Computational Intelligence Techniques for Predicting Earthquakes

[1] Department of Computer Science, Pablo de Olavide University of Seville, Spain
{fmaralv,ali}@upo.es
[2] Department of Continuum Mechanics, University of Seville, Spain
ame@us.es
[3] Department of Computer Science, University of Seville, Spain
riquelme@us.es

Abstract. Nowadays, much effort is being devoted to develop techniques that forecast natural disasters in order to take precautionary measures. In this paper, the extraction of quantitative association rules and regression techniques are used to discover patterns which model the behavior of seismic temporal data to help in earthquakes prediction. Thus, a simple method based on the k–smallest and k–greatest values is introduced for mining rules that attempt at explaining the conditions under which an earthquake may happen. On the other hand patterns are discovered by using a tree-based piecewise linear model. Results from seismic temporal data provided by the Spanish's Geographical Institute are presented and discussed, showing a remarkable performance and the significance of the obtained results.

Keywords: time series, quantitative association rules, regression.

1 Introduction

A time series is a sequence of values observed over time and, therefore, chronologically ordered. Given this definition, it is usual to find data that can be represented as time series in many research fields.

The study of the past behavior of a variable may be extremely valuable to predict its future behavior. Assuming that the nature of the earthquakes time series is stochastic, clustering techniques have shown that these time series exhibit some temporal patterns, making the modeling and subsequent prediction possible [11].

This paper analyzes and forecasts earthquakes time series by means of the application of two classical techniques: Quantitative association rules (QAR) and regression.

A revision of the latest published works reveals that the amount of metaheuristics and search algorithms related to association rules with continuous attributes is limited. Nevertheless, a classifier was presented in [13] to extract quantitative association rules from unlabeled data streams. The main novelty

of this approach lied on its adaptability to on-line gathered data. Also, a metaheuristic based on rough particle swarm techniques was presented in [1]. In this case, the special feature was the obtention of the values determining the intervals of the association rules. They also evaluated and tested several new operators in synthetic data. A multi-objective pareto-based genetic algorithm was presented in [2]. The fitness function was formed by four different objectives: support, confidence, comprehensibility of the rule (aimed at being maximized) and the amplitude of the intervals that forms the rule (intended to be minimized). The work published in [17] presented a new approach based on three novel algorithms: Value-interval clustering, interval-interval clustering and matrix-interval clustering. Their application was found especially useful when mining complex information. Another genetic algorithm was used in [16] in order to obtain numeric association rules. However, the unique objective to be optimized in the fitness function was the confidence. To fulfill this goal, the authors avoided the specification of the actual minimum support, which is the main contribution of this work. Finally, an extension of the well-known binary-coded CHC algorithm is presented in [10] for finding existing relations between atmospheric pollution and climatological conditions.

Regression techniques have been widely used for forecasting time series [5]. Thus, an empirical study on sea water quality prediction can be found in [7]. Hatzikos et al. faced the problem of forecasting water quality based on underwater sensors measurements, by means of a large variety of both linear and nonlinear methods. Also, a new methodology to build regression trees was introduced in [3]. The authors transformed quantitative data into statistical moments, and constructed a tree to estimate the forecasting interval of the target variable. Last, the problem of predicting the machinery degradation and trending of fault propagation before reaching the alarm was studied in [12]. In particular, the authors proposed an approach based on regression trees to forecast such time series.

The rest of the paper is divided as follows. Section 2 provides the methodology used in this work. The results of the approach are reported in Section 3. Finally, Section 4 discusses the achieved conclusions.

2 Methodology

The methods used to extract knowledge from earthquakes time series are described in this section. The goal is to find patterns in data that precede the appearance of earthquakes with a given magnitude.

2.1 Association Rules Mining

Let $F = \{F_1, ..., F_n\}$ be a set of features with values in $\mathbb{R}$ describing an earthquake. The desired rules are defined by the following equation:

$$\bigwedge_{i=1,\ldots,n-1} F_i \in [l_i, u_i] \Rightarrow F_n \in [l_n, u_n] \quad (1)$$

where l_i and u_i represents the lower and upper limits of the interval for F_i, respectively and the limits l_n and u_n are given depending on the objective of the problem to be solved. In the context of seismic time series, F_n represents the earthquake magnitude to be predicted and the limits l_n and u_n depend on the required size of the earthquakes to be forecasted.

