Assignment 9

Dipesh Poudel

1/8/2022

# Assignment 9: Association Rule Mining

**Instructions** Use the “AdultUCI” data available in the “arules” package and do as follows in R Script.

## 1. Attach the AdultUCI data in R

library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

data(AdultUCI)

1. Check the class, structure, dimension, head and tail of the attached data and write interpretations

### Checking the Class of the dataset

class(AdultUCI)

## [1] "data.frame"

### Checking the Structure of the dataset

str(AdultUCI)

## 'data.frame': 48842 obs. of 15 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ workclass : Factor w/ 8 levels "Federal-gov",..: 7 6 4 4 4 4 4 6 4 4 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Ord.factor w/ 16 levels "Preschool"<"1st-4th"<..: 14 14 9 7 14 15 5 9 15 14 ...  
## $ education-num : int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital-status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...  
## $ occupation : Factor w/ 14 levels "Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...  
## $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ capital-gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital-loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours-per-week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native-country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 5 39 23 39 39 39 ...  
## $ income : Ord.factor w/ 2 levels "small"<"large": 1 1 1 1 1 1 1 2 2 2 ...

### Checking the Dimesions of the dataset

dim(AdultUCI)

## [1] 48842 15

head(AdultUCI)

## age workclass fnlwgt education education-num marital-status  
## 1 39 State-gov 77516 Bachelors 13 Never-married  
## 2 50 Self-emp-not-inc 83311 Bachelors 13 Married-civ-spouse  
## 3 38 Private 215646 HS-grad 9 Divorced  
## 4 53 Private 234721 11th 7 Married-civ-spouse  
## 5 28 Private 338409 Bachelors 13 Married-civ-spouse  
## 6 37 Private 284582 Masters 14 Married-civ-spouse  
## occupation relationship race sex capital-gain capital-loss  
## 1 Adm-clerical Not-in-family White Male 2174 0  
## 2 Exec-managerial Husband White Male 0 0  
## 3 Handlers-cleaners Not-in-family White Male 0 0  
## 4 Handlers-cleaners Husband Black Male 0 0  
## 5 Prof-specialty Wife Black Female 0 0  
## 6 Exec-managerial Wife White Female 0 0  
## hours-per-week native-country income  
## 1 40 United-States small  
## 2 13 United-States small  
## 3 40 United-States small  
## 4 40 United-States small  
## 5 40 Cuba small  
## 6 40 United-States small

tail(AdultUCI)

## age workclass fnlwgt education education-num marital-status  
## 48837 33 Private 245211 Bachelors 13 Never-married  
## 48838 39 Private 215419 Bachelors 13 Divorced  
## 48839 64 <NA> 321403 HS-grad 9 Widowed  
## 48840 38 Private 374983 Bachelors 13 Married-civ-spouse  
## 48841 44 Private 83891 Bachelors 13 Divorced  
## 48842 35 Self-emp-inc 182148 Bachelors 13 Married-civ-spouse  
## occupation relationship race sex capital-gain  
## 48837 Prof-specialty Own-child White Male 0  
## 48838 Prof-specialty Not-in-family White Female 0  
## 48839 <NA> Other-relative Black Male 0  
## 48840 Prof-specialty Husband White Male 0  
## 48841 Adm-clerical Own-child Asian-Pac-Islander Male 5455  
## 48842 Exec-managerial Husband White Male 0  
## capital-loss hours-per-week native-country income  
## 48837 0 40 United-States <NA>  
## 48838 0 36 United-States <NA>  
## 48839 0 40 United-States <NA>  
## 48840 0 50 United-States <NA>  
## 48841 0 40 United-States <NA>  
## 48842 0 60 United-States <NA>

The AdultUCI is loaded as a dataframe. There are 48842 rows(observations) and 15 columns (variables) in the dataset. Variables age fnlwgt,education-num,capital-gain,capital-loss,hours-per-week have integers values. Variables workclass, education, marital-status, occupation, race,sex, native-country, income and relationship are factors out of them income and education are ordered factors.

