drawn. First of all, the new set of parameters led to better results in all the evaluated classifiers. Secondly, the ANN remains as the classifier with the best performance. When the inputs proposed in [30] were used for M5P, the classifier was unable to predict any earthquake (69 TN's and 10 FP's means that all the predicted labels were 0). On the contrary, the use of the inputs in

[35] generated completely opposed results, since it was unable to discard any zero-level event (69 FP's and 10 TP's means that all the predicted labels were 1). However, the combination of the best inputs generated competitive results, even better than those of NB or SVM.

6. Statistical tests

This section is to show that the classifiers applied to the four Chilean and the two Iberian Peninsula zones have significant statistically different results. As described in Section 3.3, Friedman's test is going to be used, declaring a level of significance of $p < 0.05$.

6.1. Chilean data

First, the matrix r_{ij} is constructed, with a size $(n \times k) = (4, 12)$, where $n = 4$ represents the four zones under analysis and $k = 12$ the twelve different classifiers applied. Each element in the matrix are the performance's mean values retrieved from Tables 12, 13, 16, 17. For legibility reasons, Table 26 summarizes the transpose matrix, r'_{ij}.

Next step is to calculate the rankings matrix as shown in Table 27.

From the analysis of Table 27 two conclusions can clearly be drawn. First, the use of the new set of inputs improved, on average, all classifiers (ANN, NB, KNN and SVM) and, second, the ANN obtained better results than any other classifier (as wanted to be proved). However, this fact does not reject the null hypothesis yet. But the application of Friedman's test led to $p = P(\chi^2_{k-1} \geqslant Q) = 0.0432$, which satisfies the initial assumption $p < 0.05$, thus concluding that p reached a significant value and, now, rejecting the null hypothesis.

Since the p-value is less than 0.05, a post hoc analysis has been carried out in order to prove that the results obtained by the ANN with the new set of inputs are statistically different (and therefore better) than those obtained by the ANN's with the initial set of inputs. The Holm-Hochberg test has been applied to compare separately the ANN with the new set of inputs and the ANN's with the inputs proposed in [40,35], respectively. Table 28 shows the p-values obtained by the ANN in [40] and by the ANN in [35] for a level of significance $\alpha = 0.05$. The test concludes that the new set of inputs generates better results since the test rejects all hypotheses.

Table 24
ANN performance for Western Azores-Gibraltar Fault with the new set of seismicity indicators.

Parameter	ANN [30]	ANN [35]	New ANN
TP	5	5	8
TN	64	52	63
FP	5	17	6
FN	5	5	2
Sensitivity (%)	50.00	50.00	80.00
Specificity (%)	92.75	75.36	91.30
P_0 (%)	92.75	91.23	96.92
P_1 (%)	50.00	22.73	57.14
Mean (%)	71.38	59.83	81.34

Table 25
Several classifiers performance with new set of inputs in Western Azores-Gibraltar Fault.

Parameter	NB [30]	NB [35]	New NB
TP	4	10	4
TN	51	0	53
FP	18	69	16
FN	6	0	6
Sensitivity (%)	40.00	100	40.00
Specificity (%)	73.91	0.00	76.81
P_0 (%)	89.47	0.00	89.83
P_1 (%)	18.18	12.66	20.00
Mean (%)	55.39	28.16	56.66

Parameter	M5P [30]	M5P [35]	New M5P
TP	10	0	5
TN	0	69	48
FP	69	0	21
FN	0	10	5
Sensitivity (%)	100	0.00	50.00
Specificity (%)	0.00	100	69.57
P_0 (%)	0.00	87.34	90.57
P_1 (%)	12.66	0.00	19.23
Mean (%)	28.16	46.84	57.34

Parameter	SVM [30]	SVM [35]	New SVM
TP	5	6	5
TN	22	36	47
FP	47	33	22
FN	5	4	5
Sensitivity (%)	50.00	60.00	50.00
Specificity (%)	31.88	52.17	68.12
P_0 (%)	81.48	90.00	90.38
P_1 (%)	9.62	15.38	18.52
Mean (%)	43.25	54.39	56.75

Table 26
Input values for the Friedman's test (transpose matrix) for Chile.