The proposed method to obtain QAR is described as follows. First, the dataset is sorted by the feature F_n, that is, by the consequent of the rule. Once the limits $[l_n, u_n]$ are set, the range of the remaining features F_i is calculated as:

$$R(F_i) = \{F_i \text{ such that } F_n \in [l_n, u_n]\} \qquad i = 1, ..., n-1 \tag{2}$$

Let f_i^M and f_i^m with $i = 1, ..., n-1$ two functions defined by:

$$\begin{array}{rl} f_i^M : \{1, ..., \#(R(F_i))\} & \longrightarrow R(F_i) \\ k & \longrightarrow f_i^M(k) = k \text{ greatest value of } R(F_i) \end{array} \tag{3}$$

$$\begin{array}{rl} f_i^m : \{1, ..., \#(R(F_i))\} & \longrightarrow R(F_i) \\ k & \longrightarrow f_i^M(k) = k \text{ smallest value of } R(F_i) \end{array} \tag{4}$$

where $\#(R(F_i))$ is the number of elements of the set $R(F_i)$.

Let S_i be the set of pair of values such that the amplitude of the interval to be searched for the feature F_i is sufficiently small. That is,

$$S_i = \{(k_1, k_2) \text{ such that } f_i^M(k_2) - f_i^m(k_1) \leq MAX_i\} \tag{5}$$

where MAX_i is the maximum allowed amplitude for the feature F_i which is a given parameter depending on the desired rules.

Thus, for any value $(k_1^i, k_2^i) \in S_i$, the rules built by the k-greatest and k-smallest values are:

$$\bigwedge_{i=1,...,n-1} F_i \in [f_i^m(k_1^i), f_i^M(k_2^i)] \Rightarrow F_n \in [l_n, u_n] \tag{6}$$

2.2 Regression: M5P Algorithm

The second method used to obtain patterns in seismic time series is the M5P algorithm available in WEKA [4]. The M5P approach [15] extends to the M5 algorithm by adding missing values techniques and transformation of features from discrete values to binary values. The algorithm M5 [14] provides a conventional decision-tree with linear regression functions at the nodes. The tree is obtained by a classical induction algorithm but the splits are obtained by maximizing the reduction of the variance and not maximizing the gain of information.

Once the tree has been built, the method computes a linear model for each node. Later the leaves of the tree are pruned while the error decreases. For each node, the error is the mean of the absolute value of the difference between the

predicted and actual values for each example reaching such node. This error is weighted depending on the number of examples which reach that node. The process is repeated until all examples are covered for one o more rules.

Thus, M5P generates models that are compact and relatively comprehensible.

3 Results

This section presents the results obtained from the application of the approaches introduced in Section 2. In particular, Section 3.1 provides a description of the data used. Sections 3.2 and 3.3 gather all relevant results mined by means of association rules and decision-tree techniques, respectively.

3.1 Data Description

The dataset used in this work has been retrieved from the catalogue of Spanish's Geographical Institute (SGI), which contains the location and magnitude of Spanish earthquakes.

Additionally, the b–value parameter of the Gutenberg–Richter law has been calculated, as it reflects the tectonics and geophysical properties of the rocks as well as the fluid pressure variations in the characterized surface [9].

Thus, each sample forming the dataset is composed by four attributes: Current earthquake magnitude, time when the earthquake occurred, associated b-value, and magnitude of the previously occurred earthquake. Note that earthquakes with magnitude lower than 3.0 have been removed from the dataset, and both aftershocks and foreshocks have been removed to avoid dependent data, as recommended in [8].

Despite the Iberian Peninsula is divided in 27 seismogenic areas according to SGI, only areas 26 and 27 (Alboran Sea and Western Azores–Gibraltar Fault, respectively) have been studied, since they are the most active ones [11]. The considered earthquakes date from 1981 to 2008, having been analyzed a total of 873 quakes.

3.2 Quantitative Association Rules Extraction

All mined association rules to forecast earthquakes are now introduced and discussed. As the goal is to find patterns that precede quake occurrences, the magnitude of the current earthquake, M_c, has been forced to be the only attribute in the consequent.

