Trabajo Reglas Asociación

Juan José Sierra González

27 de enero de 2019

En este trabajo final sobre **reglas de asociación** se ha seleccionado un dataset sobre el que trabajar obteniendo reglas que resulten de interés y que aporten información a los datos que ya se poseen. A continuación se estudiarán los itemsets frecuentes, maximales y cerrados, se obtendrán las reglas correspondientes al mínimo soporte establecido, y en general se hará uso de las técnicas aprendidas durante el curso para así alcanzar el objetivo de la práctica.

Lectura de los datos

El dataset que se ha escogido para desarrollar este trabajo ha sido *Contraceptive Method Choice* (https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice), obtenido del repositorio de datasets de la Universidad de California-Irvine. Este conjunto de datos determina el método anticonceptivo elegido por una pareja, que puede ser ninguno, a corto plazo o a largo plazo. Las mujeres que forman parte de este conjunto son mujeres casadas que no estaban embarazadas o no lo sabían en el momento de la recopilación de los datos, según la información del propio dataset.

Para dar un nombre adecuado a cada variable se va a utilizar la información extraída del archivo `cmc.names`. De esta forma también se le dará una nomenclatura adecuada a los valores que puede tomar cada una de las características, convirtiéndolas en un factor cuando sea necesario. Se tienen en cuenta datos sobre cada uno de los integrantes de la pareja y sobre la familia en general.

Estas son las variables que contiene este dataset:

- **Wife's age** → La edad de la mujer. Se divide en los siguientes rangos: muy joven (16-21), joven (22-29), mediana edad (30-39) y adulta (40+).
- **Wife's education** → La formación académica de la mujer. Se divide en 4 rangos: baja, media-baja, media-alta y alta.
- **Husband's education** → La formación académica del hombre. Se divide en 4 rangos: baja, media-baja, media-alta y alta.
- **Children** → La cantidad de hijos que ha tenido la pareja. Se divide en estos grupos: 0, 1-2, 3-4, 5-8 y 9+.
- **Wife's religion** → La mujer es de religión musulmana o no.
- **Wife working** → La mujer se encuentra trabajando o no.
- **Husband's occupation** → Cualificación del trabajo del hombre. Cabe destacar que para esta variable **no se especifica** en la información del dataset el significado de los distintos valores. En base a los primeros resultados que se pueden visualizar acerca del dataset (a continuación) se ha optado por interpretar los valores de 1 a 4 como 1 el más cualificado y 4 el menos.
- **Standard-of-living** → Nivel de vida de la familia. Dividido en bajo, medio-bajo, medio-alto y alto.
- **Media exposure** → Exposición a los medios (tienen buena cobertura mediática o no).
- **Contraceptive method** → Método anticonceptivo utilizado. Dividido en los siguientes grupos: no utilizan, utilizan métodos a corto plazo y utilizan métodos a largo plazo.

Se puede echar un vistazo a los valores más comunes del dataset utilizando la función `summary`.

```
summary(Contraceptive)
```

```
##       Wife's age  Wife's education Husband's education Children
##  Very young:112   Low     :152     Low     : 44        0  : 97
##  Young     :494   Mid-low :334     Mid-low :178        1-2:552
##  Mid-age   :527   Mid-high:410     Mid-high:352        3-4:456
##  Adult     :340   High    :577     High    :899        5-8:276
##                                                       9+ : 92
##   Wife's religion Wife working Husband's occupation Standard-of-living
##  Non-Islam: 220   Yes: 369     High    :436         Low     :129
##  Islam    :1253   No :1104     Mid-high:425         Mid-low :229
##                                Mid-low :585         Mid-high:431
##                                Low     : 27         High    :684
##
##   Media exposure Contraceptive method
##  Good    :1364   No-use    :629
##  Not good: 109   Long-term :333
##                  Short-term:511
##
##
```

En el resumen obtenido se observan los valores más frecuentes para cada variable. En general se puede observar que se trata de una población mayoritariamente musulmana donde hay bastante propensidad a tener hijos. El resto de detalles del dataset se analizarán cuando se obtengan los itemsets frecuentes y las reglas.

Ahora los datos son más legibles e interpretables, y se puede proceder a realizar un análisis más profundo y comenzar extrayendo itemsets frecuentes.