Classifier	Talca	Santiago	Valparaíso	Pichilemu
ANN [40]	0.4986	0.7860	0.6568	0.7466
ANN [35]	0.4233	0.6467	0.5333	0.7623
New ANN	0.7884	0.8487	0.7189	0.8406
NB [40]	0.4400	0.7730	0.5966	0.7133
NB [35]	0.3950	0.7212	0.5487	0.6824
New NB	0.4986	0.5868	0.5885	0.7687
KNN [40]	0.4556	0.5434	0.6240	0.7646
KNN [35]	0.5327	0.5517	0.4897	0.6134
New KNN	0.5097	0.6818	0.5878	0.8047
SVM [40]	0.4556	0.4406	0.4713	0.7650
SVM [35]	0.4556	0.6925	0.4713	0.6346
New SVM	0.4556	0.7282	0.4713	0.6880

Table 27
Average rankings of the classifiers applied to Chile.

Classifier	Ranking
ANN [40]	2.88
ANN [35]	7.25
New ANN	1.00
NB [40]	5.25
NB [35]	7.25
New NB	4.38
KNN [40]	4.38
KNN [35]	5.63
New KNN	3.50
SVM [40]	7.63
SVM [35]	7.88
New SVM	6.88

Table 28
Holm-Hochberg test using the ANN with the new set of inputs as control algorithm for Chile.

i	Classifier	z	p-value	α/i
2	ANN [35]	4.56	5.01×10^{-6}	0.025
1	ANN [40]	3.65	2.61×10^{-4}	0.050

Table 29
Input values for the Friedman's test (transpose matrix) for the Iberian Peninsula.

Classifier	Alborán Sea	W. Azores-Gibraltar Fault
ANN [30]	0.5799	0.7138
ANN [35]	0.5028	0.5983
New ANN	0.6986	0.8134
NB [30]	0.5401	0.5539
NB [35]	0.4656	0.2816
New NB	0.4613	0.5666
M5P [30]	0.3251	0.2816
M5P [35]	0.3609	0.4684
New M5P	0.5236	0.5734
SVM [30]	0.3175	0.4325
SVM [35]	0.3688	0.5439
New SVM	0.3766	0.5675

Table 30
Average rankings of the classifiers applied to the Iberian Peninsula.

Classifier	Ranking
ANN [30]	2.0
ANN [35]	4.0
New ANN	1.0
NB [30]	5.0
NB [35]	8.5
New NB	6.5
KNN [30]	11.0
KNN [35]	9.5
New KNN	4.0
SVM [30]	11.0
SVM [35]	8.5
New SVM	6.5

Table 31
Holm-Hochberg test using the ANN with the new set of inputs as control algorithm.

i	Classifier	z	p-value	α/i
2	ANN [30]	5.63	2.72×10^{-6}	0.025
1	ANN [35]	3.98	5.03×10^{-5}	0.050

6.2. Iberian Peninsula data

First, the matrix r_{ij} is constructed, with a size $(n \times k) = (2, 12)$, where $n = 2$ represents the two zones analyzed and $k = 12$ the twelve different classifiers applied. Each element in the matrix are the performance's mean values retrieved from Tables 20, 21, 24 and 25. For legibility reasons, Table 29 summarizes the transpose matrix, r'_{ij}.

The rankings matrix is shown in Table 30.

From the analysis of Table 30 two conclusions can clearly be drawn. First, the use of the new set of inputs improved, on average, all classifiers (ANN, NB, M5P and SVM) and, second, the ANN obtained better results than any other classifier, which was the initial hypothesis. Nonetheless, this fact is not enough to reject the null hypothesis. But the application of Friedman's test led to $p = P(\chi^2_{k-1} \geqslant Q) = 0.0209$, which satisfies the initial assumption $p < 0.05$, thus concluding that p reached a significant value. Therefore, the null hypothesis is rejected as wanted to be proved.