## 3. Remove “fnlwgt” and “education-num” variables from the attached data and explain the logic you have used here

# Changing dash(-) sign in column names to underscore(\_)  
names(AdultUCI)<-gsub("-","\_",names(AdultUCI))  
# Removing the said variables  
AdultUCI<-subset(AdultUCI,select=c(-education\_num,-fnlwgt))

First of all the minus(-) sign in column name was replaced with underscore(\_) sign since r interprets the minus sign as a operator. Then the columns fnlwgt and education-num variable (changed to education\_num) was removed using subset command.

## 4. Convert “age” as ordered factor variables with cuts at 15, 25, 45, 65 and 100 and label it as “Young”, “Middle-aged”, “Senior” and “Old”

age\_labels<-c("Young","Middle-aged","Senior","old")  
AdultUCI$age<-cut(AdultUCI$age,breaks = c(15,25,45,65,100),labels = age\_labels)  
AdultUCI$age<-factor(AdultUCI$age,ordered = T,labels = age\_labels)  
str(AdultUCI$age)

## Ord.factor w/ 4 levels "Young"<"Middle-aged"<..: 2 3 2 3 2 2 3 3 2 2 ...

## 5. Convert the “hours-per-week” as ordered factor variable with cuts at 0, 25, 40, 60, 168 and label it as “Part-time”, “Full-time”, “Over-time” and “Workaholic”

hours\_labels<-c("Part-time","Full-time","Over-time","Workaholic")  
AdultUCI$hours\_per\_week<-cut(AdultUCI$hours\_per\_week,breaks = c(0,25,40,60,168),labels = hours\_labels)  
AdultUCI$hours\_per\_week<-factor(AdultUCI$hours\_per\_week,ordered = T,labels = hours\_labels)  
str(AdultUCI$hours\_per\_week)

## Ord.factor w/ 4 levels "Part-time"<"Full-time"<..: 2 1 2 2 2 2 1 3 3 2 ...

## 6. Convert the “capital-gain” as ordered factor variable with cuts at –Inf, 0, median and Inf and label it as “None”, “Low” and “High”

capital\_gain\_labels<-c("None","Low","High")  
AdultUCI$capital\_gain<-cut(AdultUCI$capital\_gain,breaks = c(-Inf,0,median(AdultUCI[AdultUCI$capital\_gain>0,]$capital\_gain),Inf),labels = capital\_gain\_labels)  
AdultUCI$capital\_gain<-factor(AdultUCI$capital\_gain,ordered = T,labels = capital\_gain\_labels)  
str(AdultUCI$capital\_gain)

## Ord.factor w/ 3 levels "None"<"Low"<"High": 2 1 1 1 1 1 1 1 3 2 ...

## 7. Convert the “capital-loss” as ordered factor variable with cuts at –Inf, 0, median and Inf and label it as “None”, “Low” and “High”

capital\_loss\_labels<-c("None","Low","High")  
AdultUCI$capital\_loss<-cut(AdultUCI$capital\_loss,breaks = c(-Inf,0,median(AdultUCI[AdultUCI$capital\_loss>0,]$capital\_loss),Inf),labels = capital\_loss\_labels)  
AdultUCI$capital\_loss<-factor(AdultUCI$capital\_loss,ordered = T,labels = capital\_loss\_labels)  
str(AdultUCI$capital\_loss)

## Ord.factor w/ 3 levels "None"<"Low"<"High": 1 1 1 1 1 1 1 1 1 1 ...

## 8. Create transactions of AdultUCI data as “Adult” and check it with “Adult” command

Adult<-transactions(AdultUCI)  
Adult

## transactions in sparse format with  
## 48842 transactions (rows) and  
## 115 items (columns)

## 9. et summary of the “Adult” and interpret it critically

summary(Adult)

