关联规则挖掘实验报告

韩学博　　2120150990

1. 要求

1. 对数据集进行处理，转换成适合关联规则挖掘的形式；
2. 找出频繁项集；
3. 导出关联规则，计算其支持度和置信度；
4. 去除冗余的规则；
5. 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；
6. 使用可视化技术，如散点图、平行坐标、泡泡图等，对规则进行展示。

2. 问题描述

本次实验选取 UCI 的”急性炎症”数据集，数据中的每一行是一位病人的病情状况，实验中我们需要：

1. 对数据进行处理，将数据转换成适合关联规则挖掘的形式，找出频繁项集。
2. 导出关联规则，并计算规则的支持度和置信度。
3. 删除冗余的关联规则。
4. 将剩下的关联规则进行评价、可视化。

3. 数据处理

3.1 数据预处理

得到的原始数据集以表格的形式存储数据，但是规则挖掘时需要每一条记录转换为项集，因此我们需要进行数据预处理。针对第一列体温数据，我们定义

35-36.9 时为正常体温，37-37.9 为低烧，38.0-39。5 为发烧，高于 39.5 为高烧。其他属性如果存在相应的症状我们就记录对应的症状反应，如果不存在不记录。

原始数据如图 1，处理后的结果保存在 PreprocessData.csv 文件中，处理后的数据格式如图 2。

35,5	no	yes	no	no	no	no	no
35,9	no	no	yes	yes	yes	yes	no
35,9	no	yes	no	no	no	no	no
36,0	no	no	yes	yes	yes	yes	no
36,0	no	yes	no	no	no	no	no
36,0	no	yes	no	no	no	no	no
36,2	no	no	yes	yes	yes	yes	no
36,2	no	yes	no	no	no	no	no
36,3	no	no	yes	yes	yes	yes	no
36,6	no	no	yes	yes	yes	yes	no
36,6	no	no	yes	yes	yes	yes	no
36,6	no	yes	no	no	no	no	no
36,6	no	yes	no	no	no	no	no
36,7	no	no	yes	yes	yes	yes	no
36,7	no	yes	no	no	no	no	no
36,7	no	yes	no	no	no	no	no
36,8	no	no	yes	yes	yes	yes	no
36,8	no	no	yes	yes	yes	yes	no
36,9	no	no	yes	yes	yes	yes	no

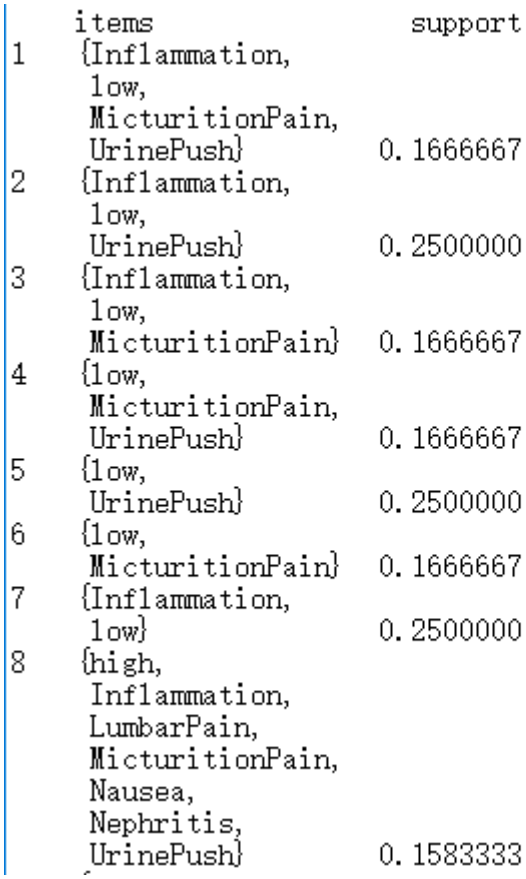
图 1

```
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,LumbarPain
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,LumbarPain
normal,LumbarPain
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,LumbarPain
normal,LumbarPain
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,UrinePush,MicturitionPain,BurningOfUrethra,Inflammation
normal,LumbarPain
low,UrinePush,MicturitionPain,Inflammation
low,UrinePush,MicturitionPain,Inflammation
```

图 2

3.2 找到频繁项集

使用 eclat() 函数可以导出数据集的频繁项集。设定支持度阈值为 0.1。频繁项集最小包括 1 项，最多 8 项。生成的频繁项集保存在 frequent_items。txt 文件中，格式如下图 3。

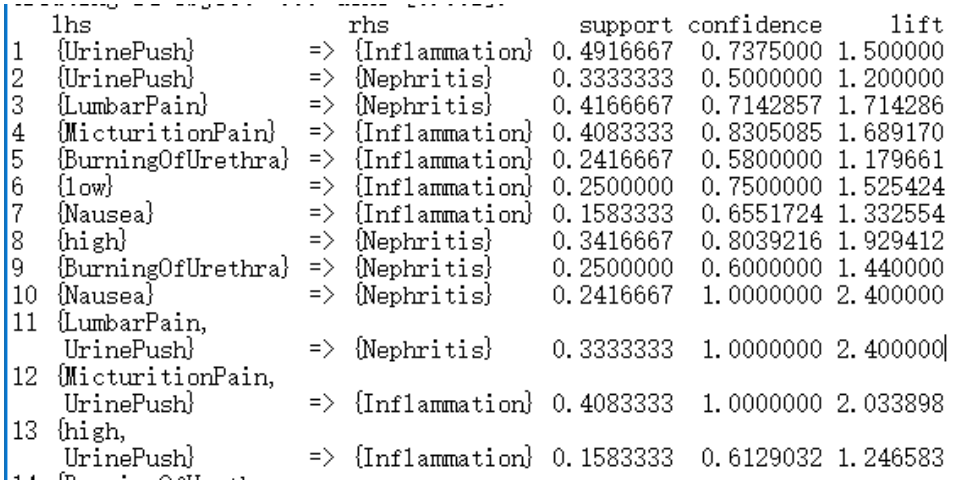


	items	support
1	{Inflammation, low, MicturitionPain, UrinePush}	0.1666667
2	{Inflammation, low, UrinePush}	0.2500000
3	{Inflammation, low, MicturitionPain}	0.1666667
4	{low, MicturitionPain, UrinePush}	0.1666667
5	{low, UrinePush}	0.2500000
6	{low, MicturitionPain}	0.1666667
7	{Inflammation, low}	0.2500000
8	{high, Inflammation, LumbarPain, MicturitionPain, Nausea, Nephritis, UrinePush}	0.1583333

图 3

3.3 导出关联规则并计算支持度和置信度

调用 aprior()函数可以导出关联规则，同时还计算出支持度和置信度。因为数据中我们最关心的是病人的疾病时肾炎还是炎症，所以我们将关联规则的右侧设置为这两个属性，同时关联规则支持度阈值设为 0.1，置信度阈值设为 0.5。结果保存在 rules.txt 中。一共产生了 72 条关联规则，格式如下图 4。

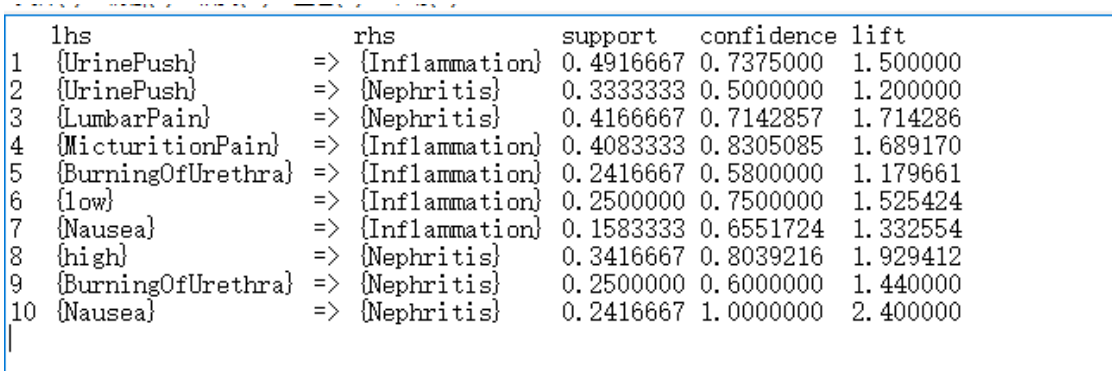


	lhs		rhs	support	confidence	lift
1	{UrinePush}	=>	{Inflammation}	0.4916667	0.7375000	1.500000
2	{UrinePush}	=>	{Nephritis}	0.3333333	0.5000000	1.200000
3	{LumbarPain}	=>	{Nephritis}	0.4166667	0.7142857	1.714286
4	{MicturitionPain}	=>	{Inflammation}	0.4083333	0.8305085	1.689170
5	{BurningOfUrethra}	=>	{Inflammation}	0.2416667	0.5800000	1.179661
6	{low}	=>	{Inflammation}	0.2500000	0.7500000	1.525424
7	{Nausea}	=>	{Inflammation}	0.1583333	0.6551724	1.332554
8	{high}	=>	{Nephritis}	0.3416667	0.8039216	1.929412
9	{BurningOfUrethra}	=>	{Nephritis}	0.2500000	0.6000000	1.440000
10	{Nausea}	=>	{Nephritis}	0.2416667	1.0000000	2.400000
11	{LumbarPain, UrinePush}	=>	{Nephritis}	0.3333333	1.0000000	2.400000
12	{MicturitionPain, UrinePush}	=>	{Inflammation}	0.4083333	1.0000000	2.033898
13	{high, UrinePush}	=>	{Inflammation}	0.1583333	0.6129032	1.246583

图 4

3.4 去除冗余规则

满足支持度阈值和置信度阈值的规则共有 72 条，里面有很多冗余规则（如果 rule2 的 lhs 和 rhs 是包含于 rule1 的，而且 rule2 的 lift 至小于或等于 rule1，则称 rule2 是 rule1 的冗余规则）。删除冗余规则后，只剩下 10 条关联规则，保存在 rules_delete_redundant.txt 文件中，如图 5。



	lhs		rhs	support	confidence	lift
1	{UrinePush}	=>	{Inflammation}	0.4916667	0.7375000	1.500000
2	{UrinePush}	=>	{Nephritis}	0.3333333	0.5000000	1.200000
3	{LumbarPain}	=>	{Nephritis}	0.4166667	0.7142857	1.714286
4	{MicturitionPain}	=>	{Inflammation}	0.4083333	0.8305085	1.689170
5	{BurningOfUrethra}	=>	{Inflammation}	0.2416667	0.5800000	1.179661
6	{low}	=>	{Inflammation}	0.2500000	0.7500000	1.525424
7	{Nausea}	=>	{Inflammation}	0.1583333	0.6551724	1.332554
8	{high}	=>	{Nephritis}	0.3416667	0.8039216	1.929412
9	{BurningOfUrethra}	=>	{Nephritis}	0.2500000	0.6000000	1.440000
10	{Nausea}	=>	{Nephritis}	0.2416667	1.0000000	2.400000

图 5

3.5 对规则进行评价

实验中我们可以使用提升度对规则进行评价，产生管理规则时一起产生了规则的提升度，所以我们可以使用提升度对规则排序评价。结果保存在 rules_delete_redundant_sorted_lift.txt 文件中，如图 6。

