关联规则挖掘

在本次作业中，选取数据量较少的数据集Building\_Permits.csv进行关联规则的挖掘；另外，根据之前数据预处理的结果以及关联规则挖掘的要求，剔除类别庞大、缺失值过多、以及取值重复会互相干扰的属性，最终选择Current Status、Fire Only Permit、Supervisor District、Zipcode四个属性建立关联规则挖掘。

1、对数据集进行处理，转换成适合关联规则挖掘的形式

此次依然使用R语言，调用arules包（library(arules) #加载arules程序包）中的关联规则挖掘函数，但需要将需要挖掘的数据集格式转化为可处理的事务集格式。

程序函数：read.transactions（' ', format = "basket", sep=','）

2、关联规则挖掘

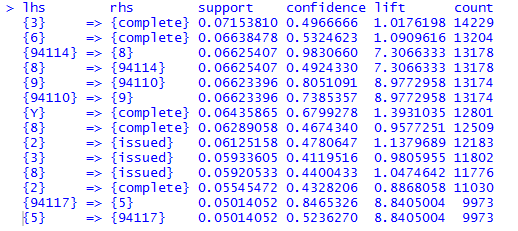
采用Apriori算法进行关联规则挖掘，该算法是一种最有影响的挖掘布尔关联规则频繁项集的算法，核心是基于两阶段频集思想的递推算法。在这里，频繁项集是指支持度大于最小阈值的项集。

程序函数：apriori(数据集,parameter=list(support= ,confidence= ,minlen = ,target="rules"))

其中，规则设置支持度最小阈值为0.05、置信度0.3.

Support（支持度）：表示同时包含A和B的事务占所有事务的比例。如果用P(A)表示使用A事务的比例，那么Support=P(A&B)

Confidence（可信度）：表示使用包含A的事务中同时包含B事务的比例，即同时包含A和B的事务占包含A事务的比例。公式表达：Confidence=P(A&B)/P(A)

3、挖掘结果和分析

如图所示为关联规则挖掘的结果，包括14条长度为2的关联规则，根据支持度对求得的频繁项集排序查看，程序函数：inspect。

通过提升度Lift对关联规则进行评价——提升度反映了关联规则中的A与B的相关性，提升度>1且越高表明正相关性越高，提升度<1且越低表明负相关性越高，提升度=1表明没有相关性。

由结果可见，属性Fire Only Permit只有一条关联规则，即“Y”与“complete”，提升度为1.3931035表明两者呈正相关；关于与属性Current Status有关的相关规则大多提升度在1左右，或呈正相关或呈负相关，但多接近于1，表明相关性不大；属性Supervisor District与Zipcode相关规则的提升度高达8左右，如“8”与“94114”、“9”与“94110”、“5”与“94117”，说明其两者间呈正相关且相关性较高，我们有理由相信两属性间有较好的相关性。