

post1-association rules

Rongyun Tang

December 11, 2018

Census income

research outlines:

1. Objective:task is to determine what kind of a person makes over 50K a year based on census data.
2. Method: statistics and association rules
3. Data Source: this “Adult” dataset was downloaded from UCI Machine Learning website: <http://archive.ics.uci.edu/ml/datasets/Adult>. It is multivariate dataset(including categorical and Integer variables) from social area. A set of reasonably clean records was extracted by the data donors. It is also split into train-test using MLC++ GenCVFiles (2/3, 1/3 random).

=====

Attribute Information:

=====

- Listing of attributes: - salaries are potentially divided into two classes: >50K, <=50K.

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.

1. Data preprocessing

```

training.data <- as.data.frame(read.csv('adult.data'))
test.data <- as.data.frame(read.csv('adult.test', skip=1))
content <- readLines('old.adult.names')

i <- grep("Attribute Information",content) + 2
var.names <- NULL
while(content[i]!="") {
  j <- gregexpr(":", content[i])[[1]][1]
  var.names <- c(var.names, substr(content[i],1,j-1))
  i <- i + 1
}
names(training.data) <- gsub("-", "", var.names)
names(test.data) <- gsub("-", "", var.names)
N.obs <- dim(training.data)[1]
N.var <- dim(training.data)[2]

# show some data information:
print("Traning data examples: ")

```

```
## [1] "Traning data examples: "
```

```
head(training.data)
```

```
##   age      workclass fnlwgt  education educationnum
## 1  50 Self-emp-not-inc 83311 Bachelors           13
## 2  38 Private 215646 HS-grad           9
## 3  53 Private 234721 11th           7
## 4  28 Private 338409 Bachelors           13
## 5  37 Private 284582 Masters           14
## 6  49 Private 160187 9th           5
##   maritalstatus      occupation  relationship  race  sex
## 1 Married-civ-spouse Exec-managerial Husband White Male
## 2 Divorced Handlers-cleaners Not-in-family White Male
## 3 Married-civ-spouse Handlers-cleaners Husband Black Male
## 4 Married-civ-spouse Prof-specialty Wife Black Female
## 5 Married-civ-spouse Exec-managerial Wife White Female
## 6 Married-spouse-absent Other-service Not-in-family Black Female
##   capitalgain capitalloss hoursperweek nativecountry  class
## 1         0         0         13 United-States <=50K
## 2         0         0         40 United-States <=50K
## 3         0         0         40 United-States <=50K
## 4         0         0         40 Cuba <=50K
## 5         0         0         40 United-States <=50K
## 6         0         0         16 Jamaica <=50K
```

```

cat("Number of observations:",N.obs,"
    Number of variables:",N.var,"\n")

```

```
## Number of observations: 32560
##
##      Number of variables: 15
```

```
#install.packages('VIM')  
library("VIM")
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
## VIM is ready to use.
```

```
## Since version 4.0.0 the GUI is in its own package VIMGUI.
```

```
##
```

```
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
```

```
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
```

```
##
```

```
##      sleep
```

