Curso: Data Science & Business Intelligence
Unidade Curricular: Data Mining
Docente: Roberto Vita

TÓPICO 3

Adult Dataset (*UCI*)

- Regras de Associação

Trabalho realizado por:

Marta Lobo

Marta Barros

O dataset selecionado *Adult* foi cedido pela UCI (http://archive.ics.uci.edu/ml/datasets/Adult) e originalmente incluía 14 variáveis num total de 48842 dados (**Figura 1**). Para simplificar a nossa análise, destas optamos por manter apenas a idade, sexo, raça, ocupação e o salário, nas quais não existem *missing values* (**Figura 2**).

```
> colnames(adult) <- c('age', 'workclass', 'fnlwgt', 'educatoin','educatoin_num', 'marital_s
tatus', 'occupation', 'relationship', 'race', 'sex','capital_gain', 'capital_loss', 'hours_p
er_week', 'native_country', 'income')
> adultrules<-adult
> adultdata<-adult
> summary(adultdata)
      age          workclass             fnlwgt          educatoin          educatoin_num
 Min.   :17.00   Length:32561       Min.   :  12285   Length:32561       Min.   : 1.00
 1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character   1st Qu.: 9.00
 Median :37.00   Mode  :character   Median : 178356   Mode  :character   Median :10.00
 Mean   :38.58                      Mean   : 189778                      Mean   :10.08
 3rd Qu.:48.00                      3rd Qu.: 237051                      3rd Qu.:12.00
 Max.   :90.00                      Max.   :1484705                      Max.   :16.00
 marital_status      occupation        relationship          race
 Length:32561       Length:32561       Length:32561       Length:32561
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



     sex             capital_gain    capital_loss    hours_per_week  native_country
 Length:32561       Min.   :    0   Min.   :   0.0   Min.   : 1.00   Length:32561
 Class :character   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00   Class :character
 Mode  :character   Median :    0   Median :   0.0   Median :40.00   Mode  :character
                    Mean   : 1078   Mean   :  87.3   Mean   :40.44
                    3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
                    Max.   :99999   Max.   :4356.0   Max.   :99.00
    income
 Length:32561
 Class :character
 Mode  :character
```

Figura 1. Sumário da estatística do dataset original.

Os dados foram posteriormente transformados em *data.frame* e podem ser inspecionados através do comando *class*() para confirmar essa mesma transformação (**Figura 2**).

O objetivo é perceber com base nas variáveis sexo, raça, ocupação e idade o salário que mais frequentemente se encontra associado a esse conjunto de caraterísticas. Para isso utilizaremos um modelo não supervisionado, mais especificamente, regras de associação, onde se aplicam as bibliotecas *arules* e *arulesViz*.

```
> class(adultdata)
[1] "data.frame"
> adultdata$educatoin <- NULL
> adultdata$fnlwgt <- NULL
> adultdata$educatoin_num <- NULL
> adultdata$workclass <- NULL
> adultdata$marital_status <- NULL
> adultdata$relationship <- NULL
> adultdata$capital_gain <- NULL
> adultdata$capital_loss <- NULL
> adultdata$hours_per_week <- NULL
> adultdata$native_country <- NULL
```

Figura 2. Transformação do dataset, através da remoção de variáveis (NULL) e verificação do tipo data.frame associado ao novo dataset.

Além do mencionado, a variável idade foi discretizada em 4 níveis via método das frequências, passando de variável numérica a categórica com quatro etiquetas de classificação atribuídas: '*Young*', '*Middle-aged*', '*Senior*' e '*Old*' (**Figura 3**).

```
> adultdata$age<-discretize(adultdata$age, method = "frequency", breaks = 4, labels = c("Young",
 "Middle-aged", "Senior", "Old"))
> adultdata$age
   [1] Senior      Old         Senior      Old         Middle-aged Senior      Old
   [8] Old         Middle-aged Senior      Senior      Middle-aged Young       Middle-aged
  [15] Senior      Middle-aged Young       Middle-aged Senior      Senior      Senior
  [22] Old         Middle-aged Senior      Old         Old         Young       Old
```

Figura 3. Discretização da variável idade.

De seguida, os nossos dados foram ajustados a classe de transações para tornar possível a aplicação de regras como se pode ver pela aplicação do algoritmo *a priori* na imagem abaixo.

O *support* define-se pelo número de transações que suportam a condição, isto é, número de eventos em que a regra de associação acontece, enquanto que a confiança é o parâmetro que define quão sólida essa condição é.

```
> Adult
transactions in sparse format with
 32561 transactions (rows) and
 28 items (columns)
> rules <- apriori(adultdata, parameter = list(support = 0.004, confidence = 0.1,minlen =2))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.1    0.1    1 none FALSE            TRUE       5   0.004      2     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 130

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[28 item(s), 32561 transaction(s)] done [0.01s].
sorting and recoding items ... [27 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [2399 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Figura 4. Criação de regras pelo algoritmo *a priori*, considerando um *support* = 0.004 e *confiança* = 0.1.

As regras foram organizadas pelo seu *lift*, indicador do número de vezes que uma condição de facto aconteceu face à chance estimada da mesma acontecer.

Conforme se pode verificar, na **Figura 5** temos que *lhs* corresponde ao 'left hand side' e *rhs* ao 'right hand side'. Utilizando a linha [1] como exemplo, percebemos que a primeira regra é definida por ocupação → sexo, ou seja, para o suporte e confiança estabelecidos, se a ocupação do sujeito for 'Priv-house-serv', primeira coluna (lhs), qual será o género mais associado (rhs).

