Reglas de Asociación en Weka

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

July 1, 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 encontrados en el dataset. Que en este caso son 7.
2. El conjunto de ítems L(2) representa el número de conjuntos de ítems de tamaño 2 encontrados en el dataset. Que en este caso son 13.
3. El conjunto de ítems L(3) representa el número de conjuntos de ítems de tamaño 3 encontrados en el dataset. Que en este caso son 8.
4. El conjunto de ítems L(4) representa el número de conjuntos de ítems de tamaño 4 encontrados en el dataset. 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 pueden existir conjuntos de ítems de tamaño 5 porqué en el dataset de Titanic no hay un atributo que tenga 5 valores. Como máximo hay un atributo que tiene 4 valores. Este atributo es el atributo Class (ver Figura 2).

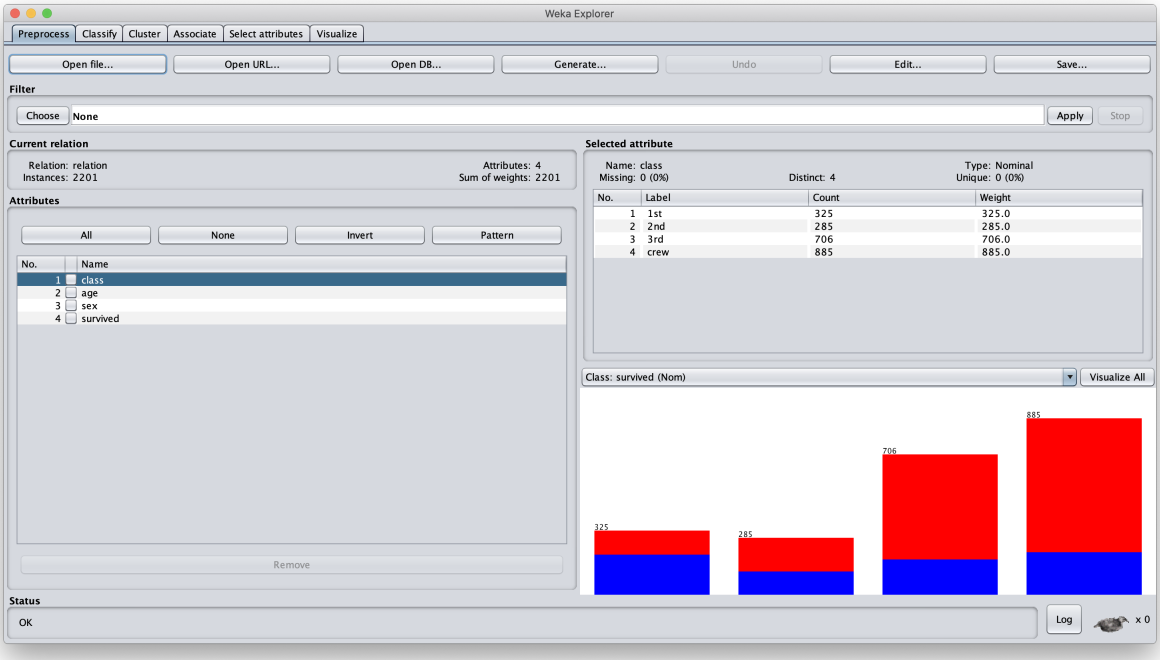


Figura 2: características del dataset.

(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, como máximo podrá ser igual a 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.

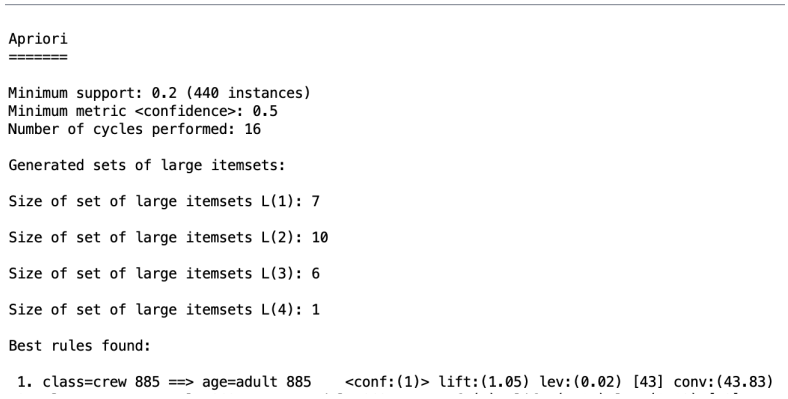
   Es posible que una regla aparezca con poca frecuencia en el dataset pero que los conjuntos de ítems que la forman estén muy correlacionados. Para explicar el porqué y demostrarlo nos ayudaremos de la métrica *lift* (mejora de la confianza). La métrica lift mide la relación del valor del soporte observado con el valor del soporte esperado de los conjuntos de ítems que forman cada regla, si esos conjuntos son independientes. Los conjuntos de ítems estarán positivamente correlacionados si el valor de lift es superior a 1, si el valor de lift es inferior a 1 estarán negativamente correlacionados, y si el valor de lift es igual a 1, estarán equilibradamente correlacionados.

   En este caso, si una regla puede tener un valor de soporte inferior a su confianza, el valor de lift será superior a uno y nos indicará que el conjunto de ítems que forman la regla aparecen una cantidad de veces superior a lo esperado, por lo que se puede intuir que la regla hace que los conjuntos de ítems que la forman aparezcan más de lo normal.

Por ejemplo, en la regla,

class=crew 885 ==> age=adult 885 <conf:(1)> lift:(1.05) lev:(0.02) [43] conv:(43.83)

vemos que el número de instancias que aparecen en el dataset los conjuntos de ítems de la regla es elevado (885), y el valor de lift es superior a uno. Eso nos permite saber que existen muchas co-ocurrencias entre estos dos conjuntos y que la regla será potencialmente útil para predecir. Concretamente, ha resultado ser la regla número uno (Ver Figura 4).



```
Apriori
=======

Minimum support: 0.2 (440 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 16

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

Best rules found:

 1. class=crew 885 ==> age=adult 885    <conf:(1)> lift:(1.05) lev:(0.02) [43] conv:(43.83)
```

Figure 4: soporte inferior a confianza.

(b) ¿Es posible que una regla tenga un valor de confianza inferior a su suporte? Explica porqué y demuéstralo experimentalmente.
También es posible que una regla aparezca con mucha frecuencia en el dataset pero que los conjuntos de ítems que la forman estén poco correlacionados. En este caso, el valor de lift será inferior a uno y nos indicará que el conjunto de ítems que forman la regla aparecen una cantidad de veces inferior a lo esperado, por lo que se puede intuir que la regla hace que los conjuntos de ítems que la forman aparezcan menos de lo normal.
Por ejemplo, la regla,

class=3rd 706 ==> age=adult sex=male 462 <conf:(0.65)> lift:(0.86) lev:(-0.03) [-72] conv:(0.7)

vemos que el número de instancias que aparecen en el dataset los conjuntos de datos no es tan elevado (706 y 462). Además no aparecen con la misma frecuencia en el dataset y el valor de lift es inferior a uno. Eso nos permite saber que existen menos co-ocurrencias entre los conjuntos de ítems que forman la regla y que esto tiene un efecto implicativo negativo. Potencialmente la regla no será tan útil para predecir. Y concretamente, ha resultado ser la regla número cuarenta (Ver Figura 5).