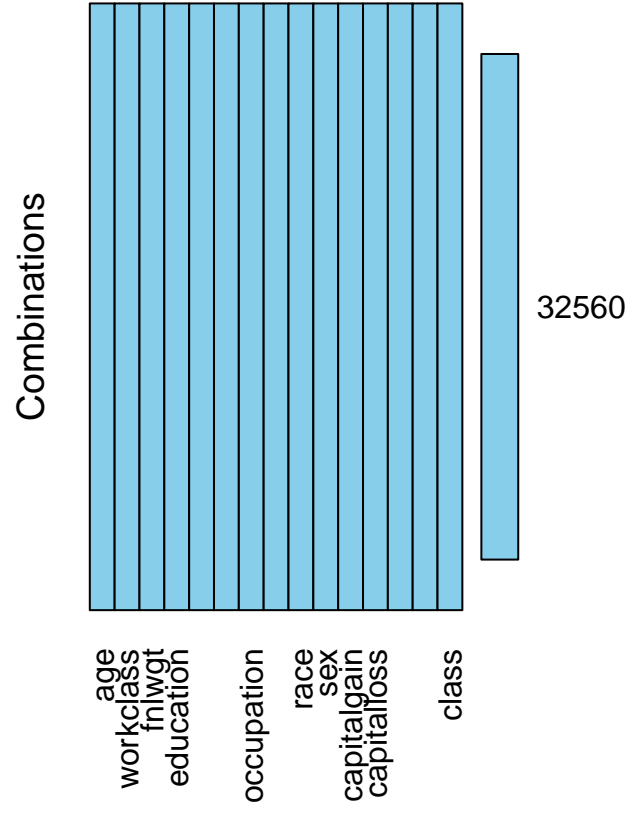
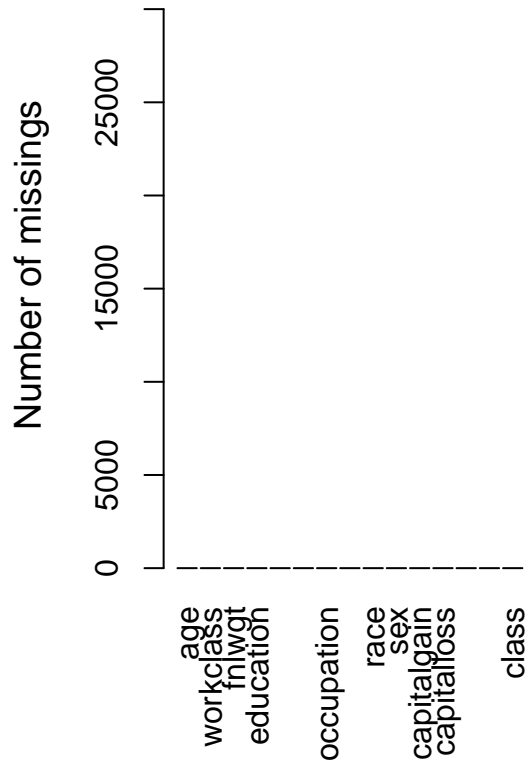
```
# missing values detection
```

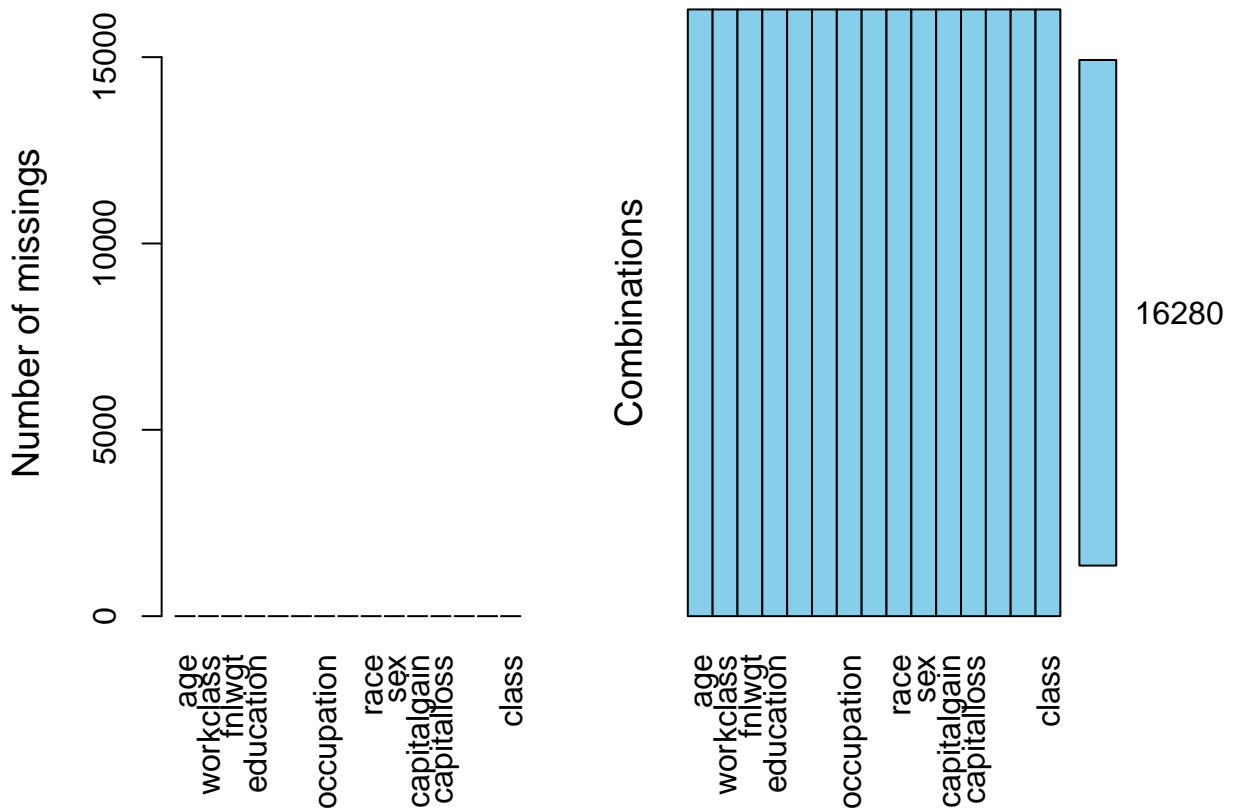
```
print("This is for missing values detection:")
```

```
## [1] "This is for missing values detection:"
```

```
aggr(training.data,prop=FALSE,numbers=TRUE)
```

```
aggr(test.data,prop=FALSE,numbers=TRUE)
```





