Escuela Superior de Informática

Sistema de prevención de accidentes de tráfico en condiciones meteorológicas adversas [Prototipo]

Andrés Montoro Montarroso

Pablo León Alcaide

Jaime Cepeda Villamayor

Pablo Ráez García-Retamero

Juan Garrido Arcos

27/10/2016

Curso: 4º Computación

ÍNDICE

1. **INTRODUCCIÓN**
2. **TAMAÑO DE LA BASE DE DATOS**
3. **SONDEO Y VIABILIDAD**
4. **PROCESO KDD**
    - **4.1. SELECCIÓN**
    - **4.2. PREPROCESO**
    - **4.3. TRANSFORMACIÓN**
5. **REPOSITORIO**
6. **BIBLIOGRAFÍA**

1. INTRODUCCIÓN

Queremos diseñar un sistema de prevención de accidentes de tráfico en caso de condiciones meteorológicas adversas en el territorio estadounidense. Dada la magnitud del proyecto completo nos vamos a centrar en un único estado representativo de este país que como explicaremos posteriormente, es un estado que nos sirve de muestra tipo para nuestro sistema ya que reúne unas condiciones y características que hemos considerado ideales para la búsqueda de patrones tormenta-accidentes para poder construir nuestro sistema. Este estado es Indiana.

Para construir este sistema de prevención vamos a utilizar el proceso KDD para extraer conocimiento de un conjunto útil de datos. Para ello como la propia definición indica necesitamos datos. Los datos que hemos conseguido se corresponden a un histórico de tormentas con sus detalles y localizaciones de todo Estados Unidos desde 1991 hasta 2016 y el caso de las bases de datos de accidentes desde 1975 hasta 2015. Analizando minuciosamente las distintas bases de datos, hemos decidido emplear el histórico de estas desde 2005 hasta 2015 ya que en las posteriores fases del proceso KDD hemos concluido que con un histórico de 10 años y una selección adecuada de los datos podemos extraer patrones significativos y representativos para la salida de nuestro sistema.

1.1. Entrada del sistema

La entrada al sistema será una predicción meteorológica en una zona geográfica concreta del estado.

1.2. Salida del sistema

La salida del sistema será la posibilidad que tiene el usuario de este de tener un accidente.

2. TAMAÑO DE LA BASE DE DATOS

Gracias a la descripción anterior podemos afirmar que nuestros datos cubren razonablemente el espacio de entrada ya que son perfectamente cotejables porque pertenecen al mismo territorio. En cuanto al espacio de salida estaría cubierto ya que tenemos

suficientes datos para conseguir nuestro objetivo en mayor o menor medida.

3. SONDEO Y VIABILIDAD

Para averiguar la viabilidad de nuestro sistema realizamos un sondeo que consistió en cruzar las bases de datos tanto de tormentas como de accidentes para establecer el número de accidentes que se habían producido en cada estado y así tener una visión del estado tipo adecuado para la resolución de nuestro problema. Los resultados fueron los siguientes:

- ALABAMA 639
- ALASKA (no matches)
- ARIZONA 319
- ARKANSAS 480
- CALIFORNIA 5536
- COLORADO 318
- CONNECTICUT 138
- DELAWARE 38
- DISTRICT OF COLUMBIA 11
- FLORIDA 1266
- GEORGIA 916
- HAWAII 32
- IDAHO 88
- ILLINOIS 1010
- INDIANA 725
- IOWA 520
- KANSAS 449
- KENTUCKY 615
- LOUISIANA 311

- MAINE 65
- MARYLANDS (error)
- MASSACHUSETTS 224
- MICHIGAN 613
- MINNESOTA 235
- MISSISSIPPI 412
- MISSOURI 989
- MONTANA 32
- NEBRASKA 199
- NEVADA 57
- NEW HAMPSHIRE 21
- NEW JERSEY 598
- NEW MEXICO 62
- NEW YORK 734
- NORTH CAROLINA 759
- NORTH DAKOTA 49
- OHIO 827
- OKLAHOMA 739
- OREGON 32
- PENNSYLVANIA 772
- PUERTO RICO (error)
- RHODE ISLAND 9
- SOUTH CAROLINA 770
- SOUTH DAKOTA 107
- TENNESSEE 706
- TEXAS 1852

- UTAH 39
- VERMONT 56
- VIRGINIA 604
- VIRGIN ISLANDS (from 2004)(error)
- WASHINGTON 26
- WEST VIRGINIA 147
- WISCONSIN 244
- WYOMING 48

Una vez obtuvimos estos datos pudimos hacernos una idea de los accidentes producidos por tormentas en los últimos 10 años.