The M_c attribute has been divided in three non-overlapped intervals: [3.0, 3.5) or small earthquakes, [3.5, 4.4) or medium earthquakes, and [4.4, 6.2] or large earthquakes (note that the largest retrieved earthquake magnitude is 6.2).

Tables 1, 2, and 3 show the rules extracted for large, medium and small earthquakes, respectively. Note that Δb and Δt represent the increment of the b–value and the time elapsed between the previous and current earthquake, respectively. Also, the magnitude of the earthquake occurred prior the current one, M_p, has been

Table 1. Association rules with consequent $M_c \in [4.4, 6.2]$

Id	Antecedent	Conf. (%)	Sup. (%)	Lift
#1	$\Delta t \in [0.02, 0.08] \wedge \Delta b \in [-0.16, -0.10] \wedge M_p \in [3.0, 3.4]$	75.0	5.7	12.4
#2	$\Delta t \in [0.00, 0.07] \wedge \Delta b \in [-0.12, -0.05] \wedge M_p \in [3.5, 4.9]$	87.5	13.2	14.4
#3	$\Delta t \in [0.00, 0.33] \wedge \Delta b \in [-0.11, -0.01] \wedge M_p \in [5.0, 6.2]$	80.0	7.6	13.2

divided in non-overlapped intervals, and therefore, transactions forming the data can be covered only by one rule. Finally, all rules have been assessed by means of three well-known and widely used indices: Confidence, support, and lift [6].

The best rules mined for large earthquakes ($M_c \in [4.4, 6.2]$) are shown in Table 1. These rules share a common feature, which is that they all present remarkable and negative Δb. Moreover, Δt is small in all rules, except for rule #3, which allows time intervals up to 0.33. From the 53 earthquakes that satisfy that $M_c \in [4.4, 6.2]$, 14 are covered by rules #1, #2 and #3, which represents a support of 26.4%. On the other hand, it is noticeable the high confidence reached by all of them: 80.8% on average. Finally, the interestingness of the rules (or lift) is 13.3 on average. Assuming that a lift greater than 1 leads to consider the rule as interesting [6], the obtained values indicate that the extracted rules provide meaningful knowledge.

Table 2 shows the QAR obtained for medium earthquakes, that is, with $M_c \in [3.5, 4.4)$. The most significative feature that share all the rules is that the b–value does not vary much (its value ranges from $\Delta b = -0.07$ to $\Delta b = 0.02$). Also remarkable is that the occurrence of these earthquakes takes place after moderately short time periods (the time elapsed between earthquakes varies from $\Delta t = 0.00$ to $\Delta t = 0.20$). As for the quality of the results, 86 earthquakes out of 344 were covered by rules #4, #5 and #6, which means a support of 25.0%. The confidence was of 76.0% on average which can be considered high. Last, the lift measure also confirms that the rules are high quality, since it has values greater than 1, in particular, 1.9 on average.

Table 3 represents the best QAR discovered for small earthquakes ($M_c \in [3.0, 3.5)$). The b–value is now characterized by moderate and positive increments (Δb ranges from 0.01 to 0.04). Moreover, in contrast to what happens with medium and large earthquakes, the time elapsed is high, varying from $\Delta t = 0.10$ to $\Delta t = 0.32$. A total of 476 small earthquakes were retrieved, from which 46 have been covered by rules #7, #8 and #9, which imply a support of 9.7%. Especially noticeable is the confidence reached by these rules which is 85.7% on average. Again, the lift measure is greater than 1 for all rules, in particular, 1.7 on average.

Table 2. Association rules with consequent $M_c \in [3.5, 4.4)$

Id	Antecedent	Conf. (%)	Sup. (%)	Lift
#4	$\Delta t \in [0.04, 0.20] \wedge \Delta b \in [-0.07, -0.01] \wedge M_p \in [3.0, 3.5]$	79.0	8.7	2.0
#5	$\Delta t \in [0.00, 0.02] \wedge \Delta b \in [-0.01, 0.00] \wedge M_p \in [3.6, 4.5]$	78.6	12.8	2.0
#6	$\Delta t \in [0.00, 0.05] \wedge \Delta b \in [-0.02, 0.02] \wedge M_p \in [4.6, 5.9]$	70.6	3.6	1.8

Table 3. Association rules with consequent $M_c \in [3.0, 3.5)$

Id	Antecedent	Conf. (%)	Sup. (%)	Lift
#7	$\Delta t \in [0.13, 0.32] \wedge \Delta b \in [0.01, 0.04] \wedge M_p \in [3.0, 3.2]$	100	2.5	1.8
#8	$\Delta t \in [0.10, 0.19] \wedge \Delta b \in [0.01, 0.03] \wedge M_p \in [3.3, 3.4]$	88.0	4.6	1.6
#9	$\Delta t \in [0.11, 0.32] \wedge \Delta b \in [0.00, 0.03] \wedge M_p \in [3.5, 5.7]$	85.7	2.5	1.6

3.3 M5P Results

This section provides the result obtained from the application of the M5P regressor. Fig. 1 illustrates the tree built by this algorithm. Thus, M5P found four linear models (LM), whose equations are listed below:

$$\text{LM 1: } M_c = -0.0160\Delta t - 11.2781\Delta b + 0.3237M_p + 2.3766 \tag{7}$$

$$\text{LM 2: } M_c = -0.0795\Delta t - 0.4022\Delta b + 0.1889M_p + 2.7826 \tag{8}$$

$$\text{LM 3: } M_c = -0.0955\Delta t - 0.4022\Delta b + 0.0213M_p + 3.2495 \tag{9}$$

$$\text{LM 4: } M_c = -0.3696\Delta t - 0.4206\Delta b + 0.0096M_p + 3.2060 \tag{10}$$

The analysis of this model reveals that the b–value is the most significative attribute, as it appears in the two first levels of the tree. Also, the coefficients corresponding to b–value have the greatest weights in the linear models.

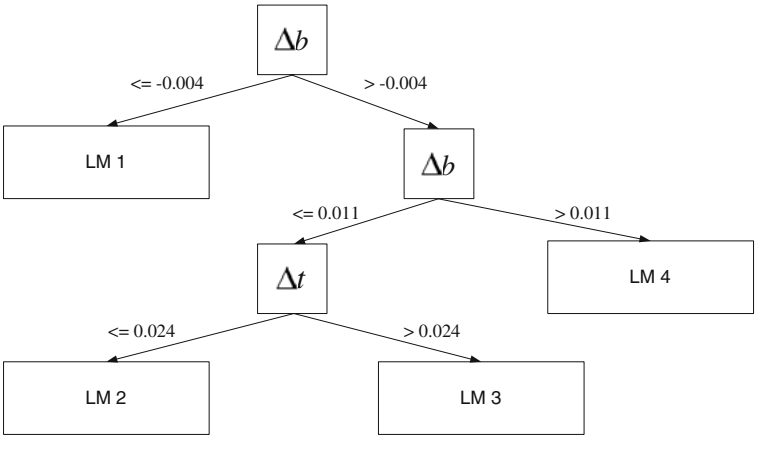

Fig. 1. Tree built with M5P algorithm

The first cutoff is set for $\Delta b = -0.004$. Thus, the first linear model, LM 1, is found when $\Delta b \leq -0.004$. This model has the biggest absolute value for the

Δb coefficient (a value of -11.2781). Moreover, as this coefficient is negative, it can be stated that the smaller is the value of Δb, the bigger is the earthquake magnitude. On the other hand, the coefficient of M_p is positive (a value of 0.3237), which leads to conclude that M_c is directly related to M_p. In other words, the magnitude of the current earthquake has a direct relation with the magnitude of the previous one.

The earthquakes occurred with $\Delta b > -0.004$ are modeled by three linear models (LM 2, LM 3 and LM 4). All of them present similar Δb coefficients, which involves inverse relation with the magnitude of the current earthquake, that is, the bigger is Δb, the smaller is M_c. Nevertheless, its influence is more moderate than that of LM 1.

The second cutoff is set for $\Delta b = 0.011$. Thus, when $\Delta b > 0.011$ the LM 4 model is provided (see equation (10). In this model, the most significative coefficient is that corresponding to Δt with a weight of -0.3696, revealing that the longer is the time elapsed, the smaller is the magnitude of the current earthquake. It is also notable that the magnitude of the previous earthquake does not influence much in this model as it is weighted by 0.0096.

When the b–value varies between -0.004 and 0.011, the model proposes two different linear models (LM 2 and LM 3), depending on the time elapsed between the previous and current earthquake. Although both linear models are quite similar, when the time elapsed is less or equal than 0.024 (LM 2 model), the magnitude of the previous earthquake influences much more than when it is greater than 0.024 (LM 3 model) as the coefficient of M_p is 0.1889 in LM 2 versus 0.0213 in LM 3.

Finally, a measure of the quality of results is now discussed. The tree presents a correlation coefficient of 0.67. The mean absolute error is 0.26 and the root mean squared error is 0.35. These errors are considered satisfactory given the stochastic nature of the problem studied.

4 Conclusions

Earthquake data from two particular areas of the Iberian Peninsula have been successfully mined by means of two different techniques: QAR and the M5P algorithm. In particular, QAR with a confidence of 83.0% and a lift of 5.6 on average have been discovered and a regression-tree with an error of 0.35 has been built. Both techniques have discovered the great influence that the b–value has in earthquakes occurrences as its variation along with the time elapsed have shown to be useful to model different earthquakes. Thus, the patterns discovered before an earthquake takes place may be useful in subsequent predictions.

Acknowledgments

The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-02, and from the Junta de Andalucía, project P07-TIC-02611, is acknowledged.

References

1. Alatas, B., Akin, E.: Rough particle swarm optimization and its applications in data mining. Soft Computing 12(12), 1205–1218 (2008)
2. Alatas, B., Akin, E., Karci, A.: MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. Applied Soft Computing 8(1), 646–656 (2008)
3. Alberg, D., Last, M., Ben-Yair, A.: Induction of mean output prediction trees from continuous temporal meteorological data. Journal of Applied Quantitative methods 4(4), 485–494 (2009)
4. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Data Mining. In: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)
5. Fu, T.-C.: A review on time series data mining. Engineering Applications of Artificial Intelligence 24(1), 164–181 (2010)
6. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3), Article 9 (2006)
7. Hatzikos, E.V., Tsoumakas, G., Tzanis, G., Bassiliades, N., Vlahavas, I.I.: An empirical study on sea water quality prediction. Knowledge-Based Systems 21, 471–478 (2008)
8. Kulhanek, O.: Seminar on b–value. Technical report, Department of Geophysics, Charles University, Prague (2005)
9. K. Lee and W. S. Yang. Historical seismicity of Korea. Bulletin of the Seismological Society of America, 71(3):846–855, 2006.
10. Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F., Riquelme, J.C.: Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution. Integrated Computer-Aided Engineering 17(3), 227–242 (2010)
11. Martínez-Álvarez, F., Morales-Esteban, A., Troncoso, A., de Justo, J.L., Rubio-Escudero, C.: Pattern recognition to forecast seismic time series. Expert Systems with Applications 37(12), 8333–8342 (2010)
12. Oh, M.S., Tan, A., Tran, V., Yang, B.S.: Machine condition prognosis based on regression trees and one-step-ahead prediction. Mechanical Systems and Signal Processing 22, 1179–1193 (2008)
13. Orriols-Puig, A., Casillas, J., Bernadó-Mansilla, E.: First approach toward online evolution of association rules with learning classifier systems. In: Proceedings of the Computation Conference Genetic and Evolutionary GECCO 2008, pp. 2031–2038 (2008)
14. Quinlan, J.R.: Learning with continous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pp. 343–348. World Scientific, Singapore (1992)
15. Wang, Y., Witten, I.H.: Induction of model trees for predicting continous classes. In: Proocedings of the poster papers of the European Conference on Machine Learning, pp. 128–137 (1997)
16. Yan, X., Zhang, C., Zhang, S.: Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Systems with Applications 36(2), 3066–3076 (2009)
17. Yin, Y., Zhong, Z., Wang, Y.: Mining quantitative association rules by interval clustering. Journal of Computational Information Systems 4(2), 609–616 (2008)