Diagnosis 数据挖掘——实验报告

1、对 diagnosis 数据集进行处理，转换其形式

导入关联规则的 R 语言包：

install.packages('arules')

加载 arules 程序包：library(arules)

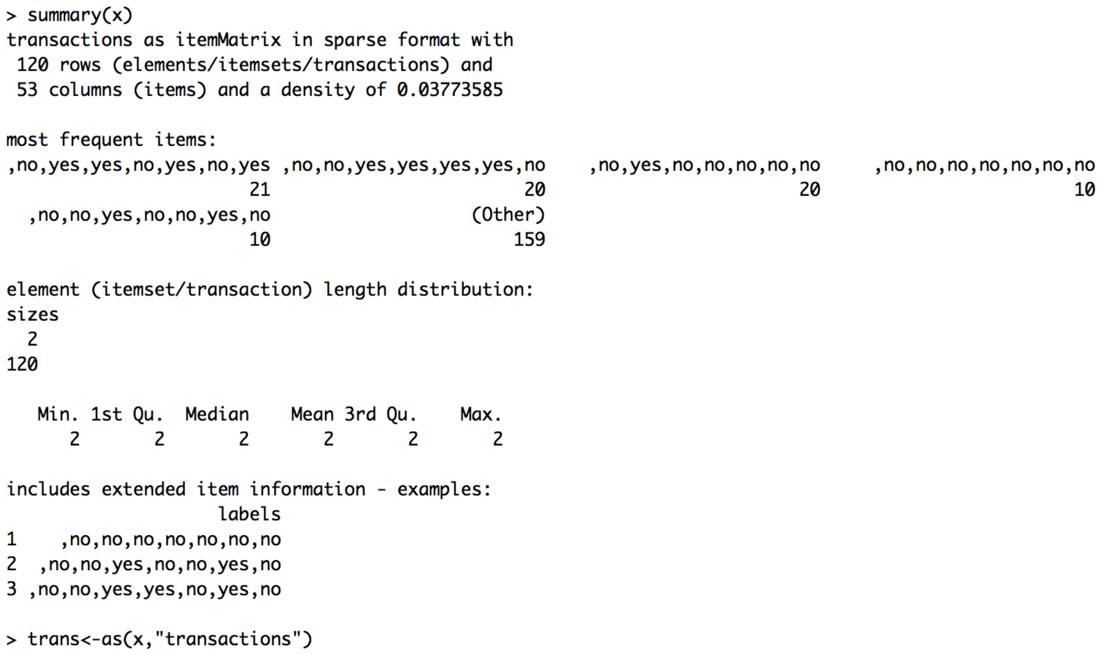
加载数据集：

把从网上下载的 diagnosis 数据(diagnosis.csv) 读入到 R 中。

x<-read.transactions("diagnosis.csv",format="basket",sep = "")

查看数据集相关的统计汇总信息，以及数据集本身

summary(x)



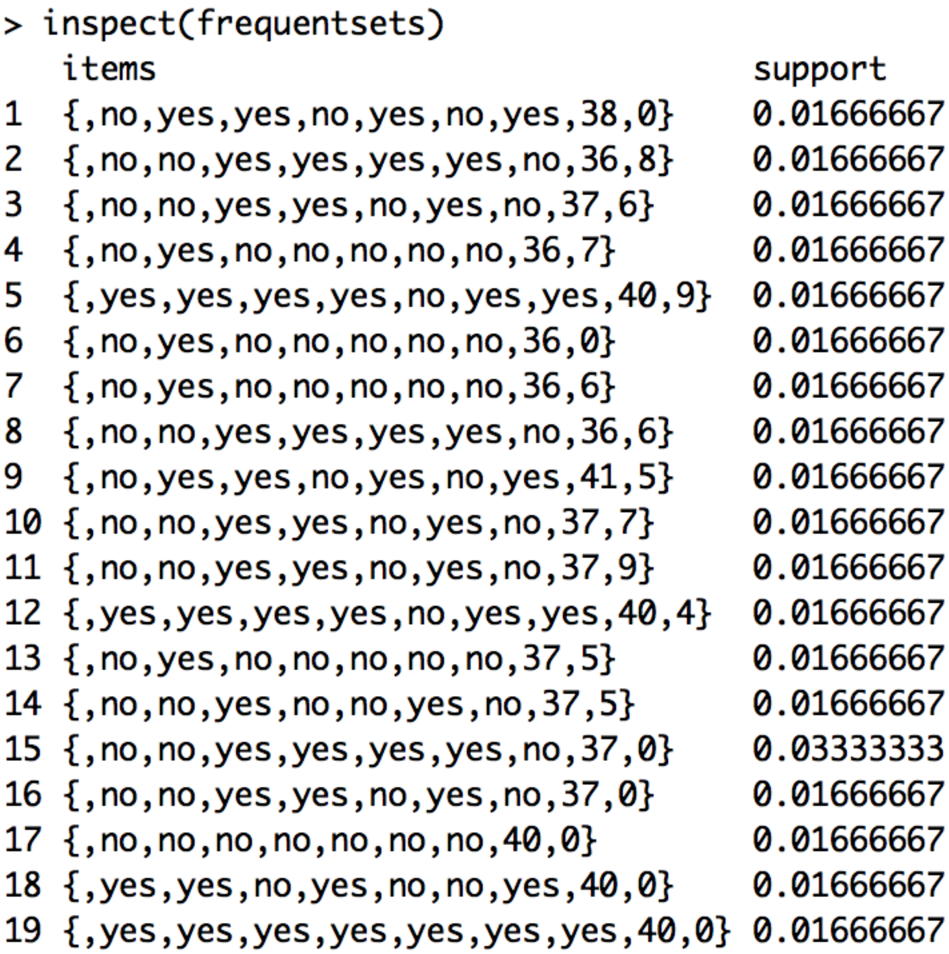
```
> summary(x)
transactions as itemMatrix in sparse format with
 120 rows (elements/itemsets/transactions) and
 53 columns (items) and a density of 0.03773585

most frequent items:
,no,yes,yes,no,yes,no,yes ,no,no,yes,yes,yes,yes,no     ,no,yes,no,no,no,no,no      ,no,no,no,no,no,no,no
                       21                        20                         20                         10
  ,no,no,yes,no,no,yes,no                   (Other)
                       10                       159

element (itemset/transaction) length distribution:
sizes
  2
120

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

includes extended item information - examples:
                    labels
1    ,no,no,no,no,no,no,no
2  ,no,no,yes,no,no,yes,no
3 ,no,no,yes,yes,no,yes,no

> trans<-as(x,"transactions")
```

2、找出频繁项集

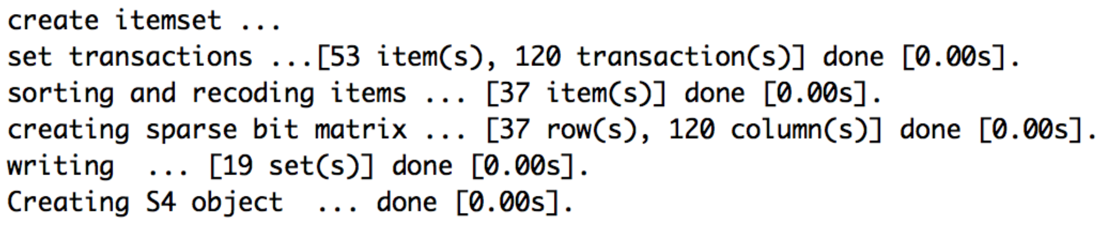
求频繁项集：

```
> #找出所有的频繁项集
> frequentsets<- eclat(trans,parameter=list(support=0.01,maxlen=10,minlen=2))
Eclat

parameter specification:
 tidLists support minlen maxlen            target   ext
    FALSE    0.01      2     10 frequent itemsets FALSE

algorithmic control:
 sparse sort verbose
      7   -2    TRUE

Absolute minimum support count: 1
```

查看求得的频繁项集 inspect(frequentsets)



```
> inspect(frequentsets)
   items                                support
1  {,no,yes,yes,no,yes,no,yes,38,0}     0.01666667
2  {,no,no,yes,yes,yes,yes,no,36,8}     0.01666667
3  {,no,no,yes,yes,no,yes,no,37,6}      0.01666667
4  {,no,yes,no,no,no,no,no,36,7}        0.01666667
5  {,yes,yes,yes,yes,no,yes,yes,40,9}   0.01666667
6  {,no,yes,no,no,no,no,no,36,0}        0.01666667
7  {,no,yes,no,no,no,no,no,36,6}        0.01666667
8  {,no,no,yes,yes,yes,yes,no,36,6}     0.01666667
9  {,no,yes,yes,no,yes,no,yes,41,5}     0.01666667
10 {,no,no,yes,yes,no,yes,no,37,7}      0.01666667
11 {,no,no,yes,yes,no,yes,no,37,9}      0.01666667
12 {,yes,yes,yes,yes,no,yes,yes,40,4}   0.01666667
13 {,no,yes,no,no,no,no,no,37,5}        0.01666667
14 {,no,no,yes,no,no,yes,no,37,5}       0.01666667
15 {,no,no,yes,yes,yes,yes,no,37,0}     0.03333333
16 {,no,no,yes,yes,no,yes,no,37,0}      0.01666667
17 {,no,no,no,no,no,no,no,40,0}         0.01666667
18 {,yes,yes,no,yes,no,no,yes,40,0}     0.01666667
19 {,yes,yes,yes,yes,yes,yes,yes,40,0} 0.01666667
```

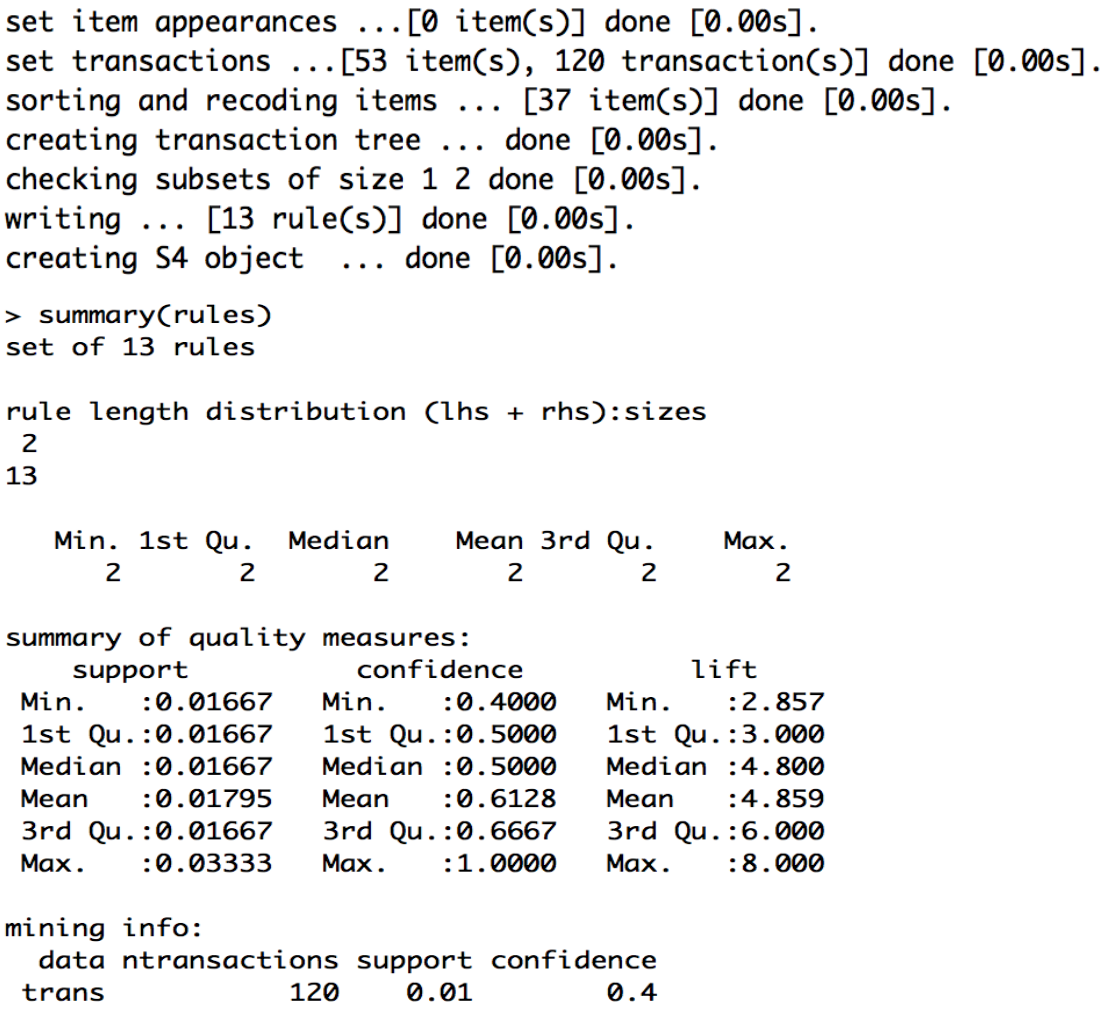


```
create itemset ...
set transactions ...[53 item(s), 120 transaction(s)] done [0.00s].
sorting and recoding items ... [37 item(s)] done [0.00s].
creating sparse bit matrix ... [37 row(s), 120 column(s)] done [0.00s].
writing  ... [19 set(s)] done [0.00s].
Creating S4 object  ... done [0.00s].
```

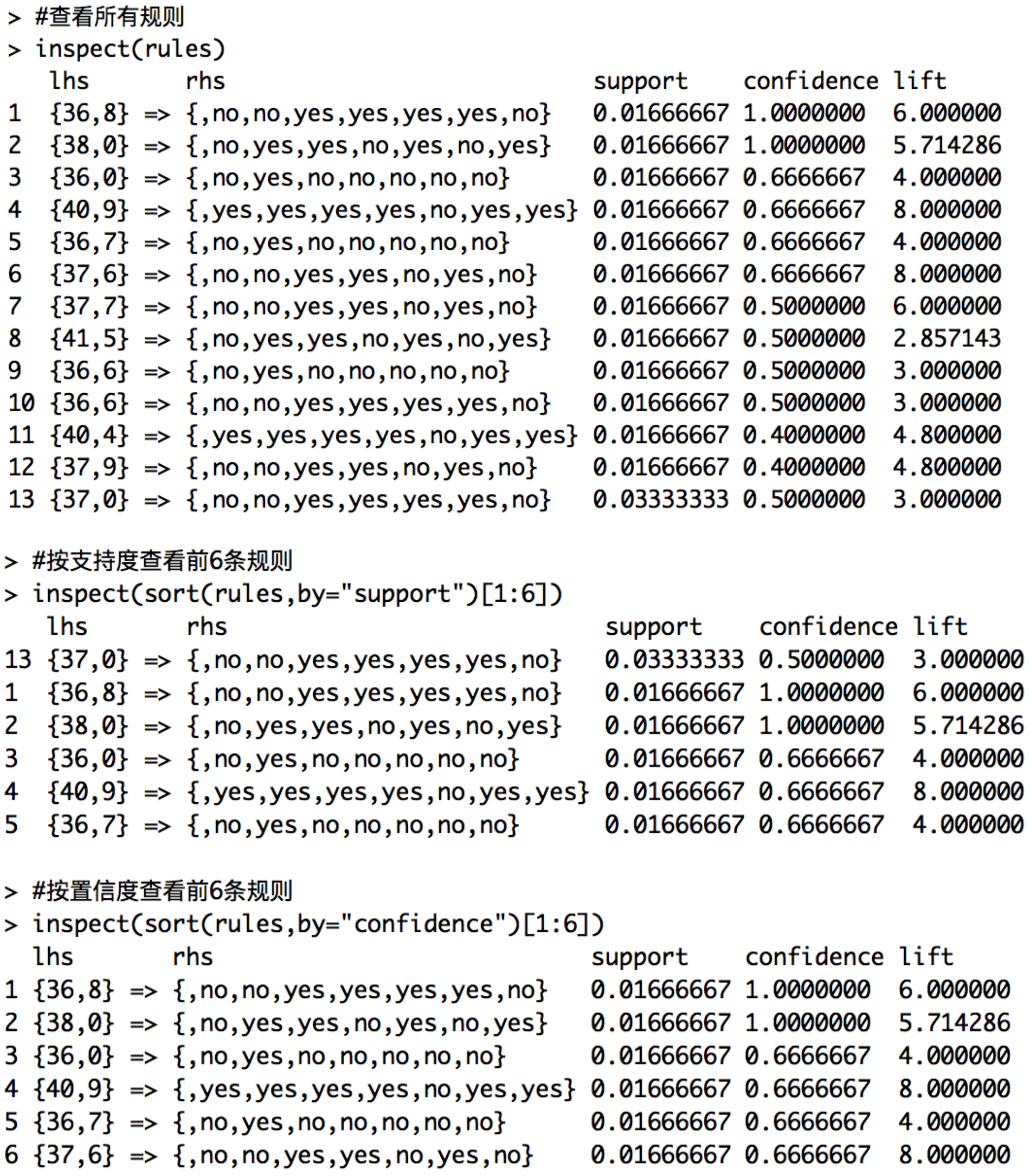
3、导出关联规则，计算其支持度和置信度

用 apriori 求关联规则，并查看相关的统计汇总信息

```
> #找出所有的关联规则
> rules <- apriori(trans,parameter=list(support=0.01,confidence=0.4,minlen=2))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.4    0.1    1 none FALSE            TRUE    0.01      2     10  rules FALSE

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

Absolute minimum support count: 1
```



```
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[53 item(s), 120 transaction(s)] done [0.00s].
sorting and recoding items ... [37 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 done [0.00s].
writing ... [13 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].

> summary(rules)
set of 13 rules

rule length distribution (lhs + rhs):sizes
 2
13

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support           confidence          lift
 Min.   :0.01667   Min.   :0.4000   Min.   :2.857
 1st Qu.:0.01667   1st Qu.:0.5000   1st Qu.:3.000
 Median :0.01667   Median :0.5000   Median :4.800
 Mean   :0.01795   Mean   :0.6128   Mean   :4.859
 3rd Qu.:0.01667   3rd Qu.:0.6667   3rd Qu.:6.000
 Max.   :0.03333   Max.   :1.0000   Max.   :8.000

mining info:
  data ntransactions support confidence
 trans           120    0.01        0.4
```

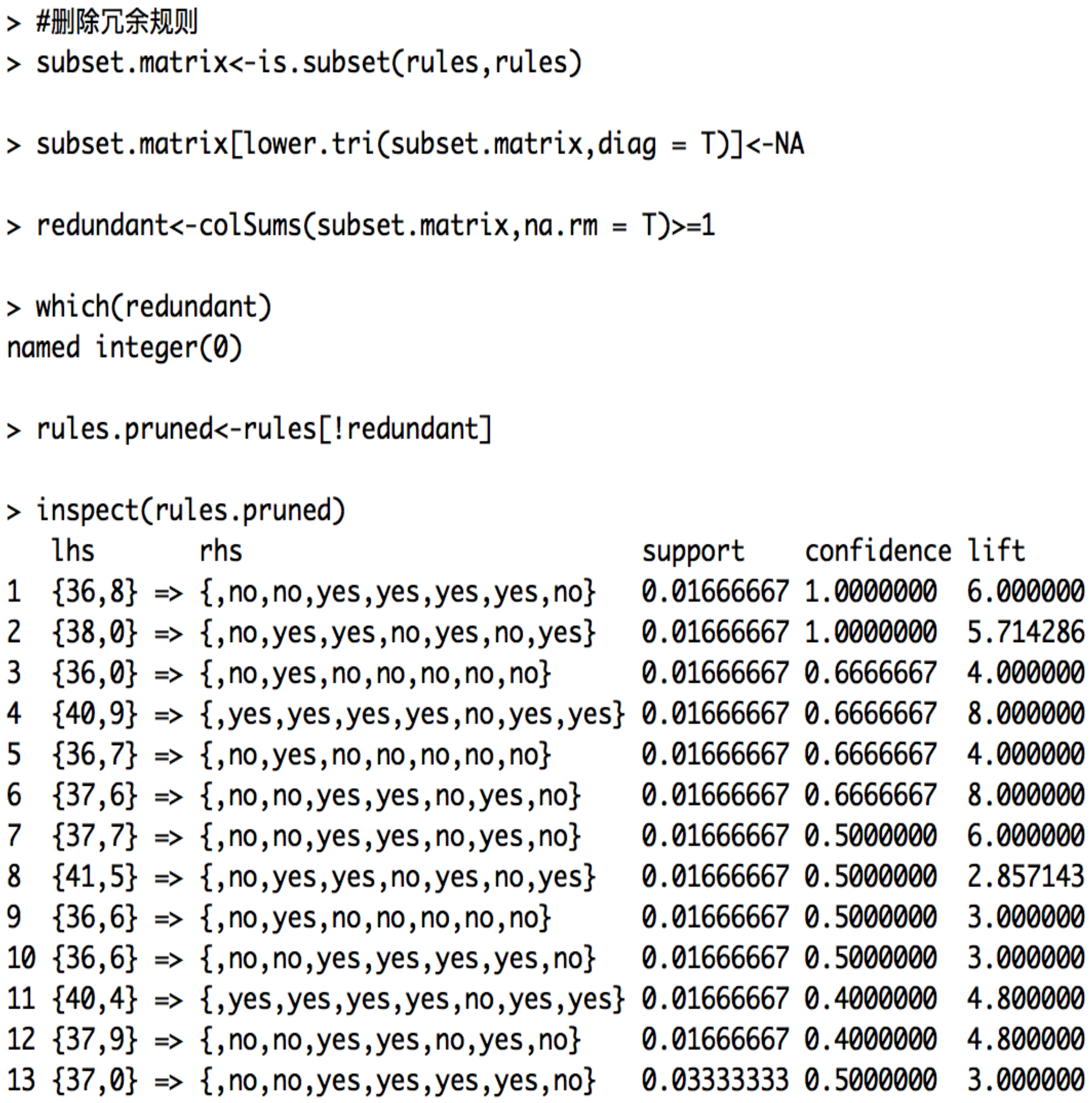


```
> #查看所有规则
> inspect(rules)
   lhs         rhs                               support    confidence lift
1  {36,8} => {,no,no,yes,yes,yes,yes,no}    0.01666667 1.0000000  6.000000
2  {38,0} => {,no,yes,yes,no,yes,no,yes}    0.01666667 1.0000000  5.714286
3  {36,0} => {,no,yes,no,no,no,no,no}       0.01666667 0.6666667  4.000000
4  {40,9} => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.6666667  8.000000
5  {36,7} => {,no,yes,no,no,no,no,no}       0.01666667 0.6666667  4.000000
6  {37,6} => {,no,no,yes,yes,no,yes,no}     0.01666667 0.6666667  8.000000
7  {37,7} => {,no,no,yes,yes,no,yes,no}     0.01666667 0.5000000  6.000000
8  {41,5} => {,no,yes,yes,no,yes,no,yes}    0.01666667 0.5000000  2.857143
9  {36,6} => {,no,yes,no,no,no,no,no}       0.01666667 0.5000000  3.000000
10 {36,6} => {,no,no,yes,yes,yes,yes,no}    0.01666667 0.5000000  3.000000
11 {40,4} => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.4000000  4.800000
12 {37,9} => {,no,no,yes,yes,no,yes,no}     0.01666667 0.4000000  4.800000
13 {37,0} => {,no,no,yes,yes,yes,yes,no}    0.03333333 0.5000000  3.000000
```

```
> #按支持度查看前6条规则
> inspect(sort(rules,by="support")[1:6])
   lhs         rhs                               support    confidence lift
13 {37,0} => {,no,no,yes,yes,yes,yes,no}    0.03333333 0.5000000  3.000000
1  {36,8} => {,no,no,yes,yes,yes,yes,no}    0.01666667 1.0000000  6.000000
2  {38,0} => {,no,yes,yes,no,yes,no,yes}    0.01666667 1.0000000  5.714286
3  {36,0} => {,no,yes,no,no,no,no,no}       0.01666667 0.6666667  4.000000
4  {40,9} => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.6666667  8.000000
5  {36,7} => {,no,yes,no,no,no,no,no}       0.01666667 0.6666667  4.000000
```

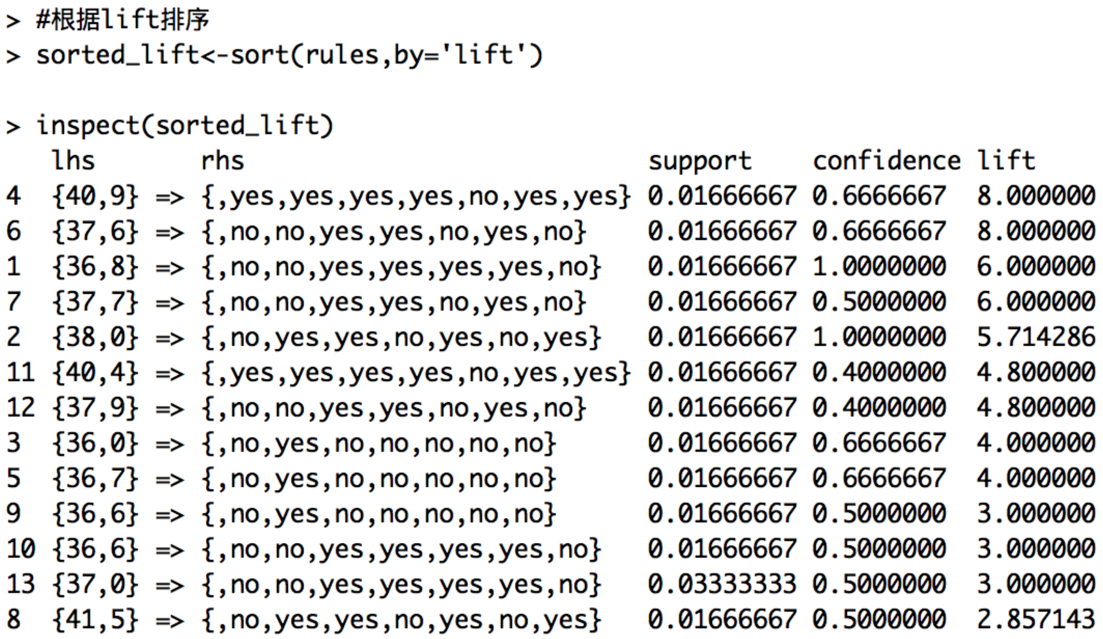
```
> #按置信度查看前6条规则
> inspect(sort(rules,by="confidence")[1:6])
  lhs         rhs                               support    confidence lift
1 {36,8} => {,no,no,yes,yes,yes,yes,no}    0.01666667 1.0000000  6.000000
2 {38,0} => {,no,yes,yes,no,yes,no,yes}    0.01666667 1.0000000  5.714286
3 {36,0} => {,no,yes,no,no,no,no,no}       0.01666667 0.6666667  4.000000
4 {40,9} => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.6666667  8.000000
5 {36,7} => {,no,yes,no,no,no,no,no}       0.01666667 0.6666667  4.000000
6 {37,6} => {,no,no,yes,yes,no,yes,no}     0.01666667 0.6666667  8.000000
```

4、删除冗余的规则



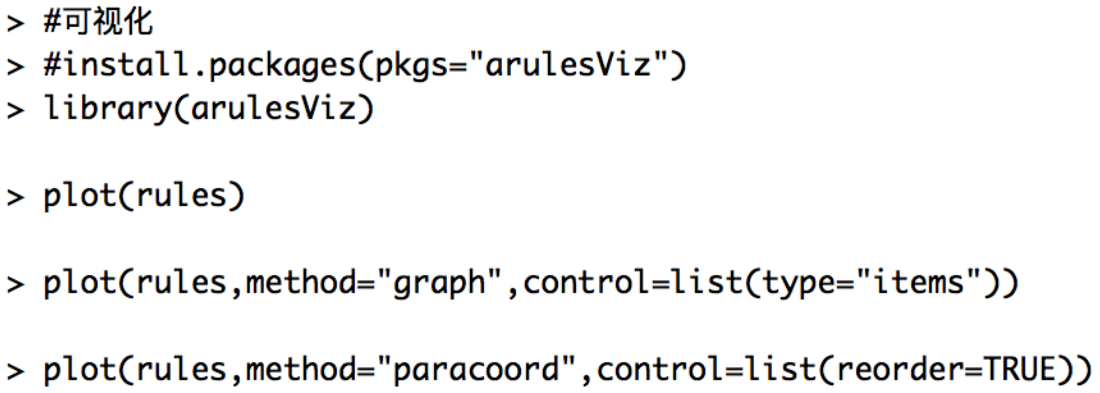
```
> #删除冗余规则
> subset.matrix<-is.subset(rules,rules)

> subset.matrix[lower.tri(subset.matrix,diag = T)]<-NA

> redundant<-colSums(subset.matrix,na.rm = T)>=1

> which(redundant)
named integer(0)

> rules.pruned<-rules[!redundant]

> inspect(rules.pruned)
   lhs        rhs                              support    confidence lift
1  {36,8}  => {,no,no,yes,yes,yes,yes,no}   0.01666667 1.0000000  6.000000
2  {38,0}  => {,no,yes,yes,no,yes,no,yes}   0.01666667 1.0000000  5.714286
3  {36,0}  => {,no,yes,no,no,no,no,no}      0.01666667 0.6666667  4.000000
4  {40,9}  => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.6666667  8.000000
5  {36,7}  => {,no,yes,no,no,no,no,no}      0.01666667 0.6666667  4.000000
6  {37,6}  => {,no,no,yes,yes,no,yes,no}    0.01666667 0.6666667  8.000000
7  {37,7}  => {,no,no,yes,yes,no,yes,no}    0.01666667 0.5000000  6.000000
8  {41,5}  => {,no,yes,yes,no,yes,no,yes}   0.01666667 0.5000000  2.857143
9  {36,6}  => {,no,yes,no,no,no,no,no}      0.01666667 0.5000000  3.000000
10 {36,6}  => {,no,no,yes,yes,yes,yes,no}   0.01666667 0.5000000  3.000000
11 {40,4}  => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.4000000  4.800000
12 {37,9}  => {,no,no,yes,yes,no,yes,no}    0.01666667 0.4000000  4.800000
13 {37,0}  => {,no,no,yes,yes,yes,yes,no}   0.03333333 0.5000000  3.000000
```

5、对规则进行评价，可使用 Lift，也可以使用教材中所提及的其它指标

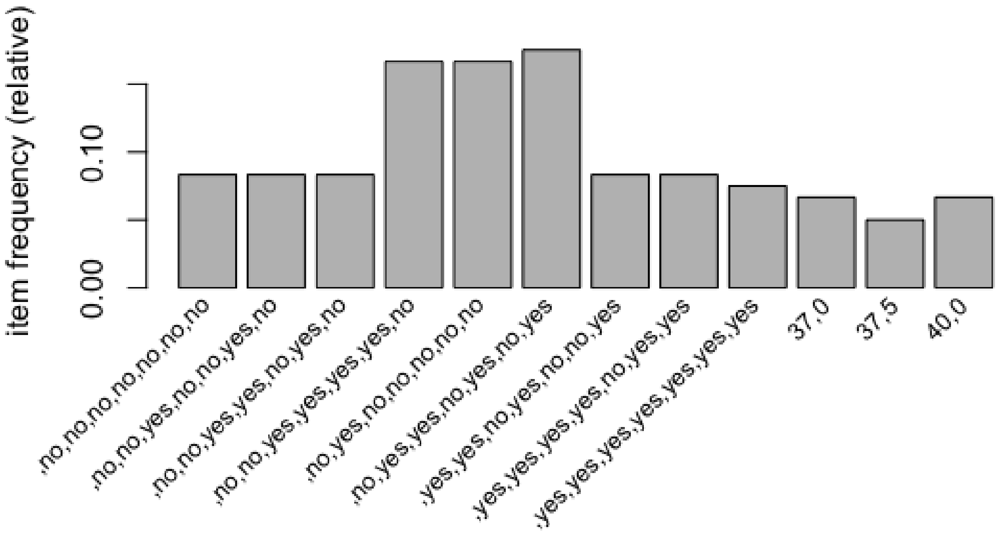


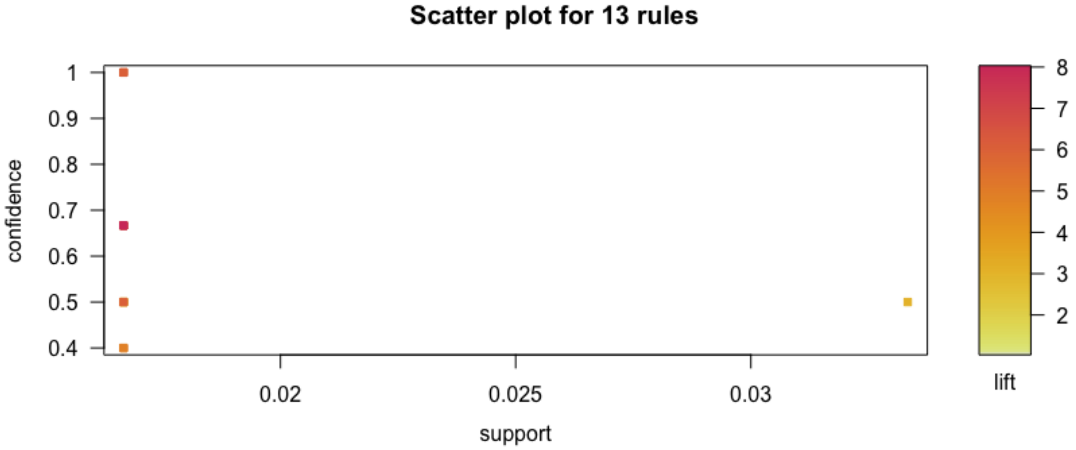
```
> #根据lift排序
> sorted_lift<-sort(rules,by='lift')

> inspect(sorted_lift)
   lhs       rhs                           support    confidence lift
4  {40,9} => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.6666667  8.000000
6  {37,6} => {,no,no,yes,yes,no,yes,no}    0.01666667 0.6666667  8.000000
1  {36,8} => {,no,no,yes,yes,yes,yes,no}   0.01666667 1.0000000  6.000000
7  {37,7} => {,no,no,yes,yes,no,yes,no}    0.01666667 0.5000000  6.000000
2  {38,0} => {,no,yes,yes,no,yes,no,yes}   0.01666667 1.0000000  5.714286
11 {40,4} => {,yes,yes,yes,yes,no,yes,yes} 0.01666667 0.4000000  4.800000
12 {37,9} => {,no,no,yes,yes,no,yes,no}    0.01666667 0.4000000  4.800000
3  {36,0} => {,no,yes,no,no,no,no,no}      0.01666667 0.6666667  4.000000
5  {36,7} => {,no,yes,no,no,no,no,no}      0.01666667 0.6666667  4.000000
9  {36,6} => {,no,yes,no,no,no,no,no}      0.01666667 0.5000000  3.000000
10 {36,6} => {,no,no,yes,yes,yes,yes,no}   0.01666667 0.5000000  3.000000
13 {37,0} => {,no,no,yes,yes,yes,yes,no}   0.03333333 0.5000000  3.000000
8  {41,5} => {,no,yes,yes,no,yes,no,yes}   0.01666667 0.5000000  2.857143
```

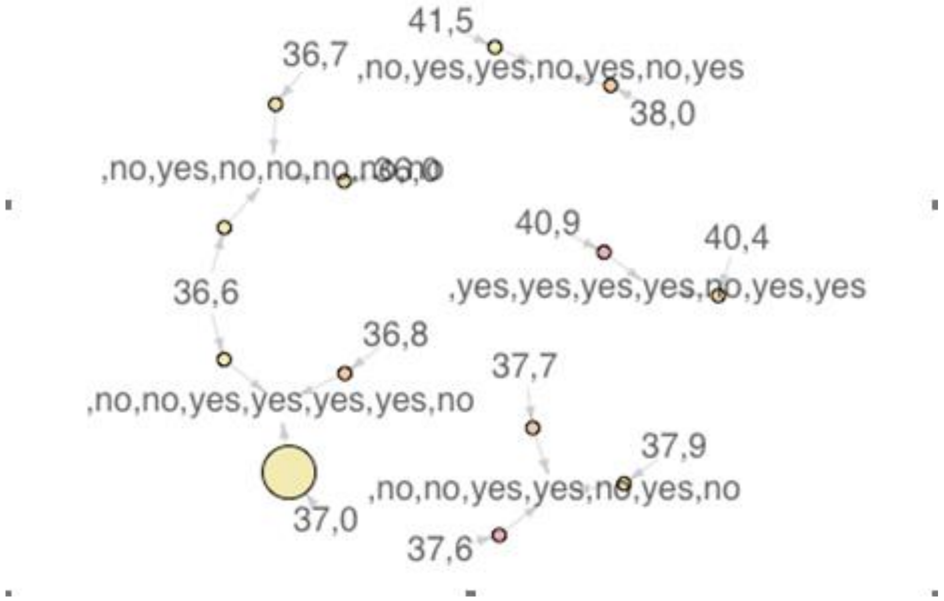
6、使用可视化技术，对规则进行展示。



```
> #可视化
> #install.packages(pkgs="arulesViz")
> library(arulesViz)

> plot(rules)

> plot(rules,method="graph",control=list(type="items"))

> plot(rules,method="paracoord",control=list(reorder=TRUE))
```


item frequency (relative)
0.00
0.10
,no,no,no,no,no,no
,no,no,yes,no,no,yes,no
,no,no,yes,yes,no,yes,no
,no,no,yes,yes,yes,yes,no
,no,yes,no,no,no,no
,no,yes,yes,no,yes,no,yes
,yes,yes,no,yes,no,no,yes
,yes,yes,yes,yes,no,yes,yes
,yes,yes,yes,yes,yes,yes,yes
37,0
37,5
40,0


Scatter plot for 13 rules
confidence
1
0.9
0.8
0.7
0.6
0.5
0.4
support
0.02
0.025
0.03
lift
8
7
6
5
4
3
2


41,5
36,7
,no,yes,yes,no,yes,no,yes
38,0
,no,yes,no,no,no,no
36,0
40,9
40,4
36,6
,yes,yes,yes,yes,no,yes,yes
36,8
37,7
,no,no,yes,yes,yes,yes,no
37,9
,no,no,yes,yes,no,yes,no
37,0
37,6

实验说明：

实验环境：使用 R 软件的 Windows 版本，运行下载文件 R-3.3.0-win.exe。

安装实验中需要的包 arules 和 arulesViz。