	lhs		rhs	support	confidence	lift
10	{Nausea}	=>	{Nephritis}	0.2416667	1.0000000	2.400000
8	{high}	=>	{Nephritis}	0.3416667	0.8039216	1.929412
3	{LumbarPain}	=>	{Nephritis}	0.4166667	0.7142857	1.714286
4	{MicturitionPain}	=>	{Inflammation}	0.4083333	0.8305085	1.689170
6	{low}	=>	{Inflammation}	0.2500000	0.7500000	1.525424
1	{UrinePush}	=>	{Inflammation}	0.4916667	0.7375000	1.500000
9	{BurningOfUrethra}	=>	{Nephritis}	0.2500000	0.6000000	1.440000
7	{Nausea}	=>	{Inflammation}	0.1583333	0.6551724	1.332554
2	{UrinePush}	=>	{Nephritis}	0.3333333	0.5000000	1.200000
5	{BurningOfUrethra}	=>	{Inflammation}	0.2416667	0.5800000	1.179661

图 6

3.6 对规则进行可视化

实验中我们使用散点图对关联规则进行可视化如图 7。图中每个点对应于相应的支持度和置信度值，分别由图形的横纵轴显示，其中关联规则点的颜色深浅由 lift 值的高低决定。

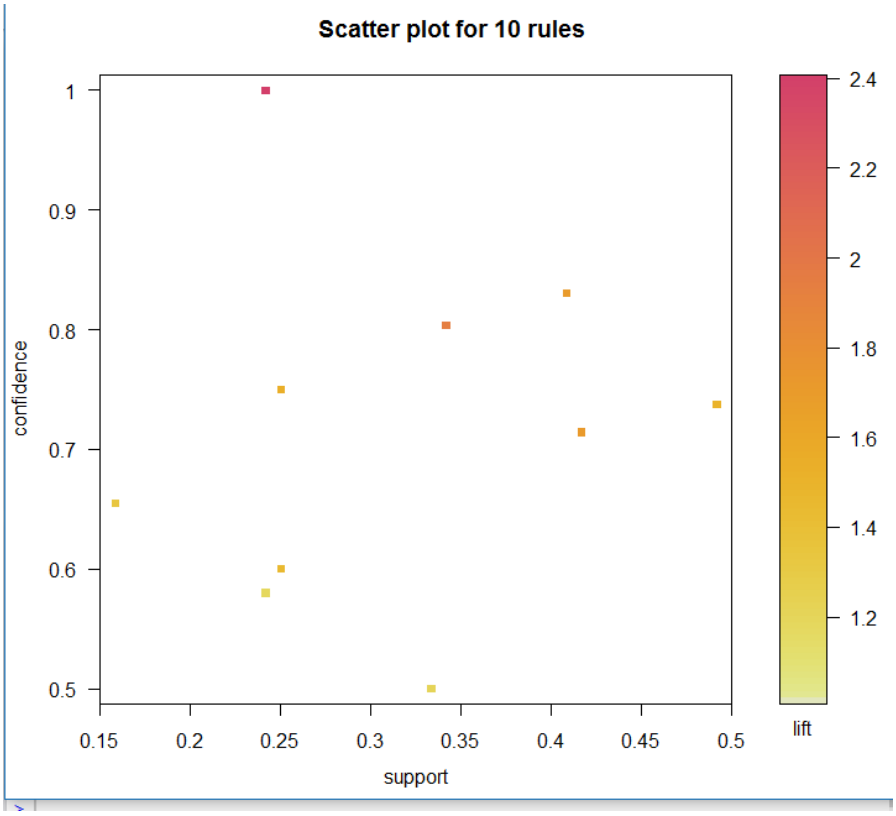


图 7

试验中我们还是用气泡图来展现关联规则，提升度 lift 是圈的颜色深浅，圈的大小表示支持度 support 的大小。LHS 的个数和分组中最频繁项集显示在列的标签里。Lift 从左上角到右下角逐渐减少，如图 8。

