数据挖掘作业二

——关联规则挖掘

姓名： 高建花

班级： 硕士 4 班

学号： 2120171010

关联规则挖掘

1. 问题描述

关联规则挖掘主要用于发现大量数据中项集之间有趣的关联或相关联系。如果两项或多项属性之间存在关联，那么其中一项的属性就可以依据其他属性值进行预测。它在数据挖掘中是一个重要的课题，最近几年已被业界所广泛研究。

关联规则挖掘的一个典型例子是购物篮分析。关联规则研究有助于发现交易数据库中不同商品（项）之间的联系，找出顾客购买行为模式，如购买了某一商品对购买其他商品的影响。分析结果可以应用于商品货架布局、货存安排以及根据购买模式对用户进行分类。

最著名的关联规则是 Apriori 算法。关联规则挖掘问题可以分为两个子问题：第一步是找出事务数据库中所有大于等于用户指定的最小支持度的数据项集，也就是频繁项集；第二步是利用频繁项集生成所需要的关联规则，根据用户设定的最小置信度进行取舍，最后得到强关联规则。识别或发现所有频繁项目集市关联规则发现算法的核心。

本实验利用 Apriori 算法对数据集 San Francisco Building Permits 进行关联规则挖掘，主要实验过程如下。

2. 关联规则的基本描述

(1) **项与项集：**这是一个集合的概念，在一篮子商品中的一件消费品即为一项（Item），则若干项的集合为项集，如{啤酒，尿布}构成一个二元项集。

(2) **关联规则：**一般记为 X=>Y 的形式，X 为先决条件，Y 为相应的关联结果，用于表示数据内隐含的关联性。如：“尿布=>啤酒”表示购买了尿布的消费者往往也会购买啤酒。其关联性强度如何，主要由三个概念——支持度、置信度、提升度来控制和评价。

(3) **支持度（Support）：**支持度是指在所有项集中{X, Y}出现的可能性，即项集中同时含有 X 和 Y 的概率。该指标作为建立强关联规则的第一个门槛，衡量了所考察关联规则在“量”上的多少。通过设定最小阈值（minsup），剔除“出镜率”较低的无意义规则，保留出现较为频繁的项集所隐含的规则。

(4) **置信度（Confidence）：**置信度表示在先决条件 X 发生的条件下，关联结果 Y 发生的概率。这是生成强关联规则的第二个门槛，衡量了所考察的关联规则在“质”上的可靠性。相似的，我们需要对置信度设定最小阈值（mincon）来实现进一步筛选。

(5) **提升度（lift）：**提升度表示在含有 X 的条件下同时含有 Y 的可能性与没有 X 这个条件下项集中含有 Y 的可能性之比：公式为 confidence(artichok => cracker)/support(cracker) = 80%/50% = 1.6。该指标与置信度同样衡量规则的可靠性，可以看作是置信度的一种互补指标。

3. 实验环境

Item	Description
Language	R
IDE	RGui
Package	arules; arulesViz

4. 关联规则挖掘

4.1 数据转换

利用 read.transactions 函数在读入数据的同时，将数据转换为 transactions 格式：

trans_data <- read.transactions("Building_Permits.csv")

4.2 频繁项集

利用 Apriori 算法得到其支持度大于等于 0.1，置信度大于 0.8，最大长度为 10 的频繁项集如下所示：

```
> freq_sets <- apriori(trans_data, parameter=list
+ (support=0.1, maxlen=10, minlen=2, target="frequent itemsets"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
         NA    0.1    1 none FALSE            TRUE       5     0.1      2     10
            target   ext
 frequent itemsets FALSE

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

Absolute minimum support count: 19890

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[929925 item(s), 198901 transaction(s)] done [1.49s].
sorting and recoding items ... [20 item(s)] done [0.12s].
creating transaction tree ... done [0.12s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [134 set(s)] done [0.00s].
creating S4 object  ... done [0.12s].
```

可见设定阈值后，满足条件的交易数据共有 134 项，对频繁项集排序后查看前五项：

```
> inspect(sort(freq_sets, by="support")[1:5])
    items                              support   count
[1] {alterations,frame}                0.5715909 113690
[2] {(5),5,wood,frame}                 0.5627222 111926
[3] {(5),5,wood,alterations}           0.5627171 111925
[4] {(5),5,wood,alterations,frame}     0.5627171 111925
[5] {family,frame}                     0.3471627  69051
```

4.3 关联规则

利用 Apriori 算法导出满足支持度大于等于 0.1，置信度大于等于 0.8，最大长度为 10 的关联规则如下所示：

```
> rules = apriori(trans_data, parameter=list(support=0.1, minlen=2))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE       5     0.1      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: 19890

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[929925 item(s), 198901 transaction(s)] done [1.29s].
sorting and recoding items ... [20 item(s)] done [0.13s].
creating transaction tree ... done [0.10s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [223 rule(s)] done [0.00s].
creating S4 object  ... done [0.12s].
```

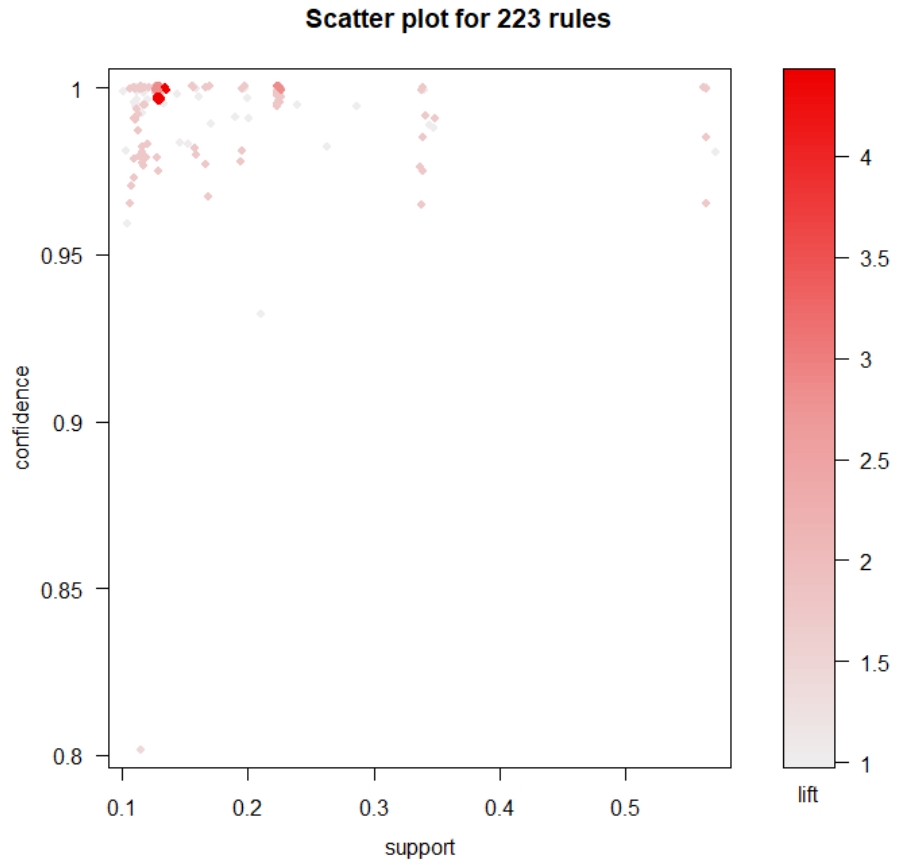
以置信度为标准对其排序后，查看 rules 的前五项

```
> inspect(sort(rules, by="support")[1:10])
     lhs                              rhs             support   confidence lift     count
[1]  {frame}                       => {alterations}  0.5715909 0.9808641  1.008206 113690
[2]  {(5),5,wood}                  => {frame}        0.5627222 1.0000000  1.716025 111926
[3]  {frame}                       => {(5),5,wood}   0.5627222 0.9656452  1.716025 111926
[4]  {(5),5,wood}                  => {alterations}  0.5627171 0.9999911  1.027866 111925
[5]  {(5),5,wood,frame}            => {alterations}  0.5627171 0.9999911  1.027866 111925
[6]  {(5),5,wood,alterations}      => {frame}        0.5627171 1.0000000  1.716025 111925
[7]  {alterations,frame}           => {(5),5,wood}   0.5627171 0.9844753  1.749487 111925
[8]  {family}                      => {frame}        0.3471627 0.9909730  1.700534  69051
[9]  {family}                      => {alterations}  0.3462879 0.9884759  1.016030  68877
[10] {family,frame}                => {alterations}  0.3432260 0.9886606  1.016219  68268
```

利用 summary 函数查看 rules 的情况，可看到其支持度、置信度和提升度等值的统计信息：

```
> summary(rules)
set of 85238 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9    10
  796  5362 12596 16698 16163 13797 10563  6443  2820

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   5.000   6.000   6.051   7.000  10.000

summary of quality measures:
    support          confidence          lift             count
 Min.   :0.01000   Min.   :0.5000   Min.   : 0.734   Min.   :  1990
 1st Qu.:0.01097   1st Qu.:0.9298   1st Qu.: 1.709   1st Qu.:  2182
 Median :0.01352   Median :0.9860   Median : 2.855   Median :  2690
 Mean   :0.01671   Mean   :0.9232   Mean   :10.327   Mean   :  3323
 3rd Qu.:0.01861   3rd Qu.:1.0000   3rd Qu.:14.785   3rd Qu.:  3702
 Max.   :0.57159   Max.   :1.0000   Max.   :67.834   Max.   :113690

mining info:
       data ntransactions support confidence
 trans data        198901    0.01        0.5
```

可视化 rules 如下（注：此处的置信度和支持度标准较高，所以图上显示的不是很密集）



4.4 评价

利用 subset 函数以及条件逻辑表达式来得到所需要的关联规则子集，举例如下（此处是在右端项中包含 frame 且提升度大于 1 的子集）：

```
> sub.rules = subset(rules, subset=rhs%in%"frame"&lift>1)
> inspect(sort(sub.rules, by="lift")[1:10])
     lhs                                          rhs     support   confidence lift     count
[1]  {dwelling,1,0,,5,wood}                    => {frame} 0.1290391 1          1.716025  25666
[2]  {(5),5,wood}                              => {frame} 0.5627222 1          1.716025 111926
[3]  {dwelling,1,0,,5,wood,dwelling,1,1}       => {frame} 0.1286167 1          1.716025  25582
[4]  {dwelling,1,0,,5,wood,family}             => {frame} 0.1290391 1          1.716025  25666
[5]  {(5),5,wood,dwelling,1,0,,5,wood}         => {frame} 0.1288882 1          1.716025  25636
[6]  {alterations,dwelling,1,0,,5,wood}        => {frame} 0.1290391 1          1.716025  25666
[7]  {(5),5,wood,with}                         => {frame} 0.1138054 1          1.716025  22636
[8]  {(5),5,wood,in}                           => {frame} 0.1112614 1          1.716025  22130
[9]  {(5),5,wood,dwelling,1,1}                 => {frame} 0.2246394 1          1.716025  44681
[10] {(5),5,wood,new}                          => {frame} 0.1193207 1          1.716025  23733
```

最后将挖掘结果写入文件：

```
df.rules = as(rules, "data.frame")
write.csv(df.rules, "Rules_BP.csv")
```

4.5 举例

数据集较大的时候往往不好把握项集之间的关系，所以本小节以 Permit Type, Street Suffix, Current Status 三项的关联规则挖掘为例来说明一下关联规则挖掘的一般流程。

```r
path = "D:/BIT/课程/2017-2018下/数据挖掘2018/作业/HomeWork2/"
setwd(path)

#读取并转换数据
data <- read.csv("Building_Permits.csv")
data1 <- data.frame(
    permittype = as.factor(data$Permit.Type),
    streetsuffix = as.factor(data$Street.Suffix),
    currentstatus = as.factor(data$Current.Status)
)
trans_data1 <- as(data1, "transactions")
inspect(trans_data1[1:3])

#频繁项集
freq1 <- apriori(trans_data1, parameter=list
    (support=0.1, maxlen=10, minlen=2, target="frequent itemsets"))
inspect(sort(freq1, by="support")[1:10])

#关联规则
rules = apriori(trans_data1, parameter=list(support=0.1, minlen=2))
inspect(sort(rules, by="support")[1:7])
```

5-12 行代码用于从整个数据中提取出 Permit Type, Street Suffix 和 Current Status 三列的数据，然后将其转换为 transactions 类型。查看其前三行如下所示：

```
> > inspect(trans_data1[1:3])
    items                                                  transactionID
[1] {permittype=4,streetsuffix=St,currentstatus=expired}   1
[2] {permittype=4,streetsuffix=St,currentstatus=issued}    2
[3] {permittype=3,streetsuffix=Av,currentstatus=withdrawn} 3
```

15-17 行代码用于得到满足相应阈值的频繁项集，共得到 10 项，按支持度从高到低排序，查看如下所示

```
> inspect(sort(freq1, by="support")[1:10])
     items                                                  support   count
[1]  {permittype=8,streetsuffix=St}                         0.6255807 124428
[2]  {permittype=8,currentstatus=complete}                  0.4546606  90432
[3]  {permittype=8,currentstatus=issued}                    0.3876370  77101
[4]  {streetsuffix=St,currentstatus=complete}               0.3336199  66357
[5]  {permittype=8,streetsuffix=St,currentstatus=complete}  0.3098240  61624
[6]  {streetsuffix=St,currentstatus=issued}                 0.2980845  59289
[7]  {permittype=8,streetsuffix=St,currentstatus=issued}    0.2762092  54938
[8]  {permittype=8,streetsuffix=Av}                         0.1968024  39144
[9]  {streetsuffix=Av,currentstatus=complete}               0.1095324  21786
[10] {permittype=8,streetsuffix=Av,currentstatus=complete}  0.1037707  20640
```

20-21 行代码用于得到满足相应阈值的关联规则，共得到 7 项，按支持度从高到低排序，查看如下所示：

```
> inspect(sort(rules, by="support")[1:7])
    lhs                                                rhs            support   confidence lift     count
[1] {streetsuffix=St}                               => {permittype=8} 0.6255807 0.8993192  1.000171 124428
[2] {currentstatus=complete}                        => {permittype=8} 0.4546606 0.9315492  1.036015  90432
[3] {currentstatus=issued}                          => {permittype=8} 0.3876370 0.9227133  1.026189  77101
[4] {streetsuffix=St,currentstatus=complete}        => {permittype=8} 0.3098240 0.9286737  1.032817  61624
[5] {streetsuffix=St,currentstatus=issued}          => {permittype=8} 0.2762092 0.9266137  1.030526  54938
[6] {streetsuffix=Av}                               => {permittype=8} 0.1968024 0.9057128  1.007282  39144
[7] {streetsuffix=Av,currentstatus=complete}        => {permittype=8} 0.1037707 0.9473974  1.053641  20640
```

以第四条规则为例说明其意义：在所有项集中，

streetsuffix 为“St”，currentstatus 为“complete”，permmittype 为 8 同时成立的概率

为 30.98%；

当 streetsuffix 为“St”且 currentstatus 为“complete”时，permmittype 为 8 的概率为 92.87%；

Lift(提升度)大于 1，说明该条规则是有效的强关联规则。

5. 实验总结

本实验利用 R 语言实现了对数据集 Building Permits 的关联规则挖掘。首先简单介绍了关联规则挖掘中的一些基本且核心的概念，然后利用 R 语言对整个数据集进行关联规则挖掘，其中使用了 Apriori 算法来得到频繁项集和关联规则。最后以数据集中 Permit Type, Street Suffix 和 Current Status 为例说明了关联规则挖掘的一般流程，并对得到的规则进行了分析和评价。