```
> sort.rule<-sort(rules,by="lift")
> sort.rule
set of 2399 rules
> top5rules <- head(rules, n = 5, by = "lift")
> inspect(head(rules,5))
    lhs                                rhs              support     confidence coverage
[1] {occupation= Priv-house-serv} => {sex= Female}    0.004330334 0.9463087  0.004576027
[2] {occupation= Priv-house-serv} => {income= <=50K} 0.004545315 0.9932886  0.004576027
[3] {race= Other}                 => {sex= Male}     0.004975277 0.5977860  0.008322840
[4] {race= Other}                 => {income= <=50K} 0.007555051 0.9077491  0.008322840
[5] {race= Amer-Indian-Eskimo}    => {sex= Male}     0.005896625 0.6173633  0.009551304
    lift      count
[1] 2.8607147 141
[2] 1.3083523 148
[3] 0.8932772 162
[4] 1.1956803 246
[5] 0.9225318 192
```

Figura 5. Definição de regras de associação, com um total de 2399.

A **Figura 6** é a exploração visual da tabela obtida anteriormente (**Figura 5**). Deste modo, numa análise geral, percebemos que indivíduos cuja ocupação se enquadre na categoria '*Exec-managerial*' ou '*Prof-specialty*', de raça '*White*', idade entre os 44-90 anos e sexo '*Male*' auferem salários mais altos (>50K). Por outro lado, a associação de raça='*Black*', sexo='*Female*' e ocupação='*Other-service*' para a mesma faixa etária age = [44,90] refletem salários mais baixos (<=50K).

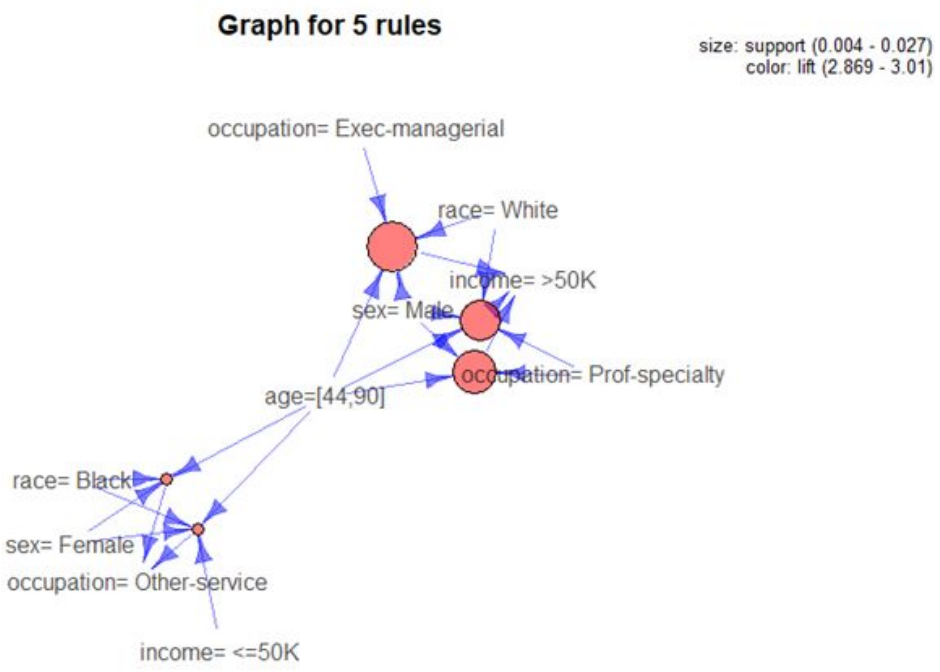

Figura 6. Gráfico de associações mais frequentes entre as variáveis para as 5 regras com *lift* mais elevado.

Utilizando as regras de associação, agora aplicadas a um exemplo pré-definido e não ao dataset completo, estabelecemos *rhs* sex= '*Female*' no sentido de compreender as oito interações mais frequentes (*lhs*) no sentido inverso *rhs* ← *lhs*, organizadas por *lift* (**Figura 7**).

