Metodología basada en minería de datos para el descubrimiento de patrones precursores de terremotos de magnitud media y elevada

Fecha de comienzo: septiembre de 2013

E. Florido-Navarro[1]
Directores de tesis: F. Martínez-Álvarez[1] y J. L. Aznarte[2]

[1]Departamento de Lenguajes y Sistemas Informáticos, Universidad Pablo de Olavide
[2]Departamento de Inteligencia Artificial, UNED

Key words: Minería de datos; series temporales; predicción; sismología.

1. Resumen

1.1. Motivación

Este trabajo de tesis doctoral aporta una nueva metodología basada en técnicas de minería de datos para descubrir patrones en series temporales de origen sísmico y su posterior aplicación en la predicción de terremotos.

Una serie temporal es una secuencia de valores en el tiempo y, consecuentemente, ordenados cronológicamente. Dada esta definición, es bastante común encontrar en diferentes campos de investigación datos que puedan ser representados como series temporales. El estudio del comportamiento pasado de una variable, puede resultar extremadamente útil para ayudar a predecir sus comportamientos futuros. Si, dado un conjunto de valores pasados de una variable, no es posible predecir con fiabilidad su comportamiento futuro, se dice que se trata de una serie temporal caótica.

Este trabajo de investigación se incluye en este contexto, ya que los eventos relacionados con los terremotos son aparentemente imprevisibles. Suponiendo que la naturaleza de una serie temporal sobre datos de terremotos es estocástica, se pretende desarrollar una novedosa metodología que, basándose en técnicas de agrupamiento, busca la existencia de ciertos patrones temporales previos a la ocurrencia de terremotos de magnitud media y elevada. El fin último de estos patrones es el de ser utilizados para poder predecir terremotos. Para evitar la bien conocida dependencia de datos, tanto las réplicas como los temblores precursores se eliminarán de las series de terremotos analizadas [5].

1.2. Hipótesis de partida

En este trabajo de investigación se parte de una serie de hipótesis, que son enumeradas a continuación:

- La minería de datos es adecuada para la predicción de terremotos, como ya se ha demostrado en otras ocasiones [7,6,10,11,14].
- La ocurrencia de un terremoto no sigue una distribución de ocurrencia totalmente caótica.
- Existen algunos patrones precedentes a grandes terremotos que se pueden medir de manera indirecta.
- El valor *b-value* es un indicador de gran alcance que proporciona mucha información acerca de futuras ocurrencias de terremotos.

1.3. Objetivos

Una breve enumeración de los objetivos que se pretenden con esta tesis doctoral son:

- La adquisición de gran conocimiento sobre las técnicas de minería de datos.
- El desarrollo de una metodología novedosa para el descubrimiento de patrones que precedan a la ocurrencia de terremotos.
- Mejorar la predicción de terremotos mediante la aplicación de los patrones descubiertos.
- Aplicación a zonas de gran actividad sísmica, como por ejemplo, Chile.

2. Metodología

En esta sección se describe la metodología propuesta para conseguir extraer conocimiento de las series temporales de terremotos. En primer lugar, es imprescindible indicar cómo se construye el conjunto de datos de terremotos. Cada terremoto está representado por tres características: su magnitud, su valor $b - value$ y su fecha de ocurrencia.

El valor $b - value$ se determina teniendo en cuenta los 50 eventos anteriores al terremoto estudiado [13]. Igualmente, como se ha hecho en [12], la magnitud de corte se establece en el valor 3 y además, los datos se agrupan en conjuntos de cinco terremotos cronológicamente ordenados de acuerdo a la metodología propuesta en Nuannin et al. [13]. Así se proporciona una ley de fácil interpretación. Cada grupo está representado por la media de la magnitud de los cinco terremotos; el tiempo transcurrido desde el primer terremoto al quinto y la variación, con signo, de los $b - value$ en este intervalo de tiempo para los cinco terremotos.

El objetivo es encontrar patrones en los datos que preceden a la aparición de terremotos con una magnitud mayor o igual a 4.5. Entonces, se aplica el algoritmo de *K-means* al conjunto de datos, con el fin de clasificar las muestras en diferentes grupos. Como paso previo, hay que determinar el número óptimo de agrupaciones ya que el algoritmo *K-means* necesita este número como parámetro de entrada. Para este propósito, se aplica a los datos el bien conocido índice Silhouette para agruparlos en diferentes números de clusters, como se propone en [8]. Por lo tanto, en los análisis posteriores cada muestra es considerada exclusivamente por la etiqueta asignada por el algoritmo *K-means*. Es decir, el

conjuntos de datos (DS) se transforma en una secuencia de etiquetas. Una vez obtenidas estas etiquetas, se buscan secuencias específicas de etiquetas como patrones precursores de terremotos de gran tamaño.

