

第二次作业——关联规则挖掘

韩梦乔 2120150989

要求

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；
去除冗余的规则；
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；
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。

数据描述

UCI 的”急性炎症”数据集

a1 Temperature of patient { 35C-42C }
a2 Occurrence of nausea { yes, no }
a3 Lumbar pain { yes, no }
a4 Urine pushing (continuous need for urination) { yes, no }
a5 Micturition pains { yes, no }
a6 Burning of urethra, itch, swelling of urethra outlet { yes, no }
d1 decision: Inflammation of urinary bladder { yes, no }
d2 decision: Nephritis of renal pelvis origin { yes, no }

Eg:

	a1	a2	a3	a4	a5	a6	d1	d2
35,5	no	yes	no	no	no	no	no	no
35,9	no	no	yes	yes	yes	yes	no	no
35,9	no	yes	no	no	no	no	no	no
36,0	no	no	yes	yes	yes	yes	no	no
36,0	no	yes	no	no	no	no	no	no
36,0	no	yes	no	no	no	no	no	no
36,2	no	no	yes	yes	yes	yes	no	no
36,2	no	yes	no	no	no	no	no	no
36,3	no	no	yes	yes	yes	yes	no	no
36,6	no	no	yes	yes	yes	yes	no	no
36,6	no	no	yes	yes	yes	yes	no	no
36,6	no	yes	no	no	no	no	no	no

实验

数据预处理

1. 为了进行关联规则挖掘，需要对实验的数据进行预处理。由于 a1 为数值属性，为此将 a1 离散化为{35,36,37,38,39,40,41}，对于其他属性，将该属性后加上该属性的取值，得到关联规则挖掘的预处理的数据。处理后的数据格式如下：

```
a1_35,a2_no,a3_yes,a4_no,a5_no,a6_no,d1_no,d2_no
a1_35,a2_no,a3_no,a4_yes,a5_yes,a6_yes,d1_yes,d2_no
a1_35,a2_no,a3_yes,a4_no,a5_no,a6_no,d1_no,d2_no
a1_36,a2_no,a3_no,a4_yes,a5_yes,a6_yes,d1_yes,d2_no
a1_36,a2_no,a3_yes,a4_no,a5_no,a6_no,d1_no,d2_no
a1_36,a2_no,a3_yes,a4_no,a5_no,a6_no,d1_no,d2_no
a1_36,a2_no,a3_no,a4_yes,a5_yes,a6_yes,d1_yes,d2_no
a1_36,a2_no,a3_yes,a4_no,a5_no,a6_no,d1_no,d2_no
a1_36,a2_no,a3_no,a4_yes,a5_yes,a6_yes,d1_yes,d2_no
a1_36,a2_no,a3_no,a4_yes,a5_yes,a6_yes,d1_yes,d2_no
a1_36,a2_no,a3_no,a4_yes,a5_yes,a6_yes,d1_yes,d2_no
a1_36,a2_no,a3_yes,a4_no,a5_no,a6_no,d1_no,d2_no
a1_36,a2_no,a3_yes,a4_no,a5_no,a6_no,d1_no,d2_no
```

2. 数据预处理的脚本为 preProcess.py，产生的结果放在了 preDiagnosis.data 中。

获得频繁项集

1. 获取频繁项集时，设置支持度为 0.3，采用 R 语言实现代码如下：

```
frequentsets=eclat(tr,parameter=list(support=0.3,maxlen=4))
```

2. 频繁项集存放在 frequencySet.txt 中，并且排序取其中的前 20 个，存放在了 frequencySetSortBySup.txt，如下：

items	support
64 {a2_no}	0.7583333
65 {a4_yes}	0.6666667
56 {a2_no,d2_no}	0.5833333
66 {a3_yes}	0.5833333
67 {a6_no}	0.5833333
68 {d2_no}	0.5833333
50 {a2_no,a5_no}	0.5083333
63 {a2_no,a4_yes}	0.5083333
69 {a5_no}	0.5083333
70 {d1_no}	0.5083333
35 {a4_yes,d1_yes}	0.4916667
71 {d1_yes}	0.4916667
72 {a5_yes}	0.4916667

```

38 {a2_no,a5_no,d1_no} 0.4250000
41 {a2_no,d1_no}      0.4250000
42 {a3_yes,d1_no}     0.4250000
44 {a5_no,d1_no}      0.4250000
14 {a2_no,a3_no,d2_no} 0.4166667
17 {a2_no,a3_no}      0.4166667
19 {a3_no,d2_no}      0.4166667

```

关联规则

1. 关联规则挖掘算法采用 apriori 算法，设置支持度为 0.3，置信度为 0.3，采用 R 语言实现如下：

```
rules = apriori(tr,parameter = list(support = 0.3,confidence = 0.3))
```

2. 挖掘出来的关联规则保存在 rules.txt 中，rulesSortByCon.txt 和 rulesSortBySup.txt 中分别是保存这按照置信度和支持度排序后的规则的前 10 条。如下所示：

lhs	rhs	support	confidence	lift
14 {}	=> {a2_no}	0.7583333	0.7583333	1.000000
13 {}	=> {a4_yes}	0.6666667	0.6666667	1.000000
10 {}	=> {a6_no}	0.5833333	0.5833333	1.000000
11 {}	=> {d2_no}	0.5833333	0.5833333	1.000000
12 {}	=> {a3_yes}	0.5833333	0.5833333	1.000000
73 {d2_no}	=> {a2_no}	0.5833333	1.0000000	1.318681
74 {a2_no}	=> {d2_no}	0.5833333	0.7692308	1.318681
8 {}	=> {d1_no}	0.5083333	0.5083333	1.000000
9 {}	=> {a5_no}	0.5083333	0.5083333	1.000000
63 {a5_no}	=> {a2_no}	0.5083333	1.0000000	1.318681

lhs	rhs	support	confidence	lift
15 {a1_37}	=> {d2_no}	0.3333333	1	1.714286
17 {a1_37}	=> {a2_no}	0.3333333	1	1.318681
19 {a4_no}	=> {d1_no}	0.3333333	1	1.967213
21 {a4_no}	=> {a6_no}	0.3333333	1	1.714286
23 {d2_yes}	=> {a3_yes}	0.4166667	1	1.714286
27 {a6_yes}	=> {a4_yes}	0.4166667	1	1.500000
33 {a3_no}	=> {d2_no}	0.4166667	1	1.714286
37 {a3_no}	=> {a2_no}	0.4166667	1	1.318681
45 {d1_yes}	=> {a4_yes}	0.4916667	1	1.500000
63 {a5_no}	=> {a2_no}	0.5083333	1	1.318681

去除冗余规则

Lift 对规则进行评价

- 1. 在对规则进行评价的过程中，使用了 Lift 指标。
- 2. rulesSortByLift.txt 中保存这采用 Lift 排序后的关联规则的前 10 条，如下所示：

lhs	rhs	support	confidence	lift
86 {a6_no,d1_no}	=> {a4_no}	0.3333333	1	3.0
89 {a3_yes,a4_yes}	=> {d2_yes}	0.3333333	1	2.4
95 {d1_yes,d2_no}	=> {a3_no}	0.3333333	1	2.4
101 {a2_no,d1_yes}	=> {a3_no}	0.3333333	1	2.4
104 {a4_yes,d2_no}	=> {a3_no}	0.3333333	1	2.4
153 {a4_yes,d1_yes,d2_no}	=> {a3_no}	0.3333333	1	2.4
157 {a2_no,d1_yes,d2_no}	=> {a3_no}	0.3333333	1	2.4
161 {a2_no,a4_yes,d1_yes}	=> {a3_no}	0.3333333	1	2.4
165 {a2_no,a4_yes,d2_no}	=> {a3_no}	0.3333333	1	2.4
182 {a2_no,a4_yes,d1_yes,d2_no}	=> {a3_no}	0.3333333	1	2.4

规则可视化

- 1. 关联规则可视化如下图 1 所示，横坐标表示支持度，纵坐标表示置信度，颜色表示 lift 值：
- 2. 关联规则可视化如下图 2 所示，横坐标表示支持度，纵坐标表示 lift 值，颜色表示置信度：
- 3. 泡泡图如图 3 所示：
- 4. 平行坐标图如图 4 所示：

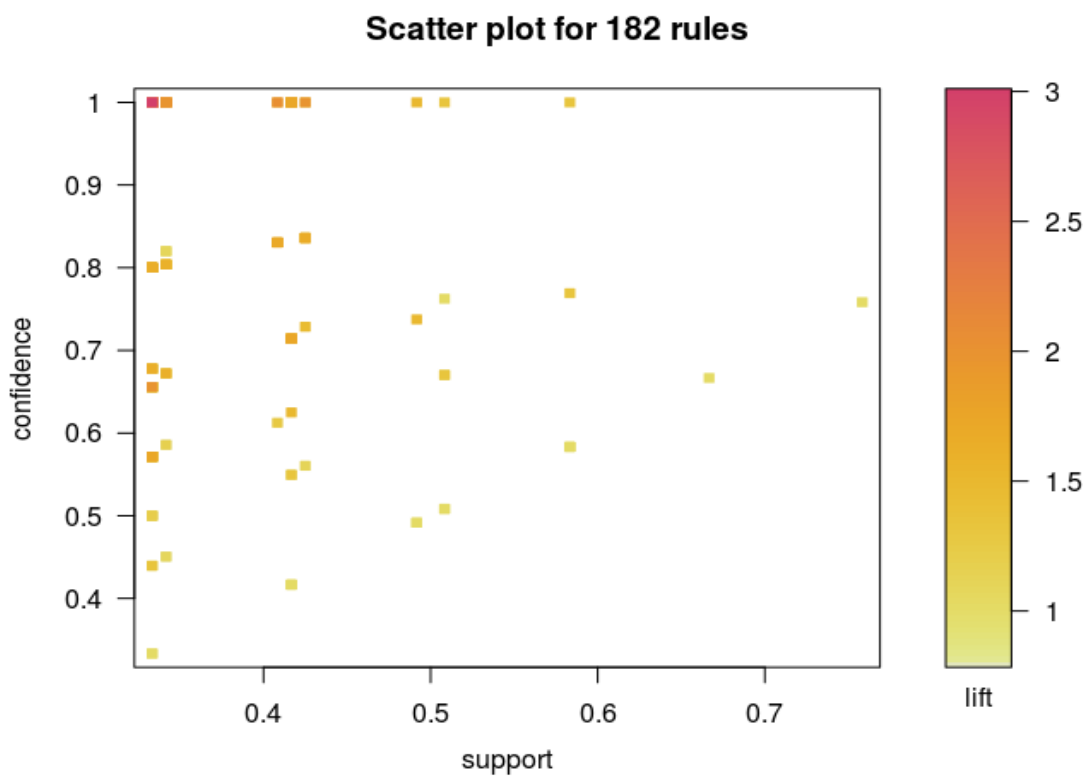


图 1

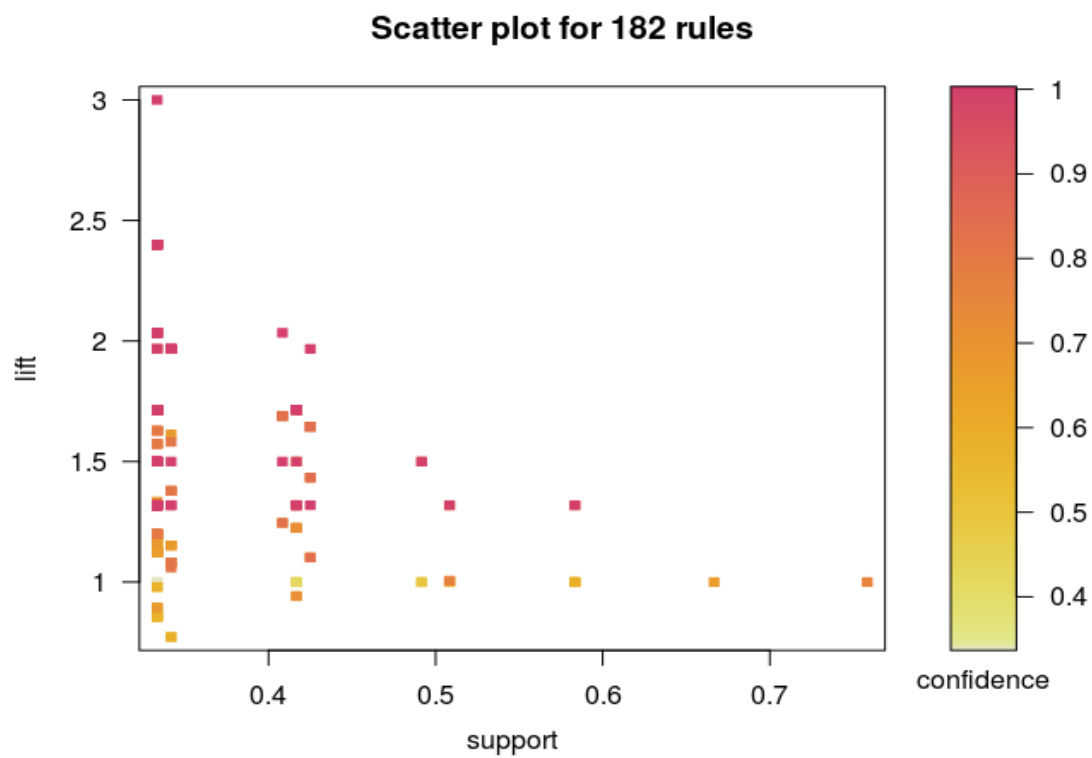


图 2

Grouped matrix for 182 rules

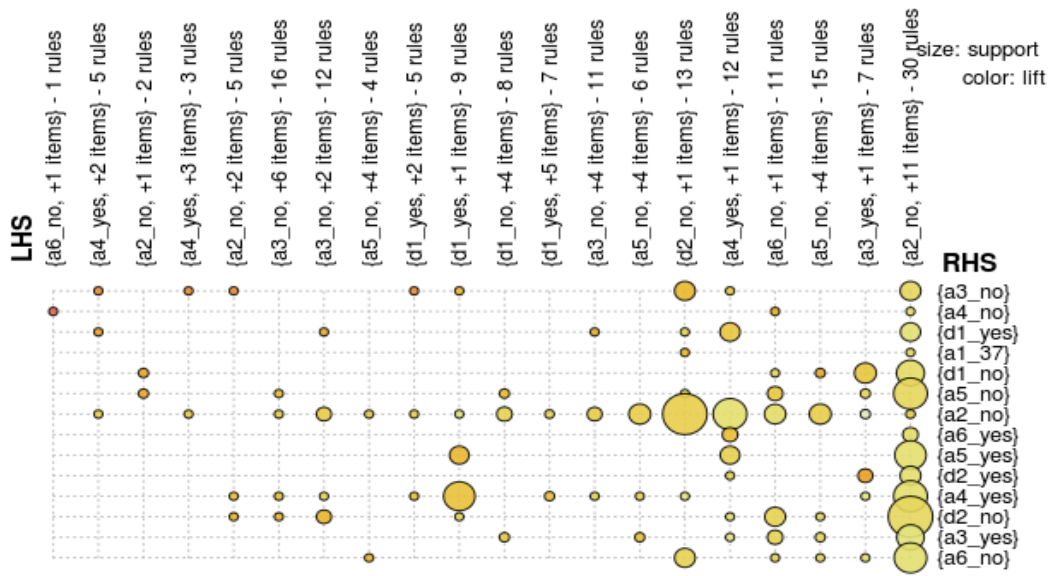


图 3

Parallel coordinates plot for 182 rules

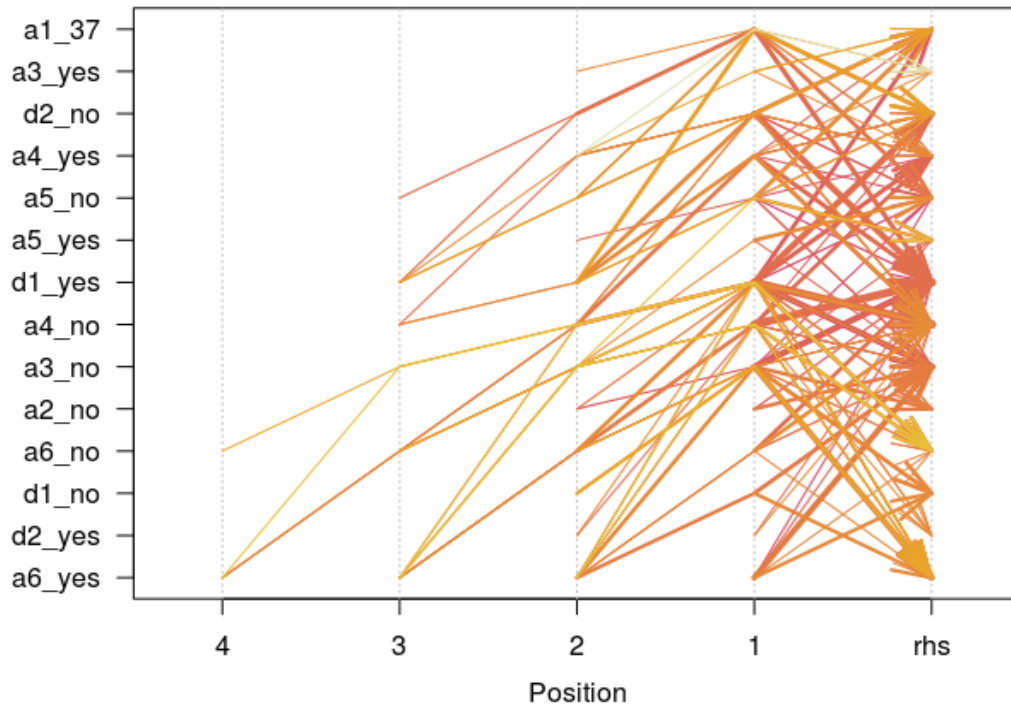


图 4