Again, the p-value is less than 0.05 and a post hoc analysis has been performed to prove that the ANN with the new set of inputs generated results statistically different (and therefore better) than those obtained by the ANN's with the original set of inputs. The Holm-Hochberg test has been applied to compare separately the ANN with the new set of inputs and the ANN's with the inputs proposed in [30,35], respectively. Table 31 shows the p-values obtained by the ANN in [30] and by the ANN in [35] for a level of significance $\alpha = 0.05$. The test concludes that the new set of inputs generates better results since the test rejects all hypotheses.

7. Discussion on the use of features

Previous sections have provided a large amount of data and tables. This section is to summarize all the information quantitatively presented, so that, general conclusions can be easily drawn.

Table 32 lists the information gain for the six zones analyzed. It can be observed that the inputs that have the highest information gain are $\{x_6, x_7, T, M_{mean}, \mu, \sigma, dE^{1/2}\}$, as they present the highest sum of information gain. The interpretation of the inclusion of these particular features in the new set of seismicity indicators is discussed now.

1. The information provided by the dynamic Gutenberg-Richter's law (information included in x_7) as well as that of Omori-Utsu's law (information codified in variable x_6) are the most valuable indicators to predict short-term earthquakes, as shown in Table 32, where they obtained the two best positions.
2. The knowledge of the average time elapsed between earthquakes as well as its mean value and associated standard deviation (features T, μ, and σ, respectively) also seems to be crucial for an accurate prediction process.

Table 32
Summary of information gain for all the different datasets studied. Ranks for features in Chile and the Iberian Peninsula (IP) are separately shown.

Feature	Talca	Santiago	Valparaíso	Pichilemu	Alborán Sea	Azores	Sum	Rank (Chile)	Rank (IP)
x_1	0.010	0.012	0.000	0.010	0.000	0.000	0.032	12	10
x_2	0.000	0.003	0.000	0.010	0.000	0.000	0.013	13	10
x_3	0.009	0.009	0.000	0.101	0.009	0.000	0.128	8	8
x_4	0.000	0.009	0.000	0.113	0.000	0.000	0.122	9	10
x_5	0.000	0.009	0.000	0.103	0.000	0.000	0.112	10	10
x_6	0.017	0.000	0.175	0.499	0.237	0.140	1.068	2	2
x_7	0.277	0.000	0.244	0.512	0.217	0.232	1.482	1	1
T	0.203	0.113	0.270	0.100	0.103	0.203	0.992	3	3
μ	0.177	0.000	0.193	0.102	0.000	0.177	0.649	5	6
M_{mean}	0.160	0.000	0.009	0.415	0.133	0.160	0.850	4	4
σ	0.084	0.029	0.188	0.152	0.094	0.084	0.631	6	5
$dE^{1/2}$	0.000	0.000	0.045	0.298	0.012	0.000	0.355	7	7
ΔM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	14	10
η	0.009	0.000	0.000	0.060	0.000	0.009	0.078	11	8
β	0.000	0.000	0.000	0.000	0.000	0.000	0.000	14	10
a	0.000	0.000	0.000	0.000	0.000	0.000	0.000	14	9

3. The mean magnitude of the last earthquakes occurred, M_{mean}, has to be also considered to improve the accuracy of any prediction system.
4. The rate of release of square root of energy completes the seven most significative features.

Note that this information was intuitively used in [31], where the authors discovered precursory patterns in the Iberian Peninsula by using only x_7, T and M_{mean} (first, third and fourth best features, respectively). Equally remarkable is the fact that from the seven best features, five of them were used in [35] and the remaining two but most significant ones were used in [40]. This confirms that both set of inputs had, at least, a significative number of meaningful features. However, some of them had null contribution and could have decreased the accuracy of the classifiers (x_1, x_2, x_4, x_5 for the Iberian Peninsula; ΔM, β, a for both Chile and the Iberian Peninsula).

