Tormentas solares

MEMORIA

Juan Manuel Pino Pintado.

TFM Data Science Kschool 2019-2020.

ÍNDICE

1.Introducción

El 1 de septiembre de 1859 se produjo lo que después se ha llamado el "evento Carrington". A las 11:18 el astrónomo Richard Carrington estaba haciendo bocetos de las manchas solares cuando observó un estallido de luz que parecía salir de dos puntos del grupo de manchas que estaba observando. Unas 17 horas después las auroras boreales (normalmente circunscritas a las regiones polares) "convirtieron la noche en día", llegando a verse en latitudes tan bajas como Colombia. Se produjeron fallos en los sistemas de telégrafo de América del Norte y Europa, los hilos sufrieron cortocircuitos y se produjeron incendios. Lo que vio Carrington fue el inicio de una de las tormentas solares (o, para ser exactos, un conjunto de ellas) más intensas de la historia.

¿Pero que es una tormenta solar? Empecemos por las *manchas solares*, que son zonas oscuras en la superficie del Sol. Estas zonas están más frías que otras partes de la superficie, debido a que se forman en sitios en los que el campo magnético del Sol es particularmente fuerte. Tan fuerte que evita que parte del calor llegue a la superficie. Las líneas de energía de dichos campos magnéticos están continuamente generándose, entrecruzándose y cambiando, y a veces esto provoca una *llamarada solar*, que no es otra cosa que una explosión de energía, que libera inmensas cantidades de radiación. A veces, las llamaradas solares van acompañas de eyecciones de masa coronal (*CMEs*), explosiones en la corona del Sol que liberan partículas sólidas de las capas exteriores y radiación al espacio. Una tormenta solar, en pocas palabras, es lo que pasa cuando una CME alcanza la tierra e interacciona con el campo magnético terrestre. Normalmente dicho campo magnético es capaz de desviar la energía, que solo se manifiesta en las auroras boreales, pero si dicha energía es muy intensa el campo magnético puede ceder, provocando fenómenos como el descrito en el evento Carrington. Obviamente estas tormentas solares también pueden afectar a los cientos de satélites artificiales que orbitan la Tierra, dejándolos inservibles temporal o permanentemente, y a las redes eléctricas en tierra. Si actualmente una CME similar a la de 1859 alcanzara la Tierra se producirían con toda seguridad grandes apagones por sobretensión, así como daños en aparatos eléctricos y satélites artificiales

El objetivo de este TFM es, a partir de datos de manchas solares, llamaradas y CMEs, poder predecir la energía que tendrá una de estas CMEs, lo que podría servirnos para tomar medidas en la Tierra y en el espacio que controlamos antes de que nos alcance.

2.Estado del arte

El llamado ciclo solar es bien conocido desde hace años. La distribución de manchas a lo largo del tiempo muestra variaciones cíclicas de aproximadamente 11 años, combinadas con “superciclos” que abarcan varios periodos.

Parece que tanto las llamaradas solares como las CMEs están relacionadas también con dicho ciclo, siendo ambas más intensas en tiempos de máximos solares.

No todas las llamaradas solares llevan aparejada la erupción de una CME, sin embargo se sabe que cuando es así, ambos fenómenos están relacionados. Se ha sugerido por ejemplo que la duración de la llamarada es clave a la hora de generar o no una CME, y se han encontrado también relaciones estadísticas de distinto nivel entre la velocidad de una CME y el flujo de rayos x provocado por una llamarada.

En febrero de 2020 se lanzó el satélite Solar Orbiter, que permitirá estudiar la atmósfera exterior del Sol así como el flujo dinámico del viento solar, las erupciones y el campo magnético solar. En su máxima aproximación al Sol estará a “solo” unos 42 millones de kilómetros, siendo la nave que más se ha acercado hasta ahora a nuestra estrella.

3.Descripción de los datos.

Nos apoyaremos en tres fuentes de datos: manchas solares, llamaradas y CMEs.