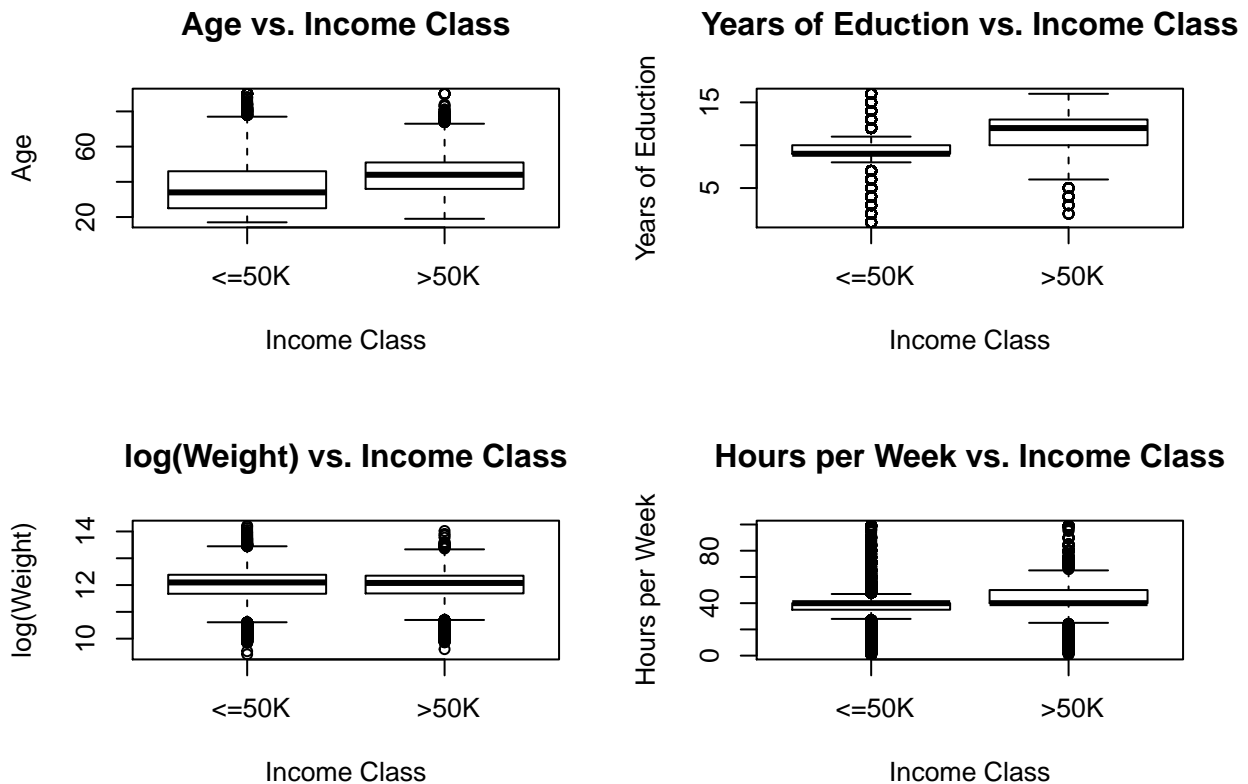
```
print("Numbers of missing values in training data and test dat: 0 , 0 ")
```

```
## [1] "Numbers of missing values in training data and test dat: 0 , 0 "
```

```
# since most of the variables are categorial, so we didn't do outliers detection here
```

