

Análisis de Predicción de Terremotos

Sandra Álvarez Teruelo
100038973

Universidad Carlos III Madrid
100038973@alumnos.uc3m.es

Iván Alejandro Fernández Pacheco
100038982

Universidad Carlos III Madrid
100038982@alumnos.uc3m.es

ABSTRACT

Este documento presenta el análisis de datos acerca de terremotos a lo largo del tiempo y sus diferentes características y como afectan en ellos.

Términos generales

Algoritmos, Experimentación, Teoría, Verificación.

Palabras clave

Minería de datos, terremotos, predicción, aprendizaje.

1. INTRODUCCIÓN

¿Por qué se valora tanto la información? Son muchas las razones, ya que puede servir para controlar, optimizar, administrar, investigar, planificar, predecir o tomar decisiones de cualquier ámbito según donde nos desarrollemos.

La información por sí misma está considerada un bien patrimonial. Alguien decía que la información es poder. Es razonable pensar que se debe cuidar, pero también hay que sacarla provecho.

La cantidad de datos que se genera y almacena hoy en día en cualquier área de conocimiento es tan vasta, que rebasa las capacidades de asimilación de cualquier ser humano.

Este hecho ha ocasionado el nacimiento de nuevas disciplinas como la denominada "Descubrimiento de Conocimiento en Bases de Datos" (KDD), que se enfoca en la extracción de información a partir del procesamiento de grandes cantidades de datos.

"Se ha estimado que cada 20 meses se duplica la cantidad de información en el mundo. ¿Qué se supone que se puede hacer con esa explosión de datos crudos? Claramente, pocos van a ser observados por los ojos humanos." [KDD Conference 1999]

La mayoría de las organizaciones han acumulado una enorme cantidad de datos que normalmente se almacenan en Sistemas de Gestión de Bases de Datos dispersas, no comunicadas e incongruentes entre sí.

Además el exagerado sobredimensionamiento de esas bases de datos excede con mucho nuestra capacidad de analizar e interpretar sus contenidos. Sin embargo, estas mismas organizaciones persiguen un objetivo distinto; de manera

conceptual y consensuada buscan obtener información, buscan que dicha información sea útil en sus procesos organizacionales y de negocio, e incluso porqué no, buscan convertir la información en conocimiento relacionado con su actividad.

El cómo llegar a trasladar datos en información e información en conocimiento es parte de una nueva aportación del mundo de las Tecnologías de la Información.

Frente a este escenario, resulta apremiante contar con métodos y herramientas computacionales capaces de analizar de forma automática y eficiente la gran cantidad de información acumulada en cualquier disciplina.

Investigar técnicas y algoritmos de la inteligencia computacional para basar en ellos el desarrollo de sistemas de software que faciliten el análisis de información y descubrimiento de conocimiento en grandes bases de datos. Particularmente se investigarán técnicas que permitan la visualización automática de información digital.

Por todo ello es necesario disponer de una serie de aplicaciones que permitan procesar la información para obtener una serie de datos de los cuales se pueda obtener un beneficio.

Estas técnicas pueden ayudar a confirmar cualquier sospecha sobre el comportamiento del sistema en un particular contexto.

Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico y mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación. La minería de datos tiene dos líneas de análisis:

-> Descripción: La misión fundamental de la minería de datos es el descubrimiento de reglas. A partir de ellas se podrán establecer una serie de relaciones entre las variables. Esto beneficiará el análisis y la descripción del modelo de estudio.

-> Predicción: Una vez se hayan establecido una serie de reglas importantes del modelo a través de la descripción, estas pueden ser usadas para predecir algunas variables de salida. Puede ser en el caso de secuencias en el tiempo, de futuras fluctuaciones de la bolsa, de prevenir sucesos catastróficos como pueden ser los terremotos que se estudian en este documento...

En esta tarea, se complementan las técnicas estadísticas tradicionales con aquellas provenientes de la inteligencia artificial. Conceptos adaptativos como los algoritmos genéticos y las redes neuronales, permiten realizar predicciones más acertadas, especialmente en casos de gran complejidad y con relaciones internas no-lineales

No solo las empresas o las instituciones son generadoras de nuevos problemas que afrontar, otros campos científicos también generan nuevos problemas donde la minería de datos se convierte en imprescindible, tales como las investigaciones originadas a raíz del proyecto Genoma, ¿que secuencias de genes motivan la aparición de enfermedades?, ¿lo hacen de forma determinista o en probabilidad? También la información transmitida por satélite puede proporcionar avances a fenómenos hasta hoy difíciles de explicar, tales como la vulcanología, los terremotos o el clima, etc.

Esta será nuestra línea a seguir: la predicción de terremotos.

2. OBJETIVO

El objetivo principal de la mayoría de los proyectos de minería de datos es usar un modelo de minería de datos para crear predicciones. En este caso, se va a realizar **predicciones acerca de los terremotos** en cualquier parte del mundo.



Figura 1. El terremoto de México, 1985

Un **terremoto** es el movimiento brusco de la Tierra causado por la brusca liberación de energía acumulada durante un largo tiempo.

En general se asocia el término terremoto con los movimientos sísmicos de dimensión considerable, aunque rigurosamente su etimología significa "movimiento de la Tierra".