Manchas solares.

El fichero original (SN_d_tot_V2.0.csv) está en la carpeta "Datos" del repositorio. Son datos recopilados por varios observatorios a lo largo de todo el mundo, y van desde 1818 a la actualidad. El formato es .csv con valores separados por ';'. Reflejaremos aquí la descripción de los campos (se marcan en rojo las usadas en el tratamiento de datos):

- Año (Year)
- Mes (Month)
- Dia (Day)
- Fecha en formato fracción de año (Fraction)
- Número total diario de manchas solares* (Spots)
- Desviación estándar (Stdev)
- Número de observaciones usadas para el cómputo (Observ)
- Indicador definitivo/provisional (Def/Prov)

*El número diario total de manchas solares se calcula según la fórmula de Rudolph Wolf:

$$R = k * (10g + s)$$

donde R es la cantidad de manchas solares, g la cantidad de grupos de manchas solares visibles, s la cantidad total de manchas individuales de todos los grupos, y k es un factor de escala (generalmente < 1) que da cuenta de las condiciones de observación.

Esta información puede consultarse en el fichero *Datos/ SN_d_tot_V2.0_description*.

Llamaradas solares.

Lo primero que hay que decir es que hemos querido coger los datos "en bruto", tal cual llegan a la Tierra, pues luego se transforman por ejemplo para definir la clase de las llamaradas, que según su intensidad actualmente se clasifican en A, B, C, M o X. Como decía nuestros datos son primarios, lo que tenemos es radiación en el espectro de rayos X recogida por el satélite GOES de la NOAA (National Oceanic and Atmospheric Administration) de 1975 a 2016. En este caso tenemos varios ficheros, generalmente uno por año, pero también hay años que tenemos 2. Son ficheros posicionales, detallamos aquí los campos:

- Año (Year)
- Mes (Month)
- Dia (Day)
- Identificador de la estación receptora (Gxx) (Station)
- Hora de inicio (Start_time)
- Hora de "pico" (energía de pico) (Max_time)

- Hora de fin (End_time)
- Energía de fondo (Background_energy)
- Energía de pico (Peak_energy)
- Energía integrada (Integrated_energy)

No usaremos los campos *Background_energy* e *Integrated_energy*, ya que la principal magnitud (en energía) de una llamarada solar es la energía de pico.

Esta información puede consultarse más extensamente en el fichero *Datos/Flares/xray-flares_description*.

CMEs.

La información de CMEs la tenemos integrada en un solo fichero, con registros separados por posiciones. Dicha información procede de las mediciones del satélite SOHO, también de la NOAA, de 1996 a 2020. Detallamos los campos:

- Fecha (AAAA/MM/DD) (Date)
- Hora (HH:MM:SS) (Time)
- Ángulo de posición central (Central_PA)
- Anchura aparente (Width)
- Velocidad lineal (calculada por interpolación lineal) (L_speed)
- Velocidad de 2º orden inicial (calculada mediante un polinomio de 2º orden) (2_I_speed)
- Velocidad de 2º orden final (calculada mediante un polinomio de 2º orden) (2_F_speed)
- Velocidad a 20 radios solares (2_20R_speed)
- Aceleración (Acceleration)
- Masa (Mass)
- Energía cinética (K_energy)
- Ángulo de posición de la medida (MPA)
- Observaciones (Remarks)

Esta información puede consultarse más extensamente en el fichero *Datos/univ_all_description.txt*.

Diremos ya aquí que nuestro "target" será la energía cinética de la CME, pero no nos interesa tanto averiguar el valor exacto, como hacernos una idea de la magnitud. Por ello "discretizaremos" la variable, asignándole números del 1 en adelante que indicarán intensidad creciente.

Uniremos estos tres datasets en uno, tomando como dato de unión la fecha.

Pretendemos así relacionar los distintos eventos, de cara a maximizar la relación CMEs – Llamaradas – Manchas, que es lo que estudiaremos.