Transacciones e itemsets de interés

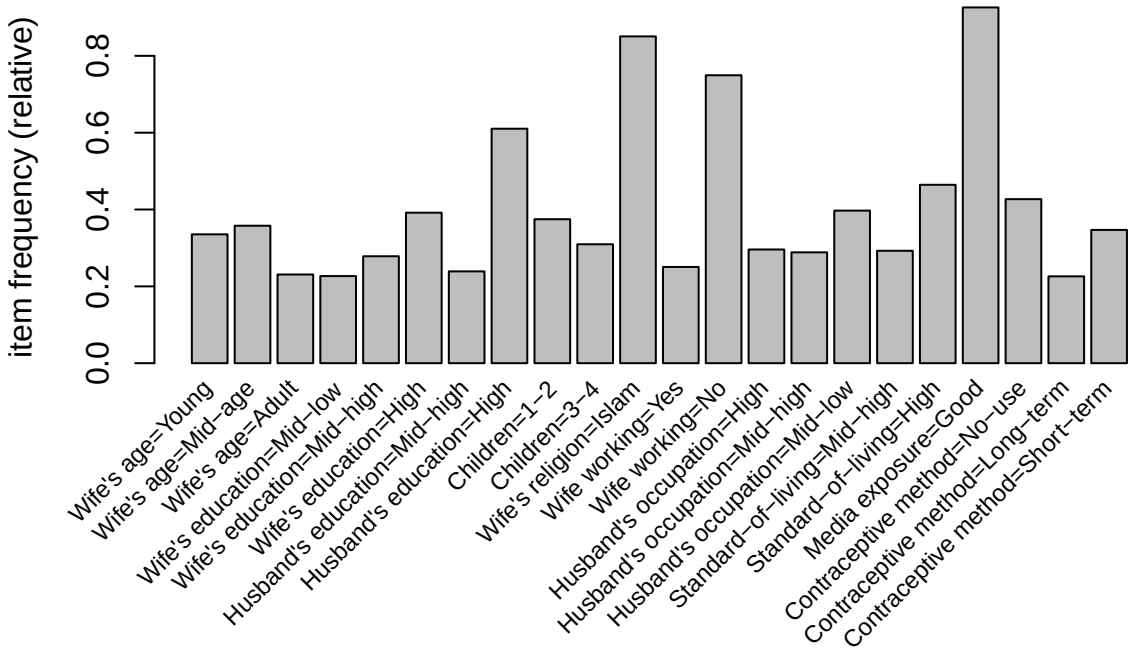
En primer lugar se van a crear las transacciones para la base de datos ya modificada. Para ello se utiliza la función `as`, indicando `"transactions"` como parámetro a convertir.

Lo primero que se observa con el resumen de las transacciones es una lista de los itemsets más frecuentes. En este caso lo que más predomina sobre todo es una buena exposición a los medios, es decir, casi ninguna familia vive alejada de la actualidad y es conocedora de lo que ocurre a su alrededor y en otras partes del mundo. Como se comentaba antes, la mayor parte de las mujeres encuestadas es de religión musulmana y no trabajan, seguramente porque se trate de un núcleo de población conservadora que relega a las mujeres al trabajo de casa.

Aparte, los hombres son los que tienen una formación académica alta, dado que se trata de una sociedad que ofrece mayores oportunidades a los hombres que a las mujeres. El siguiente itemset más frecuente es un nivel de vida alto, que ya afecta a menos del 50% de la población, pero aunque no supone una mayoría abismal sí que se puede afirmar que en este núcleo poblacional se perpetúan las tradiciones de la religión musulmana y hay pocas familias con dificultades económicas.

A continuación se muestra un gráfico con los itemsets frecuentes (por encima de un 20% de soporte).

```
itemFrequencyPlot(ContraceptiveT, support = 0.2, cex.names = 0.75)
```



De este gráfico se pueden confirmar algunas de las teorías anteriores. Una de ellas es que el nivel de vida de la población es en general alto; el rango "alto" y el "medio-alto" se reparten aproximadamente un 80% de las familias de las mujeres encuestadas.

Otra podría ser que los hombres acceden a una mejor educación que las mujeres, o al menos de forma más habitual; se puede observar que el 80% de los hombres han recibido una nivel de educación alto o medio-alto mientras que en las mujeres apenas llega al 70%. Dado que el nivel de formación es en general bastante alto tanto para hombres como para mujeres se puede asegurar que la encuesta se ha realizado en una zona de buenas condiciones económicas, lo que también es respaldado por el alto nivel de vida que aparece en el gráfico.