A la hora de elegir el estado tipo podría parecer obvio coger el estado con más accidentes por tormentas, pero para la elección del estado tipo tuvimos en cuenta también otros factores. Nos dimos cuenta que en estados como Florida la cantidad de accidentes por tormentas se concentraba en su gran mayoría en Miami, con esto queremos explicar que descartamos los estados con grandes ciudades con un gran número de habitantes porque no era una muestra representativa. Eso nos llevó a seleccionar el estado de Alabama, ya que tenía un número de accidentes considerable y la población era más homogénea y estaba distribuida por todo el estado. Pero luego observando los tipos de tormentas que había en este estado nos dimos cuenta que en los estados del sur aparecían pocos casos, por ejemplo de nevadas, y no los podemos considerar como una muestra adecuada ya que necesitábamos un estado que aparte de reunir las características ya mencionadas, necesitábamos una variedad de condiciones meteorológicas adversas que no nos podían brindar los estados del sur. Por tanto, procedimos a su descarte y de los estados del norte que cumplían todas las características que nosotros consideramos importantes encontramos el estado de Indiana y decidimos que sería nuestro estado tipo para desarrollar el prototipo de sistema de prevención de accidentes en caso de condiciones meteorológicas adversas.

4. PROCESO KDD

4.1. Selección

4.1.1. Base de datos de accidentes

- State: Este elemento identifica el estado en el que el accidente ocurrió. Los códigos son extraídos de la publicación de los GLC (Geographic Location Codes) del GSA (General Services Administration).
- YEAR: Este elemento indica el año en el que el accidente ocurrió.
- Month: Este elemento indica el mes en el que el accidente ocurrió.
- Day: Este elemento indica el día en el que el accidente ocurrió.
- Hour: Este elemento indica la hora en el que el accidente ocurrió.
- Minute: Este elemento indica el minuto en el que el accidente ocurrió.
- Ve-forms: Este elemento indica el número de vehículos en movimiento involucrados en el accidente. Por tanto, vehículos aparcados legalmente no son incluidos.
- Persons: Número de formularios enviados por personas en vehículos de motor.
- Route: Este elemento indica la señalización de la ruta de la vía de tráfico en la que el accidente ocurrió.
- M_harm: Este elemento describe el evento que resultó en la lesión más severa o, en caso de no haber ninguna lesión, el mayor daño a la propiedad involucrando ese vehículo.
- Rel_juuc RELJCT2: Este elemento identifica la localización del choque con respecto a la presencia o proximidad a componentes típicamente en cruces. La codificación de este elemento está hecha en dos sub-

campos y está basada en la localización del primer evento perjudicial del choque.

- Sp_limit: Este elemento identifica el atributo que mejor representa el límite de velocidad justo antes del evento crítico previo al choque, basado en las pruebas del caso.
- ALIGNMNT: Este elemento identifica el atributo que mejor representa la alineación de la vía previa al evento crítico previo al choque, basado en las pruebas del caso.
- LGT_COND: Este elemento recoge el tipo o nivel de luz que había en el momento del choque.
- WEATHER:Este elemento recoge las condiciones atmosféricas predominantes dadas en el momento del choque.
- CF1 CF2 CF3: Estos elementos guardan factores relacionados con el choque expressados por el investigador oficial.
- FATALS: Este elemento guarda el número de personas fatalmente perjudicadas en el choque.
- DRUNK_DR: Este elemento refleja el número de conductores borrachos involucrados en el choque.
- MILEPT: Este elemento guarda el kilómetro más cercano al punto en el que el choque ocurrió.
- TWAY_ID or TWAY_ID2: Este elemento guarda la vía en la que ocurrió el choque.
- LATITUDE: Este elemento indica la latitud a la que se provocó el choque.
- LONGITUD: Este elemento indica la longitud a la que se provocó el choque.