La Figura 1 ilustra el proceso de transformación al que se somete el conjunto de datos original, formando agrupamientos de tres ($K = 3$) y secuencias de dos etiquetas consecutivas ($w = 2$). Cabe destacar que, las cajas grises indican que un evento con magnitud mayor o igual que 4.5 (o cualquier otro umbral establecido), ha sido detectado en esta etiqueta. En la figura la E representa a los eventos o terremotos; DS representa al conjunto de datos; SL es la secuencia de etiquetas y, por último, G representa el agrupamiento de eventos.

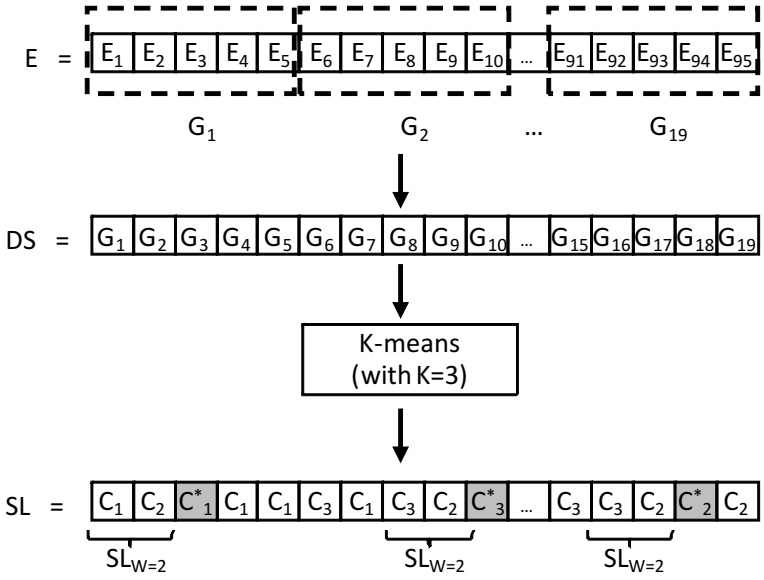

Figura 1. Metodología propuesta

Una versión inicial de esta metodología ya ha sido utilizada con éxito en distintas zonas de Chile [2,3]. No obstante, quedan aún ciertos aspectos que mejorar, que serán objeto de futuros trabajos de investigación, como son:

- Permitir que las agrupaciones de eventos se hagan para un número trivial de éstos.
- Flexibilizar el número de clusters permitidos para el agrupamiento de los datos.
- Permitir longitudes de ventana de cualquier valor.
- Realización de búsqueda automática de la mejor combinación de número de clusters y longitud de la ventana.

3. Plan de trabajo

Se detalla, a continuación, el plan de trabajo propuesto para la elaboración de esta tesis doctoral:

- Lectura de bibliografía y adquisición de conocimientos necesarios sobre la materia de estudio de la tesis doctoral (9 meses)
- El desarrollo de nuevas metodologías para mejorar la predicción de terremotos basado en técnicas de minería de datos (12 meses).
- La aplicación de la metodología de los conjuntos de datos del mundo real y evaluar los resultados (6 meses).
- Publicación de los resultados en revistas de alto impacto y redacción de la tesis doctoral (9 meses)

4. Relevancia

Esta tesis doctoral analizará y pronosticará series temporales de terremotos por medio de la aplicación de técnicas de agrupamiento. Para ser precisos, como orígenes de datos se utilizarán áreas sismogénicas. Un área sismogénica se define como una zona de terremotos con características sísmicas y tectónicas homogéneas. Esto significa que el proceso de generación de terremotos en cada área es homogénea tanto en el espacio como en el tiempo. Puede ser lineal, como por ejemplo una falla, una línea de fallas o incluso un conjunto de fallas paralelas que se encuentran cerca y a una distancia considerable desde el sitio donde se produce el terremoto. Sin embargo, una fuente de datos también puede ser un área donde las fallas son muy numerosas y se encuentran distribuidas de una manera aleatoria o no claramente definida. Desde un punto de vista tectónico, un área o zona sismogénica, puede incluir una o varias estructuras tectónicas, y su geometría está basada en información histórica, sísmica o puramente tectónica.