Resulta demasiado presuntuoso decir "predicción" al hablar de terremotos con el nivel actual de conocimientos sobre el tema. Es más realista referirse al "**riesgo**" de terremotos ya que no existe una certeza mayor que decir que en cierta zona hay una probabilidad estadística de que se registre un evento sísmico de magnitud variable desconocida. Variaciones en el comportamiento del clima o conductas anormales en algunos animales no tienen solidez científica como para ser considerados "predictivos".

Por lo demás, si alguien avisara que con certeza se producirá un terremoto en los siguientes minutos u horas, ¿se imagina el

pánico en la población, las huídas, crisis de histeria, caos, pillaje, etc? ¿Y si NO ocurre?

El objetivo, entonces, de asignar un grado de riesgo no es otro que atenuar los efectos de un terremoto. Si se presume la ocurrencia de un seísmo y se imagina cuál sería su peor consecuencia se podrían tomar las **precauciones** adecuadas para evitar un daño mayor.

En el caso de los terremotos, tanto el riesgo social como el económico son los más altos entre todos los desastres naturales. Para cuantificar el tamaño de un seísmo, se utilizan dos referencias: **intensidad y magnitud del mismo**.

Pero no sólo es importante prevenir, sino que también tener la **capacidad de reaccionar** de forma inmediata después de ocurrido el terremoto, para informar sobre su magnitud, la ubicación de su epicentro, etc., y organizar de la mejor manera la ayuda a los afectados.

Una sola alarma falsa puede hacer que se pase por alto una predicción válida, con lo cual se produce la catástrofe que se trata de evitar con la predicción.

Por dar alguna información de cantidad, se puede afirmar que sobre el 80% de ellos ocurren en áreas deshabitadas mientras que algunas decenas provocan daño en ciudades (población o construcciones). Cada año hay varios millones de temblores en el mundo pero solo algunos miles son registrados por los sismógrafos a lo ancho y largo del mundo y algunos cientos son percibidos por la población general.

Los datos que se van a manejar a lo largo de este estudio fueron compilados por la Administración Nacional Oceánica y Atmosférica Nacional de Datos de Satélites Ambientales y el Servicio de Información / Centro Nacional de Datos Geofísicos (NOAA / NESDIS / NGDC) del Catalog of Significant Earthquakes y la base de datos para Instrumentally Recorded Earthquakes.

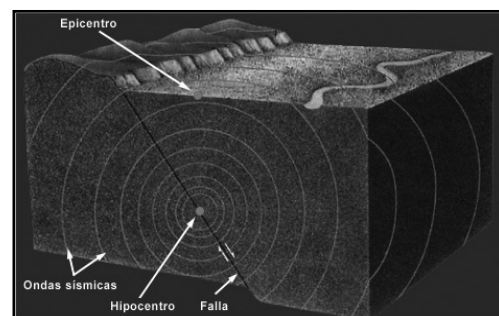


Figura 2. Partes de un terremoto

3. A PRIMERA VISTA

Una de las primeras etapas del análisis de datos puede ser el mero análisis visual de éstos, en ocasiones de gran utilidad para desvelar relaciones de interés utilizando nuestra capacidad para comprender imágenes. Por ello, antes de centrarse en el análisis exhaustivo de los datos con el objetivo de extraer conclusiones, se van a hacer reflexiones acerca de los datos que se han obtenido y la relación entre ellos.

En esta parte se emplean cuatro fases independientemente de la técnica específica de extracción de conocimiento usada.

1. Filtrado de datos.
2. Selección de Variables.
3. Extracción de Conocimiento.
4. Interpretación y Evaluación.

Estas fases se explican a continuación.

Filtrado de datos: El formato de los datos no suele ser el idóneo, y en la mayoría de las ocasiones no es posible aplicarles ningún tipo de algoritmo para estudiarlos. Mediante el preprocesado se filtran los datos se eliminan valores no válidos, se obtienen muestras de los mismos (mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, agrupamiento, etc.).

Selección de variables: Después de haber sido preprocesados, se sigue teniendo una gran cantidad de datos. La selección de características reduce el tamaño de los mismos, eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Extracción de Conocimiento: Mediante distintas técnicas se obtienen ciertos modelos de conocimiento, que representan patrones de comportamiento observados en los valores de las variables del problema o ciertas relaciones de asociación entre dichas variables.

Interpretación y evaluación: Para finalizar se procede a su validación, verificando que las conclusiones son correctas. En el caso de haber obtenido distintos modelos mediante el uso de diversas técnicas, se deben comprobar los modelos en busca de aquel que se ajuste mejor al problema. Si no se obtienen con ninguno los resultados esperados, se modificarán algunos de los procesos anteriores en busca de nuevos modelos. (Realizado en el punto 4 de la memoria)

En nuestro caso, se presentan como atributos

- **Profundidad** de la zona terrestre donde se han producido terremotos-
- **Latitud** que permite situar la zona con más riesgo de sufrir este tipo de hechos.
- **Longitud.**
- **Magnitud**, es decir, el valor en la escala de Richter, que dará una idea de la fuerza del terremoto.
- **Día** en que se produjo el suceso. Tanto este atributo como el siguiente resultan curiosos de analizar.
- **Hora.**
- **País** donde tuvo lugar. Se verá claramente como hay países mucho más propensos.
- **Muertes** que ha dejado el fenómeno a su paso.
- **Heridos.**
- **Daños** materiales que se han producido
- **Tsunami.** En que ocasiones el terremoto ha traído como consecuencia la existencia de un Tsunami.