## transactions as itemMatrix in sparse format with  
## 48842 rows (elements/itemsets/transactions) and  
## 115 columns (items) and a density of 0.1089939   
##   
## most frequent items:  
## capital\_loss=None capital\_gain=None   
## 46560 44807   
## native\_country=United-States race=White   
## 43832 41762   
## workclass=Private (Other)   
## 33906 401333   
##   
## element (itemset/transaction) length distribution:  
## sizes  
## 9 10 11 12 13   
## 19 971 2067 15623 30162   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9.00 12.00 13.00 12.53 13.00 13.00   
##   
## includes extended item information - examples:  
## labels variables levels  
## 1 age=Young age Young  
## 2 age=Middle-aged age Middle-aged  
## 3 age=Senior age Senior  
##   
## includes extended transaction information - examples:  
## transactionID  
## 1 1  
## 2 2  
## 3 3

**From Summary:** We have 48842 rows (elements/itemsets/transactions) and 115 columns (items). The most frequent most items have capital\_loss=None, capital\_gain=None.

## 10. Inspect head and tail of the “Adult” and interpret them carefully

inspect(head(Adult))

## items transactionID  
## [1] {age=Middle-aged,   
## workclass=State-gov,   
## education=Bachelors,   
## marital\_status=Never-married,   
## occupation=Adm-clerical,   
## relationship=Not-in-family,   
## race=White,   
## sex=Male,   
## capital\_gain=Low,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States,   
## income=small} 1  
## [2] {age=Senior,   
## workclass=Self-emp-not-inc,   
## education=Bachelors,   
## marital\_status=Married-civ-spouse,   
## occupation=Exec-managerial,   
## relationship=Husband,   
## race=White,   
## sex=Male,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Part-time,   
## native\_country=United-States,   
## income=small} 2  
## [3] {age=Middle-aged,   
## workclass=Private,   
## education=HS-grad,   
## marital\_status=Divorced,   
## occupation=Handlers-cleaners,   
## relationship=Not-in-family,   
## race=White,   
## sex=Male,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States,   
## income=small} 3  
## [4] {age=Senior,   
## workclass=Private,   
## education=11th,   
## marital\_status=Married-civ-spouse,   
## occupation=Handlers-cleaners,   
## relationship=Husband,   
## race=Black,   
## sex=Male,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States,   
## income=small} 4  
## [5] {age=Middle-aged,   
## workclass=Private,   
## education=Bachelors,   
## marital\_status=Married-civ-spouse,   
## occupation=Prof-specialty,   
## relationship=Wife,   
## race=Black,   
## sex=Female,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=Cuba,   
## income=small} 5  
## [6] {age=Middle-aged,   
## workclass=Private,   
## education=Masters,   
## marital\_status=Married-civ-spouse,   
## occupation=Exec-managerial,   
## relationship=Wife,   
## race=White,   
## sex=Female,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States,   
## income=small} 6

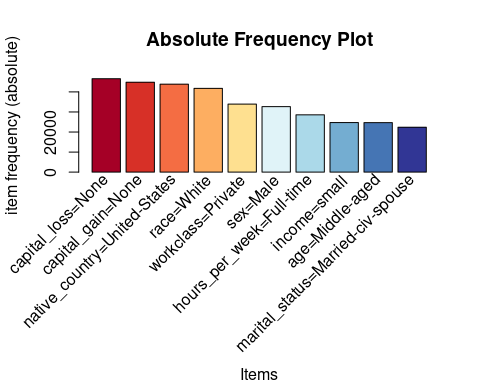
inspect(tail(Adult))

