DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Sistemas Inteligentes de Gestión

Guión de Prácticas de Minería de Datos

Práctica 2

Reglas de Asociación

© *Juan Carlos Cubero & Fernando Berzal*

FICHEROS DE DATOS

```
Titanic.arff
Datos de empleados.sav
agaricus-lepiota.csv
```

ENTREGA DE LA PRÁCTICA
Ficheros de reglas y ficheros de los proyectos KNIME

```
Asociación_Titanic
      Asociación_Titanic.txt
      Asociacion_Titanic.knime.zip
Asociación_DatosEmpleados
      DatosEmpleados_discretizados.csv
      Asociación_DatosEmpleados.txt
      Asociacion_DatosEmpleados.knime.zip
Asociación_Mushroom
      Asociacion_Mushroom.txt
      Asociación_Mushroom.knime.zip
```

Para la realización de esta práctica, se recomienda la creación de una carpeta en la que se vayan incluyendo todos los ficheros que se han de entregar al finalizarla (organizados, a su vez, en tres subcarpetas correspondientes a los 3 conjuntos de datos utilizados en este guión).

Ejercicios tipo C

Titanic

El fichero `Titanic.arff` contiene datos sobre las características de los 2201 pasajeros del Titanic. Estos datos son reales y provienen del *"Report on the Loss of the 'Titanic' (S.S.)"* (British Board of Trade , Inquiry Report (reprint), Gloucester, UK, Allan Sutton Publishing, 1990).

El formato `arff` (Attribute-Relation File Format) es el formato "oficial" de Weka y consiste, simplemente, en un fichero de texto en el que se almacena una tabla de datos, con una línea por tupla y los valores de una misma tupla separados por comas (en la misma línea del fichero de texto). Adicionalmente, los ficheros `arff` incluyen una cabecera con información adicional acerca de los nombres y tipos de datos asociados a los distintos atributos atributos de la relación, tal como se muestra a continuación:

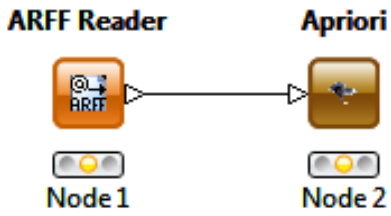
```
% Comentarios

@RELATION Persona

@ATTRIBUTE Ingresos  NUMERIC
@ATTRIBUTE Nombre   string
@ATTRIBUTE FechaNacimiento  date
@ATTRIBUTE CategoriaLaboral {Administrativo, Seguridad, Directivo}

@DATA

18000.34 , Juan , 1979-03-31 , Administrativo
22300.05 , Inma, 1967-02-25 , Directivo
........
```

Más información sobre el formato `arff` en http://weka.wiki.sourceforge.net/ARFF.

En el caso del fichero de datos correspondiente a los datos de los pasajeros del Titanic, sólo consideraremos los siguientes cuatro atributos, que ya aparecen codificados en el fichero `Titanic.arff`:

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

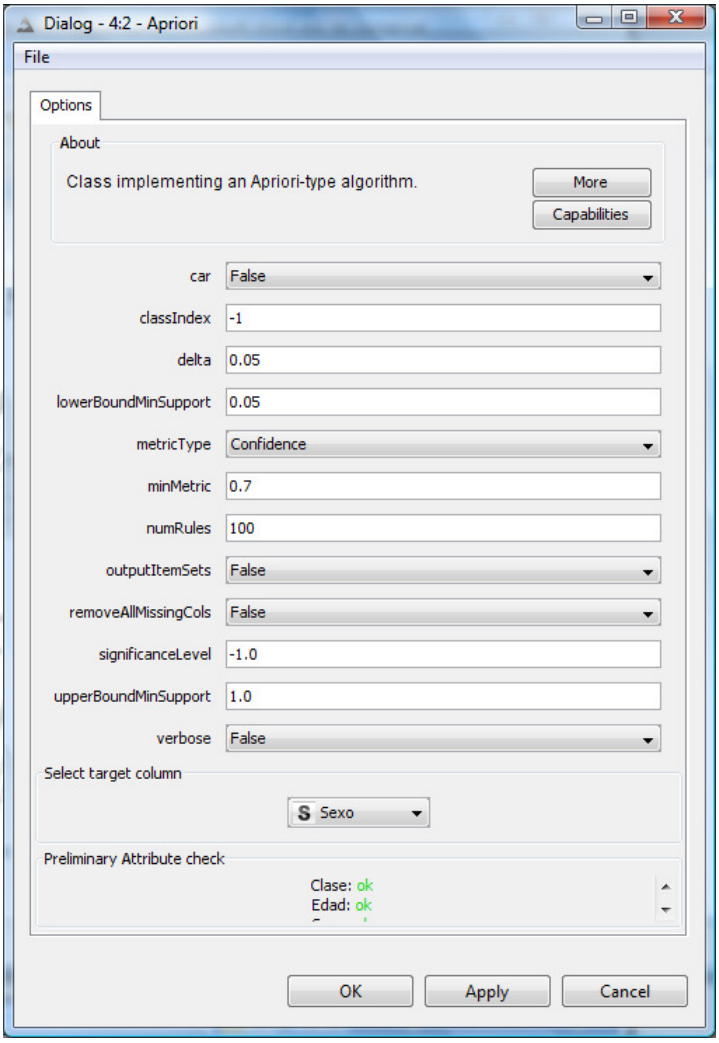
Crearemos un proyecto nuevo en KNIME llamado AsociacionTitanic, al que le añadiremos un nodo *IO > Read >ARFF Reader* (para leer datos desde un fichero en formato ARFF). Este nodo lo conectaremos con otro nodo, de tipo *Weka > Association Rules > APriori*, lo que dará lugar a un fllujo como el siguiente:



RECORDATORIO: Para poder utilizar los nodos de Weka en KNIME, hay que seleccionar *File > Update KNIME* y elegir las extensiones para Weka.

A continuación, configuraremos el lector ARFF para que acceda a los datos de nuestro fichero de datos () y estableceremos las siguientes opciones para Apriori:

- Un umbral de soporte mínimo del 5% (0.05 representa un 5% en `lowerBoundMinSuppport`)
- Confianza del 70% (0.7 en `minMetric`).



Observaciones

- `More` muestra información adicional sobre el método empleado (`Capabilities`, restricciones y tipos sobre los que opera).
- En `metricType` podemos escoger otras medidas de evaluación de las reglas (p.ej. lift).
- `OutputItemsetItems` indica si deseamos obtener los patrones frecuentes.

Ejecute el flujo, seleccionando el nodo lector y pinchando en  .

Desde el nodo Apriori, pinche con el botón izquierdo del ratón y, en su menú contextual, seleccione *View: Weka Node View*.

Cree un fichero de texto, `Asociación_Titanic.txt`, y comente el significado de, al menos, 4 reglas que le parezcan de interés.

Repita el proceso anterior cambiando la configuración del nodo Apriori:

- Soporte 1%, confianza 70%.
- Soporte 1%, confianza 85%

De todas las reglas obtenidas para cada una de las configuraciones anteriores, comente lo que le haya parecido más interesante (en el fichero `Asociación_Titanic.txt`).

Añada un nodo del tipo *Statistics* > *Statistics* y compruebe que, para el atributo *Edad*, hay 2092 tuplas con valor 1 (adulto) y sólo 109 con valor 0 (no adulto). En clase de teoría hemos comentado los problemas que surgen con la presencia de ítems demasiado frecuentes. Para eliminar su influencia en nuestro análisis, prueba las siguientes estrategias:

a) Añada un nodo de filtro para eliminar la columna *Edad* entera (*Data Manipulation* > *Column* > *Filter* > *Column Filter*) y vuelva a generar las reglas (con umbral de soporte 1% y confianza mínima 70%). Compare los resultados con los obtenidos anteriormente (con los mismos umbrales de soporte y confianza) y comente, al menos, 4 reglas.

b) Si hace lo indicado en la opción anterior, obviamente, no se generará ninguna regla relativa a *edad=0*. Lo ideal, no obstante, sería generar las reglas utilizando todos los datos disponibles y, posteriormente, filtrar las reglas obtenidas. Desgraciadamente, ni KNIME ni Weka proporcionan herramientas para hacerlo. Vuelva a la generar las reglas con todos los datos (soporte 1% y confianza 70%) y comente al menos alguna regla interesante que involucre a *edad=0*.

c) ¿Sería adecuado insertar entre el nodo de lectura de datos y el de generación de reglas un nuevo nodo del tipo *Data Manipulation* > *Row* > *Filter* > *Row Filter* > *Exclude Rows by Attribute Value*, de tal manera que se excluyan las tuplas que contengan *edad=1*? Razone su respuesta.

Mushroom (Agaericus Lepiota)

Utilice ahora el fichero `agaricus-lepiota.csv`, que contiene datos sobre la morfología de un conjunto de setas [mushrooms] y un atributo [Class] que nos indica si la seta es comestible o no.

Cree un nuevo proyecto KNIME y diseño el flujo necesario para generar reglas de asociación probando, al menos, dos combinaciones distintas de soporte y medidas de evaluación (p.ej. confianza o lift).

Compruebe el número enorme de reglas que se obtienen y, en el fichero `Asociacion_Mushroom.txt` , indique qué parámetros se ha utilizado en sus distintas configuraciones y el número de reglas obtenidas en cada caso.

Ejercicios tipo B

Datos de Empleados

A continuación, queremos extraer reglas de asociación desde el fichero `Datos de empleados.sav` de SPSS.

En primer lugar, como este fichero involucra atributos de tipo numérico, vamos a discretizar sus variables continuas utilizando el algoritmo equi-depth con 6 intervalos. Para ello usamos SPSS, ya que KNIME no tiene ningún nodo para hacerlo :-(

NOTA: KNIME sí que incluye un discretizador llamado CAIM, pero este nodo sólo es aplicable en problemas de clasificación (es un algoritmo de discretización supervisada).Weka sí tiene discretizadores, pero no disponibles desde KNIME.

Las variables que discretizaremos son salario inicial, salario actual, experiencia previa y meses desde el contrato (salario, salini, tiempemp, expprev).

Desde SPSS, una vez seleccione "*Guardar como*" y elijar como formato "*Delimitado por comas (csv)*". Entre las opciones disponibles, seleccione "*Escribir nombres de variables en hoja de cálculo*" y "*Guardar etiquetas de valor dónde se hayan definido en vez de valores de datos*". Esto nos dará como resultado un fichero que llamaremos `DatosEmpleados_discretizados.csv`.

A continuación, crearemos un proyecto en KNIME con los siguientes nodos:

- *IO > Read > FileReader*
  (para obtener datos de `DatosEmpleados_discretizados.csv`).
  Al configurar este nodo, seleccione la cabecera de cada columna para establecer un tipo de dato adecuado para los datos del fichero (en nuestro caso, como todas las variables son de medida nominal, usaremos el tipo *String*).

- *Data Manipulation >Column > Column Filter*
  (para seleccionar únicamente los atributos *Categoría Laboral*, *Sexo* y *Clasificación de Minorías*) , nodo que situaremos entre la lectura de datos y la extracción de reglas de asociación.

- *Weka > Association Rules > APriori*
  (para ejecutar el algoritmo Apriori,
  con umbral de soporte del 1% y confianza mínima del 75%).

En un fichero de texto llamado `Asociación_DatosEmpleados.txt`, comente dos reglas cualesquiera de las que haya obtenido y analice los resultados indicando si hay reglas que pudiesen considerarse redundantes.

Repita el ejercicio con los mismos parámetros, pero incluyendo también la variable *Salario Actual* (discretizada, obviamente).

Ejercicios tipo A

Titanic

Obtenga las reglas de asociación con un umbral de soporte del 1%, pero usando como medida de evaluación el interés (lift) de las reglas, para el cual se recomienda un umbral mínimo igual a 2. La primera regla que aparece tiene como antecedente *edad=0*, un valor de lift por encima de 4 y una confianza muy baja (apenas 0.2). ¿Tiene sentido esta situación? ¿Por qué? ¿Cuál sería en este caso la interpretación adecuada del valor del lift? ¿Qué interpretación semántica tiene la regla obtenida?

¿Podría destacar alguna otra regla que fuese interesante?

Datos de Empleados

Utilice Apriori con un umbral de soporte mínimo del 5% y un valor de lift de 2.

Comente dos reglas cualesquiera de las obtenidas.

Ampliación

Para extraer reglas de asociación usando otras herramientas de minería de datos:
http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Assoc_Rules_Comparison.pdf

EVALUACIÓN DE LAS PRÁCTICAS

Los ficheros de texto asociados a cada uno de los conjuntos de datos utilizados en esta práctica han de incluir todos sus comentarios y respuestas a las distintas preguntas que aparecen en el guión.

PD: No olvide incluir también sus proyectos KNIME.