Es importante destacar antes de seguir con el análisis que hemos decidido centrarnos en los terremotos de entre 4 y 9 en la escala Richter ya que son los más interesantes y devastadores (los menores de 4 no son apenas percibidos por la población) así como los ocurridos entre 1898 y 2007, figura 3, ya que se entiende que los datos de los años anteriores no son muy fiables debido a que en esas épocas no había los medios suficientes.

3.1 Panorámica general

Como un primer dato del conjunto de datos de análisis, se puede decir que el país con mayor número de fenómenos es China con 262, Irán con 208 y después Turquía, Indonesia, Japón y USA.

Para hacerse una idea de los daños materiales se puede decir que en el último siglo ha habido una media de 158.88 millones de dólares y una media de 297 casas destruidas. Lo que es aún más importante, es lo referente a pérdidas de vidas humanas donde ha habido una media de 1306.

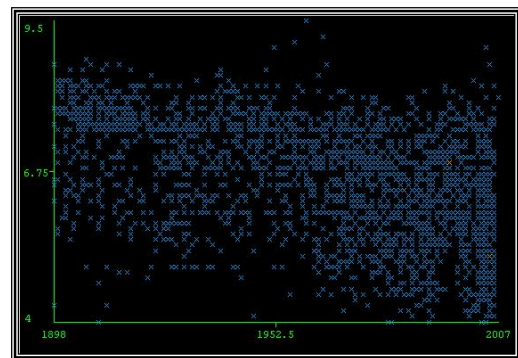


Figura 3. Magnitud vs Año

Las desviaciones de todos estos datos son bastante grandes porque lo se deduce que ha habido terremotos que han provocado muchos daños mientras que otros apenas.

En cuanto a la profundidad, los terremotos suelen producirse a una media de 37 metros siendo los más profundos los que suelen ser los más fuertes (figura 4)

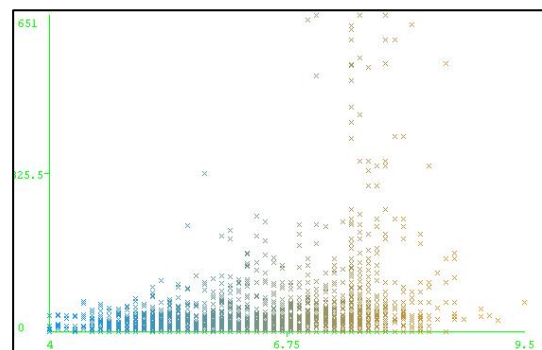


Figura 4. Profundidad vs Magnitud

3.2 Distribuciones y relaciones

Si una zona ha sufrido muchos terremotos de gran intensidad en el pasado, lo más probable es que tal cosa ocurra de nuevo. También se dice que después de uno grande, al disiparse la energía, el riesgo de un nuevo evento es más bajo. Lamentablemente esto no siempre se ha cumplido y en muchas zonas declaradas de bajo riesgo han ocurrido terremotos de tal magnitud que dejaron perplejos a sus predictores.

Valorando ambos puntos de vista, se va a analizar a lo largo de este estudio entre otros aspectos, la zona geográfica (latitud y longitud) donde suelen tener lugar los terremotos.

Observando ambas magnitudes por separado (figura 5) podemos ver como el mayor número de terremotos se concentran entre las latitudes 36° y 42° y, en cuanto a longitud, en los rangos $[-90,-70]$, $[22,44]$ y $[112,134]$.

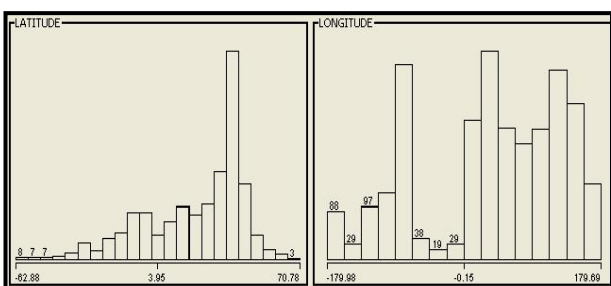


Figura 5. Latitud y Longitud

Se puede observar algo que a priori ya se podía presuponer a la vista de las noticias que llegan desde los distintos medios de comunicación.

Para verlo mejor, se tiene la figura 6 donde se puede ver como la mayoría ocurrirá dentro del "Cinturón de Fuego" (Océano Pacífico y sus márgenes, comenzando por Chile, subiendo hacia el norte por la costa sudamericana hasta llegar a Centroamérica, México, Costa Oeste de EE.UU., Alaska, Japón, Filipinas, Nueva Guinea, Islas del Pacífico Sur hasta Nueva Zelanda).

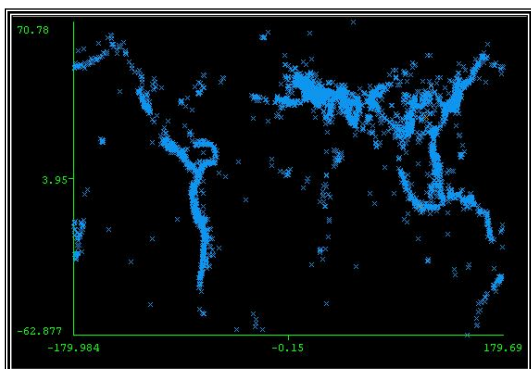


Figura 6. Latitud vs Longitud

Además, se sabe que por los estados de California, Oregón y Washington pasa la falla de San Andrés, una de las más activas del planeta y que por término medio suele producir un gran terremoto cada 100 años en la península de California.