2. statistics on numerical variables

```
par(mfrow=c(2,2)) ## Arrange plots in a 4x4 grid
boxplot(training.data[, 'age'] ~ training.data[, 'class'], main="Age vs. Income Class",
        xlab="Income Class", ylab="Age")
boxplot(training.data[, 'educationnum'] ~ training.data[, 'class'], main="Years of Eduction vs. Income Class",
        xlab="Income Class", ylab="Years of Education")
boxplot(log(training.data[, 'fnlwgt']) ~ training.data[, 'class'], main="log(Weight) vs. Income Class",
        xlab="Income Class", ylab="log(Weight)")
boxplot(training.data[, 'hoursperweek'] ~ training.data[, 'class'], main="Hours per Week vs. Income Class",
        xlab="Income Class", ylab="Hours per Week")
```



```
par(mfrow=c(1,1))
```

numerical data distribution analysis:

- In group that income >50k, people are more likely to have higher average age, higher average education years and higher weekly work hours than whose income is <=50K.
- Weight shows no difference in these two classes.

3. association rules of categorical variables

```
library(arules)
```

```
## Loading required package: Matrix
```

```
##
```

```
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
## abbreviate, write
```

```
training<-training.data[,c(2,4,6:10,14,15)]
summary(training)
```

```
##           workclass           education
## Private      :22696   HS-grad      :10501
## Self-emp-not-inc: 2541   Some-college: 7291
## Local-gov     : 2093   Bachelors   : 5354
## ?             : 1836   Masters       : 1723
## State-gov     : 1297   Assoc-voc    : 1382
## Self-emp-inc   : 1116   11th        : 1175
## (Other)       : 981    (Other)      : 5134
##           maritalstatus           occupation
## Divorced      : 4443   Prof-specialty :4140
## Married-AF-spouse : 23   Craft-repair  :4099
## Married-civ-spouse :14976 Exec-managerial:4066
## Married-spouse-absent: 418 Adm-clerical   :3769
## Never-married    :10682 Sales            :3650
## Separated        : 1025 Other-service    :3295
## Widowed          : 993   (Other)       :9541
##           relationship           race           sex
## Husband         :13193 Amer-Indian-Eskimo: 311 Female:10771
## Not-in-family    : 8304 Asian-Pac-Islander: 1039 Male :21789
## Other-relative   : 981   Black           : 3124
## Own-child        : 5068 Other              : 271
## Unmarried        : 3446 White              :27815
## Wife             : 1568
##
##           nativecountry           class
## United-States:29169   <=50K:24719
## Mexico        : 643   >50K : 7841
## ?             : 583
## Philippines    : 198
## Germany        : 137
## Canada         : 121
## (Other)        : 1709
```

use apriori rules to find association rules on people whose income class is >50K

```
rules <- apriori(training,
  control = list(verbose=F),
  parameter = list(minlen=2, supp=0.005, conf=0.8),
  appearance = list(rhs=c("class= >50K"),
    default="lhs"))

inspect(sort(rules, by="lift", decreasing = TRUE)[1:5])
```

