**关联规则挖掘**

姓名：王子硕

学号：2120151044

# 实验要求

对数据集进行处理，转换成适合关联规则挖掘的形式；找出频繁项集；导出关联规则，计算其支持度和置信度；去除冗余的规则；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；使用可视化技术，如散点图、平行坐标、泡泡图等，对规则进行展示。

# 实验数据

数据集: Titanic存活数据

来源：<https://www.kaggle.com/c/titanic/data>

目标：寻找能够预测Survival值的关联规则。

# 数据预处理

## 数据属性选择

数据分析针对泰坦尼克号旅客的存活状况，包括十二个属性：id、survival、name、age、sex、pclass、ticket、carbin、fare、emvarked。预处理过程中将其中一部分无关属性剔除掉，只剩pclass、survival、sex、age、sibsp、parch、embarked。

## 数据预处理

为了提取关联规则，我将选取后的数据离散化，数字化。预处理过程中，主要完成以下工作；

1 提出无用属性；

2 数字化离散型属性；

3 对于Age等连续性属性，将其离散化、数字化。

## 数据预处理结果

最后会在输出目录下生成两个文件：map.txt 和 data.csv.

# 关联规则提取

## Apriori介绍

Apriori算法使用频繁项集的先验知识，使用一种称作逐层搜索的迭代方法，k项集用于探索 (k+1) 项集。首先，通过扫描事务（交易）记录，找出所有的频繁1项集，该集合记做L1，然后利用L1找频繁2项集的集合L2，L2找L3，如此下去，直到不能再找到任何频繁k项集。最后再在所有的频繁集中找出强规则，即产生用户感兴趣的关联规则。

其中，Apriori算法具有这样一条性质：任一频繁项集的所有非空子集也必须是频繁的。 因为假如P(I)< 最小支持度阈值，当有元素A添加到I中时，结果项集（A∩I）不可能比I出现次数更多。因此A∩I也不是频繁的。

## 程序设计与实现

在实验中，规则挖掘的过程是使用R语言实现的。

（1）导入数据并格式化

R代码如下：

titanic<- read.csv("/Users/bitrsky/BitrSky/Study/Student/DataMining/Second/Data/data.csv", header = T,colClasses = "factor")

df <- titanic

titanic.raw<-data.frame(factor(df$Survived),factor(df$Pclass),factor(df$Sex), factor(df$Age),factor(df$SibSp),factor(df$Parch),factor(df$Embarked))

names(titanic.raw)<-names(df)

（2）不设定目标类提取关联规则

R代码如下：

rules.all <- apriori(titanic.raw)

summary(rules.all)

inspect(rules.all)

结果如下：

summary of quality measures:

support confidence lift

Min. :0.1021 Min. :0.8045 Min. :1.065

1st Qu.:0.1212 1st Qu.:0.8472 1st Qu.:1.241

Median :0.2593 Median :0.8717 Median :1.285

Mean :0.2569 Mean :0.8834 Mean :1.299

3rd Qu.:0.3547 3rd Qu.:0.9192 3rd Qu.:1.370

Max. :0.6027 Max. :1.0000 Max. :2.522

mining info:

data ntransactions support confidence

titanic.raw 891 0.1 0.8

titanic.raw 891 0.1 0.8

lhs rhs support confidence lift

1 {Age=5} => {Parch=0} 0.1032548 0.8679245 1.140591

2 {Age=4} => {Parch=0} 0.1066218 0.8333333 1.095133

3 {Age=-1} => {Parch=0} 0.1762065 0.8870056 1.165667

4 {Pclass=2} => {Embarked=0} 0.1840629 0.8913043 1.233156

5 {Survived=0} => {Sex=0} 0.5252525 0.8524590 1.316362

6 {Sex=0} => {Survived=0} 0.5252525 0.8110919 1.316362

7 {Survived=0} => {Parch=0} 0.4994388 0.8105647 1.065211

8 {Sex=0} => {Parch=0} 0.5432099 0.8388215 1.102345

9 {SibSp=1} => {Parch=0} 0.6026936 0.8832237 1.160697

10 {Survived=0,

Age=-1} => {Pclass=0} 0.1144781 0.8160000 1.480766

11 {Pclass=0,

Age=-1} => {Parch=0} 0.1313131 0.8602941 1.130564

12 {Survived=0,

Age=-1} => {Sex=0} 0.1212121 0.8640000 1.334184

13 {Sex=0,

Age=-1} => {Survived=0} 0.1212121 0.8709677 1.413538

14 {Survived=0,

Age=-1} => {SibSp=1} 0.1144781 0.8160000 1.195816

15 {Survived=0,

Age=-1} => {Parch=0} 0.1234568 0.8800000 1.156460

……

一共产生123条关联规则。

## 以Survival为目标类提取关联规则

R代码如下：

rules2 <- apriori(titanic.raw, parameter=list(minlen=2, supp=0.005, conf=0.8),

appearance = list(rhs=c("Survived=0","Survived=1"),

default ="lhs")

)

rules2.sorted <-sort(rules2,by="lift")

subset.matrix<-is.subset(rules2.sorted,rules2.sorted)

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

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

which(redundant)

rules.pruned <-rules2.sorted[!redundant]

inspect(rules.pruned)

提取的关联规则按lift排序后前20如下：

lhs rhs support confidence lift

1 {Pclass=2,

Age=1} => {Survived=1} 0.005611672 1.0000000 2.605263

2 {Age=1,

SibSp=1} => {Survived=1} 0.006734007 1.0000000 2.605263

3 {Age=0,

Embarked=1} => {Survived=1} 0.007856341 1.0000000 2.605263

4 {Pclass=2,

Age=0} => {Survived=1} 0.013468013 1.0000000 2.605263

5 {Sex=1,

SibSp=4,

Parch=1} => {Survived=1} 0.005611672 1.0000000 2.605263

6 {Sex=1,

SibSp=4,

Embarked=1} => {Survived=1} 0.005611672 1.0000000 2.605263

7 {Sex=1,

Age=10,

SibSp=0} => {Survived=1} 0.006734007 1.0000000 2.605263

8 {Sex=1,

Age=10,

Embarked=0} => {Survived=1} 0.007856341 1.0000000 2.605263

9 {Age=0,

SibSp=1,

Parch=2} => {Survived=1} 0.005611672 1.0000000 2.605263

10 {Age=0,

SibSp=0,

Parch=1} => {Survived=1} 0.012345679 1.0000000 2.605263

11 {Pclass=0,

Age=0,

SibSp=0} => {Survived=1} 0.005611672 1.0000000 2.605263

12 {Sex=0,

Age=0,

SibSp=0} => {Survived=1} 0.010101010 1.0000000 2.605263

13 {Pclass=1,

Sex=1,

Age=9} => {Survived=1} 0.006734007 1.0000000 2.605263

14 {Sex=1,

Age=8,

Parch=1} => {Survived=1} 0.005611672 1.0000000 2.605263

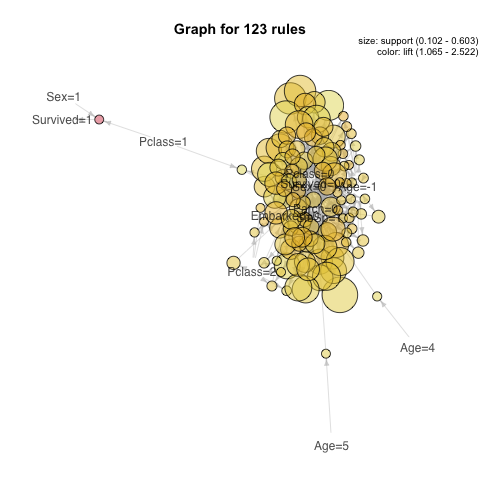
15 {Sex=1,

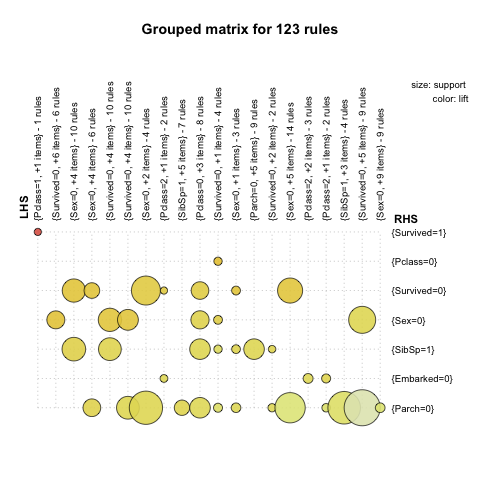
Age=8,

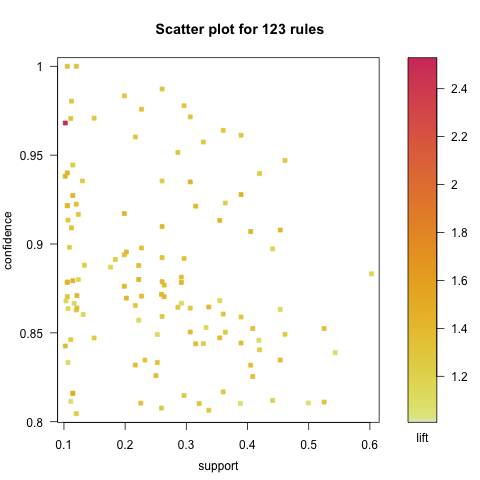
Embarked=1} => {Survived=1} 0.005611672 1.0000000 2.605263

# 关联规则结果可视化

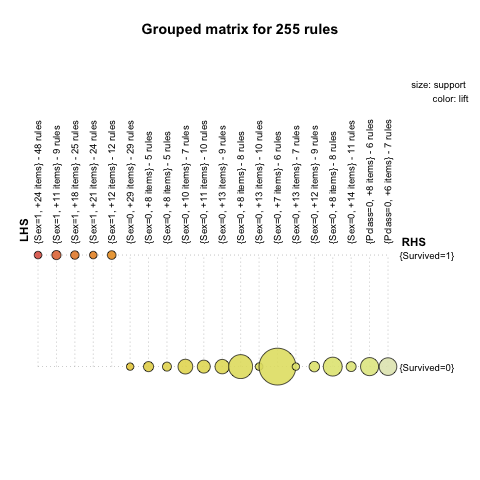
## 可视化无目标类关联规则

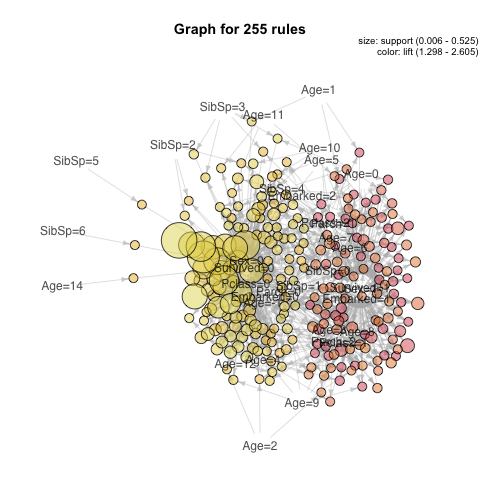






## 可视化以Survival为目标类的关联规则





# 挖掘结果分析

## 关联规则

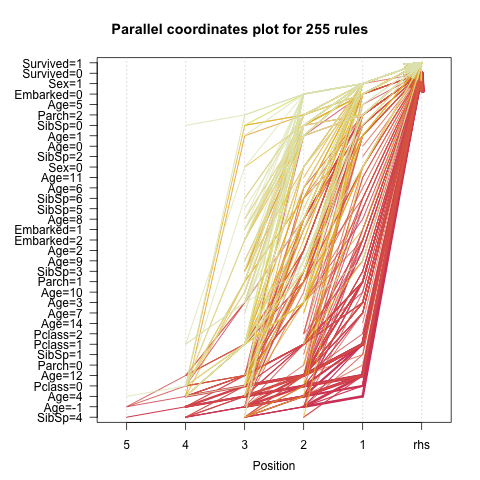


图3 关联规则 Parallel coordinates plot

目标属性为Survived，黄线表示Survived=1的关联规则，红线表示Survived=0的规则，Age=-1表示Age为空的数据。对于Survived=1的关联规则，如下为Lift为前5的规则：

1 {Pclass=2,

Age=1} => {Survived=1} 0.005611672 1.0000000 2.605263

2 {Age=1,

SibSp=1} => {Survived=1} 0.006734007 1.0000000 2.605263

3 {Age=0,

Embarked=1} => {Survived=1} 0.007856341 1.0000000 2.605263

4 {Pclass=2,

Age=0} => {Survived=1} 0.013468013 1.0000000 2.605263

5 {Sex=1,

SibSp=4,

Parch=1} => {Survived=1} 0.005611672 1.0000000 2.605263

经分析，Pclass,Age,Sibsp,Embarked,Sex等属性对于Survived的判断有相关性。其中，Age、Pclass和Sex的相关性较强。

## 强相关属性

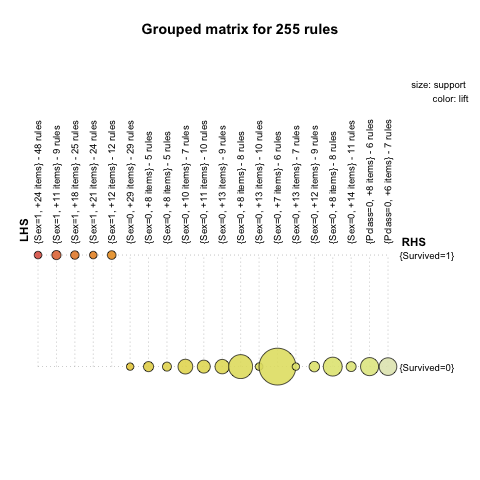


图4 Grouped matrix for 255 rules

图4表明Sex与Survived之间存在很强的相关性。在Survived=1关联规则中， Sex=1占很大的比例，表示女性的存活率大于男性。