Otro de las zonas de mayor actividad sísmica se centra en la parte del pacífico que comprende a Japón (es más que sabido que en tienen terremotos día sí y al siguiente también, y que por ello sus infraestructuras están preparadas para soportar los temblores (he aquí lo que se decía al principio de la previsión y reacción), lo que no quita para que haya algunos temblores devastadores que ni las mejores de las previsiones sean capaces de soportar).

Otro porcentaje importante ocurrirá en Los Alpides, zona que nace en Java y se extiende hacia Sumatra, Los Himalayas, y toda la zona del sudeste asiático (indonesia, la isla de java...), basta recordar el terremoto que produjo un tsunami hace 3 años estas navidades, que dejó más de 300000 víctimas..., el Mar Mediterráneo y se pierde en el Océano Atlántico. Turquía e Irán están en esta zona.

En cuanto a la latitud, mirando la figura 7, podemos ver como la mayoría de los terremotos se producen entre los 20° y 50° , y que en los polos casi no hay movimientos sísmicos.

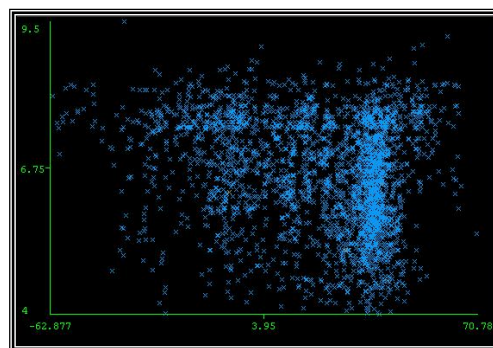


Figura 7. Magnitud vs latitud

Bien es cierto que no existe ningún lugar que se pueda considerar completamente libre de temblores (aunque la Antártida registra pocos y de baja magnitud).

Por otra parte, menos de una decena son de magnitud suficiente como para ser considerados terremotos y llamar la atención de los medios de comunicación y sólo uno o dos serán de magnitud mayor a 8 en Escala de Richter.

En la figura 8 podemos ver como aquí se trabajará con terremotos que van de 4 a 9 en la escala de Richter concentrándose la mayoría de ellos entre 6.7 y 7.7.

El valor medio de los terremotos elegidos para este estudio es de 6.613 en la escala Richter. Este valor nos da una idea de la magnitud del fenómeno.

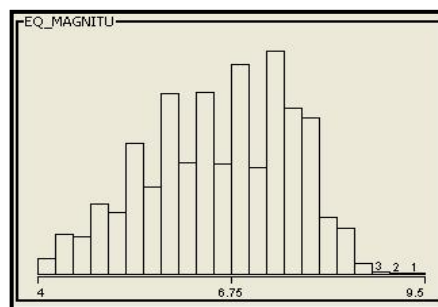


Figura 8. Valor escala Richter

Una gráfica que da una idea relevante es la que relaciona este concepto de magnitud con el de la profundidad que tiene el terreno donde se ha producido el terremoto (figura 9).

Se puede adelantar que una mayor magnitud en la escala se produce cuando hay una menor profundidad en el terreno.

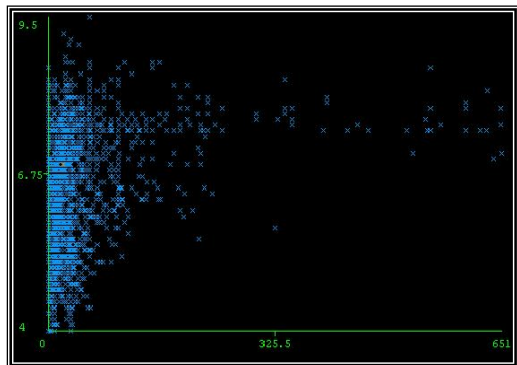


Figura 9. Valor escala Richter vs Profundidad

También resulta curioso hacer mención a los días y horas más propicios para que se produzca. En la figura 10 se ve claramente como los días 1, 16 y 31 de cada mes así como que son las primeras horas del día las más propensas a que ocurra.

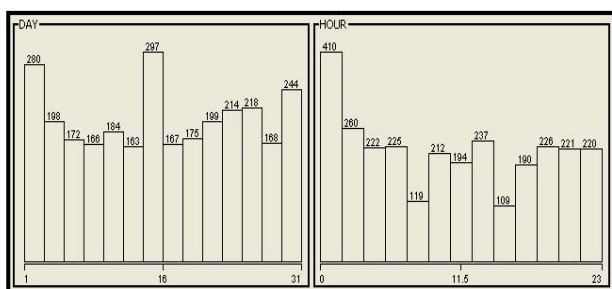


Figura 10. Días del mes y Horas del día

No hay ninguna razón científica para que se produzcan de madrugada. De hecho se considera un especie de mito que queremos que sea verdad. Varios terremotos fuertes han sido por la mañana, habiendo cogido a las personas durmiendo en sus casa, lo que si se producen derrumbes, produce muchas más víctimas.

Así que mucha gente cree que todos los terremotos grandes ocurren a esa hora pero los terremotos pueden ocurrir a cualquier hora del día. Es fácil notar los terremotos que encajan dentro del modelo y olvidar los que no.

A continuación, se ve en la figura 11 como se distinguen claramente 4 líneas horizontales lo que implican que hay cierto grupo de países donde se producen terremotos de forma continua a lo largo del último siglo.

Los países a destacar es esas líneas son China, Japón, Indonesia, Turquía y USA. Países que si relacionamos con los datos anteriores, vemos que son justo los que mayor numero de terremotos registrados.

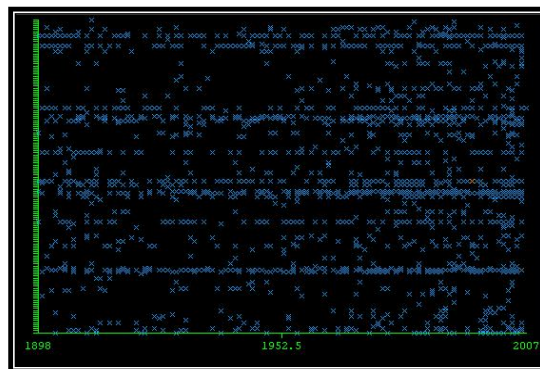


Figura 11. Países vs años

3.3 Influencia de un atributo

Hasta ahora, se ha podido ver la distribución de cada variable individualmente así como las graficas más relevantes de relaciones dos a dos entre ellas. Por último, es bueno ver otra perspectiva donde se puede ver la influencia de un rango de valores de un atributo en todos los demás.

Para esta parte se ha usado el filtro “discretize” que transforma los atributos numéricos seleccionados en atributos simbólicos, con una serie de etiquetas que resultan de dividir la amplitud total del atributo en intervalos.

A continuación, se va a analizar la figura 12, donde se ha representado los datos de la magnitud en 8 intervalos de igual frecuencia, cada uno identificado con un color diferente. De esta forma, en los otros tres atributos, se puede ver la influencia de ese intervalo de forma directa.

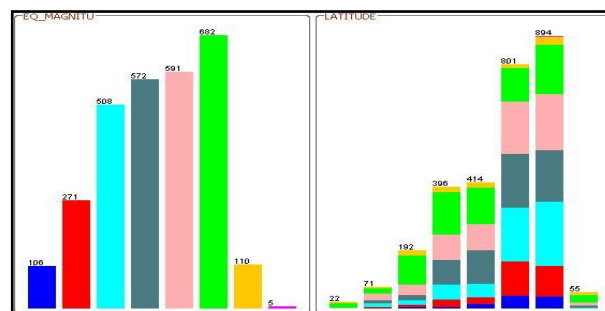


Figura 12. Magnitud y latitud

Se puede ver como terremotos de gran magnitud se pueden producir en cualquier latitud ya que el verde aparece en cualquier intervalo. Los de categoría mayor de 8'8 están todos entre 37° y 54°.

Se puede ver ahora el mismo grafico pero por países donde es China el que mas variedad de categorías tiene y donde se producen los mas devastadores, seguida de Irán, Indonesia y Turquía

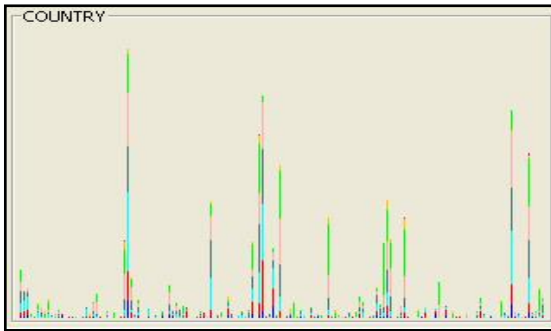


Figura 13. Países en función de la magnitud

Finalmente, se puede analizar el tema de los tsunamis provocados por los terremotos. Para ello, se ha establecido el color azul cuando no hay y el color rosa cuando si aparecen. Así de forma global se puede decir que a medida que han ido pasando los años ha aumentado el número de tsunamis registrados y se reparten prácticamente de igual manera a lo largo del día, mes y año.

De manera más particular, se ve en la figura 14 como los terremotos de mayor magnitud provocan mayor número de tsunamis, hasta el punto de que los de categoría mayor de 8.7 siempre lo hacen y los de 8.2 hasta 8.7 se provocan al 50%.

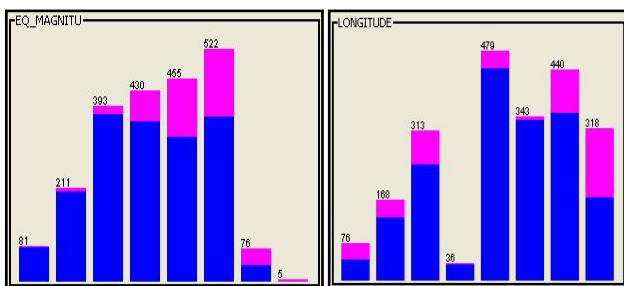


Figura 14. Magnitud y Longitud en función de Tsunamis

Se puede observar claramente como entre las longitudes 134° y 180° el riesgo de que se produzca un Tsunami es del 50% mientras que en -45° y 0° no hay casi probabilidad de que ocurra. Esta idea también se vera reflejada si vemos la grafica por países (figura 15), en la cual se distingue que donde mayor numero de terremotos hay, que es China e Irán, pocos provocan un Tsunami mientras que en Japón o Rusia se hablaría de un 50% de los casos.