```
> inspect(sort(adult_rules_rhs, by = 'lift')[1:8])
    lhs                                rhs                support confidence    coverage     lift count
[1] {occupation= Priv-house-serv} => {sex= Female} 0.004330334  0.9463087 0.004576027 2.860715   141
[2] {occupation= Priv-house-serv,
     income= <=50K}                => {sex= Female} 0.004299622  0.9459459 0.004545315 2.859618   140
[3] {age=[31,44),
     occupation= Adm-clerical,
     race= Black,
     income= <=50K}                => {sex= Female} 0.004023218  0.8036810 0.005005989 2.429548   131
[4] {occupation= Adm-clerical,
     race= Black,
     income= <=50K}                => {sex= Female} 0.010810479  0.7857143 0.013758791 2.375234   352
[5] {age=[31,44),
     occupation= Adm-clerical,
     race= Black}                  => {sex= Female} 0.004238199  0.7582418 0.005589509 2.292184   138
[6] {occupation= Adm-clerical,
     race= Black}                  => {sex= Female} 0.011271153  0.7489796 0.015048678 2.264184   367
[7] {age=[17,31),
     occupation= Adm-clerical,
     race= Black}                  => {sex= Female} 0.004115353  0.7486034 0.005497374 2.263046   134
[8] {age=[44,90],
     occupation= Adm-clerical,
     income= <=50K}                => {sex= Female} 0.020853168  0.7477974 0.027886121 2.260610   679
```

Figura 7. Regras de associação previstas em *lhs*, tendo por base um *rhs* pré-definido (sex='Female') para um support = 0.004 e confiança = 0.1.

```
> inspect(sort(adult_rules_lhs, by = 'lift')[1:8])
    lhs                rhs                          support    confidence coverage  lift     count
[1] {sex= Female} => {occupation= Adm-clerical}   0.07791530 0.2355399  0.3307945 2.034327 2537
[2] {sex= Female} => {occupation= Other-service}  0.05528086 0.1671154  0.3307945 1.651425 1800
[3] {sex= Female} => {race= Black}                0.04775652 0.1443691  0.3307945 1.504739 1555
[4] {sex= Female} => {age=[17,31)}                0.13080065 0.3954136  0.3307945 1.217845 4259
[5] {sex= Female} => {income= <=50K}              0.29458555 0.8905394  0.3307945 1.173012 9592
[6] {sex= Female} => {occupation= Prof-specialty} 0.04652806 0.1406555  0.3307945 1.106252 1515
[7] {sex= Female} => {occupation= Sales}          0.03878873 0.1172593  0.3307945 1.046049 1263
[8] {}            => {occupation= Other-service}  0.10119468 0.1011947  1.0000000 1.000000 3295
>
```

Figura 8. Aplicação do mesmo exemplo sex='*Female*', no sentido inverso lhs → rhs.

Para aumentar a eficiência do nosso modelo é importante remover informação excessiva e/ou redundante. Assim, todas as regras que se repitam podem ser reduzidas fazendo uso de `subsets <- which(colSums(is.subset(rules, rules)) > 1)` e `maximal_rules <- rules[-subsets]`. Aplicando o código referido, é possível uma redução de 2076 para 87 regras (**Figura 9**).

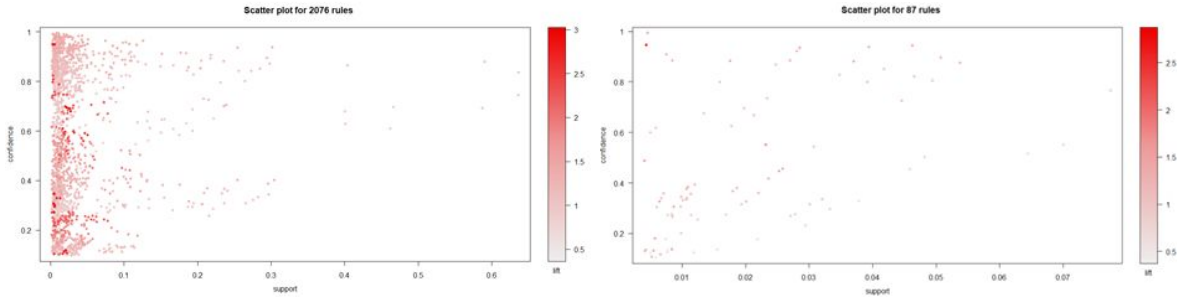

Figura 9. Gráfico de dispersão de regras atendendo ao *support*, *lift* e confiança.

Conclusões

Com base na análise efetuada, podemos afirmar com 95% de confiança que partindo de uma *occupation=Priv-house-serv,* esta estará provavelmente associada a uma mulher, existindo uma relação de dependência muito forte (lift>1).

O facto de termos valores de *confidence* tão altos permite-nos tirar conclusões bastante precisas relativamente a estas variáveis.

Por outro lado se *occupation=Adm-clerical, income<=50k, race=Black* e *age* entre 31 e 44 anos, então afirmamos com 80% de confiança que esta observação corresponde a uma mulher.

Fazendo o raciocínio inverso, ou seja, ao perceber a relação existente entre ser mulher com as restantes variáveis, apesar de termos bons valores de suporte, temos baixos níveis de confiança.

Por exemplo, se o indivíduo se trata de uma mulher cuja *occupation=Adm-clerical*, obtemos um valor de suporte de 0,08, com um valor de confiança de 24%. O lift=2,03 indica uma relação de dependência muito elevada.

Outros fatores como ser mulher e auferir *income<=50k* exibem uma relação de dependência com valor de confiança de aproximadamente 90%.

Concluímos com base neste exemplo, que aplicando as regras *"right hand side"* encontramos relações mais fortes, que nos permitem perceber se aquela observação corresponde a determinadas características do indivíduo.