## items transactionID  
## [1] {age=Middle-aged,   
## workclass=Private,   
## education=Bachelors,   
## marital\_status=Never-married,   
## occupation=Prof-specialty,   
## relationship=Own-child,   
## race=White,   
## sex=Male,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States} 48837  
## [2] {age=Middle-aged,   
## workclass=Private,   
## education=Bachelors,   
## marital\_status=Divorced,   
## occupation=Prof-specialty,   
## relationship=Not-in-family,   
## race=White,   
## sex=Female,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States} 48838  
## [3] {age=Senior,   
## education=HS-grad,   
## marital\_status=Widowed,   
## relationship=Other-relative,   
## race=Black,   
## sex=Male,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States} 48839  
## [4] {age=Middle-aged,   
## workclass=Private,   
## education=Bachelors,   
## marital\_status=Married-civ-spouse,   
## occupation=Prof-specialty,   
## relationship=Husband,   
## race=White,   
## sex=Male,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Over-time,   
## native\_country=United-States} 48840  
## [5] {age=Middle-aged,   
## workclass=Private,   
## education=Bachelors,   
## marital\_status=Divorced,   
## occupation=Adm-clerical,   
## relationship=Own-child,   
## race=Asian-Pac-Islander,   
## sex=Male,   
## capital\_gain=Low,   
## capital\_loss=None,   
## hours\_per\_week=Full-time,   
## native\_country=United-States} 48841  
## [6] {age=Middle-aged,   
## workclass=Self-emp-inc,   
## education=Bachelors,   
## marital\_status=Married-civ-spouse,   
## occupation=Exec-managerial,   
## relationship=Husband,   
## race=White,   
## sex=Male,   
## capital\_gain=None,   
## capital\_loss=None,   
## hours\_per\_week=Over-time,   
## native\_country=United-States} 48842

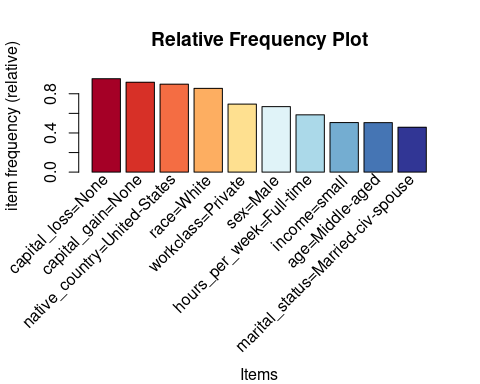
Each row is given a transactionID and values of each row is converted into a list of items in a transaction.

## 11 Create absolute and relative item frequency plot and color it with RColorBrewer package

library(RColorBrewer)  
palette = brewer.pal(10,'RdYlBu');  
# Absolute Frequency Plot  
itemFrequencyPlot(Adult,  
 type="absolute",  
 topN=10,  
 col=palette,  
 main="Absolute Frequency Plot",  
 xlab="Items"  
 )



itemFrequencyPlot(Adult,  
 type="relative",  
 topN=10,  
 col=palette,  
 main="Relative Frequency Plot",  
 xlab="Items")



## 12. Create an apriori rule as “association.rule” with support = 1%, confidence = 80% and maximum length of the rule as 10. Get summary of this rule and interpret it carefully.

association.rule<-apriori(Adult,  
 parameter = list(supp=0.01,conf=0.8,maxlen=10,target= "rules")  
)

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.01 1  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 488   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[115 item(s), 48842 transaction(s)] done [0.05s].  
## sorting and recoding items ... [67 item(s)] done [0.01s].  
## creating transaction tree ... done [0.04s].  
## checking subsets of size 1 2 3 4 5 6 7 8 9 10

## Warning in apriori(Adult, parameter = list(supp = 0.01, conf = 0.8, maxlen  
## = 10, : Mining stopped (maxlen reached). Only patterns up to a length of 10  
## returned!

## done [0.90s].  
## writing ... [197371 rule(s)] done [0.04s].  
## creating S4 object ... done [0.06s].

summary(association.rule)

## set of 197371 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 1 2 3 4 5 6 7 8 9 10   
## 4 266 3303 15219 37015 53616 48402 27754 9827 1965   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 5.000 6.000 6.318 7.000 10.000   
##   
## summary of quality measures:  
## support confidence coverage lift   
## Min. :0.01001 Min. :0.8000 Min. :0.01001 Min. : 0.8677   
## 1st Qu.:0.01251 1st Qu.:0.8953 1st Qu.:0.01353 1st Qu.: 1.0059   
## Median :0.01708 Median :0.9372 Median :0.01847 Median : 1.0398   
## Mean :0.02726 Mean :0.9283 Mean :0.02949 Mean : 1.2899   
## 3rd Qu.:0.02766 3rd Qu.:0.9669 3rd Qu.:0.02995 3rd Qu.: 1.2160   
## Max. :0.95328 Max. :1.0000 Max. :1.00000 Max. :20.6826   
## count   
## Min. : 489   
## 1st Qu.: 611   
## Median : 834   
## Mean : 1331   
## 3rd Qu.: 1351   
## Max. :46560   
##   
## mining info:  
## data ntransactions support confidence  
## Adult 48842 0.01 0.8  
## call  
## apriori(data = Adult, parameter = list(supp = 0.01, conf = 0.8, maxlen = 10, target = "rules"))