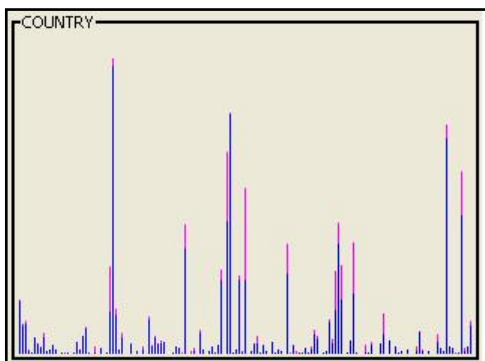


Figura 15. Países en función de Tsunamis

3.4 Asociación

Mediante algoritmos de asociación se puede realizar la búsqueda automática de reglas que relacionan conjuntos de atributos entre sí. Son algoritmos no supervisados, ya que no existen relaciones conocidas a priori con las que contrastar la validez de los resultados, sino que se evalúa si esas reglas son estadísticamente significativas.

El principal algoritmo implementado en WEKA es el algoritmo *A priori*, el cual sólo busca reglas entre atributos simbólicos, por lo cual todos los atributos numéricos deberían ser discretizados previamente. A modo de ejemplo se va a discretizar todos los atributos numéricos en 3 intervalos de igual frecuencia. Si se aplica el algoritmo de asociación con los parámetros por defecto, aparecen una serie de reglas que se muestran a continuación y que se han ordenado por eventos para hacer más sencillo su análisis.

$$FLG_TSUN=(-inf-0.5]' 2283 \Rightarrow INJURIES=(-inf-0.5]' 2153 (0.94)$$

$$INJURIES=(-inf-0.5]' 2652 \Rightarrow DAMAGE=(-inf-0.05]' 2254 (0.85)$$

El 94% de los terremotos que no provocan Tsunamis dejan menos de 1 herido así como el 85% de los mismos no dejan apenas daños. Ambas conclusiones lógicas.

$$DAY=(21.5-inf)' 949 \Rightarrow FLAG_TSUNA=(-inf-0.5]' 787 (0.83)$$

El 83% de los terremotos que se producen a partir del día 21 de cada mes, no tienen riesgo de provocar un Tsunami,

$$HOUR=(-inf-5.5]' 892 \Rightarrow FLAG_TSUNA=(-inf-0.5]' 729 (0.82)$$

El 82% de los terremotos que se producen desde la medianoche hasta a las 5:30 de la madrugada, no producen tsunamis.

$$YEAR=(1983.5-inf)' 941 \Rightarrow FLAG_TSUNA=(-inf-0.5]' 769 (0.82)$$

$$YEAR=(-inf-1949.5]' 951 \Rightarrow FLAG_TSUNA=(-inf-0.5]' 767 (0.81)$$

$$YEAR=(1983.5-inf)' 941 \Rightarrow EQ_DEPTH=(10.5-33.5]' 461 (0.49)$$

$$YEAR=(-inf-1949.5]' 951 \Rightarrow EQ_MAGNITU=(7.25-inf)' 557 (0.59)$$

$$YEAR=(1983.5-inf)' 941 \Rightarrow EQ_MAGNITU=(-inf-6.15]' 503 (0.53)$$

$$MONTH=(-inf-4.5]' 951 \Rightarrow FLAG_TSUNA=(-inf-0.5]' 768 (0.81)$$

El 82% de los terremotos que se produjeron en los 5 primeros meses del año no produjeron tsunamis.

$$EQ_MAGNIT=(7.25-inf)' 940 \Rightarrow YEAR=(-inf-1949.5]' 557 (0.59)$$

$$EQ_MAGNIT=(-inf-6.15]' 954 \Rightarrow YEAR=(1983.5-inf)' 503 (0.53)$$

$$EQ_MAGNIT=(-inf-6.15]' 954 \Rightarrow FL_TSUN=(-inf-0.5]' 905 (0.95)$$

$$EQ_MAGNIT=(7.25-inf)' 940 \Rightarrow EQ_DEPT=(33.5-inf)' 448 (0.48)$$

$$EQ_MAGNIT=(7.25-inf)' 940 \Rightarrow LATITU=(-inf-15.791]' 434 (0.46)$$

El 95% de los terremotos con una magnitud inferior a 6.15 no producen tsunamis.

$$EQ_DEPT=(-inf-10.5]' 1020 \Rightarrow FL_TSUN=(-inf-0.5]' 899 (0.88)$$

$$EQ_DEPT=(33.5-inf)' 812 \Rightarrow EQ_MAGNIT=(7.25-inf)' 448 (0.55)$$

$$EQ_DEPT=(33.5-inf)' 812 \Rightarrow EQ_MAGNIT=(7.25-inf)' 448 (0.55)$$

El 88% de los terremotos que se producen a una profundidad inferior a 10.5 m, no producen tsunamis.

$$LONGITU=(-inf-19.98]' 947 \Rightarrow LATITU=(-inf-15.791]' 461 (0.49)$$

$$LONGITU=(19.98-96.485]' 949 \Rightarrow FL_TSUN=(-inf-0.5]' 899 (0.95)$$

$$LONGITU=(19.98-96.485]' 949 \Rightarrow LATITU=(37.393-inf)' 474 (0.5)$$

$$LATITU=(-inf-15.791]' 948 \Rightarrow LONGITU=(-inf-19.98]' 461 (0.49)$$

$$LATITU=(37.393-inf)' 948 \Rightarrow LONGITU=(19.98-96.485]' 474 (0.5)$$

$$LATITU=(-inf-15.791]' 948 \Rightarrow EQ_MAGNIT=(7.25-inf)' 434 (0.46)$$

LATITU=(15.791-37.393)' 949 => FL_TSUN=(-inf-0.5)' 810 (0.85)
 LATITU=(37.393-inf)' 948 => FL_TSUN=(-inf-0.5)' 787 (0.83)
 LATITU=(37.393-inf)' 948 => LONGITU=(19.98-96.485)' 474 (0.5)

4. DESARROLLO ANALITICO

A continuación se muestra un estudio de distintos algoritmos de clasificación, entrenados con una cierta cantidad de datos, y verificados con los correspondientes datos de test.

