Big Data para el análisis de tormentas severas

Santiago Banchero[1,2], Marcelo Soria[3], and Romina N. Mezher[1]

[1] Instituto de Clima y Agua (INTA). De los Reseros y N. Repeto s/n, Hurlingham.
[2] Universidad Nacional de Luján.
[3] Facultad de Agronomía. UBA
{banchero.santiago, mezher.romina}@inta.gob.ar
soria@agro.uba.ar

Resumen Se presenta para discusión los primeros resultados del trabajo de una tesis de maestría en minería de datos y descubrimiento de conocimiento que tiene el objetivo de evaluar cuál es la capacidad de predicción de ocurrencia de granizo de un conjunto de índices de inestabilidad utilizando técnicas de aprendizaje automático en un entorno de Big Data. Además deja constancia de los procesos de ETL para la integración de un conjunto de fuentes heterogéneas con variedad de escalas de relevamiento y los primeros resultados del análisis multivariado sobre algunos eventos destacados. Una tormenta severa es un fenómeno atmosférico con capacidades destructivas, como pueden ser tormentas eléctricas intensas, tormentas de granizo y tornados. El granizo se considera un riesgo natural y los daños provocados por este fenómeno meteorológico extremo causan en Argentina graves pérdidas en algunas regiones y afecta a diferentes sectores económicos, tanto en las zonas urbanas como rurales. La precipitación de granizo se caracteriza por tener una alta variabilidad espacial y temporal lo que representa un gran desafío para el análisis y desarrollo de modelos de pronóstico a corto plazo.

Keywords: Granizo, ETL, Clustering, Big Data

1. Introducción

Una tormenta severa es un fenómeno atmosférico con capacidades destructivas, como pueden ser tormentas eléctricas intensas, tormentas de granizo y tornados. El granizo se considera un riesgo natural [9] y los daños provocados por este fenómeno meteorológico extremo causan en Argentina graves pérdidas en algunas regiones y afecta a diferentes sectores económicos, tanto en las zonas urbanas como rurales [14]. La precipitación de granizo se caracteriza por tener una alta variabilidad espacial y temporal que presenta un gran desafío para el análisis y desarrollo de modelos de pronóstico a corto plazo.

En este trabajo se utilizan índices atmosféricos (CAPE, Lift, K-Index, Total Totals, etc) [8] que se calculan a partir de variables atmosféricas (presión, altura, temperatura, temperatura de rocío, entre otras). Estas variables son provistas por Global Forecast System (GFS) [17], [12], [11], que es un sistema de

predicción numérica del tiempo y que es corrido diariamente por el US National Weather Service (NWS) para generar pronósticos a corto plazo. Estos datos están disponibles cuatro veces al día con cobertura global en formato de grilla con una resolución espacial de 0.25 grados. Además se utilizaron datos de la Red de Radares de INTA [10] puntualmente se utiliza la variable reflectividad (o dBZ) que permite conocer con cierta precisión si una nube es portadora o no de granizo [16].

En este trabajo se presentan los resultados preliminares de un análisis de casos para el año 2015 donde se buscaron las mejores alternativas de reducción de dimensionalidad para un conjunto de datos con altos niveles de colinealidad. Aquí se utilizó una técnica simple basada en el coeficiente de correlación de Pearson que permite eliminar pares con una fuerte correlación.

Y se avanzó en el análisis de clusters con el objetivo de buscar patrones de comportamiento en los diferentes eventos ocurridos en el intervalo analizado. Se utilizó la técnica basada en distribuciones Gaussianas Mixtas (o GMM del inglés Gaussian Mixture Model) para determinar el número correcto de poblaciones o grupos.

