数据挖掘第二次作业

**关联规则挖掘**

曹倩雯

3120170497

关联规则挖掘结果及分析报告

**1 基本任务**：

* 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；
* 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；

**2** **数据集描述**

* 数据集描述：

本数据集是ASCLL文件，属性间以tab分离，每一行数据以双引号开头。每个实例代表了一种潜在的病人。

例如： '35,9 no no yes yes yes yes no' （正常体温35℃-42℃）

'35,9' 患者体温

'no' 呕吐

'no' 腰痛

'yes' 尿频

'yes' 尿痛

'yes' 尿道口瘙痒灼痛

'yes' 膀胱炎

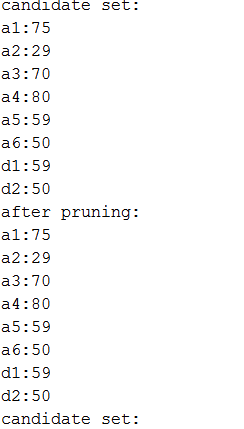
'no' 肾炎

**3 数据预处理结果**

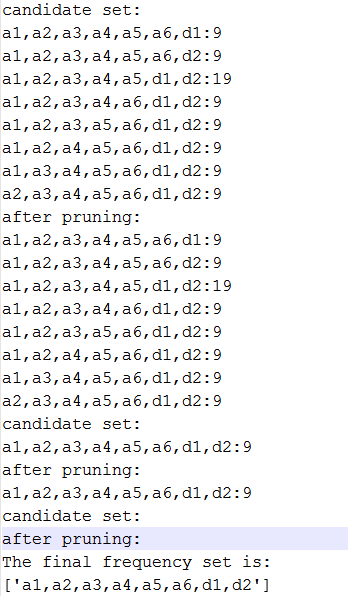
首先将原始数据利用excel表格转换成csv格式方便Python的CSV模块进行后续处理，处理过程如下：

**4 频繁规则统计与计算**

本次数据挖掘使用了Apriori算法，如果要发现强关联规则，就必须先找到频繁集。所谓频繁集，即支持度大于最小支持度的项集。本次数据挖掘令最小支持度等于0.08，即120条记录中至少出现10次。最小置信度设为0.8以下是每一轮的候选集和剪枝的结果：



……



**更多中间数据参见附件“OUTPUT.md“**

这些频繁项集的可视化目标主要是频数，3频数如下，**更多可视化图参见附件figures**：

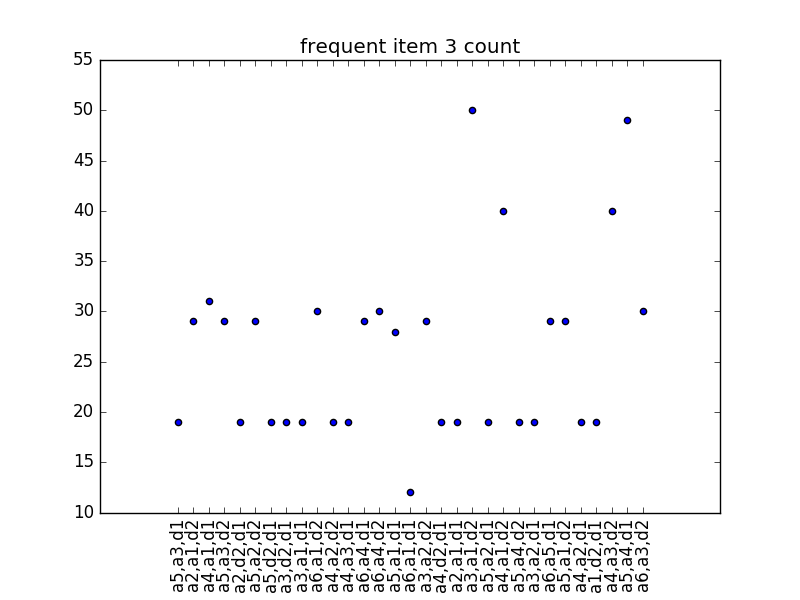
****

图4 三频繁集散点图

**5 规则、可信度、支持度**

首先我们看一下所有的高置信度的规则：考虑到去冗余的需求(以下所有规则都是满最低支持度和最低置信度的强关联规则)

未去冗余的所有规则参见OUTPUT.md

The associative rule is:

('a2^a4->a1^a3^a5^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^d1->a1^a3^a5^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a3^d1->a1^a2^a5^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('d1^d2->a1^a3^a2^a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a4->a3^a5^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^d1->a3^a5^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a3^d1->a2^a5^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^d1^d2->a3^a2^a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a3^a4->a1^a5^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a3^d1->a1^a5^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a4^a5->a1^a3^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a4^d1->a1^a3^a5^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a2^a4^d2->a1^a3^a5^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a5^d1->a1^a3^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.0)

('a2^d1^d2->a1^a3^a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a3^a4^a5->a1^a2^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a3^a4^d1->a1^a2^a5^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a3^a5^d1->a1^a2^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a3^d1^d2->a1^a2^a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a4^a5^d2->a1^a3^a2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a4^d1^d2->a1^a3^a2^a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a5^d1^d2->a1^a3^a2^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a3^a4->a5^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a3^d1->a5^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a4^a5->a3^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a4^d1->a3^a5^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a2^a4^d2->a3^a5^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a5^d1->a3^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.0)