Finally, it is worth noting that the seven best features in terms of information gain for predicting earthquakes in Chile were also the seven best ones for the Iberian Peninsula, with great difference with the remaining nine seismicity parameters analyzed. It is known that Chile and the Iberian Peninsula have different geophysical properties and it is reasonable to think that the selection of this new set of features could be a good starting point to make predictions in other areas of the world.

8. Conclusions

An optimized set of seismicity parameters for earthquake prediction has been obtained in this work. To enhance prediction, the analysis of how different seismicity indicators influence the model generation has been conducted. In particular, a feature selection, based on the information gain provided by each indicator individually, has been done in order to ensure that the ANN's inputs are those with maximum correlation with the output. This strategy has been evaluated on four Chilean areas and on two areas of the Iberian Peninsula, all of them with different geophysical properties to show the generality of the proposed method. A comparison with other well-known techniques has been provided. It is remarkable that the same set of inputs was obtained for Chile and the Iberian Peninsula. This may involve that similar patterns could be found for different seismic areas. Then, the authors propose this combination of seismicity indicators as initial choice for conducting further research in other coteries. The statistical analysis carried out shows that the results are not only better in terms of the quality parameters assessed but also present different statistical distributions. From the statistical tests conducted, it can be concluded that the new approach outperformed every algorithm to which it was compared. In short, the use of the new set of seismicity parameters as ANN's inputs generated the best results in all evaluated cases.

Acknowledgments

The authors want to thank TGT-www.geofisica.cl for the support through Grants Number 2210 and 2211. The financial support given by the Spanish Ministry of Science and Technology, Projects BIA2004-01302 and TIN2011-28956-C02-01 are equally acknowledged. This work has also been funded by Project P11-TIC-7528 by the Junta de Andalucía. Finally, this work has been partially funded by a JPI 2012 Banco Santander's grant.

References

[1] H. Adeli, A. Panakkat, A probabilistic neural network for earthquake magnitude prediction, Neural Networks 22 (2009) 1018–1024.
[2] A.S.N. Alarifi, N.S.N. Alarifi, S. Al-Humidan, Earthquakes magnitude predication using artificial neural network in northern Red Sea area, Journal of King Saud University – Science 24 (2012) 301–313.
[3] C.R. Allen, Responsibilities in earthquake prediction, Bulletin of the Seismological Society of America 66 (1982) 2069–2074.
[4] E.I. Alves, Earthquake forecasting using neural networks: results and future work, Nonlinear Dynamics 44 (1–4) (2006) 341–349.
[5] M. Anad, A. Dash, M.S.J. Kumar, A. Kesarkar, Prediction and classification of thunderstorms using artificial neural network, International Journal of Engineering Science and Technology 3 (5) (2011) 4031–4035.
[6] A. Arauzo-Azofra, J.L. Aznarte, J.M. Benítez, Empirical study of feature selection methods based on individual feature evaluation for classification problems, Expert Systems with Applications 38 (7) (2007) 8170–8177.
[7] M. Bose, F. Wenzel, M. Erdik, PreSIS: a neural network-based approach to earthquake early warning for finite faults, Bulletin of the Seismological Society of America 98 (1) (2008) 366–382.
[8] G. Box, G. Jenkins, Time Series Analysis: Forecasting and Control, John Wiley and Sons, 2008.
[9] G. Chattopadhyay, S. Chattopadhyay, Dealing with the complexity of earthquake using neurocomputing techniques and estimating its magnitudes with some low correlated predictors, Arabian Journal of Geosciences 2 (3) (2009) 247–255.
[10] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.
[11] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.
[12] M. D'Arco, A. Liccardo, N. Pasquino, ANOVA-based approach for DAC diagnostics, IEEE Transactions on Instrumentation and Measurement 61 (7) (2012) 1874–1882.
[13] C.S. Dhir, N. Iqbal, Y. Soo-Young, Efficient feature selection based on information gain criterion for face recognition, in: Proceedings of the IEEE International Conference on Information Acquisition, 2007, pp. 523–527.
[14] J.E. Ebel, D.W. Chambers, A.L. Kafka, J.A. Baglivo, Non-poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California, Seismollogical Research Letters 78 (1) (2007) 57–65.
[15] D.A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press, 2005.
[16] S. García, A. Fernández, J. Luengo, F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft Computing 13 (10) (2009) 959–977.
[17] Y. Han, L. Yu, A variance reduction framework for stable feature selection, Statistical Analysis and Data Mining 5 (5) (2012) 428–445.
[18] D.J. Hand, K. Yu, Idiot's Bayes – not so stupid after all?, International Statistical Review 69 (3) (2001) 385–399
[19] N. Houlié, J.C. Komorowski, M. de Michele, M. Kasereka, H. Ciraba, Early detection of eruptive dykes revealed by normalized difference vegetation index (NDVI) on Mt. Etna and Mt. Nyiragongo, Earth and Planetary Science Letters 246 (3–4) (2006) 231–240.
[20] Spanish's National Geographical Institute, <http://www.ign.es>.
[21] A. Jiménez, A.M. Posadas, K.F. Tiampo, Describing seismic pattern dynamics by means of using cellular automata, Lecture Notes in Earth Sciences 112 (2008) 273–290.
[22] P. Kamatchi, K.B. Rao, N.R. Iyer, S. Arunachalam, Neural network-based methodology for inter-arrival times of earthquakes, Natural Hazards 64 (2) (2012) 1291–1303.
[23] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1986) 59–69.
[24] I. Koprinska, Feature selection for brain-computer interfaces, Lecture Notes in Artificial Intelligence 5669 (2010) 100–111.
[25] F. Kulahci, M. Inceoz, M. Dogru, E. Aksoy, O. Baykara, Artificial neural network model for earthquake prediction with radon monitoring, Applied Radiation and Isotopes 67 (1) (2009) 212–220.
[26] M. Moustra, M. Avraamides, C. Christodoulou, Artificial neural networks for earthquake prediction using time series magnitude data or seismic electric signals, Expert Systems with Applications 38 (12) (2011) 15032–15039.
[27] R. Madahizadeh, M. Allamehzadeh, Prediction of aftershocks distribution using artificial neural networks and its application on the May 12, 2008 Sichuan earthquake, Journal of Seismology and Earthquake Engineering 11 (3) (2009) 111–120.
[28] F. Martínez-Álvarez, A. Troncoso, A. Morales-Esteban, J.C. Riquelme, Computational intelligence techniques for predicting earthquakes, Lecture Notes in Artificial Intelligence 6679 (2) (2011) 287–294.
[29] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, An evolutionary algorithm to discover quantitative association rules in multidimensional time series, Soft Computing 15 (10) (2011) 2065–2084.
[30] A. Morales-Esteban, F. Martínez-Álvarez, J. Reyes, Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence, Tectonophysics 593 (2013) 121–134.
[31] A. Morales-Esteban, F. Martínez-Álvarez, A. Troncoso, J.L. de Justo, C. Rubio-Escudero, Pattern recognition to forecast seismic time series, Expert Systems with Applications 37 (12) (2010) 8333–8342.
[32] K.Z. Nanjo, J.R. Holliday, C.C. Chen, J.B. Rundle, D.L. Turcotte, Application of a modified pattern informatics method to forecasting the locations of future large earthquakes in the central Japan, Tectonophysics 424 (2006) 351–366.
[33] University of Chile, National Service of Seismology, <http://ssn.dgf.uchile.cl/seismo.html>.

[34] WEKA The University of Waikatu, Data mining with open source machine learning software in Java, <http://www.cs.waikato.ac.nz/ml/weka/>.