2. Materiales y Métodos

2.1. Conjunto de datos

Para la realización de pronósticos de granizo es necesario integrar diferentes fuentes de datos, por lo que se ha requerido el diseño de distintas estrategias de ETL (*Extract-Transform-Load*) para conformar un datawarehose que permita continuar con las tareas de análisis y modelado típico de un proceso de minería de datos. En este trabajo se diseño una arquitectura de ETL para la etapa de preprocesamiento de los diferentes orígenes de datos (Figura 3). El núcleo de esta arquitectura se compone por los procesos: GFS, RADAR y Extracción, donde cada uno está comunicado por medio de colas AMQP (*Advanced Message Queuing Protocol*) [19] para el pasaje de mensajes y tareas. De esta manera se integran los datos del *Global Forecast System* (GFS) [4] con los de la Red de Radares de INTA [10].

La validación de los eventos se realiza con datos de campo que se obtienen también de diferentes fuentes. Una de las principales son empresas aseguradoras de granizo que tienen una gran cobertura de las zonas rurales. También se utilizaron puntos extraídos de reportes en redes sociales, puntualmente se trabajó con Twitter ya que es posible conseguir información en tiempo casi real para mejorar los pronósticos [5]. Además se dispone de los puntos relevados por el Proyecto Alert.Ar y su aplicación Alertamos [2] que ha permitido conseguir una buena cantidad de datos para validación. La región de estudio es la comprendida por la Red de Radares de INTA que posee una amplia cobertura en la región Pampeana (Figura 1).

El *Global Forecast System* (GFS) [17] [12] [11], es un sistema de predicción numérica del tiempo a escala global para generar pronósticos a corto plazo.

Figura 1. Área cubierta por la Red de Radares INTA

GFS provee un abanico importante de variables, entre las principales podemos encontrar datos de temperaturas, vientos, precipitaciones, humedad del suelo y concentración de ozono atmosférico entre otros [1]. Todo el mundo está cubierto por GFS con una resolución horizontal de 28 kilómetros entre puntos de la grilla, que se utiliza por los pronosticadores operacionales que predicen el tiempo hasta 16 días.

Los datos de GFS son utilizados para el cálculo de los índices y también se utilizan algunas variables atmosféricas adicionales provistas por este modelo. Los índices de inestabilidad, como: CAPE, Lift, K-Index, Total Totals, etc. [8] se obtienen a partir de variables atmosféricas como presión, altura, temperatura, temperatura de rocío, entre otras. Su principal utilidad es brindar una herramienta de exploración de condiciones de convección.

En el trabajo se utilizaron tres pronósticos (figura 2) para determinar las condiciones atmosféricas a las 12 hs UTC de un día. Las horas previas utilizadas son: 12, 18 y 24 de las corridas 00, 18 y 12 respectivamente. Así el conjunto de datos provenientes de GFS se compone de 90 variables entre las tres horas pronosticadas.

Como variable objetivo de nuestros modelos se utilizaran las observaciones de la red de radares del INTA. Este instrumental conforma una red híbrida con

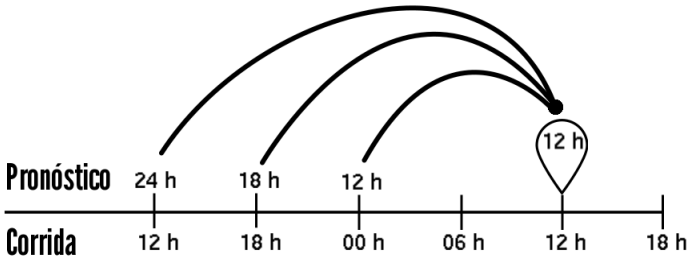

Figura 2. Pronósticos previos para las 12 utc

sensores de simple (Radar de Pergamino) y doble (Radares de Paraná y Anguil) polarización que permiten la observación detallada de la atmósfera. Puntualmente se utiliza la variable dBZ (o reflectividad) ya que se conocen los umbrales [16] de dicha variable para determinar fuertes condiciones de convección.

