Búsqueda de patrones sobre grandes volúmenes de datos temporales

D.D. Carpintero[1,2], E. Gularte[3], & G. Baume[1,2]

[1] *Facultad de Ciencias Astronómicas y Geofísicas, UNLP, La Plata, Argentina*
[2] *Instituto de Astrofísica de La Plata, UNLP-Conicet, Argentina*
[3] *Geodesia Espacial y Aeronomía, Facultad de Ciencias Astronómicas y Geofísicas, UNLP, La Plata, Argentina*

Contacto / ddc@fcaglp.unlp.edu.ar

Resumen / Se ha realizado un estudio comparativo de diferentes técnicas de minería de datos, aplicadas al caso particular de dependencia temporal, alta dimensionalidad y muestreo irregular. Para realizar el reconocimiento de patrones en el conjunto de datos se utilizaron tanto técnicas supervisadas (árboles de decisión) como no supervisadas (reglas de asociación y técnicas de agrupamiento). Se presenta un ejemplo de aplicación para caracterizar la ionósfera terrestre usando datos a lo largo de un ciclo solar y provistos por sondadores ubicados en latitudes geográficas medias. Nuestro análisis ha permitido describir y predecir el comportamiento ionosférico en base a los distintos enfoques provistos por las técnicas implementadas.

Abstract / We have carried out a comparative study of different data mining techniques, applied to the case of time dependence, high dimensionality and irregular sampling. We have used supervised (decision trees) and unsupervised (association rules and clustering) techniques to recognize patterns within the data. We present an implementation example in which the terrestrial ionosphere is characterized by means of data from a solar cycle, obtained at mid-latitudes. Our analysis allows to describe and predict the ionospheric behaviour based on the different approaches provided by the implemented techniques.

Keywords / methods: data analysis — techniques: miscellaneous — Earth

1. Introducción

En la actualidad, en Astronomía se dispone de bases de datos estructuradas de gran volumen y con una elevada tasa de crecimiento. Dichos datos pueden ser espaciales, temporales, secuenciales o multimedia. Este panorama requiere el uso de técnicas automáticas y objetivas para procesarlos y analizarlos, exigiendo un replanteo de los métodos tradicionales.

En cada base, los datos poseen ciertas características que los definen. Estas características pueden ser valores de alguna variable física continua (velocidad, presión, etc.), discreta (día del mes, dimensión del espacio, etc.) o incluso no cuantitativa (forma, color, etc.). Así, se abre la posibilidad de una búsqueda sistemática de patrones en función de las características disponibles.

En este trabajo se presenta un estudio preliminar de diferentes técnicas de minería usando datos temporales de muchas dimensiones y con un muestreo irregular. En particular, se aplicaron dichas técnicas para caracterizar el comportamiento de la ionósfera terrestre en latitudes medias.

2. Datos

Se utilizaron datos correspondientes a las siguientes estaciones ionosféricas de Sudáfrica: Grahamstown, Hermanus, Louisvale y Madimbo. Dichos datos fueron obtenidos de DIDBase (*Digital Ionogram Database*) a través del portal *Global Ionospheric Radio Observatory* (GIRO, `http://giro.uml.edu/`), abarcando prácticamente todo el ciclo solar 24 (período 2009-2018).

Las características utilizadas para cada observación fueron el instante en que se llevó a cabo (tiempo local LT y fecha) y diferentes parámetros ionosféricos: frecuencias máximas fo para las cuales las señales de radio de un sondador son reflejadas y sus correspondientes alturas h para las capas E, Es, F, F1 y F2, y el factor de propagación MD de la región $F2$ que representa la frecuencia óptima para transmitir una señal a una distancia de 3000 km (Kelley, 1989). También se utilizaron otros parámetros físicos como índices geomagnéticos (ap y $DST = Disturbance\ Storm\text{-}Time$) e índices de actividad solar.

