homework2

对数据集进行处理

本次作业选择数据集Wine Reviews。对于数据集中的“country”,“points”,“price”三个属性进行关联规则挖掘。“country”为标称属性，代表每种酒的生产国家。“points”为数值属性，表示每种酒的评价得分。“price”为数值属性，代表每种酒的价格。

```
library(arules)
data<-read.csv('C:/Users/YQ/Desktop/datamine/wine-reviews/winemag-data-130k-v2.csv',head=T,sep=',',na.strings = "",nrows=3000)
data1<-data[c("country","points","price")]
wineData=data1[complete.cases(data1),]
head(wineData)
```

```
##    country points price
## 2 Portugal     87    15
## 3       US     87    14
## 4       US     87    13
## 5       US     87    65
## 6    Spain     87    15
## 7    Italy     87    16
```

为了使得该数据集适用于关联规则挖掘，我们对数据进行了一下简单的重构。对于数值属性“points”,“price”进行离散化处理。根据数值的大小，将他们分别分为四个等级，表示评分的高低及价格的高低。并将离散后的数据转换为适用于Apriori算法处理的transactions类型。

```
# wineData<-wineData[c(-3119),]
for(i in 1:nrow(wineData)){
  if(wineData[i,2]>=95){
    wineData[i,2]<-"1st"
  }else if(wineData[i,2]>=90){
    wineData[i,2]<-"2nd"
  }else if(wineData[i,2]>=85){
    wineData[i,2]<-"3rd"
  }else{
    wineData[i,2]<-"4th"
  }

  if(wineData[i,3]>=2000){
    wineData[i,3]<-"1st"
  }else if(wineData[i,3]>=1000){
    wineData[i,3]<-"2nd"
  }else if(wineData[i,3]>=500){
    wineData[i,3]<-"3rd"
  }else{
    wineData[i,3]<-"4th"
  }
}
wineData$points<-as.factor(wineData$points)
wineData$price<-as.factor(wineData$price)
wineTrans<-as(wineData,"transactions")
inspect(wineTrans[1:10])
```

```
##      items                                    transactionID
## [1]  {country=Portugal,points=3rd,price=4th} 2
## [2]  {country=US,points=3rd,price=2nd}       3
## [3]  {country=US,points=3rd,price=2nd}       4
## [4]  {country=US,points=3rd,price=1st}       5
## [5]  {country=Spain,points=3rd,price=2nd}    6
## [6]  {country=Italy,points=3rd,price=2nd}    7
## [7]  {country=France,points=3rd,price=1st}   8
## [8]  {country=Germany,points=3rd,price=2nd}  9
## [9]  {country=France,points=3rd,price=1st}   10
## [10] {country=US,points=3rd,price=2nd}       11
```

关联规则挖掘

导出关联规则，计算其支持度及置信度。下面执行apriori算法，将支持度support的阈值设置为0.01，置信度confidence的阈值设置为0.7。

```
library(arules)
rules<-apriori(wineTrans,parameter =list(minlen=2,support=0.01,confidence=0.7))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.7    0.1    1 none FALSE            TRUE       5    0.01      2
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 28
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[28 item(s), 2805 transaction(s)] done [0.00s].
## sorting and recoding items ... [17 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [15 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
inspect(rules)
```

```
##      lhs                                 rhs          support    confidence
## [1]  {price=4th}                      => {points=3rd} 0.02994652 0.7304348
## [2]  {country=Chile}                  => {points=3rd} 0.03244207 0.7000000
## [3]  {points=2nd}                     => {price=1st}  0.26417112 0.8251670
## [4]  {price=2nd}                      => {points=3rd} 0.27486631 0.7185461
## [5]  {country=US}                     => {price=1st}  0.29910873 0.7003339
## [6]  {country=Argentina,price=2nd}    => {points=3rd} 0.01497326 0.7636364
## [7]  {country=Argentina,points=3rd}   => {price=2nd}  0.01497326 0.7118644
## [8]  {country=Portugal,price=2nd}     => {points=3rd} 0.01497326 0.8235294
## [9]  {country=Chile,price=2nd}        => {points=3rd} 0.02495544 0.8045977
## [10] {country=Chile,points=3rd}       => {price=2nd}  0.02495544 0.7692308
## [11] {country=Spain,price=2nd}        => {points=3rd} 0.01746881 0.7000000
## [12] {country=France,points=2nd}      => {price=1st}  0.03885918 0.7569444
## [13] {country=Italy,points=2nd}       => {price=1st}  0.04171123 0.8602941
## [14] {country=Italy,price=2nd}        => {points=3rd} 0.05062389 0.8352941
## [15] {country=US,points=2nd}          => {price=1st}  0.13368984 0.8503401
##      lift     count
## [1]  1.286170  84
## [2]  1.232580  91
## [3]  1.431412 741
## [4]  1.265237 771
## [5]  1.214865 839
## [6]  1.344633  42
## [7]  1.860932  42
## [8]  1.450094  42
## [9]  1.416759  70
## [10] 2.010897  70
## [11] 1.232580  49
## [12] 1.313067 109
## [13] 1.492347 117
## [14] 1.470810 142
## [15] 1.475080 375
```

根据关联结果中的提升度(life)进行降序排序。

```
sorted_lift<-sort(rules,by='lift')
inspect(sorted_lift)
```

```
##      lhs                                rhs          support    confidence
## [1]  {country=Chile,points=3rd}      => {price=2nd}  0.02495544 0.7692308
## [2]  {country=Argentina,points=3rd}  => {price=2nd}  0.01497326 0.7118644
## [3]  {country=Italy,points=2nd}      => {price=1st}  0.04171123 0.8602941
## [4]  {country=US,points=2nd}         => {price=1st}  0.13368984 0.8503401
## [5]  {country=Italy,price=2nd}       => {points=3rd} 0.05062389 0.8352941
## [6]  {country=Portugal,price=2nd}    => {points=3rd} 0.01497326 0.8235294
## [7]  {points=2nd}                    => {price=1st}  0.26417112 0.8251670
## [8]  {country=Chile,price=2nd}       => {points=3rd} 0.02495544 0.8045977
## [9]  {country=Argentina,price=2nd}   => {points=3rd} 0.01497326 0.7636364
## [10] {country=France,points=2nd}     => {price=1st}  0.03885918 0.7569444
## [11] {price=4th}                     => {points=3rd} 0.02994652 0.7304348
## [12] {price=2nd}                     => {points=3rd} 0.27486631 0.7185461
## [13] {country=Chile}                 => {points=3rd} 0.03244207 0.7000000
## [14] {country=Spain,price=2nd}       => {points=3rd} 0.01746881 0.7000000
## [15] {country=US}                    => {price=1st}  0.29910873 0.7003339
##      lift     count
## [1]  2.010897  70
## [2]  1.860932  42
## [3]  1.492347 117
## [4]  1.475080 375
## [5]  1.470810 142
## [6]  1.450094  42
## [7]  1.431412 741
## [8]  1.416759  70
## [9]  1.344633  42
## [10] 1.313067 109
## [11] 1.286170  84
## [12] 1.265237 771
## [13] 1.232580  91
## [14] 1.232580  49
## [15] 1.214865 839
```

上面满足支持度阈值和置信度阈值的规则存在冗余规则，冗余规则的定义是：如果rules2的lhs和rhs是包含于rules1的，而且rules2的lift小于或者等于rules1，则称rules2是rules1的冗余规则。下面对冗余规则进行删除，最终关联规则精简到11条。

```
subset.matrix<-is.subset(sorted_lift,sorted_lift,sparse = F)
subset.matrix[lower.tri(subset.matrix,diag=T)]<-NA
redundant<-colSums(subset.matrix,na.rm = T)>=1
rules.pruned<-sorted_lift[!redundant]
inspect(rules.pruned)
```

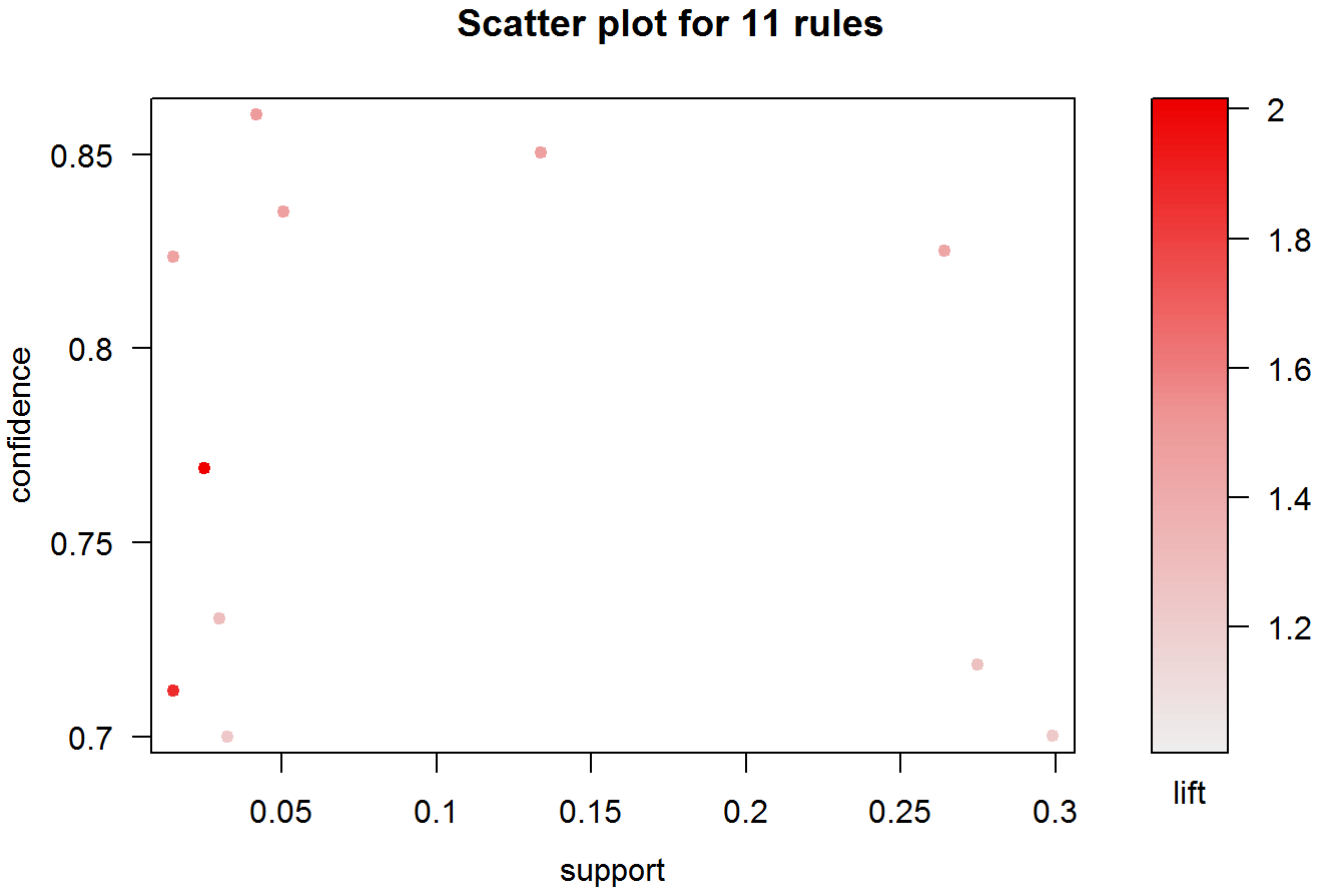
```
##      lhs                                rhs          support    confidence
## [1]  {country=Chile,points=3rd}      => {price=2nd}  0.02495544 0.7692308
## [2]  {country=Argentina,points=3rd} => {price=2nd}  0.01497326 0.7118644
## [3]  {country=Italy,points=2nd}      => {price=1st}  0.04171123 0.8602941
## [4]  {country=US,points=2nd}         => {price=1st}  0.13368984 0.8503401
## [5]  {country=Italy,price=2nd}       => {points=3rd} 0.05062389 0.8352941
## [6]  {country=Portugal,price=2nd}    => {points=3rd} 0.01497326 0.8235294
## [7]  {points=2nd}                    => {price=1st}  0.26417112 0.8251670
## [8]  {price=4th}                     => {points=3rd} 0.02994652 0.7304348
## [9]  {price=2nd}                     => {points=3rd} 0.27486631 0.7185461
## [10] {country=Chile}                 => {points=3rd} 0.03244207 0.7000000
## [11] {country=US}                    => {price=1st}  0.29910873 0.7003339
##      lift     count
## [1]  2.010897  70
## [2]  1.860932  42
## [3]  1.492347 117
## [4]  1.475080 375
## [5]  1.470810 142
## [6]  1.450094  42
## [7]  1.431412 741
## [8]  1.286170  84
## [9]  1.265237 771
## [10] 1.232580  91
## [11] 1.214865 839
```

挖掘结果分析

由关联规则4，11可以看出，由美国生产的酒，价格较高在2000以上的可能性较大。 由关联规则5可以看出，意大利生产的价格在1000到2000的酒，评价得分在80到85分的置信度大约在83.5%左右。 由关联规则7可以看出，评分在90到95分的酒，价格高于2000的可能性较大，置信度约为82.5%。

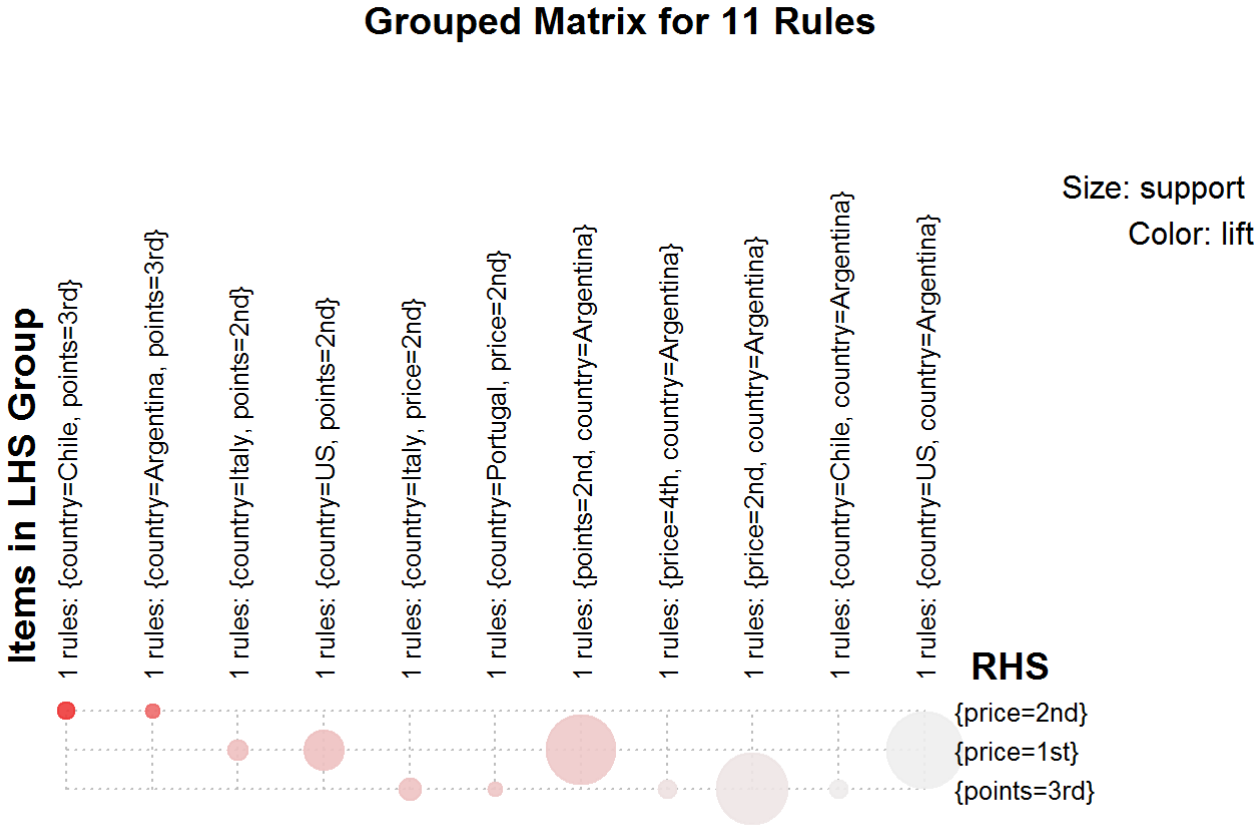
对关联规则可视化

```
library(arulesViz)
plot(rules.pruned)
```



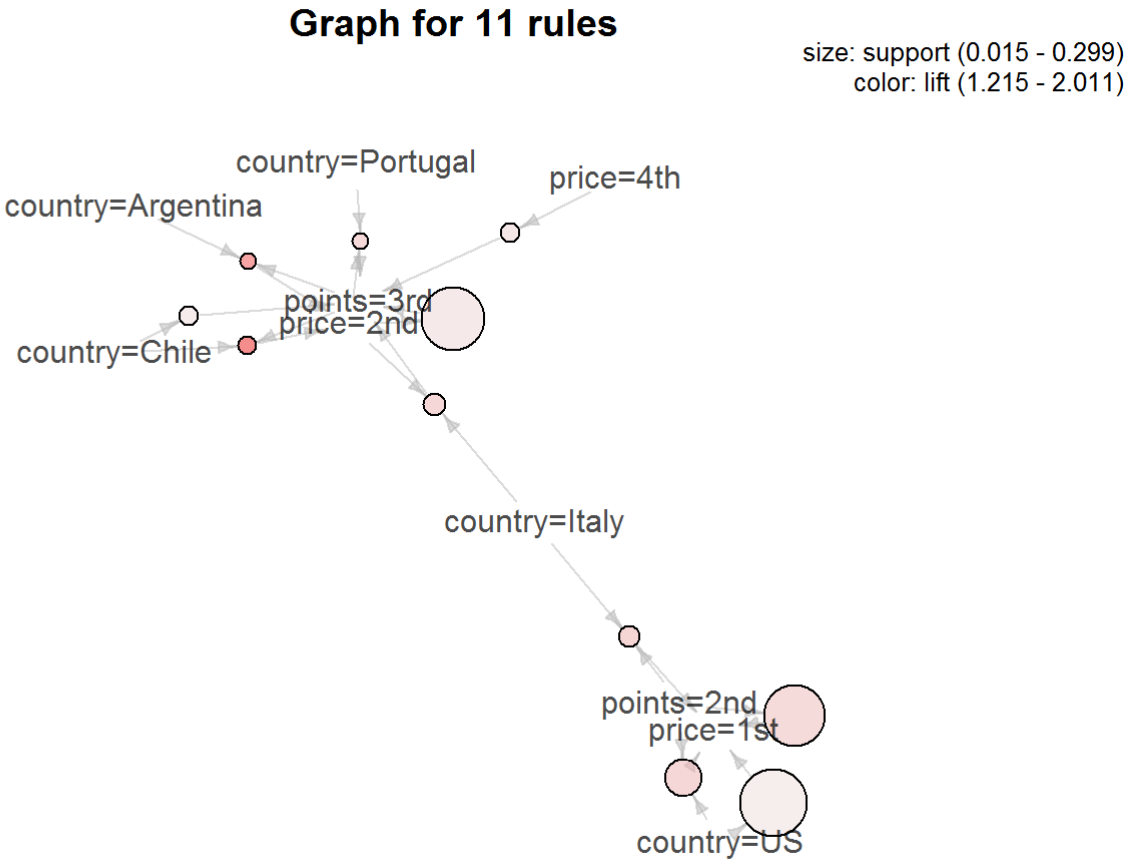
图中的点颜色越深，表示lift值越大，可以看到lift值高的点集中在低support上。

```
plot(rules.pruned, method = "grouped")
```



提升度lift是圈的颜色深浅，圈的大小表示支持度support的大小。LHS的个数和分组中最重要（频繁）项集显示在列的标签里。lift从左上角到右下角逐渐减少。

```
plot(rules.pruned, method = "graph")
```



通过箭头和圆圈来表示关联规则，利用顶点代表项集，边表示规则中关系。圆圈越大表示支持度support越大，颜色越深表示提升度lift越大。