2.2. Integración de los datos

La integración de estos datos para el análisis se realizó a través de diferentes módulos de procesamientos integrados y gestionando los trabajos utilizando dos colas de mensajes AMQP (implementadas con rabbitMQ [18]) como se muestro más arriba en la figura 3. Se eligió esta opción de implementación ya que facilitó sobremanera poder escalar horizontalmente agregando módulos en diferentes equipos disponibles. Los siguientes párrafos detallan las tareas de cada uno de estos módulos.

El módulo GFS es el encargado de preprocesar las grillas GRIB/GRIdded Binary [6] para poder calcular los índices atmosféricos y realizar las transformaciones espaciales correspondientes para que coincidan con los datos de Radar. Además realiza los recortes correspondientes para delimitar solo el área de interés.

En el siguiente módulo de RADAR los que se realiza es el cálculo del producto CMAX para el día del evento. Esto consiste en tomar un día completo de imágenes para cada radar (son 144 barridos diarios, uno cada diez minutos) y para las dos primeras elevaciones integrar los valores obteniendo la mayor reflectividad durante el día. Esto va a determinar los lugares por donde pasaron las nubes con mayor probabilidad de haber alcanzado condiciones de convección.

Por último, el módulo de extracción que permite calcular para cada celda de GFS la densidad de píxeles de radar con valores de dBZ superiores a un umbral (en este caso 30 dBZ). Para esto se utiliza un ajuste no paramétrico para cada celda GFS utilizando densidad Kernel y calculando la integral del área bajo la curva de para valores de dBZ mayores a treinta [3].

Estos módulos permiten contar con una matriz de datos donde para cada corrida de GFS tenemos la probabilidad de que una celda se hayan dado condiciones de convección. Si bien no significa con seguridad que las nubes observadas

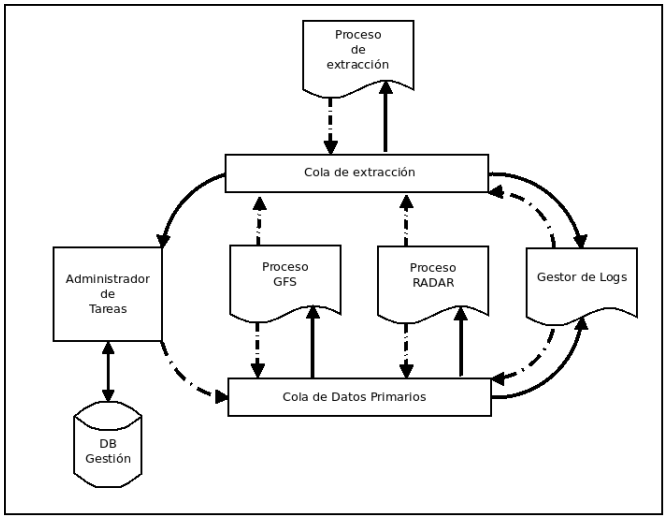

Figura 3. Arquitectura del sistema de preprocesamiento e integración de GFS y RADAR

hayan precipitado granizo ante valores muy altos de probabilidad existe una fuerte correlación entre los valores calculados y los puntos testigos que se vieron a campo.

Con respecto a los puntos de validación (o verdades de campo) como se mencionó antes, se trabaja principalmente con datos de empresas aseguradoras de granizo. En este sentido se han firmado convenios de vinculación y cooperación para en intercambio de datos.

También se ha obtenido una significativa cantidad de puntos de reportes de caída de granizo de la red social Twitter. Para esto se realizó un desarrollo utilizando la API REST de Twitter, haciendo una búsqueda por *granizo* como palabra clave y luego se los separa entre los que vienen con información de ubicación de los que no. Cuando contienen la posición son validados con datos de radar y los que pasan el filtro son incorporados a la base de datos de eventos. Por otro lado, los tweets que no poseen información de ubicación se analiza el texto y a través de una técnica de extracción de información como es reconocimiento de entidades (o NER - *Named-entity recognition*) se trata de determinar la ubicación utilizando un buscador de topónimos. Para esto se utiliza la herramienta MITIE: MIT Information Extraction [15].