3. Metodología y resultados

La primera etapa del procedimiento fue la identificación y eliminación de datos anómalos (*outliers*), definidos como aquellos datos que difieren en más de 3σ de los valores medios obtenidos a través de un suavizado con filtro gaussiano. A continuación se realizó, según el caso, la normalización o la categorización de las características. Como paso siguiente, se aplicaron reglas de asociación y técnicas de agrupamiento para describir el comportamiento ionosférico, así como también la construcción de árboles de decisión para vincular los parámetros ionosféricos con la actividad geomagnética.

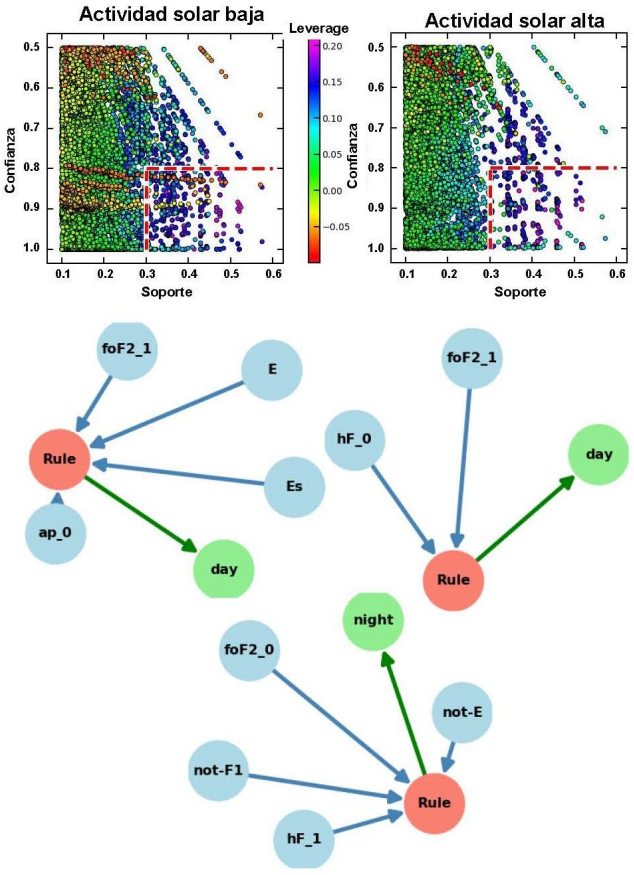

Figura 1: Reglas de asociación correspondientes a la estación Grahamstown. Paneles superiores: Localización de las reglas de asociación de acuerdo a diferentes métricas durante un años de baja y alta actividad solar. Grafos inferiores: Representación esquemática de algunas de las reglas.

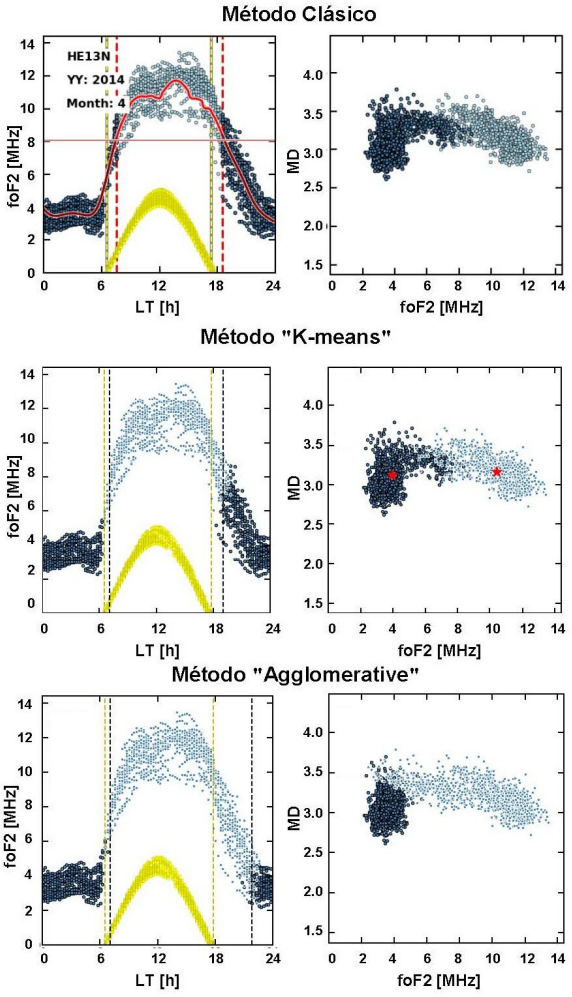

Figura 2: Resultados de diferentes técnicas de agrupamiento sobre los datos de la estación Hermanus en un dado mes de un año de alta actividad solar. En amarrillo se denota la trayectoria del sol durante el día, mientras que la curva roja en el panel superior izquierdo representa valores medios.