Tendremos por tanto un conjunto de datos final de 1975 a 2016.

4.Metodología.

Hemos dividido el trabajo en dos partes:

4.1. Adquisición, limpieza y preparación de los datos.

Manchas solares.

Se han interpolado los valores que teníamos como NaN en la columna *Spots* (la única que nos interesa realmente), para perder los menos registros posibles. Se ha hecho el tratamiento de outliers, usando como medida (más o menos) la media + 3*(Desviación estándar). Se ha creado una columna nueva (*Date*) con la fecha en formato datetime para poder unir posteriormente los datos.

Llamaradas solares.

Aquí lo primero que hacemos es unir todos los ficheros en uno, así la información es más fácil de tratar. Se han eliminado los caracteres extraños e integrado la información en un dataframe. Se ha transformado la variable *Year* a cuatro posiciones y se ha creado la correspondiente columna de fecha en formato datetime (*Date*). También se crea una columna nueva, la duración del evento, a partir de las horas inicial y final (*Duration*). Como tenemos varios registros para cada día, se ha agrupado por la columna creada de fecha, utilizando la media para todos los valores. Se hace el tratamiento de outliers, utilizando gráficas de distribución (histogramas y boxplots) para elegir en cada caso los valores de corte.

CMEs.

Se integra el fichero y se hace una limpieza de caracteres extraños en algunos campos. Al tratar los NaN interpolamos para la mayoría de columnas, una vez más para perder los menos valores posibles, pero al tratar la columna *K_energy* no tenemos más remedio que eliminar dichos NaN, pues no podemos interpolar en nuestra columna objetivo, sería adulterar el resultado. Esta es posiblemente la ***decisión de diseño clave*** del TFM, pues al hacer esto perdemos un tercio de los registros de este dataframe. Se crea la consabida columna de fecha en formato datetime (*Date*) y se agrupa por ella tomando la media para todos los valores, puesto que tenemos más de un registro por fecha. A la hora de tratar los outliers nos basamos como en la ocasión anterior en un análisis gráfico y estadístico para tomar las decisiones. Como se ha comentado, discretizamos la variable *K_energy* creando otra, *Class*, que tomará valores de 1 a 6 en función de la intensidad creciente de *K_energy*.

Se integran los tres dataframes en uno utilizando la columna *Date*. Después de todas las depuraciones, tratamientos, agrupaciones etc nos quedamos con un fichero de tan sólo 4294 registros. No es buena noticia, aunque hemos intentado por todos los medios limitar las pérdidas son muy pocos registros.

4.2. Modelado.

Comenzamos por ver la correlación de las distintas variables entre sí y con la target, *Class*. *Central_PA*, *MPA* y *Acceleration* tienen una correlación muy baja con *Class*, menor que 0,1, por tanto las excluiremos de nuestro estudio. Se observa también que *2_I_speed* y *2_F_speed* están muy correlacionadas entre si (era esperable) por tanto nos quedamos solo con una de ellas, *2_I_speed,* y desechamos *2_F_speed*. Nuestras "features" finales serán por tanto *Spots*, *Duration*, *Peak_energy*, *Width*, *2_I_speed* y *2_20R_speed*.

A partir de aquí empezamos a ejecutar modelos, la técnica para cada uno será siempre la misma:

1. Crear la instancia y entrenar el modelo inicial
2. Comprobar métrica inicial en el conjunto de train
3. Ejecutar validación cruzada
4. "Tuning" del modelo mediante GridSearchCV o RandomizedSearchCV.
5. Entrenamiento del modelo con los parámetros óptimos obtenidos en el apartado anterior.
6. Comprobación de la métrica final del mejor modelo entrenado con los datos del conjunto de test. Ésta es la que tomaremos como referencia.

Respecto a la métrica, hemos elegido el indicador f1, pues es una combinación de precisión y recall, las dos métricas que nos interesan. También hemos usado el valor 'weighted' para *average*, para calcular las métricas para cada etiqueta balanceadas.