We got 197371 rules. We can also see number of rules with different numbers of items. The rules with support count less than 488 was discarded.

## 13. Inspect the first 10 rules and interpret it critically.

inspect(head(association.rule,10))

## lhs rhs support confidence coverage lift count  
## [1] {} => {race=White} 0.85504279 0.8550428 1.00000000 1.0000000 41762  
## [2] {} => {native\_country=United-States} 0.89742435 0.8974243 1.00000000 1.0000000 43832  
## [3] {} => {capital\_gain=None} 0.91738668 0.9173867 1.00000000 1.0000000 44807  
## [4] {} => {capital\_loss=None} 0.95327792 0.9532779 1.00000000 1.0000000 46560  
## [5] {education=5th-6th} => {capital\_loss=None} 0.01009377 0.9685658 0.01042136 1.0160372 493  
## [6] {education=Doctorate} => {race=White} 0.01076942 0.8855219 0.01216166 1.0356463 526  
## [7] {education=Doctorate} => {capital\_loss=None} 0.01076942 0.8855219 0.01216166 0.9289231 526  
## [8] {marital\_status=Married-spouse-absent} => {capital\_gain=None} 0.01218214 0.9474522 0.01285779 1.0327730 595  
## [9] {marital\_status=Married-spouse-absent} => {capital\_loss=None} 0.01240735 0.9649682 0.01285779 1.0122632 606  
## [10] {education=12th} => {native\_country=United-States} 0.01140412 0.8477930 0.01345154 0.9446958 557

We can see that there are empty rules as well which is not of interest to us so we can remove them.

## 14. Remove the empty rules from the “association.rule” and inspect the first 10 rules with interpretations.

association.rule<-apriori(Adult,  
 parameter = list(supp=0.01,conf=0.8,maxlen=10,minlen=2,target= "rules")  
)

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.01 2  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 488   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[115 item(s), 48842 transaction(s)] done [0.06s].  
## sorting and recoding items ... [67 item(s)] done [0.01s].  
## creating transaction tree ... done [0.04s].  
## checking subsets of size 1 2 3 4 5 6 7 8 9 10

## Warning in apriori(Adult, parameter = list(supp = 0.01, conf = 0.8, maxlen  
## = 10, : Mining stopped (maxlen reached). Only patterns up to a length of 10  
## returned!

## done [0.91s].  
## writing ... [197367 rule(s)] done [0.04s].  
## creating S4 object ... done [0.06s].

inspect(head(association.rule[1:10]))

## lhs rhs support confidence coverage lift count  
## [1] {education=5th-6th} => {capital\_loss=None} 0.01009377 0.9685658 0.01042136 1.0160372 493  
## [2] {education=Doctorate} => {race=White} 0.01076942 0.8855219 0.01216166 1.0356463 526  
## [3] {education=Doctorate} => {capital\_loss=None} 0.01076942 0.8855219 0.01216166 0.9289231 526  
## [4] {marital\_status=Married-spouse-absent} => {capital\_gain=None} 0.01218214 0.9474522 0.01285779 1.0327730 595  
## [5] {marital\_status=Married-spouse-absent} => {capital\_loss=None} 0.01240735 0.9649682 0.01285779 1.0122632 606  
## [6] {education=12th} => {native\_country=United-States} 0.01140412 0.8477930 0.01345154 0.9446958 557

We see that there are no more empty rules. We are only seeing rules with one item on LHS and one on RHS.

## 15. Create a new rule as “capital.gain.rhs.rule” with “capital-gain=None” in the RHS with support of 1%, confidence of 80%, maximum length of 10 and minimum length of 2.

capital.gain.rhs.rule<-apriori(Adult,  
 parameter=list(supp=0.01,conf=0.8,maxlen=10,minlen=2),  
 appearance=list(default="lhs",rhs="capital\_gain=None")  
 )

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.01 2  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 488   
##   
## set item appearances ...[1 item(s)] done [0.00s].  
## set transactions ...[115 item(s), 48842 transaction(s)] done [0.05s].  
## sorting and recoding items ... [67 item(s)] done [0.01s].  
## creating transaction tree ... done [0.05s].  
## checking subsets of size 1 2 3 4 5 6 7 8 9 10

## Warning in apriori(Adult, parameter = list(supp = 0.01, conf = 0.8, maxlen  
## = 10, : Mining stopped (maxlen reached). Only patterns up to a length of 10  
## returned!

## done [1.03s].  
## writing ... [35433 rule(s)] done [0.01s].  
## creating S4 object ... done [0.02s].

## 16. Get summary of this rule and interpret it critically.

summary(capital.gain.rhs.rule)

## set of 35433 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 2 3 4 5 6 7 8 9 10   
## 60 706 3062 7110 9790 8377 4537 1508 283   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 5.000 6.000 6.212 7.000 10.000   
##   
## summary of quality measures:  
## support confidence coverage lift   
## Min. :0.01001 Min. :0.8000 Min. :0.01009 Min. :0.8720   
## 1st Qu.:0.01265 1st Qu.:0.9015 1st Qu.:0.01359 1st Qu.:0.9827   
## Median :0.01744 Median :0.9453 Median :0.01882 Median :1.0304   
## Mean :0.02819 Mean :0.9287 Mean :0.03051 Mean :1.0124   
## 3rd Qu.:0.02864 3rd Qu.:0.9654 3rd Qu.:0.03092 3rd Qu.:1.0523   
## Max. :0.87066 Max. :1.0000 Max. :0.95328 Max. :1.0901   
## count   
## Min. : 489   
## 1st Qu.: 618   
## Median : 852   
## Mean : 1377   
## 3rd Qu.: 1399   
## Max. :42525   
##   
## mining info:  
## data ntransactions support confidence  
## Adult 48842 0.01 0.8  
## call  
## apriori(data = Adult, parameter = list(supp = 0.01, conf = 0.8, maxlen = 10, minlen = 2), appearance = list(default = "lhs", rhs = "capital\_gain=None"))

We filtered out the rules where value of capital\_gain is None in RHS. We have 35433 such rules.

## 17. Create a new rule as “hours.per.week.ft.rule” with “hour-per-week=Full-time” in the RHS with support of 1%, confidence of 80%, maximum length of 10 and minimum length of 2.

hours.per.week.ft.rule<-apriori(Adult,  
 parameter=list(supp=0.01,conf=0.8,maxlen=10,minlen=2),  
 appearance=list(default="lhs",rhs="hours\_per\_week=Full-time")  
 )

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.01 2  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 488   
##   
## set item appearances ...[1 item(s)] done [0.00s].  
## set transactions ...[115 item(s), 48842 transaction(s)] done [0.09s].  
## sorting and recoding items ... [67 item(s)] done [0.01s].  
## creating transaction tree ... done [0.09s].  
## checking subsets of size 1 2 3 4 5 6 7 8 9 10

## Warning in apriori(Adult, parameter = list(supp = 0.01, conf = 0.8, maxlen  
## = 10, : Mining stopped (maxlen reached). Only patterns up to a length of 10  
## returned!