2.3. Reducción de dimensionalidad

Uno de los principales problemas encontrados en el análisis de datos fue la gran dependencia existente entre las variables. Esto se debe en gran medida a que los índices atmosféricos se calculan a partir de las mismas variables GFS en la mayoría de los casos. En la figura 4 se muestra la matriz de correlaciones

para un conjunto de variables GFS e índices derivados del modelo. Estas 29 variables para el pronóstico de las 12 hs UTC inicialmente tenían valores de correlación que variaban entre -0.9592 y 0.9976. Para mejorar esta situación de manera eficiente se optó por realizar una eliminación por pares correlacionados [20].

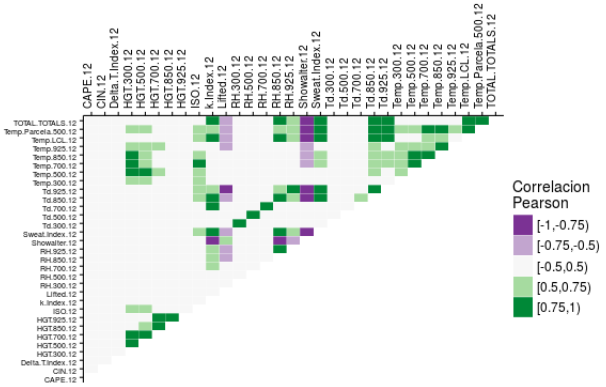

Figura 4. Matríz de correlaciones del pronóstico -12 hs compuesta por un total de 29 variables y valores de correlación entre -0.9592 y 0.9976

Luego de aplicar este método de eliminación quedaron solo doce variables y la correlación global disminuyó a valores entre -0.57460 y 0.53890.

2.4. Análisis de conglomerados

Se realizó un análisis de conglomerados utilizando las variables e índices que fueron extraídas en el proceso de reducción de dimensionalidad. De esta manera se segmentaron las celdas GFS, tratando de buscar comportamientos homogéneos entre los sectores afectados por granizo. Para realizar esta tarea se utilizó la técnica de GMM del paquete R mclust [7] para determinar el número correcto de poblaciones o clusters existentes en cada pronóstico GFS. De esta manera se evaluó un conjunto de 29 fechas con eventos positivos, es decir, que hay evidencia de la caída de granizo. En la figura 5, se muestra dos ajustes para casos particulares donde hay evidencia focalizada en pocos clusters (puntos de siniestros verificados) y en otra los eventos están dispersos en gran parte de la región de estudio.

Luego con los valores de K ya estimados se procedió a realizar un agrupamiento utilizando PAM (*Partitioning Around Medoids*) provisto por el paquete cluster [13] de R. Los agrupamientos obtenidos presentaron valores débiles de coeficiente de silueta para todos los agrupamientos realizados, no se superó el

0.35 como ancho promedio de silueta. Pero en el análisis detallado de los clusters, se obtuvieron valores superiores a 0.5 para algunos grupos.

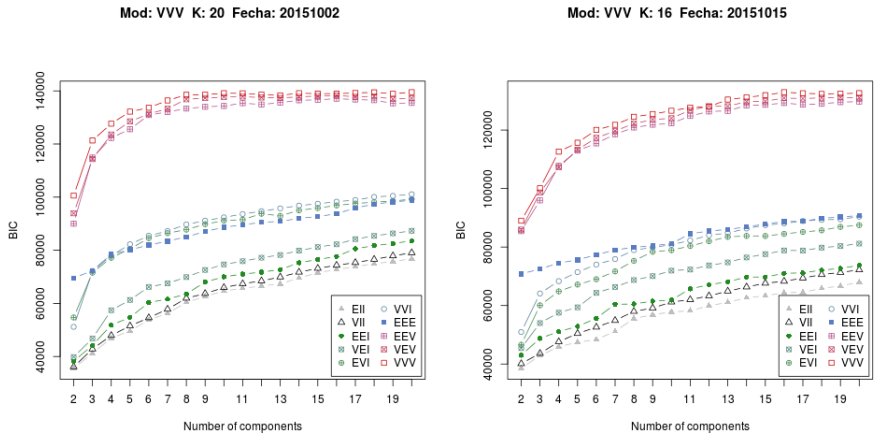

Figura 5. Ajuste de GMM con mclust. En el gráfico de la izquierda se muestra el ajuste para el día 02/10/2015 donde el algoritmo que mejor ajusta es el VVV (*ellipsoidal, varying volume, shape, and orientation*) que converge para un k=20; a derecha, se muestra el ajuste para el día 15/10/2015 y se observa que el mejor algoritmo también es VVV pero converge con k=16.

Los agrupamientos obtenidos fueron cruzados con puntos de validación donde ya estaba constatada la caída de granizo. Los datos testigo de los eventos se concentraron en unos pocos clusters, un total de 3 en el caso del 02/10/2015. Mientras que en el caso del 15/10/2015 hay una gran dispersión de los eventos y en gran parte de los clusters (un total de 7) se ha observado granizo.

Se realizó una comparación visual de las características de estos grupos donde se tenía evidencia de caída de granizo. En la figura 6 puede observarse cómo son los clusters con eventos positivos a través de la comparación de sus firmas en un diagrama de coordenadas paralelas. Allí es posible observar que las variables observadas para el caso del 02/10/2015 en los tres grupos tienen comportamientos bien diferentes. En el otro evento, el del 15/10/2015 hay algunos clusters que son mucho más homogéneos. En los casos de los clusters 7, 8, 10 y 11 las similitudes son muy marcadas mientras que en el resto son diferentes. Lo interesante es que en casi todos los grupos se observa que los valores del CAPE de superficie es bajo y que la humedad relativa (RH) en las capas bajas también es alta.

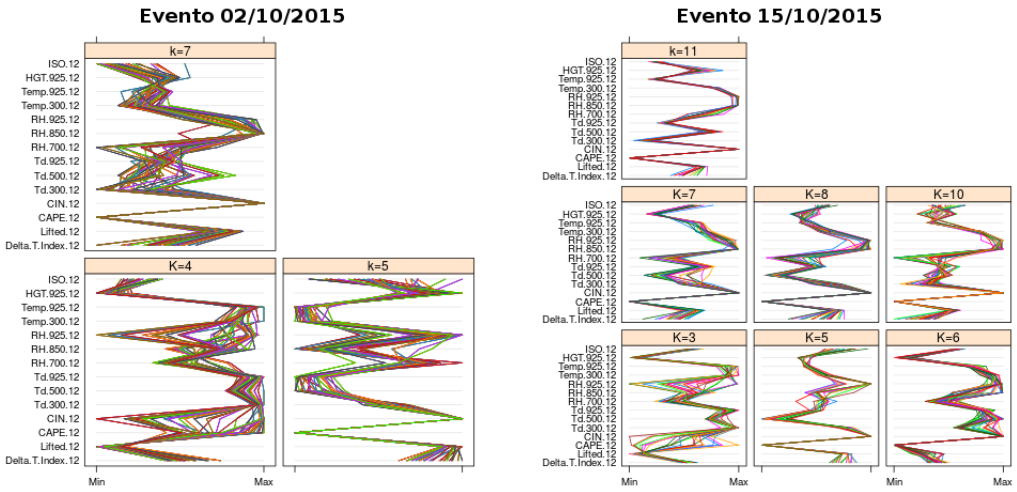

Figura 6. Gráficos de coordenadas paralelas donde se muestran los clusters donde se observó granizo para las dos eventos de referencia estudiados

3. Comentarios Finales

Los desafíos principales del trabajo han sido poder realizar una integración de una cantidad importante de fuentes de información que tienen una gran variabilidad espacio-temporal. La forma en que se atacó ese problema tiene la ventaja de poder escalar horizontalmente con lo cual el techo esta dado por las capacidades de hardware con que se cuente. En relación al análisis, se presentaron alternativas válidas para reducción de dimensionalidad y segmentación. Los resultados son preliminares resta realizar un análisis más exhaustivo para poder caracterizar de manera más precisa a los clusters donde se verificó la caída de granizo.

Referencias

1. Global Forecast System (GFS) | National Centers for Environmental Information (NCEI) formerly known as National Climatic Data Center (NCDC). `https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs`
2. Alert.ar: Alertamos. `http://alertamos.smn.gov.ar/`, accedido el 13 de Mayo de 2016
3. Banchero, S., Soria, M.A., Mezher, R.: Predicción de granizo utilizando índices atmosféricos. In: Simposio Argentino de GRANdes DAtos (AGRANDA 2015)-JAIIO 44 (Rosario, 2015) (2015)

4. Center, E.M.: The Global Forecast System (GFS) - Global Spectral Model (GSM) (gfs version 11.0.6). Tech. rep., Environmental Modeling Center [Disponible en http://www.emc.ncep.noaa.gov/GFS/doc.php] (2003)
5. Cox, J., Plale, B.: Improving automatic weather observations with the public twitter stream. IU School of Informatics and Computing (2011)
6. Dey, C., et al.: Guide to the wmo table driven code form used for the representation and exchange of regularly spaced data in binary form: Fm 92 grib. Tech. rep., WMO Tech. Rep., 98 pp.[Available online at http://www. wmo. int/pages/prog/www/WMOCodes/Guides/GRIB/GRIB2_062006. pdf.] (2007)
7. Fraley, C., Raftery, A.E., Scrucca, L.: mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation (2016), `https://CRAN.R-project.org/package=mclust`, r package version 5.2
8. Gottlieb, R.: Analysis of stability indices for severe thunderstorms in the northeastern united states. Ph.D. thesis (2009)
9. Hermida, L., Sánchez, J.L., López, L., Berthet, C., Dessens, J., García-Ortega, E., Merino, A.: Climatic trends in hail precipitation in france: spatial, altitudinal, and temporal variability. The Scientific World Journal 2013 (2013)
10. INTA - Instituto Nacional de Tecnología Agropecuaria: Red de Radares INTA. `http://radar.inta.gov.ar`, accedido el 13 de Mayo de 2016
11. Kalnay, E., Kanamitsu, M., Baker, W.: Global numerical weather prediction at the national meteorological center. Bulletin of the American Meteorological Society 71(10), 1410–1428 (1990)
12. Kanamitsu, M.: Description of the nmc global data assimilation and forecast system. Weather and Forecasting 4(3), 335–342 (1989)
13. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: cluster: Cluster Analysis Basics and Extensions (2016), r package version 2.0.4 — For new features, see the 'Changelog' file (in the package source)
14. Mezher, R.N., Doyle, M., Barros, V.: Climatology of hail in argentina. Atmospheric research 114, 70–82 (2012)
15. mit-nlp: mit-nlp/MITIE. `https://github.com/mit-nlp/MITIE`, accedido el 14 de Mayo de 2016
16. Rinehart, R.: Radar for meteorologists (1999)
17. Sela, J.G.: Spectral modeling at the national meteorological center. Monthly Weather Review 108(9), 1279–1292 (1980)
18. Videla, A., Williams, J.J.: RabbitMQ in action. Manning (2012)
19. Vinoski, S.: Advanced message queuing protocol. IEEE Internet Computing (6), 87–89 (2006)
20. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML. vol. 3, pp. 856–863 (2003)