4.1.2. Base de datos tormentas (locations)

- EEVENT_ID: ID asignada por NWS para indicar una parte específica dada en un episodio tormentoso; liga

el episodio entre tres archivos descargados de la página web de SPC.

- EPISODE_ID: ID asignada por NWS para indicar un episodio tormentoso; liga el episodio con la información en los archivos de los detalles de eventos. Un episodio puede tener varios eventos diferentes.
- RANGE: Un evento meteorológico será referenciado, como mínimo, al décimo de milla más cercano a su centro geográfico (no desde los límites de la población, aeropuerto, o lago interior).
- LATITUDE: La latitud donde el evento comienza (Redondeado a las centésimas en grados decimales; incluye un '-' si está al sur del ecuador).
- LONGITUDE: La longitud donde el evento comienza (Redondeado a las centésimas en grados decimales; incluye un '-' si está al Oeste del meridiano).
- LAT2: La latitud donde el evento termina (Redondeado a las centésimas en grados decimales; incluye un '-' si está al sur del ecuador).
- LON2: La longitud donde el evento termina (Redondeado a las centésimas en grados decimales; incluye un '-' si está al Oeste del meridiano).

4.1.3. Base de datos de tormentas (details)

- EVENT_ID: ID asignada por NWS para indicar una parte específica dada en un episodio tormentoso; liga el episodio entre tres archivos descargados de la página web de SPC.
- EPISODE_ID: ID asignada por NWS para indicar un episodio tormentoso; liga el episodio con la información en los archivos de los detalles de eventos. Un episodio puede tener varios eventos diferentes.
- EVENT_TYPE: El tipo elegido para el evento debe ser aquel que describa más precisamente dicho evento

basándose en bajas, lesiones, daño, etc. A pesar de esto, eventos significativos, como tornados, que no causen daño, deben ser también incluidos.

- BEGIN_DATE_TIME: Hora a la que el episodio comenzó. En formato de 24 horas.
- END_DATE_TIME: Hora a la que el episodio terminó. En formato de 24 horas.
- INJURIES_DIRECT: Número de lesiones directamente relacionadas con el evento.
- INJURIES_INDIRECT: Número de lesiones indirectamente relacionadas con el evento.
- DEATHS_DIRECT: Número de muertes directamente relacionadas con el evento.
- DEATHS_INDIRECT: Número de muertes indirectamente relacionadas con el evento.
- DAMAGE_PROPERTY: Suma estimada de daño a la propiedad producida por el evento.
- DAMAGE_CROPS: Suma estimada de daño a los campos producida por el evento.
- MAGNITUDE: Medida en el grado del tipo de magnitud (Usado con velocidades de viento y tamaño del granizo).
- TOR_F_SCALE: La escala mejorada Fujita describe la fuerza del tornado basado en la cantidad y tipo de daño causado por el tornado. La escala Fujita de daño variará en el área de daño; por tanto, el mayor valor será guardado para cada evento.
- BEGIN_LAT: La latitud donde el evento comienza (Redondeado a las centésimas en grados decimales; incluye un '-' si está al sur del ecuador).
- BEGIN_LON: La longitud donde el evento comienza (Redondeado a las centésimas en grados decimales; incluye un '-' si está al Oeste del meridiano).

- END_LAT: La latitud donde el evento termina (Redondeado a las centésimas en grados decimales; incluye un '-' si está al sur del ecuador).
- END_LON: La longitud donde el evento termina (Redondeado a las centésimas en grados decimales; incluye un '-' si está al Oeste del meridiano).

