Reglas de Asociación en Weka

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

June 30, 2020

En esta práctica se realiza un estudio acerca de los datos del hundimiento del Titanic a través de la herramienta Weka. Los datos se encuentran en la dirección `http://www.hakank.org/weka/titanic.arff` y corresponden a las características de los 2201 pasajeros del Titanic. Estos datos son reales y se han obtenido de *"Report on the Loss of the 'Titanic' (S.S.)" (1990), British Board of Trade Inquiry Report_(reprint), Gloucester, UK: Allan Sutton Publishing.*

Para realizar esta práctica, se debe cargar el dataset Titanic que se ha descargado anteriormente y contestar a las siguientes preguntas:

1. Cuando ejecutamos el algoritmo Apriori de Weka, podemos utilizar diferentes umbrales de soporte. Dependiendo de qué umbrales de soporte pongamos, nos saldrán más o menos itemsets. Como resultado, Weka nos proporciona un conjunto de ítems L(1)... L(4) cuyos números van variando conforme cambiamos el umbral de soporte.

   Responde a las siguientes preguntas, utilizando capturas de pantalla y explicando los resultados de manera clara y concisa:

   (a) ¿Qué representan cada uno de estos conjuntos de ítems?

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.15 (330 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 13

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 2
```

Figura 1: soporte = 0.15 y confianza = 0.9.

1. El conjunto de ítems L(1) representa el número de conjuntos de ítems de tamaño 1. Que en este caso son 7.
2. El conjunto de ítems L(2) representa el número de conjuntos de ítems de tamaño 2. Que en este caso son 13.
3. El conjunto de ítems L(3) representa el número de conjuntos de ítems de tamaño 3. Que en este caso son 8.
4. El conjunto de ítems L(4) representa el número de conjuntos de ítems de tamaño 4. Que en este caso son 2.

(b) ¿Puede existir L(0)? Explica porqué.

No puede existir L(0). El conjunto de ítems L(0) representa el número de conjuntos de ítems de tamaño 0, es decir, el número de conjuntos de ítems vacíos, y el conjunto vacío ($\emptyset$) no es válido como conjunto de ítems.

(c) ¿Puede existir L(5)? Explica porqué.

No puede existir L(5) porqué el dataset de Titanic solo contiene cuatro atributos diferentes. (ver Figura 2).

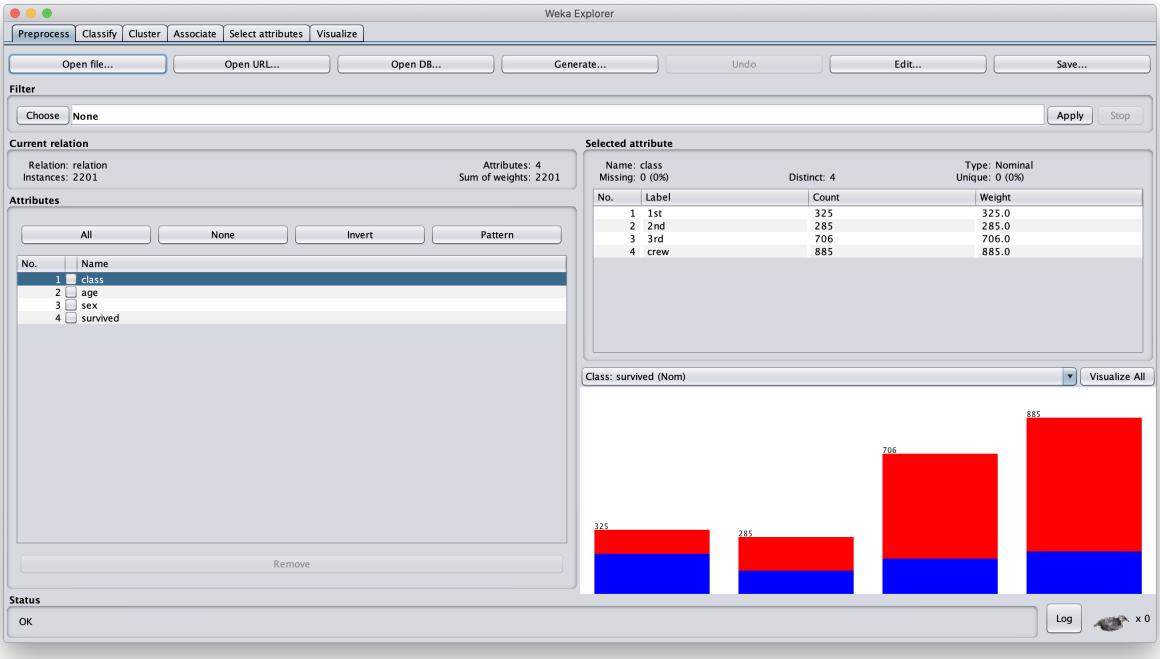

Figura 2: características del dataset.

(d) ¿Puede L(1) tomar un valor mayor que 10? Explica de manera teórica que eso no es posible y compruébalo experimentalmente.

L(1) no puede tomar un valor mayor que 10. El número de conjuntos de ítems de tamaño 1 será 9. Ya que, como máximo, el número de conjuntos de ítems de tamaño 1, cuenta con los diferentes valores de cada atributo del dataset.

En este caso, el dataset de Titanic, contiene 9 valores diferentes:

- Class ("1st", "2nd", "3rd", "Crew")
- Age "Adult", "Child"
- Sex "Male", "Female"
- Survived "Yes", "No"

Para comprobarlo experimentalmente, el umbral de confianza debe ser igual a 1. (ver Figura 3)

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 20 -T 0 -C 1.0 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (220 instances)
Minimum metric <confidence>: 1
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 15

Size of set of large itemsets L(3): 9

Size of set of large itemsets L(4): 2
```

