# DATA MINING

Project3

SHIYANG ZHANG

47319809

2017.11

# Contents



Contents of table and figures:

# 1.Execution summary:

In this paper, we want to use association rule mining to build rules between item sets. We would like to find out associations between items according to analyze each transaction. I use two data set to represent two transactions. The first one is 2005_03, the second one is 2013_03. The year of both data set are 4 years later after under new president administration.

In this project, I get some conclusion:

1. Very high frequent items do not have high lift. So is always not so useful as LHS.
2. Agency is a center of different cluster of rules.
3. Frequent items are connection of different clusters.
4. Association rule reveals some trivial principles that are easy to be ignored in past two projects.

# 2.Data Preparation

At first, do basic cleaning for invalid data set. Using NA replace unknown characters like '########', 'UNSP', '*'. Then impute NA-Age with median age of the same agency. Impute NA-Pay with median pay for the Age of the employee at that agency. Then drop NA pays. Finally, dealing with values have duplicate IDs:

1. Select rows with duplicate IDs
2. Order selection by ID, then Agency, then descending Pay
3. Select rows where the ID and Agency are duplicated (same employee at same agency),
4. Get row numbers for rows with the lowest pay for each grouping in the above selection
5. Reselect from data frame where rows are not required to removed

Secondly, I set some features into null since there are not so important. Like ID, Name, Date, Agency, Schedule and NSFTP. And use discretize(frequency) method to discretize Age, Education, LOS and Pay those numeric values. Since continuous values are inconvenient for building associate rules. And frequency method divides values into 3 categories with similar frequency value of each category, which will decrease rules count later.

The data subsets I choose is 2005-03 and 2013-03. 2005 is 4 years later after Bush became a president. 2013 is 4 years later after Obama became a president. I build two transaction set by using 'transaction' method based on two data frames. Item frequency plot are shown as follow[1]. Figure 1 represents transaction of 2005. Figure 2 represents transaction of 2013.

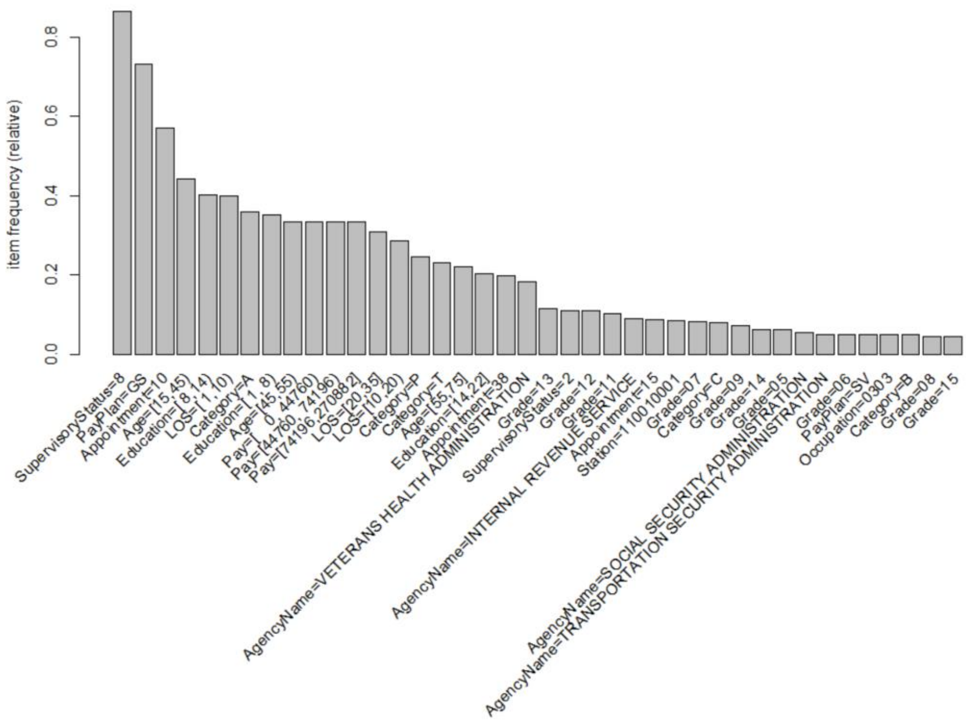


Figure 1 2005 item frequency plot

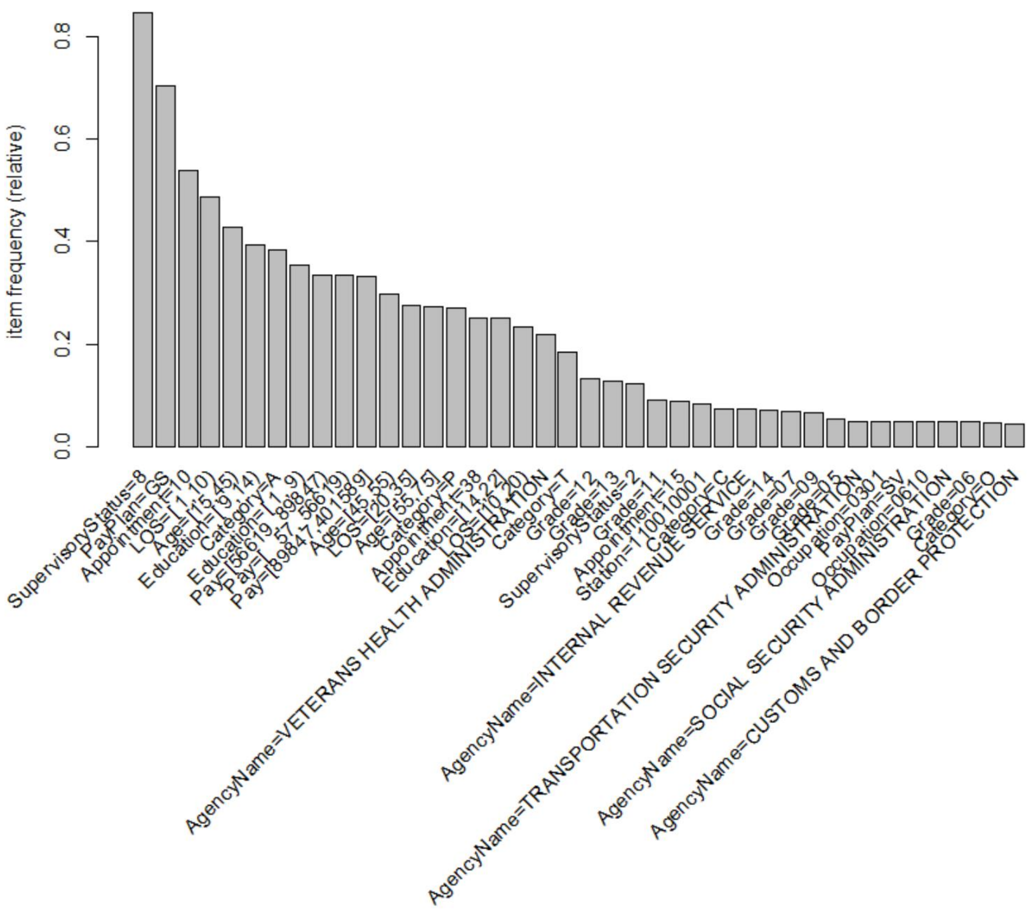


Figure 2 2013 item frequency plot

Following are some statistics for data subset. Using 2005-03 as example to show:

| Features | Scale | Description |
|---|---|---|
| Station | Nominal | e.g 110010001 |
| Age | Ordinal | e.g [15,45):525183<br>[45,55):397677 |
| Education | Ordinal | e.g [1,8):416962<br>[8,14):477887 |
| Pay plan | Nominal | e.g GS:869097 |
| Grade | Nominal | e.g 13:135643 |
| LOS | Ordinal | e.g [1,10):475577<br>[10,20):340116 |
| Occupation | Nominal | e.g 0303:58501 |
| Category | Nominal | e.g A: 427244 |
| Pay | Ordinal | e.g [0,44760):395367<br>[44760,74196):395330 |
| Supervisory Status | Nominal | e.g 8:1025856 |
| Appointment | Nominal | e.g 10:678597 |
| Agency Name | Nominal | e.g VATA:218367 |

Table 1 Statistics about transaction set

# 3.Modeling

## 3.1 Frequent, Closed and Maximal itemsets

Figure 3 and figure 4 respectively shows different itemset size number of 2005 and 2013. Form two charts, we can see that itemset size 5 is most popular, which may demonstrate that 5 items are enough to do association analysis, and pretty useful. These 5 features are also important factors that people concern most, like pay, education, los, age, etc.

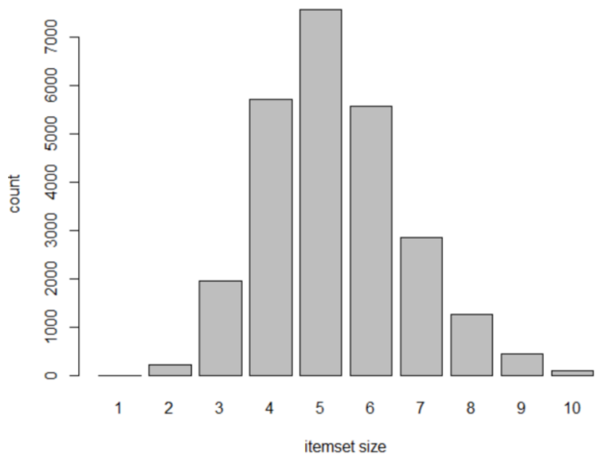


Figure 3 2005 itemset size plot

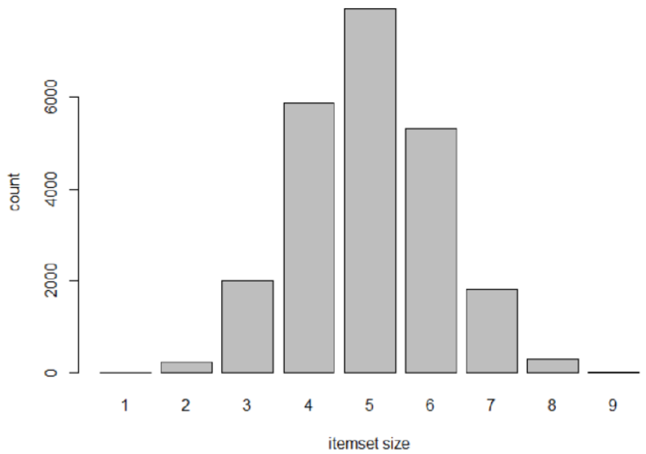


Figure 4 2013 itemset size plot

An itemset is maximal frequent if none of its immediate supersets is frequent. An itemset is closed if none of its immediate supersets has the same support as the itemset[2].

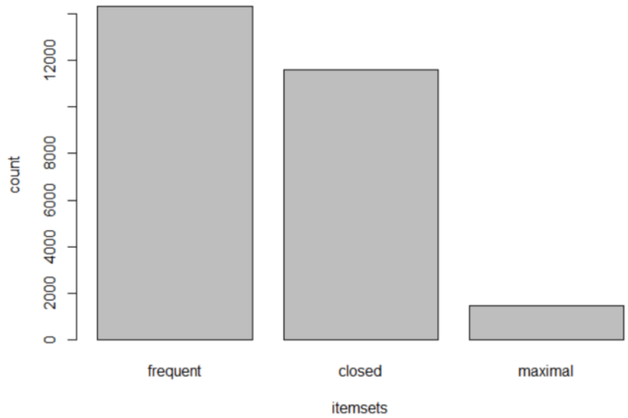


Figure 5 2005 frequent-closed-maximal plot

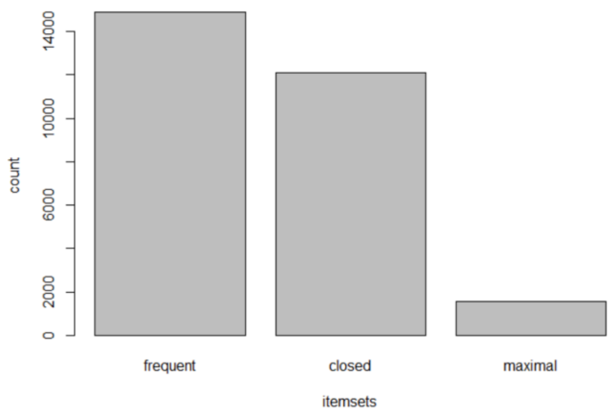


Figure 6 2013 frequent-closed-maximal plot

## 3.2 Create sets of association rules

In this case, rule1 is association rule from transaction1(2005), rule2 is association rule from transaction2(2013). I choose minimum threshold support equals to 0.01, minimum threshold confidence equals to 0.8. The summary of rule1 is shown as figure 7. The summary of rule 2 is shown as figure 8.

```
set of 25688 rules

rule length distribution (lhs + rhs):sizes
   1    2    3    4    5    6    7    8    9   10
   1  226 1961 5718 7561 5572 2847 1252  450  100

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   5.000   5.273   6.000  10.000

summary of quality measures:
    support          confidence          lift             count
 Min.   :0.01000   Min.   :0.8000   Min.   : 0.9249   Min.   :  11860
 1st Qu.:0.01314   1st Qu.:0.9078   1st Qu.: 1.2304   1st Qu.:  15583
 Median :0.02094   Median :0.9660   Median : 2.3652   Median :  24834
 Mean   :0.02376   Mean   :0.9468   Mean   : 5.8373   Mean   :  28180
 3rd Qu.:0.03119   3rd Qu.:0.9979   3rd Qu.: 4.3451   3rd Qu.:  36988
 Max.   :0.86503   Max.   :1.0000   Max.   :76.7375   Max.   :1025856

mining info:
   data ntransactions support confidence
 trans1       1185917    0.01        0.8
```

Figure 7 summary of rule1

```
> summary(rules2)
set of 23494 rules

rule length distribution (lhs + rhs):sizes
   1    2    3    4    5    6    7    8    9
   1  224 2008 5876 7931 5326 1811  297   20

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   5.000   4.972   6.000   9.000

summary of quality measures:
    support          confidence          lift             count
 Min.   :0.01000   Min.   :0.8000   Min.   : 0.9456   Min.   :  13376
 1st Qu.:0.01139   1st Qu.:0.8961   1st Qu.: 1.1940   1st Qu.:  15235
 Median :0.01395   Median :0.9620   Median : 2.0004   Median :  18650
 Mean   :0.01856   Mean   :0.9413   Mean   : 5.6811   Mean   :  24824
 3rd Qu.:0.01906   3rd Qu.:0.9981   3rd Qu.: 3.9508   3rd Qu.:  25496
 Max.   :0.84631   Max.   :1.0000   Max.   :77.7301   Max.   :1131810

mining info:
   data ntransactions support confidence
 trans2       1337350    0.01        0.8
```

Figure 8 summary of rule2

## 3.3 Discuss patterns

```
> inspect(head(rules1, by = "lift"))
    lhs                                                  rhs                  support confidence     lift count
[1] {Category=O,
     AgencyName=BUREAU OF PRISONS/FEDERAL PRISON SYSTEM} => {Occupation=0007} 0.01291827  0.9979806 76.73748 15320
[2] {PayPlan=GS,
     Category=O,
     AgencyName=BUREAU OF PRISONS/FEDERAL PRISON SYSTEM} => {Occupation=0007} 0.01291827  0.9979806 76.73748 15320
[3] {Category=O,
     SupervisoryStatus=8,
     AgencyName=BUREAU OF PRISONS/FEDERAL PRISON SYSTEM} => {Occupation=0007} 0.01181364  0.9977922 76.72299 14010
[4] {PayPlan=GS,
     Category=O,
     SupervisoryStatus=8,
     AgencyName=BUREAU OF PRISONS/FEDERAL PRISON SYSTEM} => {Occupation=0007} 0.01181364  0.9977922 76.72299 14010
[5] {Age=[15,45),
     Category=O,
     AgencyName=BUREAU OF PRISONS/FEDERAL PRISON SYSTEM} => {Occupation=0007} 0.01084646  0.9975958 76.70789 12863
[6] {Age=[15,45),
     PayPlan=GS,
     Category=O,
     AgencyName=BUREAU OF PRISONS/FEDERAL PRISON SYSTEM} => {Occupation=0007} 0.01084646  0.9975958 76.70789 12863
```

Figure 9. Six Largest lift in rule1

```
> inspect(head(rules2, by = "lift"))
    lhs                                               rhs                  support confidence     lift count
[1] {PayPlan=GS,
     Category=C,
     Appointment=38,
     AgencyName=VETERANS HEALTH ADMINISTRATION} => {Occupation=0679} 0.01049912  0.9541964 77.73007 14041
[2] {PayPlan=GS,
     Category=C,
     Pay=[   57, 56619),
     Appointment=38,
     AgencyName=VETERANS HEALTH ADMINISTRATION} => {Occupation=0679} 0.01040042  0.9540435 77.71761 13909
[3] {PayPlan=GS,
     Category=C,
     Pay=[   57, 56619),
     SupervisoryStatus=8,
     Appointment=38,
     AgencyName=VETERANS HEALTH ADMINISTRATION} => {Occupation=0679} 0.01016862  0.9534460 77.66894 13599
[4] {PayPlan=GS,
     Category=C,
     SupervisoryStatus=8,
     Appointment=38,
     AgencyName=VETERANS HEALTH ADMINISTRATION} => {Occupation=0679} 0.01022844  0.9534397 77.66843 13679
[5] {Category=C,
     SupervisoryStatus=8,
     Appointment=38,
     AgencyName=VETERANS HEALTH ADMINISTRATION} => {Occupation=0679} 0.01022844  0.9089641 74.04538 13679
[6] {Category=C,
     Pay=[   57, 56619),
     SupervisoryStatus=8,
     Appointment=38,
     AgencyName=VETERANS HEALTH ADMINISTRATION} => {Occupation=0679} 0.01016862  0.9087203 74.02553 13599
```

Figure 10 Six largest lift in rule2

The threshold for support and confidence are 0.01 and 0.8 for both rule1 and rule2.
In figure 9, first 6 rules LHS are very similar and all RHS are occupation 0007. In reference website[3], 0007 means Correctional Officer Series. And in all six LHS, agency name are shown and are the same. So maybe in prison, most job are correctional officer series, so this lift will be high.
In figure 10, RHS are still occupation, but it changes. Occupation 0679 represents Medical, Hospital, Dental, and Public Health Group. In agency VETERANS HEALTH ADMINISTRATION, occupation number may be less.

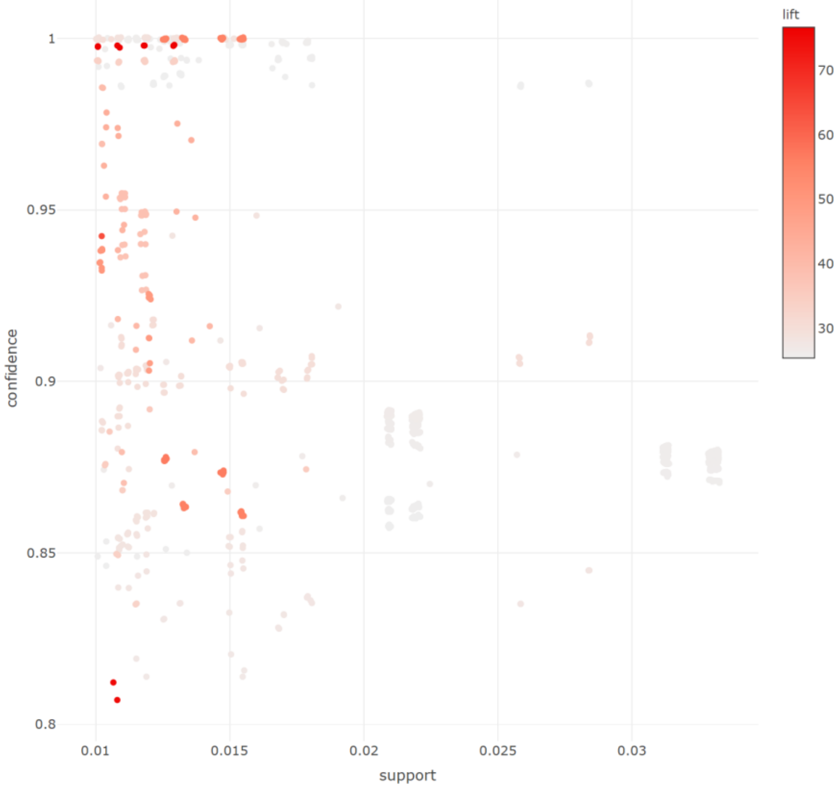

Figure 11 rule1 support-confidence plot

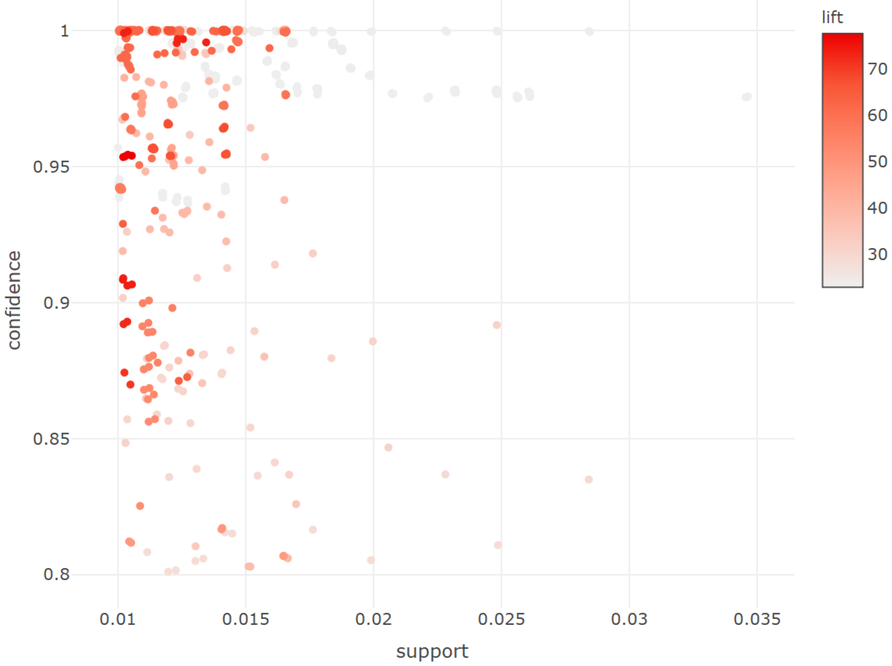

Figure 12 rule2 support-confidence plot

In Figure 11 and figure 12, support, confidence and lift are displayed. In figure 11, some points with high support, middle confidence, low lift shows that high frequency is not very useful for association rules. And some points are low support, high confidence with low lift. It may because some RHS already have high frequency, no matter the rule is. So lift is an important factor to filter some high frequency item. In figure 11 and 12, most high lift rules located in low support and middle or high confidence area. Therefore, the occur of high frequency items may not result from other items. Lift and confidence strengthen relationship between LHS and RHS, and filter some unnecessary association[4].

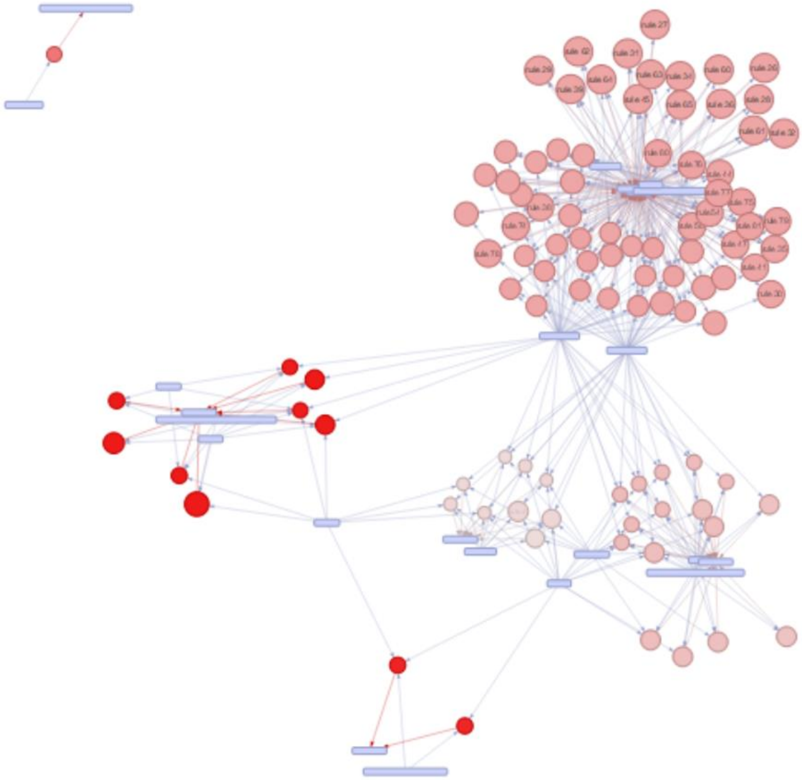

Figure 13 rule1 graph

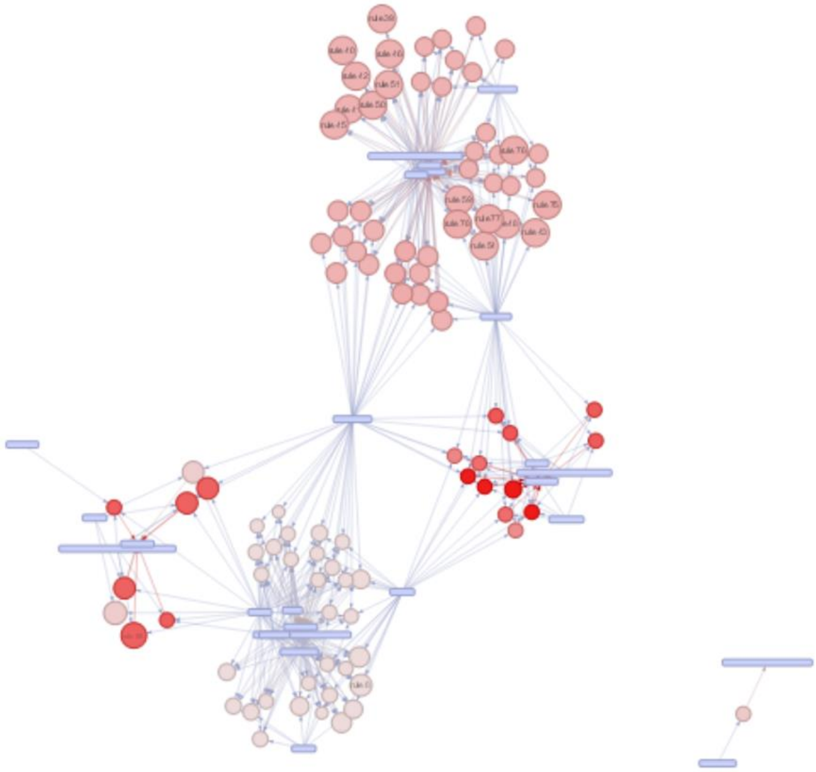

Figure 14 rule2 graph

In figure 13 and figure 14, it only shows 100 rules using lift due to space limits. I find that in association's graph figure, each cluster focus on 1 agency. So agency has large influence on other features. The connections between different cluster are high frequent items. The reddest rules have highest lift. Also, rules around each agency have similar lift value. Maybe some agency is small, occupation or pay scheme differs little that association rules are meaningful. Some agency is bigger with large variances under different conditions. Some items and rules are separate from main structure, it shows that this agency has totally different scheme from other agencies.

I change the support and confidence threshold as 0.05 and 0.9. The first 6 rules ordered by lift are shown below:

```
> inspect(head(rules1, by = "lift"))
    lhs                                                        rhs
[1] {PayPlan=SV}                                            => {AgencyName=TRANSPORTATION SECURITY ADMINISTRATION}
[2] {AgencyName=TRANSPORTATION SECURITY ADMINISTRATION}     => {PayPlan=SV}
[3] {PayPlan=GS,Grade=14}                                   => {Pay=[74196,270882]}
[4] {PayPlan=GS,Grade=11,Category=A,SupervisoryStatus=8} => {Pay=[44760, 74196)}
[5] {PayPlan=GS,Grade=11,Category=A}                        => {Pay=[44760, 74196)}
[6] {Grade=11,Category=A}                                   => {Pay=[44760, 74196)}
    support    confidence lift      count
[1] 0.05020672 1.0000000  19.861612 59541
[2] 0.05020672 0.9971864  19.861612 59541
[3] 0.05379634 0.9989666   2.997549 63798
[4] 0.05319597 0.9954713   2.986230 63086
[5] 0.05609752 0.9954363   2.986125 66527
[6] 0.05736152 0.9935589   2.980493 68026
```

Figure 15 head rule1

```
> inspect(head(rules2, by = "lift"))
    lhs                  rhs                                      support confidence     lift count
[1] {LOS=[ 1,10),
     Category=P,
     Appointment=38} => {AgencyName=VETERANS HEALTH ADMINISTRATION} 0.05148166  0.9062418 4.130399 68849
[2] {PayPlan=GS,
     Grade=14}       => {Pay=[89847,401589]}                        0.06436012  0.9989786 3.005744 86072
[3] {Grade=14,
     Appointment=10} => {Pay=[89847,401589]}                        0.05092534  0.9972764 3.000622 68105
[4] {Grade=14}       => {Pay=[89847,401589]}                        0.07067634  0.9937235 2.989932 94519
[5] {PayPlan=GS,
     Grade=11,
     Category=A}     => {Pay=[56619, 89847)}                        0.05278424  0.9984018 2.986772 70591
[6] {PayPlan=GS,
     Grade=11,
     Appointment=10} => {Pay=[56619, 89847)}                        0.05825326  0.9977204 2.984734 77905
```

Figure 16 head rule2

The lift value decrease a lot compared with support=0.01, confidence=0.9. Maybe this threshold deletes some items not so frequent but very meaningful for building association rules. In this case, RHS are not occupation, it is pay or agency. This looks like some conclusion that we made for classification. Therefore, association reveals some principle that we cannot discover in classification. It is more careful.

## 3.4 Dataset comparison

There are lots of change between two datasets. The match between rule1 and rule2 is 0. So it changes a lot between 2005 to 2013. Caluclate quality for some of rules1 in trans2 and look at the difference:

```
> inspect(r[which(diff$supp > 0.2 & diff$supp!=1)])
    lhs                                                 rhs                     support confidence      lift count
[1] {PayPlan=SV,
     LOS=[ 1,10),
     Category=T,
     Appointment=38}                                 => {SupervisoryStatus=8} 0.03566607  0.9147671  1.057495 42297
[2] {Grade=11,
     LOS=[10,20)}                                    => {PayPlan=GS}          0.03047178  0.9636790  1.314978 36137
[3] {Age=[15,45),
     LOS=[10,20),
     Category=T}                                     => {Appointment=10}      0.02177134  0.8149165  1.424149 25819
[4] {Age=[15,45),
     LOS=[ 1,10),
     Occupation=0019,
     Category=T,
     Appointment=38,
     AgencyName=TRANSPORTATION SECURITY ADMINISTRATION} => {PayPlan=SV}       0.02537108  1.0000000 19.917653 30088
[5] {Age=[15,45),
     Category=C,
     Appointment=10}                                 => {PayPlan=GS}          0.01409289  0.9714037  1.325518 16713
[6] {LOS=[ 1,10),
     Occupation=0019,
     SupervisoryStatus=8,
     AgencyName=TRANSPORTATION SECURITY ADMINISTRATION} => {PayPlan=SV}       0.03560114  1.0000000 19.917653 42220
```

Figure 17 for which rules did support increase by 10%

```
> inspect(r[which(diff$supp < -0.1)])
    lhs                                                                 rhs                 support    confidence
[1] {Age=[55,75],Category=A,Appointment=10}                          => {PayPlan=GS}        0.05393548 0.9332487
[2] {LOS=[ 1,10),Occupation=0610,SupervisoryStatus=8,Appointment=38} => {PayPlan=VN}        0.01022669 0.8857727
[3] {Occupation=0905,Category=P}                                     => {Education=[14,22]} 0.02059250 0.9186353
[4] {PayPlan=VN,LOS=[ 1,10)}                                         => {Appointment=38}    0.01155730 1.0000000
    lift      count
[1]  1.273455 63963
[2] 29.587723 12128
[3]  4.540331 24421
[4]  5.013092 13706
```

Figure 18 for which rules did support decrease by 10%

```
> inspect(r[which(diff$lift > 0.1)])
    lhs                                          rhs                  support  confidence      lift count
[1] {Age=[15,45),
     LOS=[10,20),
     Category=T}                              => {Appointment=10}    0.02177134   0.8149165  1.424149 25819
[2] {PayPlan=AT,
     Category=A,
     AgencyName=FEDERAL AVIATION ADMINISTRATION} => {Appointment=38}    0.01540327   0.9954226  4.990145 18267
[3] {Grade=09,
     Category=A,
     Appointment=10}                          => {PayPlan=GS}        0.02250579   0.9911248  1.352429 26690
[4] {LOS=[ 1,10),
     Occupation=0610,
     SupervisoryStatus=8,
     Appointment=38}                          => {PayPlan=VN}        0.01022669   0.8857727 29.587723 12128
[5] {Occupation=0905,
     Category=P}                              => {Education=[14,22]} 0.02059250   0.9186353  4.540331 24421
[6] {PayPlan=VN,
     LOS=[ 1,10)}                             => {Appointment=38}    0.01155730   1.0000000  5.013092 13706
```

Figure 19 for which rules did lift increase by 10%

```
> inspect(r[which(diff$lift < -0.1)])
    lhs                                    rhs          support    confidence lift     count
[1] {Age=[45,55),Grade=14,LOS=[20,35]} => {PayPlan=GS} 0.01288454 0.88513    1.207795 15280
```

Figure 20 for which rules did lift decrease by 10%

These difference means that huge changes happened between 2005 and 2013. Some items are frequent at LHS are not so powerful after 8 years.

# 4.Evaluation

I think the most interesting thing I found is that in 2005 and 2013, association rules are highly related with agency. And almost every rules of some specific agencies have very high lift. This shows that some agency focus on some occupation, so the rules works pretty good for them. In past project, we focus more on Education, LOS, Pay, those very high frequent items. But for this project, high frequent may be not a good thing, because its lift will be low. It cannot show strong association with other items. So this project focus on more trivial thing compared to past two project.

For employees, I will suggest them carefully analyzing different agencies because agency is a very important factor for analyzing rules. And occupation is a good choice to choose. Because it related to many practical items. They can choose a proper agency, position, pay plan to earn more money. For employers, I suggest them focus on some trivial evidence. These evidence may not easy to find based on classification method, because classification only focus on high frequent feature which most people can learn. To obtain benefit, they can try to understand the scheme of its company, and make some change to increase income or improve employee preferences.

# 5.reference

[1] http://michael.hahsler.net/SMU/EMIS7331/data/federal_employment/project3.html
[2] https://rawgit.com/mhahsler/Introduction_to_Data_Mining_R_Examples/master/chap6.html
[3]https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/#url=Standards
[4] http://michael.hahsler.net/SMU/EMIS7331/data/federal_employment/project3-2.html