## done [1.26s].  
## writing ... [159 rule(s)] done [0.01s].  
## creating S4 object ... done [0.02s].

## 18. Get summary of this rule and interpret it critically.

summary(hours.per.week.ft.rule)

## set of 159 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 3 4 5 6 7 8   
## 3 16 48 58 29 5   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.000 5.000 6.000 5.686 6.000 8.000   
##   
## summary of quality measures:  
## support confidence coverage lift   
## Min. :0.01001 Min. :0.8000 Min. :0.01216 Min. :1.367   
## 1st Qu.:0.01066 1st Qu.:0.8047 1st Qu.:0.01318 1st Qu.:1.375   
## Median :0.01179 Median :0.8086 Median :0.01456 Median :1.382   
## Mean :0.01237 Mean :0.8089 Mean :0.01530 Mean :1.383   
## 3rd Qu.:0.01331 3rd Qu.:0.8129 3rd Qu.:0.01654 3rd Qu.:1.389   
## Max. :0.01992 Max. :0.8266 Max. :0.02471 Max. :1.413   
## count   
## Min. :489.0   
## 1st Qu.:520.5   
## Median :576.0   
## Mean :604.4   
## 3rd Qu.:650.0   
## Max. :973.0   
##   
## mining info:  
## data ntransactions support confidence  
## Adult 48842 0.01 0.8  
## call  
## apriori(data = Adult, parameter = list(supp = 0.01, conf = 0.8, maxlen = 10, minlen = 2), appearance = list(default = "lhs", rhs = "hours\_per\_week=Full-time"))

We have filtered the rules which contains hours\_per\_week=Full-time. There are 159 such rules. The minimum number of items (lhs+rhs) in the selected rules is 3 we have 3 such rules. Similarly we have 16,48,58,29,5 rules with 4,5,6,7 and 8 items respectively.

## 19. Get new rule of “hours.per.week.ft.rule” as “conf.sort.rule” by sorting “hours.per.week.ft.rule” in descending order by “confidence” and inspect the head and tail rules with critical interpretation.

conf.sort.rule<-sort(hours.per.week.ft.rule,by="confidence",decreasing = TRUE)

inspect(head(conf.sort.rule))

## lhs rhs support confidence coverage lift count  
## [1] {age=Middle-aged,   
## occupation=Adm-clerical,   
## relationship=Unmarried,   
## sex=Female,   
## capital\_gain=None} => {hours\_per\_week=Full-time} 0.01005282 0.8265993 0.01216166 1.412771 491  
## [2] {age=Middle-aged,   
## occupation=Adm-clerical,   
## relationship=Unmarried,   
## capital\_gain=None} => {hours\_per\_week=Full-time} 0.01066705 0.8243671 0.01293968 1.408956 521  
## [3] {age=Middle-aged,   
## occupation=Adm-clerical,   
## relationship=Unmarried,   
## capital\_gain=None,   
## capital\_loss=None} => {hours\_per\_week=Full-time} 0.01042136 0.8236246 0.01265304 1.407687 509  
## [4] {age=Middle-aged,   
## relationship=Unmarried,   
## race=Black,   
## sex=Female,   
## capital\_gain=None} => {hours\_per\_week=Full-time} 0.01029851 0.8218954 0.01253020 1.404732 503  
## [5] {age=Middle-aged,   
## workclass=Private,   
## education=HS-grad,   
## race=Black,   
## capital\_gain=None,   
## capital\_loss=None} => {hours\_per\_week=Full-time} 0.01148602 0.8201754 0.01400434 1.401792 561  
## [6] {age=Middle-aged,   
## education=HS-grad,   
## occupation=Adm-clerical,   
## sex=Female,   
## capital\_gain=None,   
## capital\_loss=None} => {hours\_per\_week=Full-time} 0.01031899 0.8195122 0.01259162 1.400658 504

We have got 6 rules with highest confidence value above the threshold. Each rule has high confidence and lift>1 which means these rules are of high interest to us.

inspect(tail(conf.sort.rule))