En primer lugar, para tener cierta perspectiva, vamos a explicar brevemente cada uno de los algoritmos que se han usado:

BayesNet -> Red Bayesiana de aprendizaje que utiliza diferentes algoritmos de búsqueda y medidas de calidad.

Proporciona una serie de estructuras de datos (estructura de la red, distribuciones de probabilidad condicional, etc), comunes a la Red de aprendizaje de los algoritmos de Bayes como K2 y B.

ClasificationViaRegresion -> Clasificación de la clase para hacer uso de los métodos de regresión. La clase es binaria y un modelo de regresión se construye para cada clase de valor.

JRip -> Esta clase implementa una regla propositiva de aprendizaje, repetida incremental. Recurre a la poda para reducir el error.

NaiveBayesSimple -> Utiliza un clasificador bayesiano cuyos atributos numéricos son modelados como una distribución normal.

NNge -> Usa un algoritmo similar al vecino mas cercano no anidado de forma que los hiperrectangulos pueden ser vistos como condiciones IF-THEN

OrdinalClassClassifier -> Meta clasificador que permite que la clasificación estándar de algoritmos sea aplicada a problemas ordinarios de clasificación.

Hay que destacar que las reglas que rozan el umbral de 50% serán reglas de baja precisión y habrá que considerarlas simplemente como tendencias.

SimpleLogistic -> Clasificador lineal para la construcción de modelos de regresión logística. LogitBoost con funciones de regresión simple como base de aprendizaje se utiliza para la instalación de modelos logísticos. El número óptimo de iteraciones del LogitBoost para realizar un cruce validado conduce a la selección automática de atributo

REPTree -> Los arboles de decisión son algoritmos de aprendizaje por inducción supervisada que pretenden modelar los datos de ejemplo mediante un árbol. Los nodos intermedios son los atributos de entrada de los ejemplos presentados, las ramas representan valores de dichos atributos y los nodos finales son los valores de la clase.

Para elegir que atributos y en que orden aparecen en el árbol, se utiliza una función de evaluación llamada ganancia de información (reducción de entropía del conjunto al clasificar usando un determinado atributo).

Tienen como ventaja que son fáciles de programar pues se traducen en regla if-else. Los hay que trabajan con atributos nominales únicamente, como el ID3, y que trabajan también con atributos numéricos, como el C4.5 (j48 en Weka).

El REPTree en concreto es un método de aprendizaje rápido mediante árboles de decisión. Construye un árbol de decisión usando la información de varianza y lo poda usando como criterio la reducción del error. Solamente clasifica valores para atributos numéricos una vez. Los valores que faltan se obtienen partiendo las correspondientes instancias.

En la siguiente tabla se muestra el porcentaje de aciertos y errores que se obtienen con cada uno de ellos, para algunas de las variables que se han empleado para realizar el estudio.

Nota: no todos los algoritmos se han probado con todas las variables, de ahí los huecos en blanco en la tabla.

Tabla 1. Tabla comparativa que recoge los resultados de cada clasificador

	Magnitud	Longitud	Latitud	Daños	Muertes	Heridos
BayesNet	31.8342 %	86.2966 %	79.2691 %	99.6901 %	99.4835 %	99.0702 %
ClasificationViaRegresion	32.1855 %	89.74 %				
JRip	24.7365 %	88.7561 %	80.8855 %	99.6901 %	99.5868 %	99.0702 %
NaiveBayesSimple	32.045 %	60.3306 %	79.1322 %			
NNge		89.3886 %	82.1504 %	99.5868 %	99.5868 %	99.0702 %
OrdinalClassClassifier	32.8883 %	90.1616 %	81.7287 %	99.8967 %	99.5868 %	99.0702 %
REPTree	27.7893 %	90.5833 %	82.9234 %	99.8967 %	99.5868 %	99.0702 %
SimpleLogistic	31.9747 %	90.0211 %	82.7829 %	99.7934 %		

Observando la tabla se puede ver que para todos los atributos elegidos, los distintos clasificadores tienen un porcentaje de éxito bastante elevado.

Con el que peor resultado se obtiene es con la longitud, quizás porque sea el atributo que más depende de todos los demás, y por lo tanto el más variable.

Cabe destacar que los algoritmos que han obtenido mayor precisión, a excepción de la predicción de la magnitud, son el "NNge" y el "OrdinalClassClassifier".

La longitud y la latitud son bastante aseguibles de clasificar, debido a que la mayoría de los terremotos se concentran en una serie de regiones del planeta.

En cuanto a daños, muertes y heridos, también se clasifican bastante bien debido a que los estos parámetros suelen ser mas o menos parecidos para terremotos de una magnitud similar, de ahí que el porcentaje de éxito en la clasificación este cercano al 100%

EL REPTree (figura 16) también consigue unos buenos resultados, y además tiene la ventaja de que es fácil de implementar debido a que muchas de las sentencias son if-then.

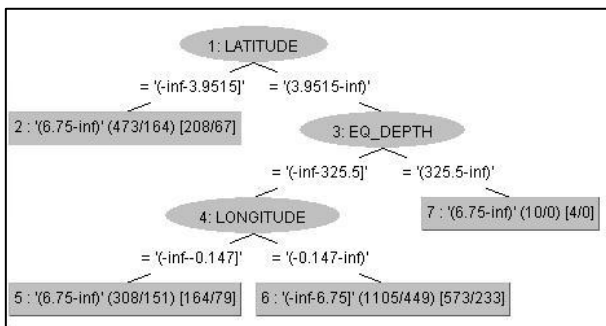


Figura 16. Árbol REPTree

Como se puede apreciar, de cada nodo salen 2 ramas, de ahí que dicha estructura de datos se pueda representar como un conjunto de sentencias if-then anidadas.

A continuación se muestra un gráfico de la predicción de la magnitud en función de los otros parámetros (figura 17).

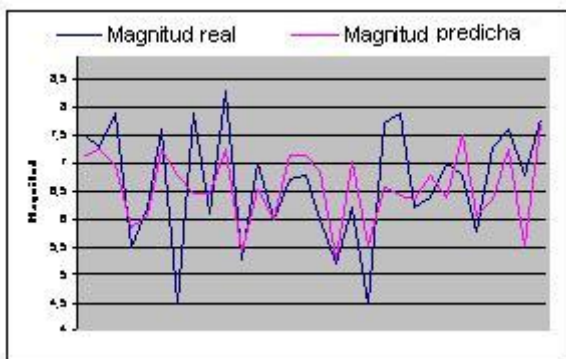


Figura 17. Comparación de magnitudes

Como se puede observar, los datos previstos no se acercan demasiado a la realidad, algo que se podía ver en la tabla, en el que el porcentaje de aciertos era demasiado bajo.

A continuación se muestra una comparativa de las longitudes reales y las predichas (figura 18).

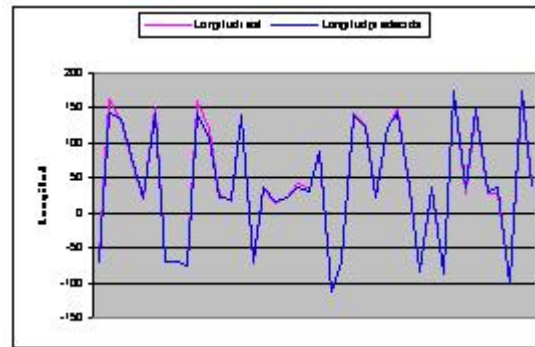


Figura 18. Comparación de longitudes

Como se podía esperar, se parecen bastante la real y la que nos da como resultado el algoritmo a la vista del porcentaje de acierto que se muestra en la tabla 1.

Cabe mencionar como último apunte, que el árbol de la figura 16 no se corresponde con las simulaciones de las graficas 17 y 18, ya que ese esta hecho con 2 grupos, y no con 8, para poder representarlo en este documento.

5. CONCLUSIONES

Estas son algunas de las conclusiones que han sacado tras todos los análisis realizados.

Con un conjunto de datos adecuados y con una herramienta adecuada, Weka en este caso, se puede obtener conclusiones que a simple vista serían complicadas de discernir.

Los resultados a los que se llega después de los distintos análisis pueden valer para múltiples funciones. Se puede facilitar la clasificación de nuevos casos y de ahí predecir las medidas que hay que tomar al respecto.

Después de las medidas oportunas para minimizar los daños de los terremotos, a partir de la predicción de donde se producen, podrían desarrollarse algunas tecnologías para aprovechar la gran cantidad de energía que estos desprenden.

El tener una muestra de datos significativa hace que se pueda acotar el rango de datos que se va a tener.

Con las conclusiones a las que se llega, se podría averiguar que otros datos serían interesantes para seguir desarrollando la capacidad de predicción.

A la vista de los distintos análisis se puede concluir que la gran mayoría de los terremotos, sobre todos los mas devastadores se reducen a 3 zonas del planeta: la costa del pacífico del contiene americano, la zona de oriente medio, y la zona pacífica del continente asiático y oceánico.

Los terremotos suelen tener una magnitud media de 6,6 en la escala Richter y los que más víctimas provocan son los que producen tsunamis, tanto de forma directa (arrasados por la ola), como indirecta (consecuencias del tsunami: destrucción de infraestructuras, propagación de epidemias...)

6. OTRAS INVESTIGACIONES POSIBLES A DESARROLLAR

Como futuros estudios a realizar, sería interesante la posibilidad que ofrece WEKA de construir y evaluar un clasificador de forma cruzada con dos ficheros de datos.

Un posible análisis de este tipo sería la generación con el filtro de instancias de dos conjuntos de datos correspondientes a terremotos de misma magnitud pero de lugares geográficos distintos.

Otro posible trabajo a realizar sería automatizar la búsqueda de atributos más apropiada para explicar un atributo objetivo, en un sentido de clasificación supervisada.

Así, podríamos explorar qué subconjuntos de atributos son los que mejor pueden clasificar la clase de la instancia.

Por último, no se ha tenido en cuenta la opción de agrupamiento o clustering y, por lo tanto, es una posibilidad que queda abierta para otro análisis.

7. BIBLIOGRAFIA

[1] http://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos

[2]

<http://www.idescat.net/sort/questiio/questiio.pdf/25.3.4.Aluja.pdf>

[3] <http://www.angelfire.com/nt/terremotosPrediccion/>

[4] <http://www.selecciones.com/acercade/art.php?id=557>

[5] <http://scalab.uc3m.es/~docweb/aa/software.html>