4.2. Preproceso

4.2.1. Base de accidentes

En este caso, lo que se ha realizado es una limpieza de los accidentes. Ya que en los accidentes los datos de fecha se encontraban en columnas distintas (DAY, MONTH, etc.) lo que se ha hecho es transformarlas a un formato de fecha válido e introducirlas en una columna nueva. Una vez hecho eso, se han eliminado las columnas que no servían. Hemos considerado que no servían tablas que en principio no están nada relacionadas con el hecho de que no haya nada relacionado con la meteorología o con la ubicación del accidente. Esto es debido a que, para poder relacionar ambas partes, se ha de encontrar en el mismo espacio y en la misma relación de tiempo (ya que por ejemplo no tendría sentido relacionar un accidente sucedido en 1990 con una tormenta que ocurrió en la otra parte del mundo en 2002). También hemos dejado datos interesantes de cara al futuro como por ejemplo la cantidad de heridos ya que hemos intentado valorar de forma objetiva que la prioridad donde se debe enfocar el sistema es donde más víctimas se hayan producido para así poder evitarlas en un futuro o centrarse en los focos más críticos para que al menos los daños sean menores.

Después, se han eliminado aquellos accidentes en los cuales la latitud y la longitud estaban mal introducidos. (se salían de la latitud y longitud del estado). Aquí estamos eliminando los datos corruptos porque solo consiguen distorsionar la propia gráfica ya que son datos incorrectos y pueden darse falsos positivos o incluso casos contrarios en los que se podría en un futuro focalizar hacia objetivos que realmente no tienen ningún tipo de riesgo.

4.2.2. Base de datos de tormentas

Por un lado, lo que se ha hecho es una limpieza de la latitud y la longitud porque estaban mal introducidos los datos. Al igual que en la base de datos de accidentes, no nos aporta nada tener datos totalmente corruptos para poder abstraer conocimiento por el hecho de que pueden producirse falsos positivos de riesgo. Para ello, simplemente había casos en los que la coma estaba mal introducida por lo que se ha intentado corregir los que solo tenían este único fallo y hemos descartado los que el número estaba directamente mal insertado.

Después, se han eliminado aquellos eventos a los que no estaban asociados a una tormenta ya que en la base de datos tenemos mucha información y no todas las filas nos resultan útiles porque no son tormentas o no de la categoría que nosotros estamos analizando.

4.3. Transformación

En esta fase, hemos preparado la tabla de tormentas para la fase de minería del proyecto. Lo que hemos hecho es asociar el tipo de evento a una nueva tabla:

- 'Heavy Rain': LLUVIA de nivel 1. Corresponde a las tormentas que han desembocado en lluvia fuerte.
- 'Flood': LLUVIA de nivel 2. Corresponde a las tormentas que han desembocado en inundación.
- 'Flash Flood': LLUVIA de nivel 3. Corresponde a las tormentas que ha desembocado en inundación rápida.
- 'Strong wind': VIENTO de nivel 1. Corresponde a las tormentas que han provocado un viento fuerte.
- 'High wind': VIENTO de nivel 2. Corresponde a las tormentas que han provocado un viento mas fuerte que el del nivel 1.
- 'Thunderstorm wind' VIENTO de nivel 3. Corresponde a un viento tormentoso.
- 'Tornado': TORNADO, con niveles que corresponden al nivel asignado en la tabla de tormentas.

- ‘Hail’: GRANIZO.
- ‘Winter Storm’: Nieve de nivel 1. Corresponde a las tormentas de nieve.
- ‘Heavy Snow’: NIEVE de nivel 2. Corresponde a las fuertes nevadas.
- ‘Lightning’: T. ELECTRICA de nivel 1. Corresponde a los relámpagos.

Los casos de NIEVE, LLUVIA, TORNADO, VIENTO, T.ELECTRICA y GRANIZO los añadimos a nuevas columnas de la tabla, y prepararlos para la fase de minería.

5. REPOSITORIO

En el siguiente enlace se puede observar todo el proceso KDD hasta la etapa de transformación incluyendo los algoritmos de clustering y aprendizaje automático empleados.

https://github.com/PabloLeon23/proyecto_mineria

6. BIBLIOGRAFÍA

ftp://ftp.nhtsa.dot.gov/fars/

http://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/

https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812315