Reglas de asociación

Raúl Carlomagno, Franco Catania

**Resumen.** El presente informe consiste en la obtencion de reglas de asociación mediante el algoritmo a priori a un dataset que detalla ataques de tiburones. El dataset es analizado y procesado para obtener reglas fuertes luego de aplicarse el algoritmo a priori. La intención es poseer un conocimiento de cómo inciden las 3 métricas (support, confidence y lift) en el mineado de reglas de asociacion.

Introducción

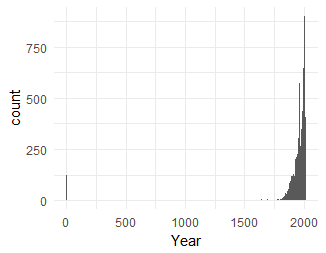
La Minería de Datos refiere al proceso de extracción no trivial de información implícita, útil y previamente desconocida, desde los datos de una base de datos. Dentro de ella, la minería de datos, en su estrategia de descubrimiento de reglas de asociación, propone encontrar conjuntos de elementos que co-ocurren juntos frecuentemente en una base de datos. La aplicación de esta tecnología, también denominada Minería de Reglas de Asociación (MRA), puede generar una gran cantidad de reglas, por lo que surge la necesidad de seleccionar aquellas reglas de asociación relevantes desde la perspectiva del contexto.  
Una regla de asociación consta de un conjunto de ítems, el cuerpo de regla (lhs), que conducen a otro ítem, la cabecera de regla (rhs). La regla de asociación relaciona el cuerpo de regla (lhs) con la cabecera de regla (rhs). Una regla de asociación puede contener las siguientes características: soporte (support), fiabilidad (confidence) e importancia (lift).  
  
Soporte: una transacción da soporte a una regla de asociación si la transacción contiene el cuerpo de regla y la cabecera de regla. El soporte de regla es la proporción de transacciones que dan soporte a la regla de asociación con respecto al número total de transacciones de la base de datos de transacciones.  
Confianza: la confianza de una regla de asociación es su firmeza o fiabilidad. La confianza se define como el porcentaje de transacciones que dan soporte a la regla con respecto a todas las transacciones que dan soporte al cuerpo de regla. Una transacción da soporte al cuerpo de regla si contiene todos los ítems del cuerpo de regla.  
Importancia: la importancia de una regla viene dada por el factor por el cual la confianza sobrepasa la confianza esperada

Materiales y Métodos

El dataset obtenido disponía de la siguiente estructura:

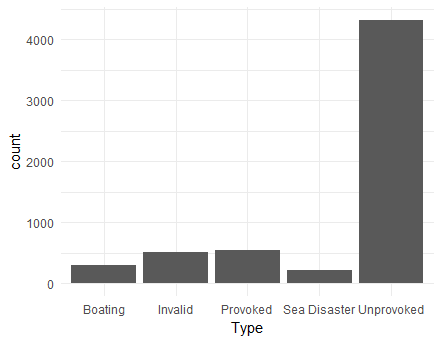
|  |  |  |
| --- | --- | --- |
| Case Number | Texto | Numero del caso, construido a partir de la fecha de ocurrencia del hecho |
| Date | Texto | Fecha de ocurrencia del hecho |
| Year | Numerico Entero | Año de ocurrencia del hecho |
| Type | Categorico (5) | Tipo de hecho del accidente |
| Country | Texto | Pais de ocurrencia del hecho |
| Area | Texto | Area del país de ocurrencia del hecho |
| Location | Texto | Lugar dentro del área de ocurrencia del hecho |
| Activity | Texto | Actividad realizada en el momento del hecho |
| Name | Texto | Nombre de la persona afectada |
| Sex | Categorica (5) | Sexo de la persona afectada |
| Age | Texto | Edad discretizada de la persona afectada |
| Injury | Texto | Lesión resultante del ataque |
| Fatal(Y/N) | Categorica (4) | Si el ataque resultó en muerte |
| Time | Texto | Horario del ataque |
| Species | Texto | Especie de tiburón involucrada |
| Investigator or Source | Texto | Fuente que registró el ataque |
| pdf | Texto | Nombre del archivo para descargar información del ataque |
| href | Texto | URL del archivo de ataque para descargar |
| original order | Numerico entero | Id del ataque |

La única variable numérica es la variable Year, que sin realizar ningún tipo de transformación tiene la siguiente distribución:



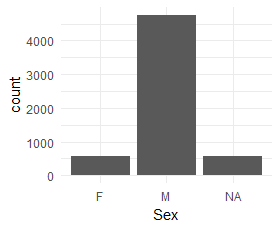
Para el preprocesamiento se trabajará sobre sus valores extremos y el valor cero como se ve en el gráfico

La variable categorica Type está distribuida de la siguiente manera:



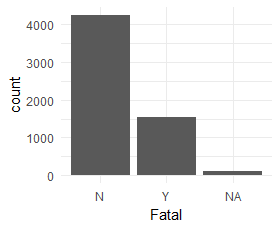
Claramente se observa que la amplia mayoría de los accidentes fueron no provocados.

La siguiente variable categórica Sex, poseía valores que no eran ni M ni F, esos valores fueron sustituidos por NA:



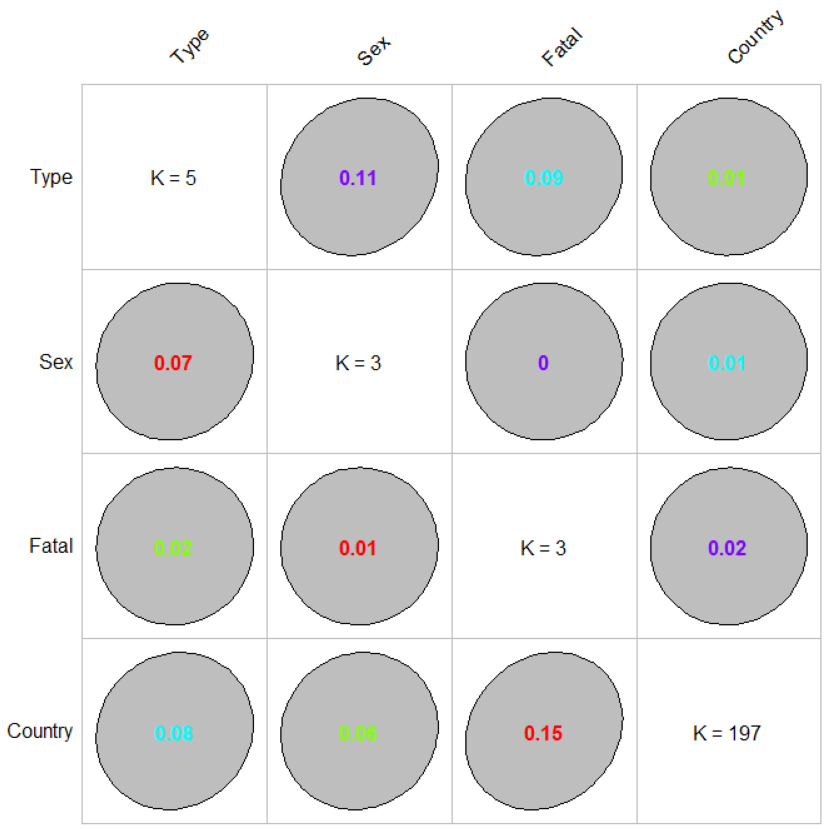
Nuevamente, se observa un amplio predominio, en este caso, el sexo masculino en la mayoría de los accidentes reportados.

La variable Fatal poseía valores que no eran ni Y ni N, esos valores, fueron imputados a NA para no ser tenidos en cuenta al elaborar las reglas, pero en este caso, esos valores son ínfimos:



El análisis de otras variables que podrían ser categóricas, pero son texto, no tiene sentido alguno, ya que por ejemplo, la variable Country posee 174 valores, muy alta cardinalidad para analizarla y graficarla.

Al disponer solamente de una variable numérica, no podemos analizar su correlación con otra variable numérica, pero si, lo que se realizó, fue analizar la asociación de las variables categóricas mediante los estadísticos de Goodman-Kruskal, obteniendo la métrica tau (τ) (también, se podría haber utilizado el test de chi cuadrado):



Cuando el valor tau entre 2 variables es 1, significa que hay una asociación alta entre esas variables, caso contrario, el valor 0 indica la nula asociación entre esas variables, el valor K de la diagonal, denota la cantidad de valores de esa variable. Cabe mencionar que el valor tau es bidireccional, por ejemplo, Sex explicada por Type tiene un valor de 0.11, pero Type explicada por Sex tiene un valor de 0.07. Claramente se puede observar que no hay asociación entre las variables categóricas, ninguna de las 4 variables es explicada por ninguna otra variable.

**Preprocesamiento**:

Las siguientes variables fueron eliminadas al final del proceso, ya que no aportaban al análisis, poseían un alto nivel de cardinalidad o eran texto puro:

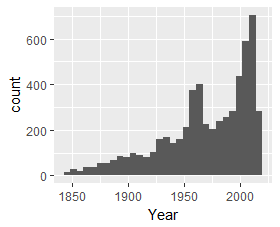
**Href, Href\_formula, pdf, investigator or source, name, injury, location, area, activity, case number, date, original order**

**Sex**: se reemplazó con NA los valores que no eran ni M ni F.

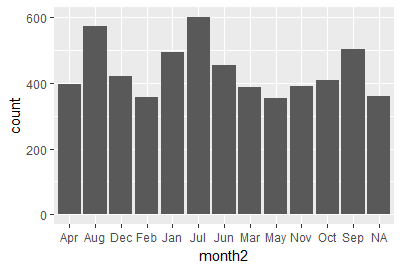
**Fatal**: se reemplazaron los valores que no eran ni Y ni N por NA.

**Country**: no se realizó ningún proceso especial

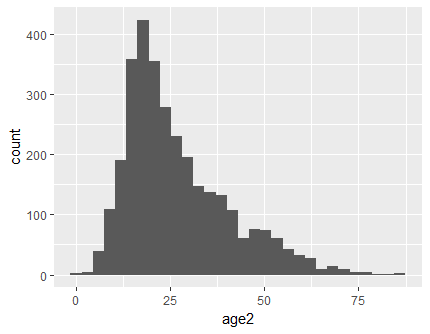
**Year**: solo nos quedamos con los registros donde Year era mayor a 1845, y no nos basamos en la variable Year original, sino que se creó una nueva Year donde se obtuvo el año a partir del parseo de case number, ya que se notaron algunas diferencias y el año dentro de case number era mas acertado. La distribución de Year, luego del proceso, fue la siguiente:



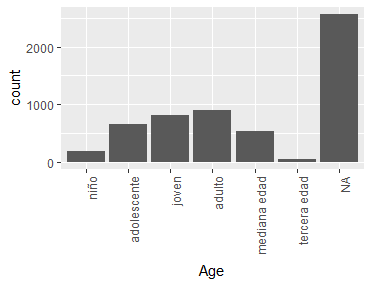
**Month**: se creó una variable month, a partir del parseo del mes en la variable case number, en los casos donde el mes no estaba disponible se imputó ese valor a NA:



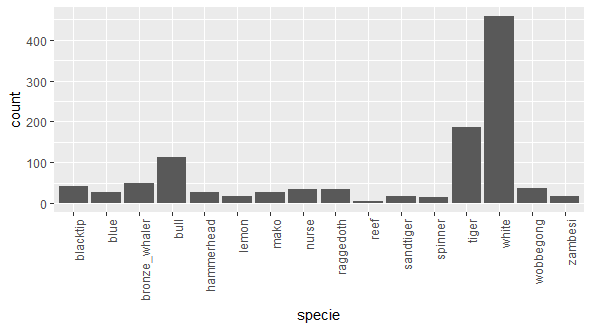
**Age**: la variable age original disponía no solo de valores numéricos, sino de otras pseudo discretizaciones, años de nacimiento, etc. Solo el 55% de los valores eran numéricos. Primero solo se extrajeron los valores numéricos, quedándonos un 45% de valores NA. Se intentó reemplazar los missing values por la media o mediana, pero dado el alto porcentaje de missing values, la distribución era altamente distorsionada. La distribución sin reemplazo era la siguiente:



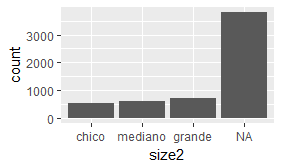
Distribución “tipica” de edad. Pero necesitábamos discretizar esta variable para poder ser utilizada como variable para el tipo de clase transactions para la minería de reglas de asociación, por ende, se realizó una discretizacion fija especificando los niveles según rangos etareos (niño [0 a 12), adolescente [13 a 18), joven [18 a 25), adulto [25 a 40), mediana edad [40 a 65) y tercera edad [65 a inf)), la discretizacion respetó lo mas posible a la distribución original. En los casos donde no se pudo calcular la edad, se asignó NA, la distribución final fue la siguiente:



**Specie**: a partir de la variable Species original se realizó un proceso muy simple de textmining para poder extraer las principales especies de tiburones. Se pasó a minúsculas el campo Species y se fueron removiendo números, stopwords (shark era una importante stopword en este caso), puntuaciones, etc. del campo, luego se extrajeron todas las palabras y se las ordenó por frecuencia, llegando a obtener 19 especies de tiburones. Luego se buscó en la variable Species original pero curada, las distintas especies de tiburones para formar la variable categórica specie donde predomina el tiburón blanco, pero con 80% de valores NA que no se muestran en el siguiente gráfico para no distorsionar a escala:



**Size**: La variable Species original disponía de los tamaños de tiburones, a veces en pies, otras veces en metros, se realizaron búsquedas con expresiones regulares dentro de la variable para extraer los tamaños en 2 formas, una por pies y otra por metros según corresponda, luego, se convirtió la de pies a metros y se creó una nueva variable unificada por metro. A esta nueva variable se le aplicó una discretizacion por kmeans y siempre discretizaba en 3 partes, paso siguiente, teniendo en cuenta los bins que asignaba kmeans, se discretizó de forma fija en 3 valores (chico [0 a 1.8m), mediano [1.8 a 3m) y grande [3m a inf)), quedando finalmente la siguiente distribución, donde obviamente tenemos gran cantidad de tamaños que no tenían valores numéricos:



Resultados obtenidos al aplicar las reglas de asociación

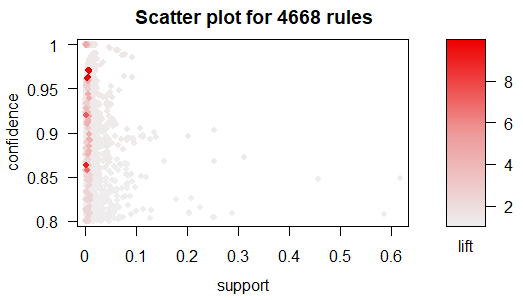
Una vez tenida la estructura preparada se procedió a crear las transacciones a partir del dataset. Todas las variables eran del tipo categóricas, por lo que la creación de las transacciones finalizó con 5685 transacciones y 396 columnas, ya que se crea una columna binaria por cada valor de la variable categórica, y en este caso, los países eran demasiados.

Al minear diferentes reglas se tuvo que analizar los parámetros para el mineado de las mismas, dado que se buscaban reglas que aporten conocimiento sobre la causa, las reglas obtenidas fueron siempre ordenadas por lift que sería la importancia de la regla, un lift igual a 1 indicaria que no hay efecto entre los elementos o conjuntos de elementos de una regla, como si fueran independientes. Lift mayor a 1 indicaria un efecto positivo de un elemento o conjunto de elementos sobre el otro elemento o conjunto de elemento; y lift menor a 1 indicaria un efecto negativo entre los elementos o conjunto de elementos, como si fueran excluyentes.

**Analisis general con soporte de 0.6%**:

* Los ataques de la especie blacktip, lemon, hammerhead, nurse, raggedoth y blacktip no fueron fatales (confidence 1 y lift 1.37)
* Los ataques provocados en sudafrica, no fueron fatales (confidence 1 y lift 1.37)
* La especie raggedoth ataca en sudafrica (confidence 0.97 y lift 9.94)
* La especie wobbegong ataca en australia (confidence 0.91 y lift 4.25)
* En USA, si atacó la especie nurse fue por una provocación (confidence 0.86 y lift 9.10)
* En USA si un accidente fue fatal del tiburón blanco, su tamaño era grande (confidence 0.85 y lift 6.79)
* La especie spinner ataca en USA (confidence 1 y lift 2.79)

El siguiente gráfico muestra las reglas generadas con el soporte de 0.6%, se pueden observar pocas reglas con lift alto y muestra como dada una regla su soporte tiende a ser muy bajo y su confidence alta para que su lift sea alto.



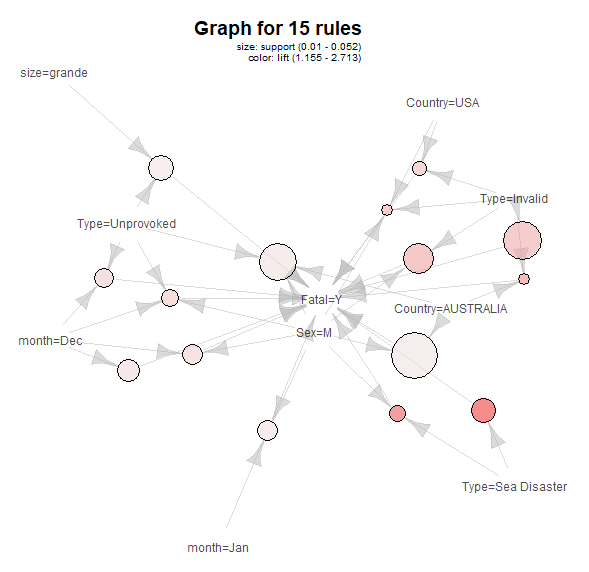
**Analisis de accidentes fatales**:

Se forzó la obtención de las reglas para que el consecuente sea fatal, ninguna de las reglas obtenidas superó el 70% de confidence y el soporte es muy bajo, salvo reglas triviales como si es hombre, el accidente fue fatal.

Del respectivo análisis se obtuvieron las siguientes condiciones para que un accidente sea fatal:

* Si el accidente fue en el océano (support 0.02, confidence 0.68 y lift 2.71)
* Si ocurrió en Australia (support 0.05, confidence 0.26 y lift 1.04)
* Si ocurrió en Sudafrica (support 0.022, confidence 0.23 y lift 0.92)
* Si la especie es blanca (support 0.017, confidence 0.21 y lift 0.86)

El siguiente grafo de las reglas nos permite observar las 15 principales reglas encontradas ordenadas por lift, pero de una manera mucho mas intuititiva. El tamaño de los circulos indica el soporte y la intensidad de color, indica mayor lift. Puede observarse como los accidentes del oceano tienen un gran lift asi como tambien el tamaño del nodo relacionado con Australia, donde se hayan gran cantidad de accidentes fatales.



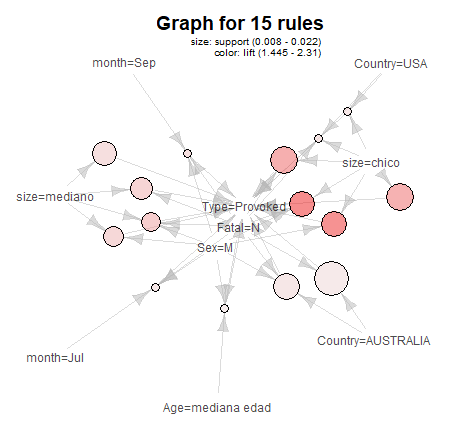
**Análisis de tamaño de tiburones**:

Mineando por características de tiburón se encontró:

* Si la especie fue mako de tamaño grande, la actividad era de navegación (support 0.001, confidence 0.85, lift 16.35)
* En USA, si el accidente fue fatal y la especie blanco, el tamaño del tiburón fue grande (support 0.002, confidence 0.85, lift 7.29). Esta regla respalda a la regla de los tiburones blancos del análisis de fatalidad de un punto anterior.
* Si la especie del ataque fue tiger, su tamaño fue grande (support 0.01, confidence 0.34, lift 2.72)
* Si el país es USA y la especie de ataque fue nurse, su tamaño era chico (support 0.001, confidence 0.85, lift 8.92)

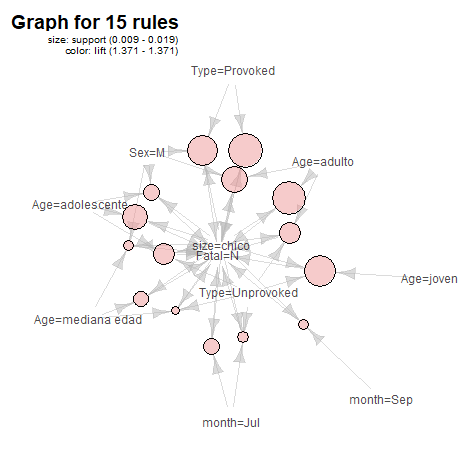
**Análisis de accidentes provocados**:

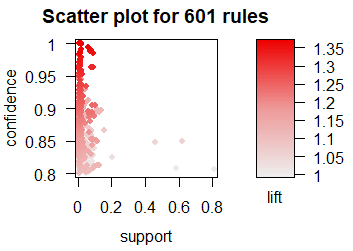
Realizando énfasis en el consecuente como un accidente provocado, se puede observar en el siguiente grafico de las primeras 15 reglas ordenadas por lift, de la incidencia del tamaño del tiburón; en mayor medida, los de tamaño chico y segunda medida, los de tamaño mediano. Esto puede deberse a subestimar a un tiburón dado su tamaño y pretender jugar con él o provocarlo.



**Análisis fuera de USA, Australia y Sudáfrica**:

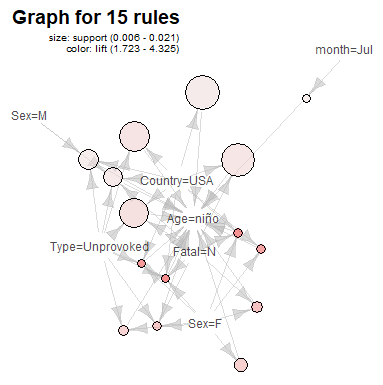
Excluyendo estos 3 países, se puede observar que el tamaño de los tiburones en gran parte de los ataques es chico. ¿Podria confirmarse que en estos 3 países se encuentran los tiburones mas grandes? Las reglas generadas tienen un lift mayor a 1 en la mayoría de los casos.





**Análisis de ataques a niños:**

Con un soporte de 0.5% se generaron reglas que tengan como consecuente a un niño y se detectó que en las principales reglas ordenadas por lift, las víctimas eran de sexo femenino. No especificamos las reglas porque no aportan conocimiento específico, pero en la gran mayoría estaba involucrado el sexo femenino.



Referencias

Hryniewicz, O. (2006). Goodman–Kruskal γ (tau) measure of dependence for fuzzy ordered categorical data. *Computational Statistics & Data Analysis*, *51*(1), 323-334.