Titanic存活数据关联规则挖掘源代码

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)