NOMBRES: Benjamín Farías Valdés, Juan Hernández Sánchez

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2433 — Minería de Datos — 2' 2019

Entrega Propuesta de Proyecto

1. Introducción

A lo largo de la historia, los desastres naturales han significado una gran amenaza para las personas y los lugares donde habitan. Éstos muchas veces traen consigo una gran cantidad de muertes, daños y pérdidas de todo tipo, las que muchas veces terminan siendo irreparables. Es por esto que a través de los años el ser humano ha buscado mejorar las formas de predecir y estudiar estos eventos, para así poder minimizar la cantidad de muertes, daños en los ecosistemas y daños en las infraestructuras generados por estos.

Los ciclones son uno de los fenómenos por los que el ser humano se ve más frecuentemente amenazado, y a la vez corresponden a uno de los desastres naturales más destructivos que existen. Gracias a las mediciones que se han realizado a lo largo de varias décadas, existe una gran cantidad de información sobre las condiciones atmosféricas previas, concurrentes y posteriores a los ciclones. En la red se encuentran disponibles bases de datos que recopilan mediciones de presión, viento y tipo de ciclón (entre muchas otras características) [1], por lo que su análisis podría ser enriquecedor para la búsqueda de patrones que nos permitan identificar con antelación qué tipo de sistema se está formando en la atmósfera y si podría o no desencadenar un ciclón devastador.

Con el uso de algoritmos y modelos clasificadores buscaremos reconocer el estado de una posible formación atmosférica peligrosa dadas las mediciones necesarias que se obtengan en dicho momento y lugar.

Figure 1: Inundaciones Huracán Katrina

2. Marco Teórico

Para comenzar será necesario integrar distintas bases de datos sobre sistemas de ciclones, de manera de obtener un set de datos con una gran cantidad de información variada sobre estos eventos, lo que es muy importante al momento de querer evitar la obtención de un modelo sesgado. Esto se realizará mediante una búsqueda de datasets en Kaggle y Google Dataset Search (principalmente los subidos por la NOAA) [2]. Posterior a esto se deberá realizar el pre-procesamiento e integración de los datos, en donde se deberán realizar las siguientes acciones:

- Buscar como tratar con valores nulos (borrarlos o asignarles algún valor basado en algún criterio).
- En el caso de los valores no definidos utilizaremos modelos de regresión para estimar valores adecuados y utilizarlos [3].
- Normalizar los datos para que queden en la misma escala y representen mejor las distintas relaciones entre sus atributos.
- Aplicar PCA para reducir las dimensiones de las distintas bases de datos sin perder mucha información (en caso de que sea necesario).
- Generar una base de datos que integre todas las encontradas, de manera que se tenga un set de atributos estándar para poder alimentar a un modelo clasificador.

Para el clasificador se utilizarán modelos de SVM (Support Vector Machines), los que serán entrenados mediante el dataset obtenido en los pasos anteriores. Se eligió este clasificador debido a que no es tan común verlo utilizado en este tipo de situaciones, permitiendo así una posterior comparación con otros modelos más típicamente utilizados en la búsqueda de patrones climáticos (como la regresión [4]). Las SVM trabajan construyendo un hiperplano (o conjunto de ellos) a partir de los datos de entrada, que luego permite separar los datos en las distintas clases posibles con una gran precisión.

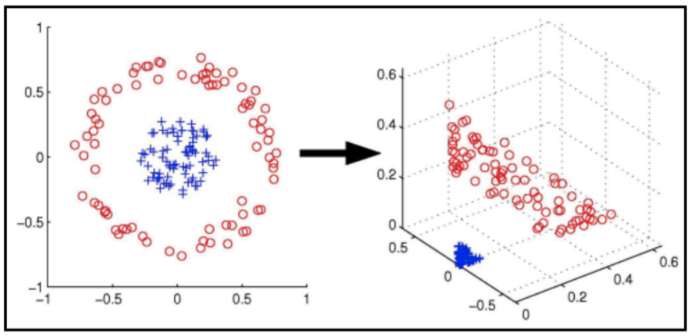

Figure 2: Support Vector Machines

Posteriormente se pasará por una etapa de validación del modelo, probándolo con una fracción del dataset NO utilizada para entrenarlo y así poder ajustar más sus parámetros. Finalmente se probará con un set de testing, determinando la confiabilidad del modelo a partir de distintas métricas, tales como 'Accuracy' [5].

A continuación se aplicará el modelo entrenado para clasificar nuevos eventos atmosféricos según su estado, permitiendo un posterior estudio de los resultados para tomar medidas de prevención ante potenciales catástrofes derivadas de dichos eventos. También se comparará con otros modelos utilizados para clasificar patrones climáticos, en busca de apartados en los que pueda ser de mayor utilidad que éstos.

Finalmente se visualizará en forma básica la información obtenida del modelo clasificador, permitiendo un mejor entendimiento inmediato de los resultados.

3. Conclusión

Una vez llevado a cabo lo mencionado en el marco teórico, se espera que seamos capaces de identificar el estado actual de un ciclón con una precisión aceptable para así reducir los daños de una posible catástrofe. Sin embargo, también se espera que el proceso de integración de las bases de datos se dificulte debido a la ausencia de un protocolo estándar para las mediciones realizadas por estaciones meteorológicas distintas, además de los diversos cambios que han sufrido los procesos de recolección de datos atmosféricos a lo largo de los años [6].

Por otra parte, una mala estimación de datos o una falta de relación entre variables puede llevarnos a un modelo poco confiable o sesgado, limitando el alcance de este proyecto, por lo que es de especial énfasis la fase inicial de recolección de datos y formación del dataset.

Finalmente, se podría desarrollar más a fondo este proyecto para ser capaz de no sólo clasificar los ciclones según estado, sino también predecir su trayectoria y vida dadas sus condiciones actuales.

4. Referencias

[1] https://www.kaggle.com/noaa/noaa-severe-weather-data-inventory
[2] https://www.noaa.gov/
[3] https://www.redalyc.org/jatsRepo/674/67452917005/html/index.html
[4]https://towardsdatascience.com/weather-forecasting-with-data-science-approaches-cb8f2afd3f38
[5] http://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/
[6] https://www.britannica.com/topic/classification-1703397