##	lhs	rhs	support	confidence	lift	count
## [1]	{workclass= Private,					
##	education= Masters,					
##	occupation= Exec-managerial,					
##	relationship= Husband,					
##	nativecountry= United-States}	=> {class= >50K}	0.00509828	0.9273743	3.850951	166
## [2]	{workclass= Private,					

```

##      education= Masters,
##      maritalstatus= Married-civ-spouse,
##      occupation= Exec-managerial,
##      relationship= Husband,
##      nativecountry= United-States}      => {class= >50K} 0.00509828 0.9273743 3.850951 166
## [3] {workclass= Private,
##      education= Masters,
##      occupation= Exec-managerial,
##      relationship= Husband,
##      sex= Male,
##      nativecountry= United-States}      => {class= >50K} 0.00509828 0.9273743 3.850951 166
## [4] {workclass= Private,
##      education= Masters,
##      maritalstatus= Married-civ-spouse,
##      occupation= Exec-managerial,
##      sex= Male,
##      nativecountry= United-States}      => {class= >50K} 0.00509828 0.9273743 3.850951 166
## [5] {workclass= Private,
##      education= Masters,
##      maritalstatus= Married-civ-spouse,
##      occupation= Exec-managerial,
##      relationship= Husband,
##      sex= Male,
##      nativecountry= United-States}      => {class= >50K} 0.00509828 0.9273743 3.850951 166

```

4. validation for association rules

```

# using test data and confusion matrix to test the results
test<-test.data[,c(2,4,6:10,14,15)]
summary(test)

```

```

##      workclass      education
## Private      :11209  HS-grad      :5283
## Self-emp-not-inc: 1321  Some-college:3587
## Local-gov      : 1043  Bachelors   :2670
## ?              :  963  Masters      : 934
## State-gov      :  683  Assoc-voc    : 679
## Self-emp-inc    :  579  11th         : 636
## (Other)        :  482  (Other)      :2491
##      maritalstatus      occupation
## Divorced      :2190  Prof-specialty :2032
## Married-AF-spouse : 14  Exec-managerial:2020
## Married-civ-spouse :7403  Craft-repair   :2013
## Married-spouse-absent: 210  Sales          :1854
## Never-married    :5433  Adm-clerical   :1841
## Separated        : 505  Other-service   :1628
## Widowed         : 525  (Other)        :4892
##      relationship      race      sex
## Husband      :6523  Amer-Indian-Eskimo: 159  Female: 5421
## Not-in-family :4278  Asian-Pac-Islander: 480  Male  :10859
## Other-relative: 525  Black              :1560
## Own-child     :2512  Other              : 135

```



```
##   Unmarried      :1679   White           :13946
##   Wife           : 763
##
##       nativecountry      class
##   United-States:14661   <=50K.:12434
##   Mexico       : 308   >50K. : 3846
##   ?            : 274
##   Philippines  : 97
##   Puerto-Rico  : 70
##   Germany      : 69
##   (Other)      : 801
```

```
test.rule1<-subset(test, (workclass==' Private') & (education==' Masters') & (occupation==' Exec-manage
accuracy.rules1=nrow(subset(test.rule1,class==' >50K.'))/nrow(test.rule1)

test.rule2<-subset(test,workclass==' Private' & education==' Masters' & maritalstatus==' Married-civ-sp
accuracy.rules2=nrow(subset(test.rule2,class==' >50K.'))/nrow(test.rule2)

test.rule3<-subset(test,workclass==' Private' & education==' Masters' & occupation==' Exec-managerial' &
accuracy.rules3=nrow(subset(test.rule3,class==' >50K.'))/nrow(test.rule3)

test.rule4<-subset(test,workclass==' Private' & education==' Masters' & maritalstatus==' Married-civ-sp
accuracy.rules4=nrow(subset(test.rule4,class==' >50K.'))/nrow(test.rule4)

test.rule5<-subset(test,workclass==' Private' & education==' Masters' & maritalstatus==' Married-civ-sp
accuracy.rules5=nrow(subset(test.rule5,class==' >50K.'))/nrow(test.rule5)

paste('accuracy of association rule1,rule2,rule3,rule4 and rule5 are:',round(accuracy.rules1,3),round(a
```

```
## [1] "accuracy of association rule1,rule2,rule3,rule4 and rule5 are: 0.909 0.909 0.885 0.909 0.909"
```

5. conclusions

- numerical statistic analysis showed that people have higher average age, higher average education years and higher weekly work hours are more likely to have income >50K.
- association rules listed in this project showed good accuracies: rule1(91%),rule2(91%),rule3(89%),rule4(91%),rule5(91%)
- association rules showed that people who have master degrees, work privately married male with an Exec-managerial occupation have large possibilities(>89%) to have salaries more than 50,000.