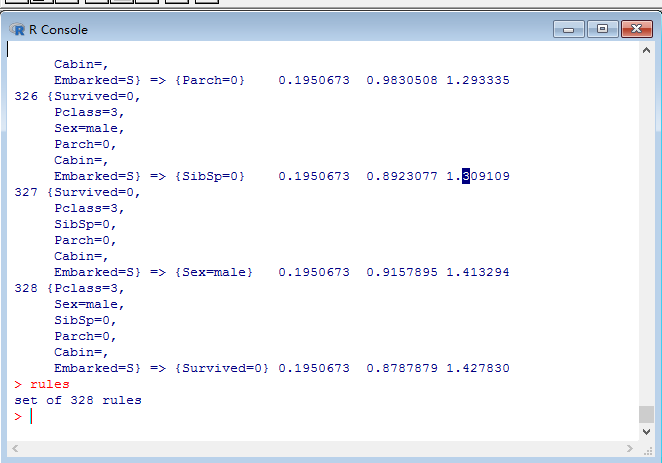
# 数据挖掘作业2

**计算机学院 2120150980 谌重喜**

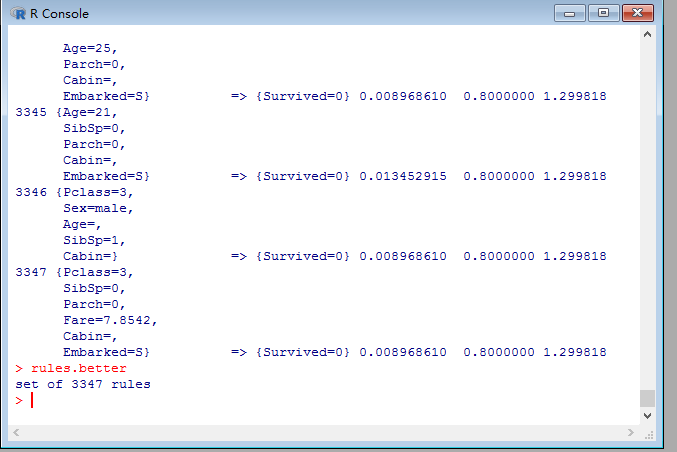
数据集：<https://www.kaggle.com/c/titanic/download/train.csv>

环境：win10+R语言3.3.1

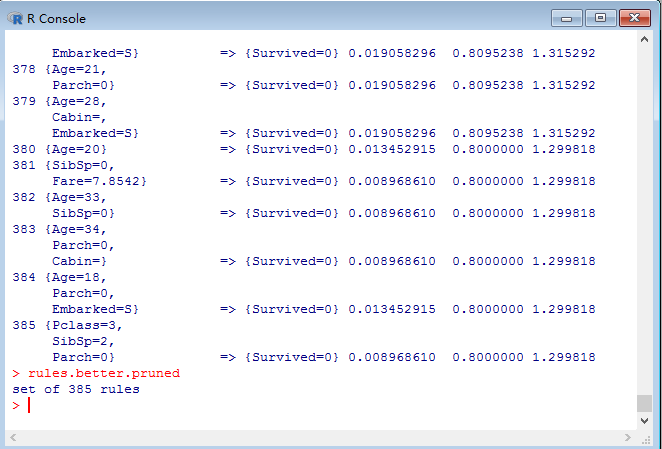
1. 加载库：arules，用apriori算法进行关联分析



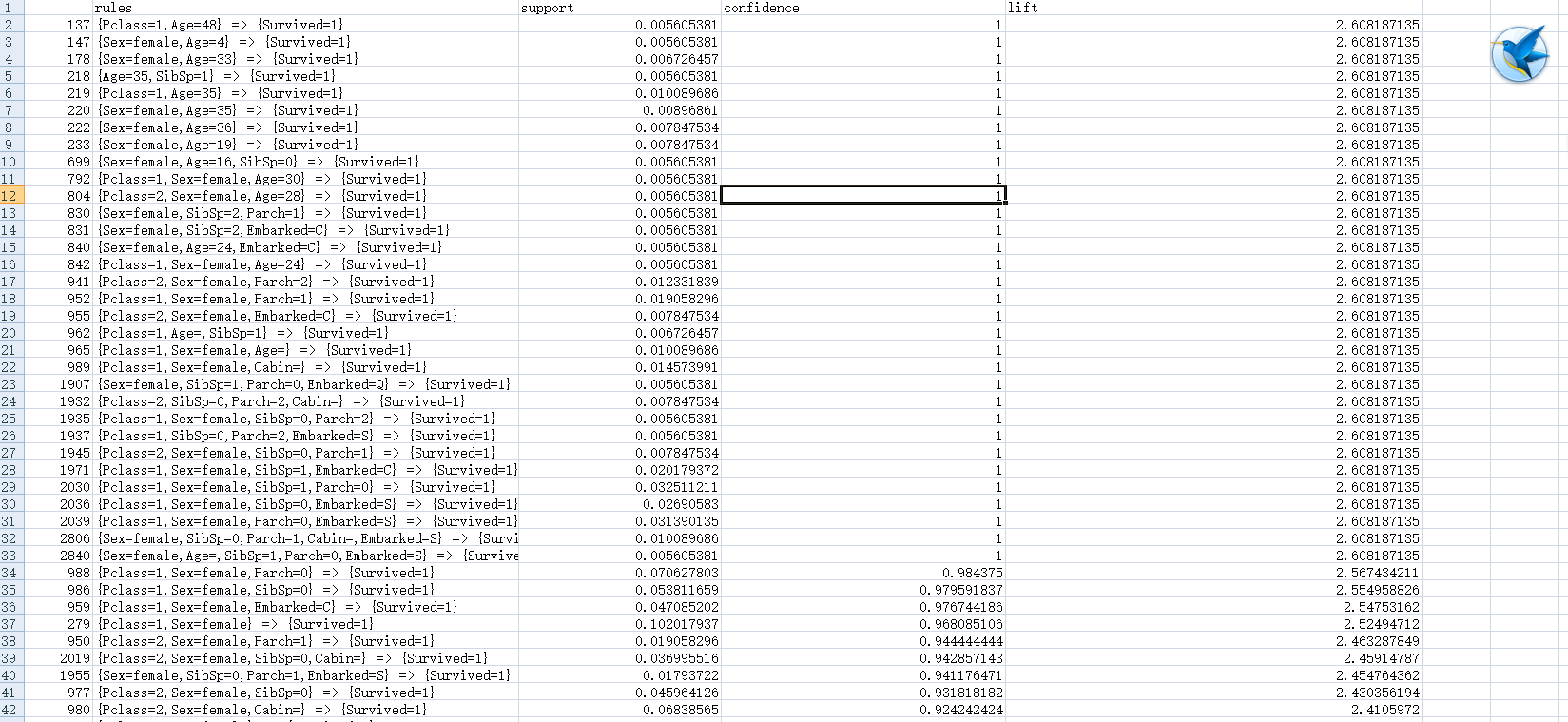
1. 设置条件，若看生存情况，条件就设置为右边的集合rhs为rhs=c("Survived=0","Survived=1")，最小关联项、支持度、置信度分别设置为2、0.005、0.8，结果如图：



1. 去除冗余的规则，结果如图：

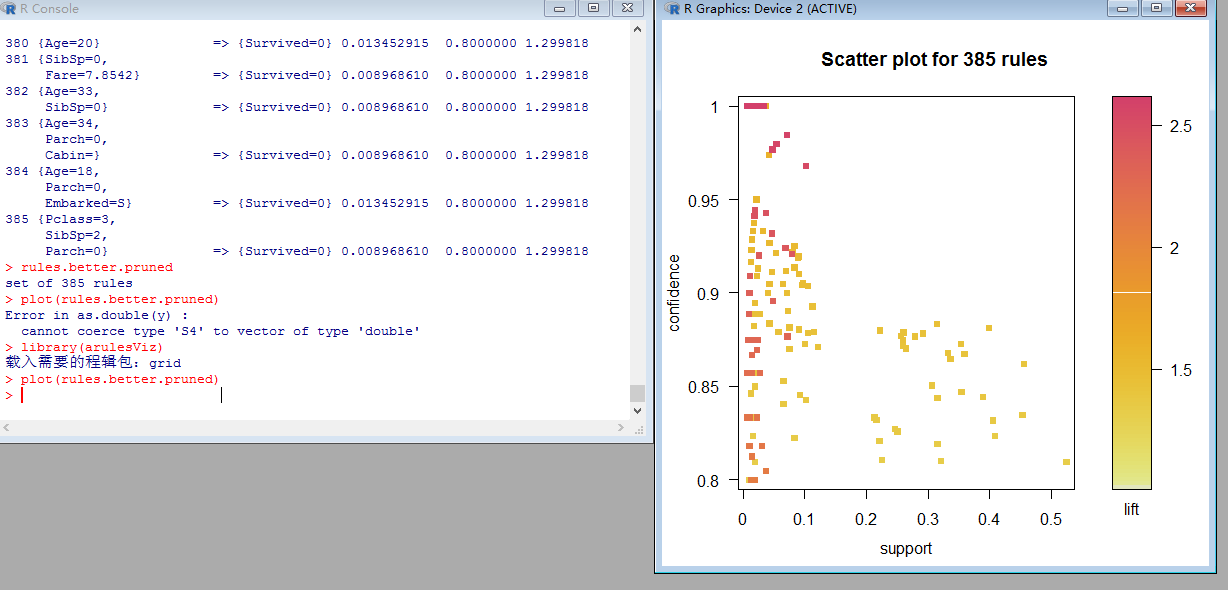


1. 对结果进行评价，因为有385条相关的规则，导出结果，从挖出的规则来分析，如图：



可以看出最明显的结果是置信度为1的情况下，生存下来的人的条件有这么几个特征：Pclass=1、Sex=female，头等舱的女乘客生存几率最大

1. 画图需要加载arulesViz库，结果如图：



**代码：**

|  |
| --- |
| setwd("C:/Users/usr/Desktop/1")  df <- read.table("train.csv", sep=",", header=T)  #为了让table的列格式正确，主要是arules库不支持数字，对纯数字的段加了点符号，让读的时候都城字符串，算法才能处理  library(arules)  rules <- apriori(df)  inspect(rules)  #保留结果中包含生存变量的关联规则  rules.better <-apriori(df, parameter=list(minlen=2,supp =0.005,conf =0.8), appearance= list(rhs=c("Survived=0","Survived=1"),default="lhs"), control= list(verbose=F))  rules.better.sorted<-sort(rules.better,by="lift")  inspect(rules.better.sorted)  #去除冗余的规则  # find redundant rules  subset.matrix <- is.subset(rules.better.sorted, rules.better.sorted)  subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA  redundant <- colSums(subset.matrix, na.rm=T) >= 1  which(redundant)  # remove redundant rules  rules.better.pruned <- rules.better.sorted[!redundant]  inspect(rules.better.pruned)  #画图  library(arulesViz)  plot(rules)  plot(rules.better)  plot(rules.better.pruned) |