('a1^a2^d1^d2->a3^a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a3^a4^a5->a2^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a3^a4^d1->a2^a5^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a3^a5^d1->a2^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a3^d1^d2->a2^a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a4^a5^d2->a3^a2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a4^d1^d2->a3^a2^a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a5^d1^d2->a3^a2^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a3^a4^a5->a1^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a3^a4^d1->a1^a5^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a2^a3^a4^d2->a1^a5^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.285714285714286)

('a2^a3^a5^d1->a1^a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.0)

('a2^a3^d1^d2->a1^a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.285714285714286)

('a2^a4^a5^d1->a1^a3^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.4)

('a2^a4^a5^d2->a1^a3^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a2^a4^d1^d2->a1^a3^a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a2^a5^d1^d2->a1^a3^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.0)

('a3^a4^a5^d1->a1^a2^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a3^a4^a5^d2->a1^a2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a3^a4^d1^d2->a1^a2^a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a3^a5^d1^d2->a1^a2^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a4^a5^d1^d2->a1^a3^a2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a2^a3^a4^a5->d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a3^a4^d1->a5^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a2^a3^a4^d2->a5^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.4489795918367347)

('a1^a2^a3^a5^d1->a4^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.0)

('a1^a2^a3^d1^d2->a5^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.4489795918367347)

('a1^a2^a4^a5^d1->a3^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.4)

('a1^a2^a4^a5^d2->a3^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a2^a4^d1^d2->a3^a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a2^a5^d1^d2->a3^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.0)

('a1^a3^a4^a5^d1->a2^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a3^a4^a5^d2->a2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a3^a4^d1^d2->a2^a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a3^a5^d1^d2->a2^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)

('a1^a4^a5^d1^d2->a3^a2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a2^a3^a4^a5^d1->a1^d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.4)

('a2^a3^a4^a5^d2->a1^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.870967741935484)

('a2^a3^a4^d1^d2->a1^a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 3.1578947368421053)

('a2^a3^a5^d1^d2->a1^a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.3076923076923075)

('a2^a4^a5^d1^d2->a1^a3', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.2641509433962264)

('a3^a4^a5^d1^d2->a1^a2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a1^a2^a3^a4^a5^d1->d2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.4)

('a1^a2^a3^a4^a5^d2->d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.0338983050847457)

('a1^a2^a3^a4^d1^d2->a5', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 2.0338983050847457)

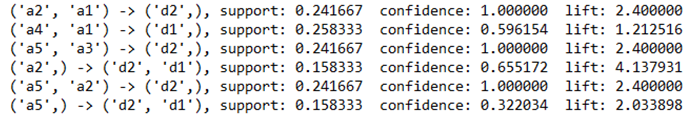
('a1^a2^a3^a5^d1^d2->a4', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 1.5)

('a1^a2^a4^a5^d1^d2->a3', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 1.7142857142857142)

('a1^a3^a4^a5^d1^d2->a2', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 4.137931034482759)

('a2^a3^a4^a5^d1^d2->a1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 1.6)

当然，我们也可以只选取前6个属性（特征属性），到后两个结果的映射规则，部分结果如下所示：



Lift支持度，也叫做提升度，它是可信度与期望可信度的比值，反映了关联规则中的相关性，其可视化如下：

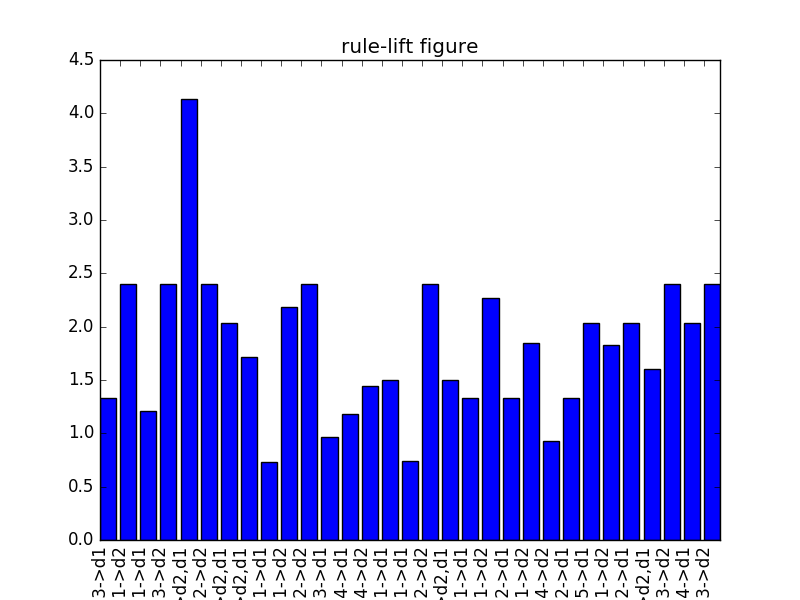


图 5 规则提升度

提升度大于1表示正相关，小于1表示负相关，等于1表示不相关，图中所示，绝大多数规则都是大于1的，并且a2->d1,d2已经超过了4，说明其正相关性非常高，我们可以发现呕吐是两种炎症的一个极为明显的先发症状。结合之前的强关联规则中有一条：('a1^a2^a4->a3^a5^d2^d1', ':', 'support->', 0.15833333333333333, 'coffidence->', 1.0, 'lift->', 6.315789473684211)这说明 呕吐，高温，尿频患者极有可能同时患有两种疾病。