El reto para encontrar métodos eficaces para el pronóstico de terremotos ha sido abordado desde hace ya más de 100 años [4]. El uso de información histórica en la predicción de terremotos es algo muy frecuente hoy en día y existe un conocido grupo, el *Regional Earthquake Likelihood Model* (RELM), que ha propuesto múltiples modelos para la estimaciones de riesgos de los mismos [1]. Por tanto, nuestro objetivo es encontrar patrones temporales y modelar comportamientos de series temporales que se ven implicados en la ocurrencia de un terremoto de tamaño medio-grande, que en este trabajo de investigación se han considerado aquellos con una magnitud igual o superior a 4.5.

Aunque algunas técnicas de agrupamiento han sido utilizadas ya anteriormente de manera satisfactoria en el estudio de series temporales [9], su apliación a la ocurrencia de terremotos supone un nuevo paso muy importante y no ha sido ampliamente desarrollado. Se debe tener en cuenta que un grupo de temblores más pequeños antes o después uno más grande se denota como una agrupación de terremotos por los sismólogos. Sin embargo, este concepto no debe confundirse con las técnicas de agrupamiento utilizadas en esta tesis doctoral, que son uno de los principales objetivos de la inteligencia artificial.

Agradecimientos

Se quiere agradecer la ayuda recibida por el Ministerio de Ciencia y Tecnología, la Junta de Andalucía y la Universidad Pablo de Olavide mediante los proyectos TIN2014-55894-C2-2-R, P12-TIC-1728 y APPB813097, respectivamente. Igualmente, esta tesis está parcialmente financiada por una beca Ramón y Cajal.

Referencias

1. E. H. Field. Overview of the working group for the development of Regional Earthquake Likelihood Models. *Seismological Research Letters*, 78(1):7–16, 2007.
2. E. Florido, F. Martínez-Álvarez, J. L. Aznarte, A. Morales-Esteban, J. Reyes, and A. Troncoso. Discovering patterns preceding earthquakes in Chilean time series. In *Proceedings of the International Work-Conference on Time Series Analysis*, pages 819–826, 2014.
3. E. Florido, F. Martínez-Álvarez, A. Morales-Esteban, J. Reyes, and J. L. Aznarte-Mellado. Detecting precursory patterns to enhance earthquake prediction in Chile. *Computers and Geosciences*, 76:112–120, 2015.
4. R. J. Geller. Earthquake prediction: a critical review. *Geophysical Journal International*, 131(3):425–450, 1997.
5. O. Kulhanek. Seminar on $b-$value. Technical report, Department of Geophysics, Charles University, Prague, 2005.
6. F. Martínez-Álvarez, J. Reyes, A. Morales-Esteban, and C. Rubio-Escudero. Determining the best set of seismicity indicators to predict earthquakes. two case studies: Chile and the Iberian Peninsula. *Knowledge-Based Systems*, 50:198–210, 2013.
7. F. Martínez-Álvarez, A. Troncoso, A. Morales-Esteban, and J. C. Riquelme. Computational intelligence techniques for predicting earthquakes. *Lecture Notes in Artificial Intelligence*, 6679(2):287–294, 2011.
8. F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. S. Aguilar-Ruiz. Energy time series forecasting based on pattern sequence similarity. *Pattern Recognition Letters*, 32(12):1662–1665, 2011.
9. F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. M. Riquelme. Discovering patterns in electricity price using clustering techniques. In *Proceedings of the International Conference on Renewable Energies and Power Quality*, pages 65–67, 2007.
10. A. Morales-Esteban, F. Martínez-Álvarez, and J. Reyes. Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence. *Tectonophysics*, 593:121–134, 2013.
11. A. Morales-Esteban, F. Martínez-Álvarez, R. Scitovski, and S. Scitovski. A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic zoning. *Computers and Geosciences*, 73:132–141, 2014.
12. A. Morales-Esteban, F. Martínez-Álvarez, A. Troncoso, J. L. de Justo, and C. Rubio-Escudero. Pattern recognition to forecast seismic time series. *Expert Systems with Applications*, 37(12):8333–8342, 2010.
13. P. Nuannin, O. Kulhanek, and L. Persson. Spatial and temporal b value anomalies preceding the devastating off coast of nw sumatra earthquake of december 26, 2004. *Geophysical Research Letters*, 32, 2005.
14. J. Reyes, A. Morales-Esteban, and F. Martínez-Álvarez. Neural networks to predict earthquakes in Chile. *Applied Soft Computing*, 13(2):1314–1328, 2013.