Además existen pocas mujeres muy jóvenes en la encuesta (la mayoría se encuentra entre los 30 y los 40) lo que también explica que haya mayor nivel de vida de forma general en la población, ya que influye favorablemente al nivel de vida que una pareja tenga edad de trabajar y haya adquirido experiencia como para recibir un buen salario.

Por último, cabe destacar el curioso caso de "Husband's occupation" ya que, como se indicó anteriormente, su significado no viene especificado en la descripción del dataset. Si se asume que sigue la nomenclatura que se ha decidido para este estudio ocurre que la mayoría de hombres de la población tienen un trabajo muy cualificado. Esto entra dentro de lo esperado considerando el alto nivel de vida que se ha observado que tiene la población.

Pero, ¿y si el significado semántico de la variable fuese justo el contrario y hubiese muchísimos hombres con trabajos poco cualificados? Existirían dos explicaciones para este hecho: la primera es que se tratase de familias adineradas poseedoras de tierras en las que el hombre ejerce un trabajo poco cualificado pero a su vez muy lucrativo. Indonesia es una tierra donde abundan los recursos naturales, y trabajar estos recursos podría permitir que estas familias gozasen de un alto nivel de vida. La segunda es que en la Indonesia de 1987 se concibiese de forma diferente a lo que se piensa hoy en día el concepto de "alto nivel de vida" y "alto nivel de

estudios", y realmente no fuera tan extraño encontrarse el caso de una persona con una buena formación académica ejerciendo un puesto de trabajo poco cualificado.

Asumiendo que el caso que se ha decidido en el apartado del análisis del dataset es el correcto, se va a seguir optando por la solución que parece la más coherente.

Reglas con Apriori

Ahora que el dataset y sus itemsets frecuentes han sido analizados y se tiene una visión general de la población bastante cercada, se puede utilizar el **algoritmo Apriori** para empezar a extraer reglas de asociación que aporten nueva información hasta ahora desconocida. Para que las reglas sean suficientemente significativas se van a fijar los parámetros mínimos de soporte y confianza a 0.1 y 0.8 respectivamente.

De estas reglas obtenidas, tiene sentido mirar aquellas que no tienen un soporte especialmente alto ni se cumplen siempre (una confianza de 1 no es interesante porque es casi seguro que esté reflejando información obvia y lógica).

```
reglas = apriori(ContraceptiveT, parameter = list(support = 0.1,
                                  confidence = 0.8, minlen = 2))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      2
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 147
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[34 item(s), 1473 transaction(s)] done [0.00s].
## sorting and recoding items ... [27 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 done [0.00s].
## writing ... [598 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
mejoresReglas = subset(reglas,
                subset = support < 0.4 & confidence < 1 & lift > 1)

# Eliminar reglas redundantes
esSubconjunto = is.subset(mejoresReglas, mejoresReglas)
esSubconjunto[lower.tri(esSubconjunto, diag = TRUE)] = FALSE
redundantes = colSums(esSubconjunto, na.rm=TRUE) >= 1
mejoresReglas = mejoresReglas[!redundantes]
```

Lo que más llama la atención de las reglas generadas (eliminando aquellas redundantes o que son subconjuntos de otras reglas) es que en casi todos los consecuentes aparecen las variables "Media Exposure" y "Wife's religion", con los respectivos valores "Good" e "Islam", ya que son los predominantes para estos atributos. Como se pudo observar en el resumen del dataset que se mostró anteriormente, estas variables son binarias y la inmensa mayoría de la población sigue una de las dos alternativas; en este caso se trata de un núcleo con

buen acceso a la cobertura mediática y con una población musulmana dominante. La presencia de estas dos características provoca que casi cualquier regla tenga una de ellas en el consecuente, relegando muchas de esas reglas a simples vueltas de tuerca al anterior gráfico mostrado volviendo a aportar la misma información.

No obstante, si uno se limita a comprobar aquellas reglas simples (sólo una condición en el antecedente) que no están afectadas por estas variables puede encontrar algunos detalles interesantes:

- Las mujeres que **no son musulmanas** están casadas con hombres con una **alta formación académica**.
- Los hombres que tienen un trabajo **altamente cualificado** tienen una **alta formación académica**.
- Las mujeres con una **alta formación académica** están casadas con hombres de igual formación.