Se han probado los modelos siguientes:

- LogisticRegression
- K_Neighbors
- SVM
- XGBoost
- DecisionTree
- RandomForest

5.Resumen de resultados.

Lo que se observa es que, salvo el modelo de LogisticRegression, los restantes dan resultados similares, en la horquilla de 0.57 – 0.60. Posteriormente se ha ensamblado otro modelo utilizando los dos que mejores valores han dado para f1, SVM y RandomForest. El resultado ha sido mejor que los otros, por lo tanto nos quedamos con este último modelo ensamblado como mejor opción para nuestros datos (aunque de cara a la puesta en producción quizá optaríamos por SVM, más sencillo y casi con el mismo resultado). A continuación, reproducimos a modo de resumen las métricas obtenidas para cada modelo, así como el classification report del modelo ganador.

Modelo	**f1_score**
LogisticRegression	0.529176
K_Neighbors	0.572684
DecisionTree	0.573106
XGBoost	0.579969
RandomForest	0.602780
SVM	0.603464
Ensembled	0.612131

	precision	recall	f1-score	support
1	0.00	0.00	0.00	3
2	1.00	0.12	0.22	24
3	0.61	0.69	0.65	124
4	0.58	0.71	0.64	237
5	0.68	0.58	0.62	219
6	0.77	0.45	0.57	38
accuracy			0.62	645
macro avg	0.61	0.43	0.45	645
weighted avg	0.64	0.62	0.61	645

6.Conclusiones.

Los resultados del trabajo predictivo evidentemente no son buenos. Con medias de f1_score de 0.6 no se puede confiar en que el modelo nos "avise" con la suficiente precisión de una posible tormenta, sobre todo para las clases altas, 5 y 6, que son las que más nos interesaría predecir. Igualmente se observa la poca precisión en valores con menos registros.

Uno de los problemas, si no el principal, está en los datos. Hemos tenido que interpolar bastantes valores en algunas variables por no disponer de ellos, y lo principal, hemos desechado un tercio de nuestros datos de CMEs porque no teníamos valores para la variable target. El resultado es que al final nos hemos quedado con menos de 5000 registros para trabajar, y eso ha influido en los modelos. Otro problema ha sido la distribución de los datos, muchos de ellos claramente asimétricos, con distribuciones long-tail, que hemos intentado corregir con el tratamiento de outliers.

El resumen es que se ha hecho una primera aproximación al problema, está claro que hay que refinar los datos (contar con mejores y más abundantes registros) y nuestros procesos, ya que por ejemplo para simplificar apenas se han hecho suposiciones o acotaciones físicas, con lo cual nuestro trabajo al final queda lejos de la rigurosidad que se le presupone a cualquier estudio científico. No obstante, en la parte positiva, los modelos muestran que existen evidentes correlaciones entre las variables, lo que indica que estamos en el buen camino.

7.Anexo.

Repositorio con los datos y los notebooks: https://github.com/jmpinop/TFM

Instrucciones para el trabajo:

1. Clonar el repositorio al entorno de trabajo.
2. Todos los datos necesarios se encuentran en la carpeta *Datos*.
3. En la carpeta *Notebooks*, ejecutar *Data_preparing.ipynb*
4. También en *Notebooks*, ejecutar *Models.ipynb*.
5. Abrir visualizaciones (*Visualizations*) en formato presentación o Tableau Story.

Bibliografía:

- https://www.ngdc.noaa.gov/stp/solar/solar-indices.html
- http://www.wdcb.ru/stp/data.html
- https://www.sciencedirect.com/science/article/pii/S2090997712000235
- https://iopscience.iop.org/article/10.3847/1538-4357/aaebfc

Los ficheros de datos usados pueden encontrarse en:

- http://sidc.be/silso/datafiles#total
- https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/x-rays/goes/xrs/input-files/
- https://cdaw.gsfc.nasa.gov/CME_list/UNIVERSAL/text_ver/