Las diferentes etapas del procedimiento fueron llevadas a cabo utilizando las bibliotecas en PYTHON provistas por SCIKIT LEARN (Pedregosa et al., 2011) y MLXTEND (Raschka, 2018).

3.1. Reglas de asociación

Para obtener las reglas de asociación se empleó el algoritmo *A priori* (Witten et al., 2011), agrupando los datos por año y por estación. Las características seleccionadas fueron: $foF2$, hF, MD, $hmF2 - hF$ (espesor de la capa F2), E, Es, F1, ap y LT. Las mismas se categorizaron de la siguiente manera: un bajo o alto valor; la presencia o no de las capas E, Es y F1; un bajo, medio o alto valor del índice ap, y la presencia o no del Sol (día y noche).

Las reglas de asociación así encontradas se evaluaron mediante diferentes métricas (Witten et al., 2011): *confianza* (*confidence*), *alza* (*lift*), *soporte* (*support*), *convicción* (*conviction*) e *influencia* (*leverage*), y se seleccionaron aquellas con mejor performance general (rectángulos rojos punteados en la Fig. 1, en la cual solo se han graficado *soporte*, *confianza* e *influencia*).

Las reglas seleccionadas son coherentes con el comportamiento conocido de la ionósfera (Kelley, 1989), reforzando la validez de la metodología utilizada. La cantidad de reglas halladas durante la actividad solar alta es menor que en la actividad solar baja, indicando menor cantidad de vínculos entre las características consideradas. Además, la Fig. 1 revela la presencia de grupos de reglas cuyas métricas se encuentran vinculadas linealmente entre sí.

3.2. Agrupamiento

Por otra parte, las técnicas de agrupamiento clásicas (Kelley, 1989) se compararon con las técnicas *promedios K* (*K-means*) y *aglomeración* (*agglomerative*) (Witten et al., 2011); los resultados pueden verse en la Fig. 2. Los datos se subdividieron por año, por estación y por mes. Las características consideradas fueron $foF2$, hF, MD y LT, las cuales fueron previamente normalizadas. Se ensayaron distintas métricas para el cálculo de distancias, y al no encontrarse mayores diferencias en los resultados se optó por utilizar la métrica euclidiana.