[35] A. Panakkat, H. Adeli, Neural network models for earthquake magnitude prediction using multiple seismicity indicators, International Journal of Neural Systems 17 (1) (2007) 13–33.

[36] A. Panakkat, H. Adeli, Recent efforts in earthquake prediction (1990–2007), Natural Hazards Review 9 (2) (2008) 70–80.

[37] A. Panakkat, H. Adeli, Recurrent neural network for approximate earthquake time and location prediction using multiple sesimicity indicators, Computer-Aided Civil and Infrastructure Engineering 24 (2009) 280–292.

[38] K. Ramar, T.T. Mirnalinee, An ontological representation for tsunami early warning system, in: Proceedings of the IEEE International Conference on Advances in Engineering, Science and Management, 2012, pp. 93–98.

[39] J. Reyes, V. Cárdenas, A Chilean seismic regionalization through a Kohonen neural network, Neural Computing and Applications 19 (2010) 1081–1087.

[40] J. Reyes, A. Morales-Esteban, F. Martínez-Álvarez, Neural networks to predict earthquakes in Chile, Applied Soft Computing 13 (2) (2013) 1314–1328.

[41] R. Romero-Záliz, C. Rubio-Escudero, I. Zwir, C. del Val, Optimization of multi-classifiers for computational biology: application to gene finding and expression, Theoretical Chemistry Accounts: Theory Computation and Modeling 125 (3) (2010) 599–611.

[42] R. Ruiz, J.C. Riquelme, J.S. Aguilar-Ruiz, Incremental wrapper-based gene selection from microarray expression data for cancer classification, Pattern Recognition 39 (12) (2006) 2383–2392.

[43] C. Shang, D. Barnes, Support vector machine-based classification of rock texture images aided by efficient feature selection, in: Proceedings of the IEEE International Joint Conference on Neural Networks, 2012, pp. 1–8.

[44] S. Srilakshmi, R.K. Tiwari, Model dissection from earthquake time series: a comparative analysis using nonlinear forecasting and artificial neural network approach, Computers and Geosciences 35 (2009) 191–204.

[45] Y. Su, J. Dai, X. Liu, Q. Xu, Y. Zhuang, W. Chen, X. Zheng, EEG channel evaluation and selection by rough set in P300 BCI, Journal of Computational Information Systems 6 (6) (2010) 1727–1735.

[46] W. Sun, S. Shan, C. Zhang, P. Ge, L. Tao, Prediction of typhoon losses in the South-East of China based on B-P network, in: Proceedings of the IEEE International Conference on Artificial Intelligence and Computational Intelligence, 2010, pp. 252–256.

[47] K.F. Tiampo, R. Shcherbakov, Seismicity-based earthquake forecasting techniques: ten years of progress, Tectonophysics 522–523 (2012) 89–121.

[48] Y. Toya, K.F. Tiampo, J.B. Rundle, C.C. Chen, W. Klein, Pattern informatics approach to earthquake forecasting in 3D, Concurrency and Computation: Practice and Experience 22 (2010) 1569–1592.

[49] C. Wang, The study on the spam filtering technology based on Bayesian algorithm, International Journal of Computer Science Issues 10 (3) (2013) 668–675.

[50] Y. Wang, I.H. Witten, Induction of model trees for predicting continious classes, in: Proceedings of the European Conference on Machine Learning, Praga, 1997, pp. 128–137.

[51] A. Zamani, M.R. Sorbi, A.A. Safavi, Application of neural network and ANFIS model for earthquake occurrence in Iran, Earth Science Informatics 6 (2) (2013) 71–85.

[52] F. Zhang, The future of hurricane prediction, Computing in Science and Engineering 13 (1) (2011) 9–12.

[53] Z. Zheng, X. Wu, R. Shihari, Feature selection for text categorization on imbalanced data, ACM SIGKDD Explorations Newsletter 6 (1) (2004) 80–89.