Figura 3: soporte = 0.1 y confianza = 1.

2. Además de los valores de soporte, el algoritmo Apriori de Weka nos permite utilizar diferentes umbrales de soporte y confianza. Responde a las siguientes preguntas, utilizando capturas de pantalla y explicando los resultados de manera clara y concisa:

   (a) ¿Es posible que una regla tenga un valor de soporte inferior a su confianza? Explica porqué y demuéstralo experimentalmente.

   Sí. Lo podemos volver a ver en la Figura 1.

   (b) ¿Es posible que una regla tenga un valor de confianza inferior a su suporte? Explica porqué y demuéstralo experimentalmente.

   Sí, ver Figura 4.

   (c) La variación del umbral de confianza (dado un umbral fijo de soporte) no afecta a los conjuntos L(1)... L(4). ¿Por qué?

   Porqué la variación del umbral de confianza es una métrica enfocada para las reglas, y mide la frecuencia con que se pueden encontrar. En cambio, el umbral de soporte es una métrica enfocada para los conjuntos de ítems, que mide la proporción de estos dentro del dataset. Y como los conjuntos L(1)... L(4) representan el número de apariciones de conjuntos de ítems según su tamaño dentro del dataset, el umbral les afectará.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 20 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.7 -S -1.0 -c -1
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.7 (1541 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 2

Size of set of large itemsets L(2): 1
```

Figure 4: Soporte = 0.7 y confianza = 0.5.

3. Usaremos ahora, 0.75 como valor mínimo de soporte y de confianza 0.00. Comprobamos que obtenemos dos reglas de asociación, sin embargo, L(2) es 1. ¿Qué quiere decir esto? ¿A qué corresponde L(2)? ¿Qué itemset representa?

   Ver Figura 5.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 20 -T 0 -C 0.0 -D 0.05 -U 1.0 -M 0.75 -S -1.0 -c -1
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.75 (1651 instances)
Minimum metric <confidence>: 0
Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 2

Size of set of large itemsets L(2): 1

Best rules found:

 1. sex=male 1731 ==> age=adult 1667    <conf:(0.96)> lift:(1.01) lev:(0.01) [21] conv:(1.32)
 2. age=adult 2092 ==> sex=male 1667    <conf:(0.8)> lift:(1.01) lev:(0.01) [21] conv:(1.05)
```

Figure 5: Soporte = 0.75 y confianza = 0.

4. Analiza el conjunto de reglas que salen al aplicar diferentes umbrales de soporte y confianza. Coge una regla, la que veas más interesante, y coméntala. Explica sus valores de métricas y qué representan, y el significado de la regla, es decir, el conocimiento que te aporta dicha regla.

   Después de aplicar diferentes umbrales de soporte y confianza la regla elegida es: