**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。

1. **实验环境**

|  |  |
| --- | --- |
| **Item** | **Description** |
| Language | R |
| IDE | RGui |
| Package | arules; arulesViz |

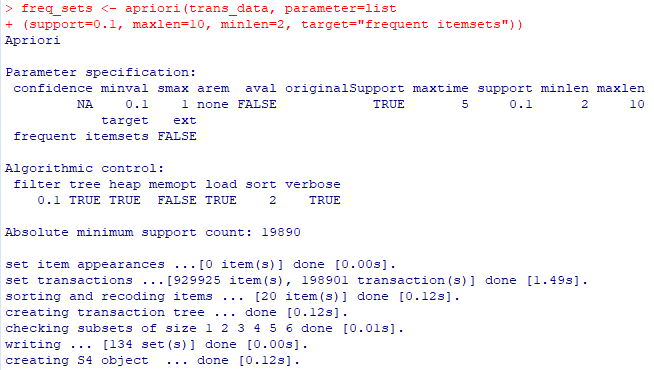
1. **关联规则挖掘**
   1. **数据转换**

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

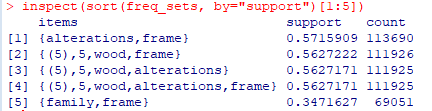
trans\_data <- read.transactions("Building\_Permits.csv")

* 1. **频繁项集**

利用Apriori算法得到满足其支持度、置信度、最大长度等阈值的频繁项集：

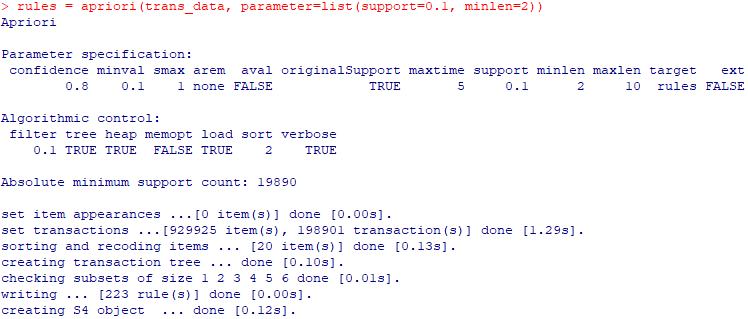


可见设定阈值后，满足条件的交易数据共有134项，对频繁项集排序后查看前五项：

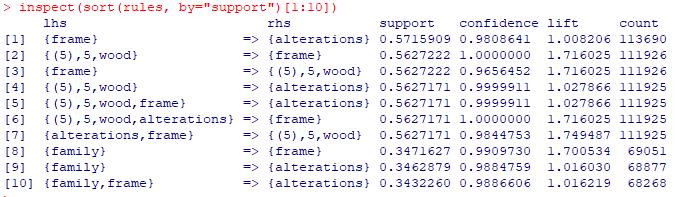


* 1. **导出关联规则并计算支持度和置信度**

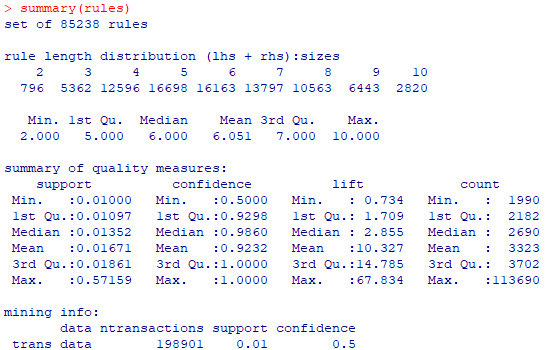
利用Apriori算法导出关联规则如下：



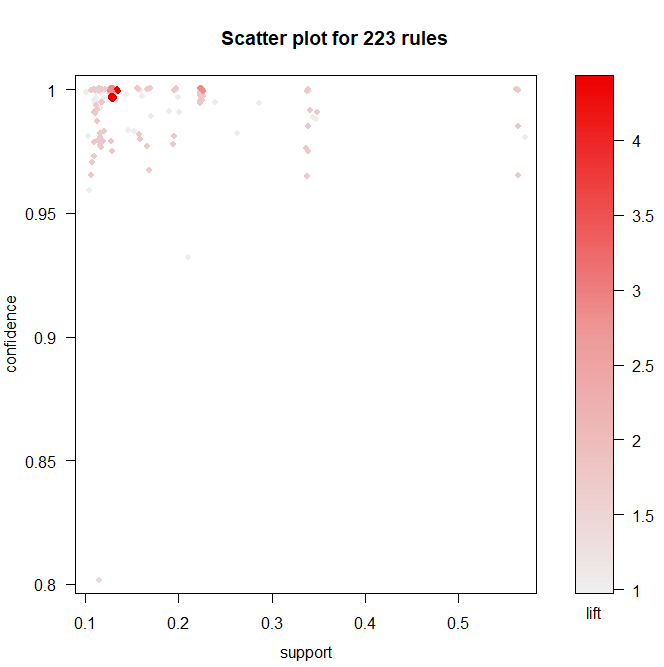
以置信度为标准对其排序后，查看rules的前五项



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：

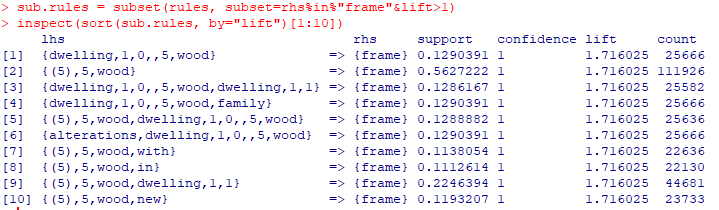


可视化rules如下



* 1. **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）：



最后将挖掘结果写入文件：