## lhs rhs support confidence coverage lift count  
## [1] {occupation=Adm-clerical,   
## relationship=Unmarried} => {hours\_per\_week=Full-time} 0.01756685 0.8003731 0.02194832 1.367947 858  
## [2] {workclass=Private,   
## occupation=Adm-clerical,   
## relationship=Unmarried,   
## sex=Female,   
## capital\_gain=None,   
## capital\_loss=None,   
## native\_country=United-States} => {hours\_per\_week=Full-time} 0.01033946 0.8003170 0.01291921 1.367851 505  
## [3] {occupation=Adm-clerical,   
## relationship=Unmarried,   
## sex=Female,   
## capital\_gain=None,   
## capital\_loss=None,   
## income=small} => {hours\_per\_week=Full-time} 0.01042136 0.8003145 0.01302158 1.367847 509  
## [4] {occupation=Machine-op-inspct,   
## sex=Female,   
## capital\_gain=None,   
## native\_country=United-States} => {hours\_per\_week=Full-time} 0.01031899 0.8000000 0.01289873 1.367309 504  
## [5] {occupation=Machine-op-inspct,   
## sex=Female,   
## capital\_loss=None,   
## native\_country=United-States} => {hours\_per\_week=Full-time} 0.01040088 0.8000000 0.01300111 1.367309 508  
## [6] {workclass=Private,   
## education=HS-grad,   
## race=Black,   
## capital\_gain=None,   
## native\_country=United-States,   
## income=small} => {hours\_per\_week=Full-time} 0.01171123 0.8000000 0.01463904 1.367309 572

We have got 6 rules with lowest confidence value above the threshold. Even though the value of confidence is lower the value of lift>1 which means these rules are also of high interest to us.

1. Plot the “hours.per.week.ft.rule” with arulesViz package with plot, plot with “two-key plot”, engine=”plotly”, method=graph & engine=htmlwidget and parallel coordinate plot and interpret each graph carefully.

library(arulesViz)  
library(plotly)

## Loading required package: ggplot2

##   
## Attaching package: 'plotly'

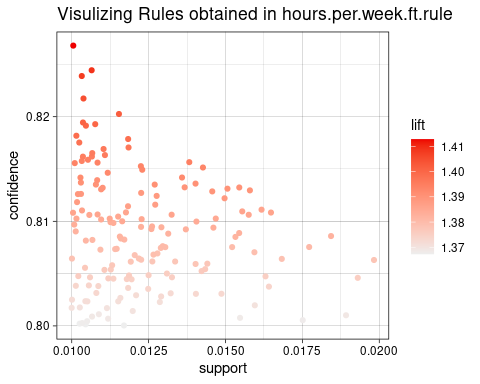
## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

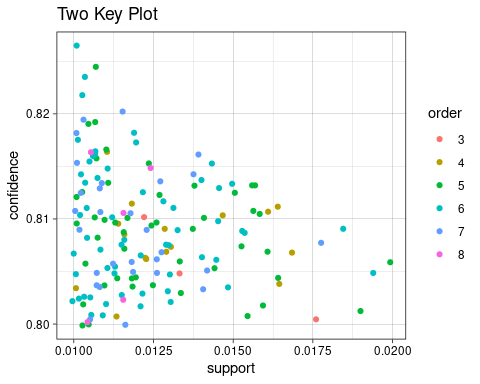
plot(hours.per.week.ft.rule,main="Visulizing Rules obtained in hours.per.week.ft.rule")

## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.

 This is a scatter plot of support and confidence values. In the graph, lighter the color of the point(dot) in the graph lower the value of lift.

plot(hours.per.week.ft.rule,method = "two-key plot",main="Two Key Plot")

## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.



This graph shows the scatter plot of support and confidence. The different color shows the number of items(lhs+rhs) in the rule.

#plot(hours.per.week.ft.rule,engine = "plotly",main="Intercative Plot")

Using engine=“plotly” we have created a interactive scatter plot of support and confidence. The lighter dot means lower value of lift.

#plot(hours.per.week.ft.rule,engine = "htmlwidget",method = "graph")

Using htmlwidget engine we get the graph of items and rules in graph format.

#plot(hours.per.week.ft.rule,method = "paracoord")

Parallel coordinate system is used for visulizing the multi-dimesion in 2d. This graph does not make much sense in association rule mining.