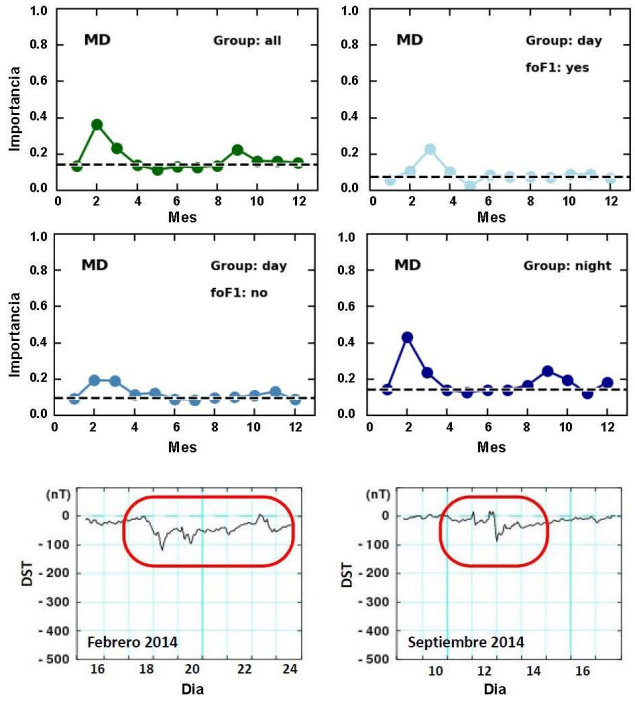

Figura 3: Resultados para la estación Grahamstown durante un año de alta actividad solar. Paneles superiores y medios: Importancia de *MD* a lo largo del año para diferentes conjuntos de datos (ver texto). Paneles inferiores: Indice *DST* en dos meses del año con actividad geomagnética. Las flechas indican los eventos significativos en *MD* correlacionados con las alteraciones registradas en el índice *DST*.

La búsqueda se centró en la distinción de dos grupos: uno bajo la acción de los efectos del Sol y el otro no. Finalmente, para la validación de cada una de las técnicas se utilizó el índice de perfilado (*silhouette*, Witten et al., 2011), que relaciona la distancia de los miembros dentro de su grupo con la distancia entre grupos.

En la Fig. 2 se presenta un ejemplo de aplicación. En general, los grupos encontrados por el método de promedios K son similares, aunque algo mejores, que los obtenidos con el método clásico. Además, el método de promedios K provee las coordenadas de los centros de los grupos en el espacio de las características utilizadas (denotado por estrellas rojas en la Fig. 2). Sin embargo, el método de aglomeración brinda los mejores resultados, incorporando el efecto residual del Sol después del atardecer. Esto es debido a que este método se basa en la distancia entre los datos de cada grupo y no en la distancia a un centro común. En general, el comportamiento de los grupos encontrados es similar en todas las estaciones de medición, siendo consistente con el hecho de que todas se encuentran en latitudes medias.

3.3. Árboles de decisión

Como técnica de predicción se utilizó la técnica supervisada *árboles de decisión* (*Random Forest Decision trees*, Witten et al., 2011), a fin de identificar la actividad geomagnética a partir de los parámetros ionosféricos. Se consideraron los índices geomagnéticos *ap* y *DST* como variables de salida. En ambos casos se adoptó tanto la modalidad numérica (regresión) como la categórica (clasificación). Los datos utilizados fueron separados por año, por mes y por estación, y agrupados por presencia o no del Sol (día - noche) y por presencia o no de la capa F1 durante el día. Las características de entrada fueron la frecuencia, el espesor y la altura de las capas, junto con el tiempo local.

El índice de *importancia* (Witten et al., 2011) de las características utilizadas sirvió para identificar las tormentas geomagnéticas. Los resultados se compararon con el comportamiento del índice geomagnético.

El índice *DST* en formato numérico arrojó los resultados con mayor sensibilidad de detección. Además, se encontró que las características *MD* y *foF*2 resultaron importantes en todos los casos estudiados. En los casos de alta actividad solar, las características vinculadas con las alturas de las capas correlacionaron con las tormentas geomagnéticas (ver Fig. 3).

4. Conclusiones

En sistemas complejos, de muchas dimensiones, las relaciones y correlaciones entre los distintos parámetros son en general difíciles de determinar, excepto que se tenga a disposición un modelo físico que permita vincular algunos parámetros entre sí. Las reglas de asociación permiten hallar relaciones entre parámetros que de otro modo pasarían inadvertidas. Las técnicas de agrupamiento aglomeran los datos en grupos que comparten ciertas características comunes, mientras que los árboles de decisión permiten establecer la importancia que cada parámetro adquiere en relación con otros, obteniendo así las dependencias entre las características.

Agradecimientos: Este trabajo ha sido parcialmente financiado por el PIP 112-201701-00055 de CONICET y los Programas de incentivos 11/G153, 11/G154 y 11/G158 de la UNLP. Los autores agradecen a South African National Space Agency (SANSA) por brindar la continua elaboración y disponibilidad de los datos ionosféricos.

Referencias

Kelley M.C., 1989, *The Earth's Ionosphere*, International Geophysics Series, Academic Press

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Raschka S., 2018, The Journal of Open Source Software, 3

Witten I.H., Frank E., Hall M.A., 2011, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA