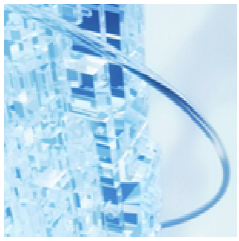


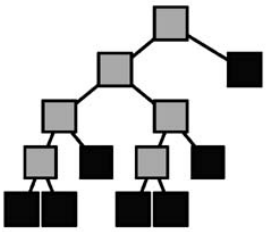
Prácticas de Minería de Datos

Grado en Ingeniería Informática

Curso 2013-14

PRÁCTICA 8

Reglas de Asociación



OBJETIVO

- Aprender a manejar un algoritmo que descubre reglas de asociación.
- Saber interpretar los resultados y manipular los datos para descubrir reglas interesantes.

1. Introducción

Weka incluye distintos métodos orientados a buscar asociaciones entre los datos, los cuales se encuentran en la pestaña “Associate”. Los algoritmos de *asociación* permiten la búsqueda automática de reglas que relacionan conjuntos de atributos entre sí. Son algoritmos no supervisados, en el sentido de que no existen relaciones conocidas a priori con las que contrastar la validez de los resultados, sino que se evalúa si esas reglas son estadísticamente significativas.

Los algoritmos de *asociación* son una tarea descriptiva que tiene como objetivo identificar relaciones no explícitas entre atributos categóricos (nominales). Las reglas de asociación expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos (expresan las combinaciones de valores de los atributos que suceden más frecuentemente).

Son utilizados cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones entre los datos.

Surgieron inicialmente para afrontar el análisis de la cesta de la compra de los comercios (*Market Basket Analysis*) para identificar productos que son frecuentemente comprados juntos, siendo muy útil la información obtenida para ajustar inventarios, organizar físicamente las estanterías en las tiendas o en campañas publicitarias.

Debido a sus características, estas técnicas tienen una gran aplicación práctica en muchos campos, aparte del comercial (interesante para comprender los hábitos de compra de los clientes). Así en el entorno sanitario estas herramientas se emplean para identificar factores de riesgo en la aparición o complicación de enfermedades.

Es importante señalar que estos métodos **sólo funcionan con datos nominales**, no con numéricos.

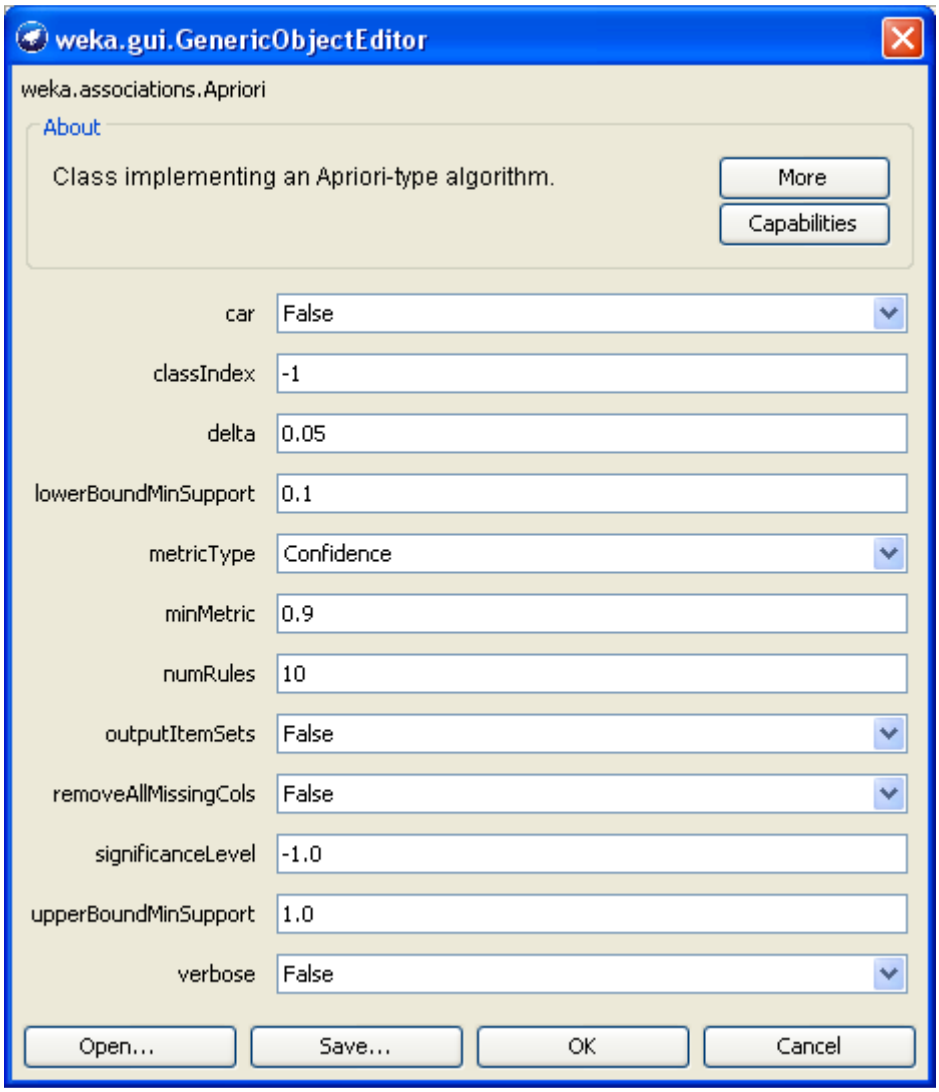
Es sin duda el apartado más sencillo y más simple de utilizar, carente de apenas opciones, en el que basta con seleccionar un método, configurarlo y verlo funcionar.

El principal algoritmo de asociación implementado en WEKA es el algoritmo *Apriori*, aunque también tiene implementado algunas variantes (*PredictiveApriori* y *Tertius*). **El algoritmo *Apriori* únicamente puede buscar reglas entre atributos simbólicos**, razón por la que se requiere haber **discretizado** todos los atributos numéricos (en caso que los hubiera).

Apriori es un algoritmo de aprendizaje de reglas de asociación muy simple y popular que se basa en la búsqueda de conjuntos de ítems con una determinada cobertura. Para evaluar las reglas se emplea la medida de soporte o cobertura (número de instancias del conjunto de datos a los que la regla se aplica) y la confianza o precisión (porcentaje de veces que la regla predice correctamente -se cumple cuando se puede aplicar-, es decir, cociente entre el número de ejemplos que contienen el antecedente y consecuente de la regla y el número de ejemplos que contienen el antecedente, siendo el consecuente el mismo u otro distinto).

Las reglas que interesan son aquellas que tienen un valor de soporte muy alto, por lo que se buscan reglas que cubran una gran cantidad de ejemplos.

La figura siguiente muestra los parámetros que se pueden configurar para el algoritmo *Apriori*.



Explicación de algunos de ellos:

- El parámetro "*car*" indica si se desea obtener reglas de clasificación (es decir, reglas en las que el consecuente siempre haga referencia a un determinado atributo que nosotros indiquemos) o reglas de asociación (reglas en las que el consecuente puede ser cualquier atributo o atributos). Para el primer caso, la opción "*classIndex*" sirve para indicar el atributo que hará de clase (obligamos que las reglas de asociación que se generen tengan en el consecuente el atributo indicado en "*classIndex*")
- Con la opción "*upperBoundMinSupport*" indicamos el límite superior de cobertura requerido (por defecto 1.0 que es el 100%) para aceptar un conjunto de ítems (los *itemsets* con un soporte mayor no son tenidos en cuenta). Si no se encuentran conjuntos de ítems suficientes para generar las reglas requeridas se va disminuyendo el límite (en tantas unidades como las indicadas en "*delta*") hasta llegar al límite inferior (opción "*lowerBoundMinSupport*") que por defecto es del 0.1 o 10%.
- Con la opción "*minMetric*" indicamos la confianza mínima (u otras métricas dependiendo del criterio de ordenación seleccionado en "*metricType*") para mostrar una regla de asociación.
- Con la opción "*numRules*" indicamos el número de reglas que deseamos que aparezcan en pantalla. La ordenación de estas reglas en pantalla puede configurarse mediante la opción "*metricType*". Algunas opciones que se pueden utilizar son: confianza de la regla, **lift** (confianza dividido por la proporción de ejemplos cubiertos por el consecuente de la regla), etc.

2. Ejemplo

Para ilustrar su funcionamiento en WEKA y a modo de ejemplo, estudiaremos algunos datos del hundimiento del Titanic. Los datos se encuentran en el fichero "titanic.arff" y corresponden a características de los 2.201 pasajeros del Titanic. Estos datos son reales y se pueden obtener de: "Report on the Loss of the 'Titanic' (S.S.)" (1990), British Board of Trade Inquiry Report (reprint), Gloucester, UK. Allan Sutton Publishing. (http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html)

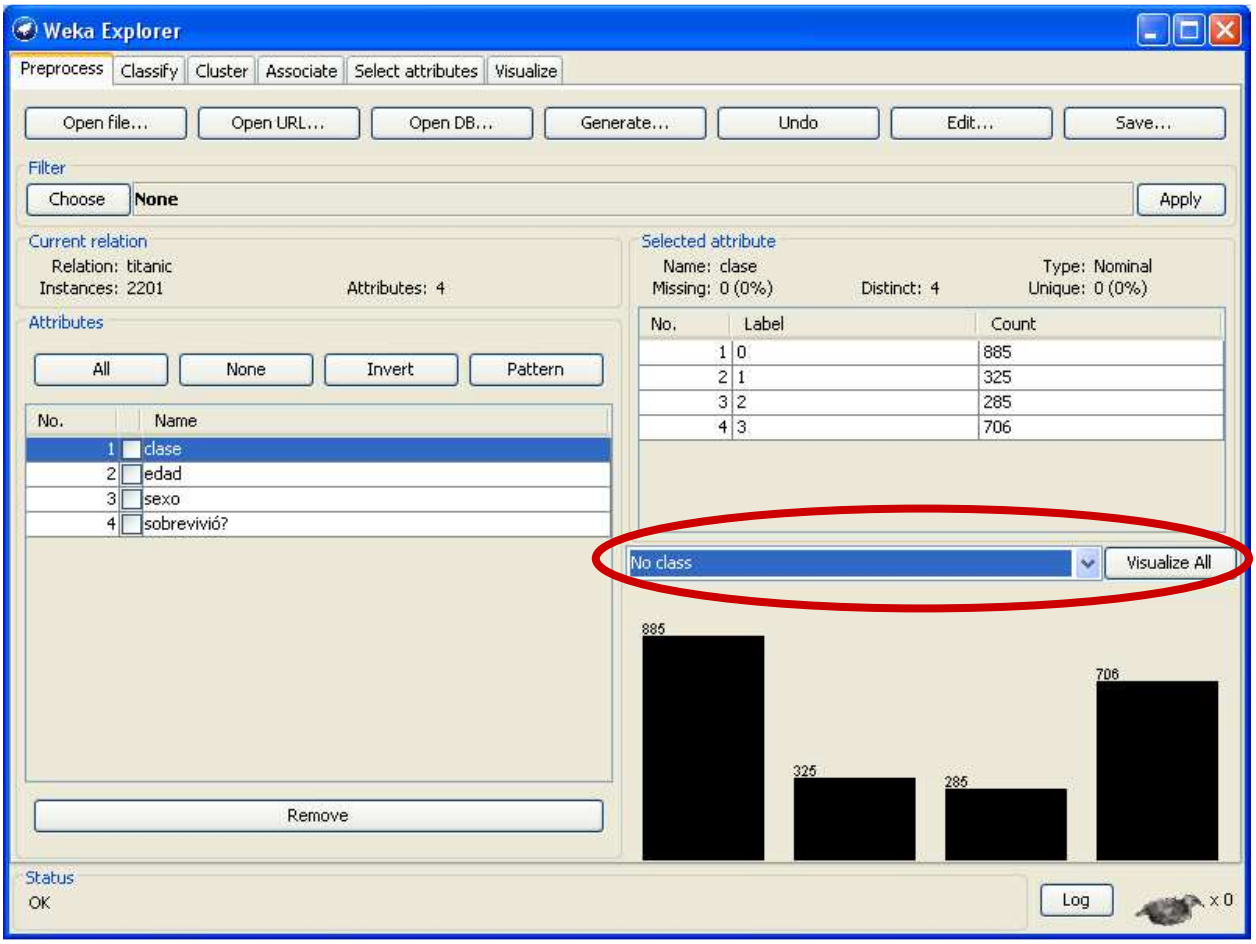
Este fichero consta de 4 atributos:

- Clase (0 = tripulación, 1 = primera, 2 = segunda, 3 = tercera)
- Edad (1 = adulto, 0 = niño)
- Sexo (1 = hombre, 0 = mujer)
- Sobrevivió (1 = sí, 0 = no)

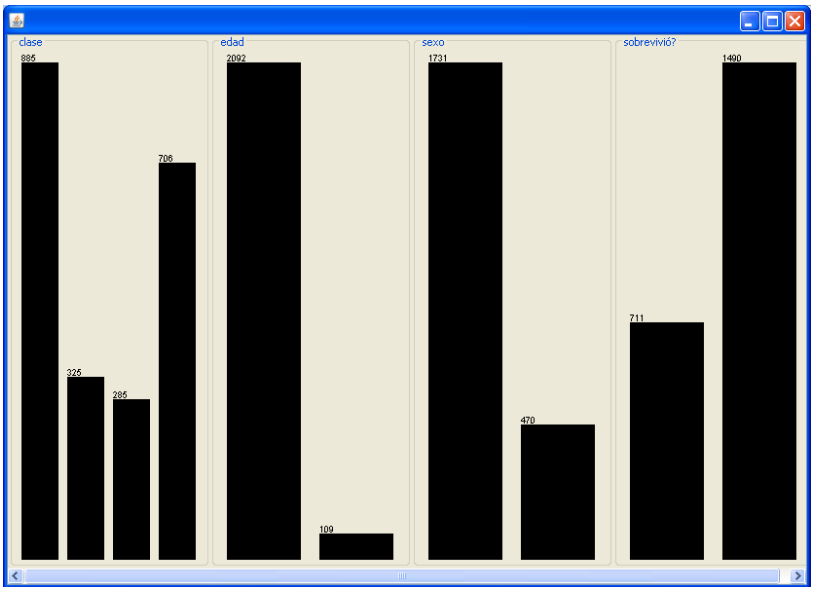
El objetivo es encontrar reglas de asociación interesantes a partir de estos datos

Para empezar abrimos la ventana *Explorer* de Weka y desde la pestaña *Preprocess* abrimos el fichero *titanic.arff* que comentábamos antes.

Cuando abrimos un fichero en Weka, por defecto nos asigna una clase (el último atributo). En este caso, como no vamos a hacer aprendizaje supervisado, tendremos que indicar que no hay clase.

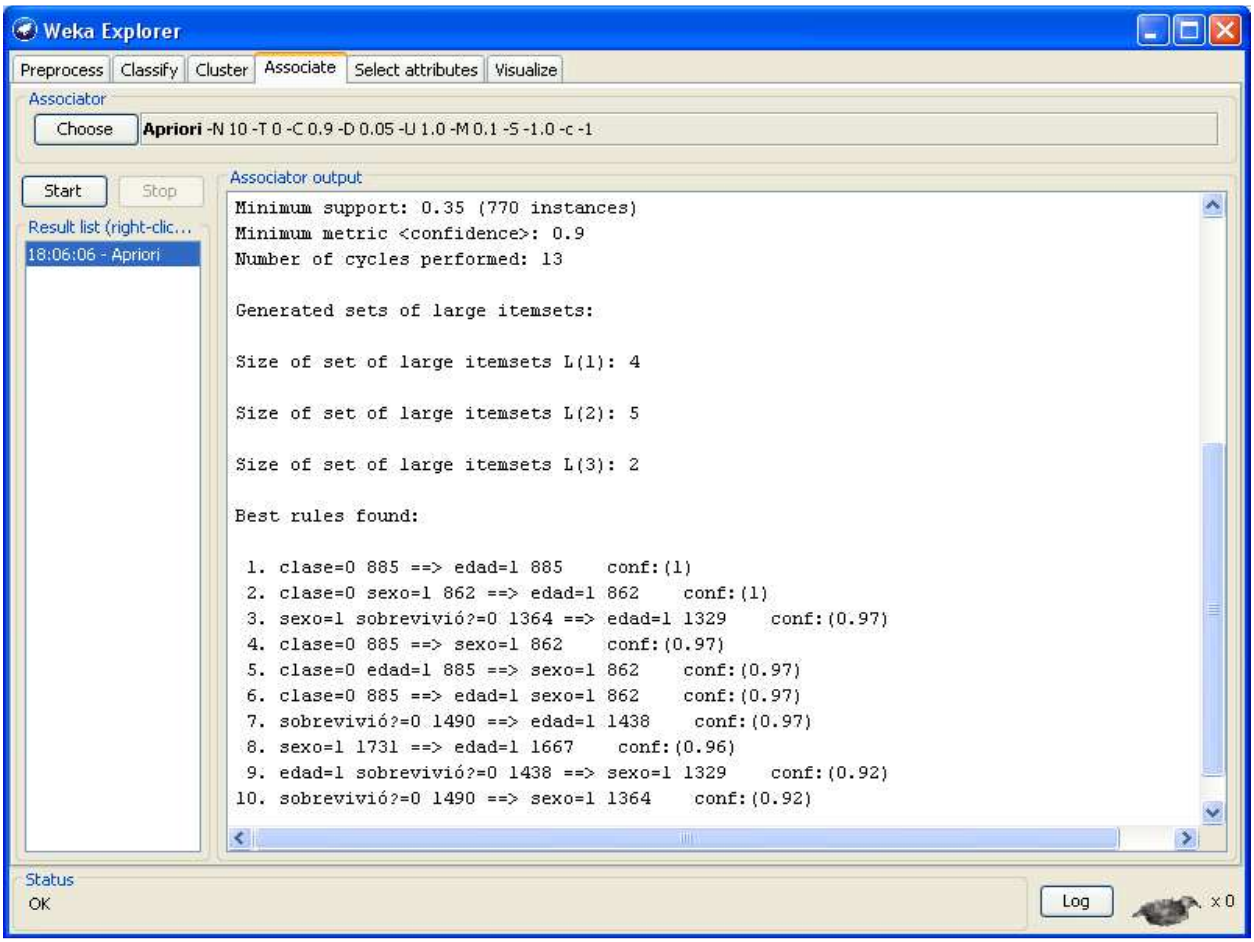


Podemos visualizar gráficamente toda la base de datos para ver cómo están distribuidos los valores de los atributos.



Ahora que ya hemos cargado los datos nos vamos directamente a la pestaña *Associate*, pinchamos en *Choose* y escogemos el algoritmo de reglas de asociación deseado y tras configurarlo pulsamos *Start*.

En el caso de usar como algoritmo *Apriori*, la salida del algoritmo con los parámetros por defecto es:



En cada regla, tenemos la cobertura (soporte) del antecedente (parte izquierda de la regla) y la de ambos conjuntamente (antecedente y consecuente de la regla, es decir, de la regla completa), así como la confianza (precisión) de la regla.

Analizando detenidamente las reglas de asociación generadas, podemos ver que algunas reglas son interesantes aunque otras no lo son tanto.

Por ejemplo, la regla 1. clase=0 885 ==> edad=1 885 conf:(1)
indica que, como era de esperar, toda la tripulación (clase=0) era adulta (edad=1).

La regla 2. clase=0 sexo=1 862 ==> edad=1 862 conf:(1)
nos indica lo mismo, pero teniendo en cuenta a los varones.

Las reglas 4, 5 y 6 son muy similares y no aportan nada interesante:
4. clase=0 885 ==> sexo=1 862 conf:(0.97)
5. clase=0 edad=1 885 ==> sexo=1 862 conf:(0.97)
6. clase=0 885 ==> edad=1 sexo=1 862 conf:(0.97)

La regla 3. sexo=1 sobrevivió?=0 1364 ==> edad=1 1329 conf:(0.97)
nos indica que los varones que no sobrevivieron fueron, en su mayoría, adultos (97%).

La regla 7. sobrevivió?=0 1490 ==> edad=1 1438 conf:(0.97)
destaca que la mayoría que murieron fueron adultos (97%),

y la regla 10. sobrevivió?=0 1490 ==> sexo=1 1364 conf:(0.92)
describe que la mayoría de los fallecidos fueron hombres (92%).

Se puede comprobar que, a veces, la calidad de las reglas de asociación viene sesgada por la presencia de atributos que estén fuertemente descompensados. Por ejemplo, en este caso la escasa presencia de niños provoca que no aparezcan en las reglas de asociación, ya que las reglas con este **ítemset** poseen una baja cobertura y son filtradas. Podríamos solucionar este efecto realizando algún proceso de selección de los datos o cambiando el método de selección de reglas.

Ejercicios

1. Utilizando la base de datos del Titanic, realizar los siguientes ejercicios:

   - Calcular el factor de interés (*lift*) de la primera regla que descubre el algoritmo con los parámetros por defecto. En función del valor calculado, ¿qué podrías decir del interés de dicha regla?
   - Ejecuta el algoritmo Apriori utilizando el factor de interés (*lift*) como métrica y mostrando un máximo de 100 reglas (*numRules*):
     - ¿Cuál es la mejor regla según esa métrica?
     - ¿Cuál es el valor de interés de dicha regla?
     - Dependiendo del consecuente de la mejor regla, explica la razón por la que dicha regla es la de mayor interés.
     - ¿Qué posición ocupa la mejor regla del Algoritmo Apriori cuando se ejecuta con los parámetros por defecto? Para realizar este ejercicio debes reducir el valor de *MinMetric*.

2. Aplicar el algoritmo Apriori a la base de datos “Cesta_compra.arff” y escribir un pequeño informe con las reglas más interesantes que descubráis. NOTA: podéis realizar todas las transformaciones que estiméis oportuno, utilizar distintas métricas y distintos valores en los parámetros?

¿Cómo entregar la práctica?

- Utilizar un documento de texto para responder a las cuestiones y subirlo a través de la plataforma Web