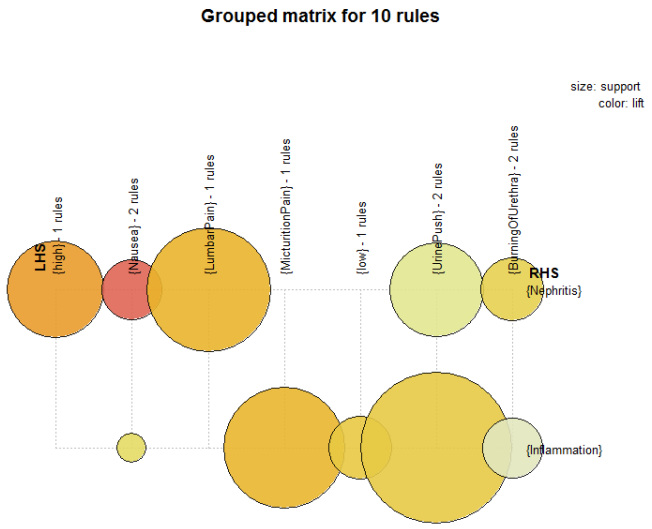
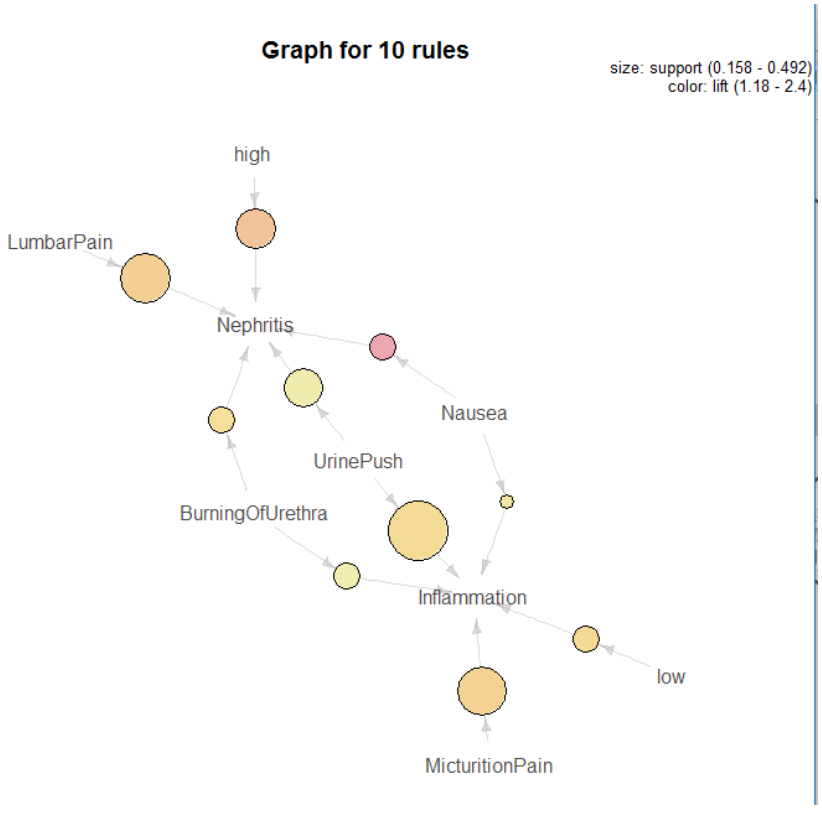


图 8

也可以通过箭头和圆圈来表示关联规则，利用项点表示项集，边表示规则中关系。圆圈越大表示支持度 support 越大，颜色越深表示提升度 lift 越大。



如图 9。

图 9

4. 实验结果分析

关联规则中，高烧可以判断为急性肾炎，腰疼时可以判断为急性肾炎，这符合数据集对急性肾炎的介绍，急性肾炎始于突然发烧，有时超过 40 度，发烧伴随着双面腰椎疼痛，不少有恶心和呕吐的症状。

关联规则排尿疼痛可以判断为急性膀胱炎症，低烧可以判断为急性膀胱炎症，这符合数据集中对急性膀胱炎症的介绍，腹部和排尿时突发的疼痛，体温升高，但是通常不会超高 38 度。

而对于有恶心的症状，尿道口的肿胀症状，尿频现象两种疾病都会产生，所以都存在关联规则。