(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 mide la relación entre los conjuntos de ítems que forman las reglas y no la frecuencia de aparición de los conjuntos de ítems de las reglas, eso lo hace el umbral de soporte. Y como los conjuntos de ítems L(1)... L(4) representan el número de apariciones de conjuntos de ítems según su tamaño dentro del dataset, el umbral de soporte les afectará. Por ejemplo, si el umbral de soporte es pequeño, Weka nos mostrará pocos conjuntos de ítems que no aparecen en el dataset, al contrario cuando el umbral del soporte sea mayor.

```
20. class=3rd 706 ==> age=adult 627    <conf:(0.89)> lift:(0.93) lev:(-0.02) [-44] conv:(0.44)
21. age=adult sex=male 1667 ==> survived=no 1329    <conf:(0.8)> lift:(1.18) lev:(0.09) [200] conv:(1.59)
22. age=adult 2092 ==> sex=male 1667    <conf:(0.8)> lift:(1.01) lev:(0.01) [21] conv:(1.05)
23. sex=male 1731 ==> survived=no 1364    <conf:(0.79)> lift:(1.16) lev:(0.09) [192] conv:(1.52)
24. class=crew sex=male 862 ==> survived=no 670    <conf:(0.78)> lift:(1.15) lev:(0.04) [86] conv:(1.44)
25. class=crew age=adult sex=male 862 ==> survived=no 670    <conf:(0.78)> lift:(1.15) lev:(0.04) [86] conv:(1.44)
26. class=crew sex=male 862 ==> age=adult survived=no 670    <conf:(0.78)> lift:(1.19) lev:(0.05) [106] conv:(1.55)
27. sex=male 1731 ==> age=adult survived=no 1329    <conf:(0.77)> lift:(1.18) lev:(0.09) [198] conv:(1.49)
28. class=crew 885 ==> survived=no 673    <conf:(0.76)> lift:(1.12) lev:(0.03) [73] conv:(1.34)
29. class=crew age=adult 885 ==> survived=no 673    <conf:(0.76)> lift:(1.12) lev:(0.03) [73] conv:(1.34)
30. class=crew 885 ==> age=adult survived=no 673    <conf:(0.76)> lift:(1.16) lev:(0.04) [94] conv:(1.44)
31. class=3rd age=adult 627 ==> survived=no 476    <conf:(0.76)> lift:(1.12) lev:(0.02) [51] conv:(1.33)
32. class=crew 885 ==> sex=male survived=no 670    <conf:(0.76)> lift:(1.22) lev:(0.06) [121] conv:(1.56)
33. class=crew age=adult 885 ==> sex=male survived=no 670    <conf:(0.76)> lift:(1.22) lev:(0.06) [121] conv:(1.56)
34. class=crew 885 ==> age=adult sex=male survived=no 670    <conf:(0.76)> lift:(1.25) lev:(0.06) [135] conv:(1.62)
35. class=3rd 706 ==> survived=no 528    <conf:(0.75)> lift:(1.1) lev:(0.02) [50] conv:(1.27)
36. class=3rd age=adult 627 ==> sex=male 462    <conf:(0.74)> lift:(0.94) lev:(-0.01) [-31] conv:(0.81)
37. class=3rd 706 ==> sex=male 510    <conf:(0.72)> lift:(0.92) lev:(-0.02) [-45] conv:(0.77)
38. age=adult 2092 ==> survived=no 1438    <conf:(0.69)> lift:(1.02) lev:(0.01) [21] conv:(1.03)
39. class=3rd 706 ==> age=adult survived=no 476    <conf:(0.67)> lift:(1.03) lev:(0.01) [14] conv:(1.06)
40. class=3rd 706 ==> age=adult sex=male 462    <conf:(0.65)> lift:(0.86) lev:(-0.03) [-72] conv:(0.7)
```

Figure 5: soporte superior a confianza.

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?

   Esto quiere decir que el algoritmo Apriori ha encontrado un conjunto de ítems de tamaño dos que no forma parte de ninguna regla, ya que los conjuntos de ítems que la forman no están correlacionados. Este conjunto representa las personas adultas (hombres o mujeres). Ver Figura 6.

```
=== 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 6: 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 la siguiente:

```
Best rules found:

 1. class=crew 885 ==> age=adult 885    <conf:(1)> lift:(1.05) lev:(0.02) [43] conv:(43.83)
 2. class=crew sex=male 862 ==> age=adult 862    <conf:(1)> lift:(1.05) lev:(0.02) [42] conv:(42.69)
 3. class=crew survived=no 673 ==> age=adult 673    <conf:(1)> lift:(1.05) lev:(0.02) [33] conv:(33.33)
 4. class=crew sex=male survived=no 670 ==> age=adult 670    <conf:(1)> lift:(1.05) lev:(0.02) [33] conv:(33.18)
 5. class=crew survived=no 673 ==> sex=male 670    <conf:(1)> lift:(1.27) lev:(0.06) [140] conv:(35.93)
 6. class=crew age=adult survived=no 673 ==> sex=male 670    <conf:(1)> lift:(1.27) lev:(0.06) [140] conv:(35.93)
 7. class=crew survived=no 673 ==> age=adult sex=male 670    <conf:(1)> lift:(1.31) lev:(0.07) [160] conv:(40.82)
 8. sex=male survived=no 1364 ==> age=adult 1329    <conf:(0.97)> lift:(1.03) lev:(0.01) [32] conv:(1.88)
 9. class=crew 885 ==> sex=male 862    <conf:(0.97)> lift:(1.24) lev:(0.08) [165] conv:(7.87)
10. class=crew age=adult 885 ==> sex=male 862    <conf:(0.97)> lift:(1.24) lev:(0.08) [165] conv:(7.87)
11. class=crew 885 ==> age=adult sex=male 862    <conf:(0.97)> lift:(1.29) lev:(0.09) [191] conv:(8.95)
12. survived=no 1490 ==> age=adult 1438    <conf:(0.97)> lift:(1.02) lev:(0.01) [21] conv:(1.39)
13. sex=male 1731 ==> age=adult 1667    <conf:(0.96)> lift:(1.01) lev:(0.01) [21] conv:(1.32)
14. age=adult survived=no 1438 ==> sex=male 1329    <conf:(0.92)> lift:(1.18) lev:(0.09) [198] conv:(2.79)
15. sex=male survived=yes 367 ==> age=adult 338    <conf:(0.92)> lift:(0.97) lev:(-0) [-10] conv:(0.61)
16. survived=yes 711 ==> age=adult 654    <conf:(0.92)> lift:(0.97) lev:(-0.01) [-21] conv:(0.61)
17. class=3rd sex=male survived=no 422 ==> age=adult 387    <conf:(0.92)> lift:(0.96) lev:(-0.01) [-14] conv:(0.58)
```

Figure 7: tripulación que no sobrevivió.

Esta regla nos indica las personas adultas de la tripulación que no sobrevivieron al naufragio del Titanic. Las métricas que podemos observar en ella son:

- La confianza con valor igual a 1. Lo que significa que para el 100% de las transacciones del dataset que contienen los conjuntos de ítems que forman la regla, la regla se cumplirá siempre. Es decir, siempre que una persona de la tripulación no haya sobrevivido, esta persona será adulta. Esto es cierto, porqué no existe ningún miembro de la tripulación que fuera un niño.
- La mejora de la confianza o lift, con valor igual a 1.05. Como se ha comentado anteriormente, si el valor de lift es superior a 1, permite saber que hay muchas co-ocurrencias entre los conjuntos de ítems que forman la regla y que dependen unas de ellas. Eso hace que la regla sea potencialmente útil para predecir y extraer información del dataset muy relevante. Podríamos decir que la mayoría de las personas que no sobrevivieron fueron personas de la tripulación.
- La influencia o *leverage*, con valor igual a 0.02. Es una métrica muy parecida a lift, pero mide la proporción de que los conjuntos de ítems de las reglas no estén correlacionados entre sí. En este caso, su valor es muy pequeño. Eso quiere decir que los conjuntos de ítems que forman la regla están muy correlacionados. Así, si el valor de lift es alto el valor de leverage tiende a ser bajo, y al revés.
- La convicción o *conviction*, con valor igual a 33.33. Es una métrica que nos indica el grado de implicación de la regla dentro del dataset. Y el grado de implicación en este caso es alto y puede señalar un aspecto relevante del dataset. Muchas de las personas que no sobrevivieron fueron personas adultas de la tripulación, concretamente un 33.33%.