Practica 4 Inducción de reglas de asociación.

Trabajaremos con los archivos supermercado.arff y empleados.arff.

empleados.arff consta de 318 individuos supuestos empelados de una compañía de los que se dispone de información de las siguientes variables ,todas ellas atributos nominales algunos de dos posibles valores y otros de más de dos posibles valores.

@attribute 'Sueldo ' {'\'(-inf-12500]\'','\'(12500-22500]\'','\'(22500-35000]\'','\'(35000-inf)\''}

@attribute Casado {Sí,No}

@attribute Coche {No,Sí}

@attribute Hijos {0,1,2,3}

@attribute Alq/Prop {Alquiler,Prop}

@attribute Sindic. {No,Sí}

@attribute Bajas/Ańo {ninguna,'\'1-5\'','\'5-10\'','\'más de 10\''}

@attribute Antigüedad {'\'(-inf-5.75]\'','\'de 5 a 10\'','\'de 10 a 15\'','\'más de 15\''}

@attribute Sexo {H,M}

@attribute departamento {comercial,administración,almacen,'recursos humanos'}

supermercado.arff es una modificación de una conocida base de datos ( usada en uno de los cursos de weka de la Universidad de Waikato) en la que se han traducido y cambiado algunos productos de los que constaban en la original extraída de los tickets de compra dun supermercado neozelandés .

Tiene 4627 registros compras donde se consigna , como atributos, la realización o no de la compra de los distintos productos . 78 productos en total a los que hay que añadir un atributo adicional “total” que informa de si el monto total de la compra ha sido alto o bajo. Los atributos de productos son de naturaleza dicotómica pudiendo tomar el valor **t** si se ha producido la compra de tal producto o no tomar ningún valor si no se ha producido la compra.

1.- Para los dos conjuntos de datos : Usando las opciones por defecto del algoritmo a priori , extraer la asociaciones más relevantes y comentarlas, incidiendo en su sentido para la situación analizada , y su confianza y cobertura

2.- Si con el conjunto de datos supermercado, marcamos en el algoritmo a priori la opción car-> true , no se obtiene ninguna regla con los requisitos por defecto



 ( no se obtiene ninguna regla cuyo consecuente sea un ítem relativo a la “clase” ( atributo “total” ) . No hay ninguna regla con confianza superior a 0.9 que nos permita asociar cierta cesta de compra a un total elevado o bajo.

Si reducimos la exigencia a una confianza inferior, (0.80 , por ejemplo) ya podemos asociar algunas reglas que pueden predecir una compra de elevada factura siempre que al menos estén presentes algunos productos .

¿Qué productos predicen una elevada factura con un 80% de confianza.?

¿De esas distintas combinaciones de productos, cuál se da con mayor frecuencia? ¿Es una frecuencia relevante?

3.- Volvemos a situar la opción **car** en **false**. Si exigimos que la cobertura supere el 50% de casos ¿Cuánto tendremos que reducir la confianza exigida para contar con alguna combinación de productos asociados en una compra mayoritaria ( recordemos que la cobertura es mayor que el 50%) ?

¿Cuáles son las dos asociaciones de mayor confiabilidad en una mayoría de compras ( más del 50% de las veces )

¿con qué confianza?

¿En función de estos resultados aconsejarías alguna disposición de expositores de artículos? ¿Cuál?

4. Seleccionamos sólo los casos que suponen una alta factura. Para ello aplicamos el filtro **RemoveWithValues** con la opción nominalindices = 1 ( ya que el primer valor del útimo atributo se corresponde con total bajo)

weka.filters.unsupervised.instance.RemoveWithValues -S 0.0 -C last -L 1



Una vez hecho esto, **Eliminamos el útimo atributo** (**Total**) ya que como ahora todos los casos son de total alto, cualquier regla incluirá como consecuente total=alto , dando una información irrelevante.

¿Cuáles son las cuatro primeras reglas inducidas si las seleccionamos con las opciones por defecto ( min. Cobertura=0.1 min confinaza=0.9)?

¿Podemos concluir que los productos más frecuentes comprados en las compras de elevado total son los implicados en estas reglas? Razonar la respuesta

5. Deshaz los cambios realizados antes : Volvemos a la base de datos original.

Ahora nos planteamos obtener las asociaciones con las opciones por defecto pero permitiendo hasta 30 reglas.

Cuál es la mínima confianza asignada a estas 30 reglas. ¿qué característica tienen en común? ¿Por qué crees que ocurre esto? ¿ Alguna idea para evitarlo y obtener quizá conclusiones más interesantes?

7. Ahora pasamos a la otra base de datos Empleados.arff. Analizamos la aplicación del algoritmo con opciones por defecto. Si no se ha hecho el punto 1 para estos datos, es el momento de hacerlo, sin olvidar los comentarios.

Observamos que la regla **Vivienda en propiedad ⇒ coche** se cumple con una confianza del 100%. Sin embargo, si usamos la vivienda en propiedad como un predictor de tener coche sólo acertaríamos en poco más del 54 de las ocasiones:

Hay 243 empleados con coches de los cuales sólo 132 tienen la vivienda en propiedad.( aproximadamente el 54%)

 Comenta este hecho.

8. La regla: **Sueldo ='(-inf-12500]' 111 ==> Sexo=H 111 <conf:(1)>** , se da con una confianza del 100%

¿Qué quiere decir? ¿todos los hombres tienen un bajo sueldo? ¿Si un empleado es hombre tendrá bajo sueldo?

9. Usa la opción de **car-> true** y con las opciones por defecto y comenta las reglas que afectan al departamento comercial. ¿El perfil que proponen para un empleado de ese departamento podría funcionar como buen predictor de que un empleado pertenece a ese departamento? ( Haz un planteamiento similar al del punto anterior )

10. Selecciona sólo los empleados del departamento de almacén. Aplica las opciones por defecto y considera que todas reglas predicen sexo=H ( Hombre).¿qué pasa aquí?

¿Tienen interés esas reglas? ¿Por qué? ¿qué podemos hacer para obtener reglas con más “sustancia”?