数据挖掘作业二——关联规则挖掘

姓名：张燕妮

学号：2120151065

指导老师：汤世平

日期：2016 年 7 月 10 日

一、 数据源

UCI 的”急性炎症”数据集

二、 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。

三、 提交的内容

1. 对数据集进行处理的源程序（preProcessing.py）
2. 关联规则挖掘的源程序（rules.R）
3. 挖掘结果及分析（结果数据及分析图）
4. 挖掘过程的报告

四、 实验过程

1. 对数据集进行处理，转换成适合关联规则挖掘的形式

对原始数据进行处理，处理方法为：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。

数据处理的源程序为：preProcessing.py

数据处理结果：preProcessingResult.data

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. 找出频繁项集

利用 R 语言获得频繁项集，设置支持度为 0.3，代码如下：

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

频繁项集结果保存在：FreqSet.txt

取出其中的前 20 项保存在：FreqSetTop20.txt

下面列出这前 20 项内容：

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

3. 导出关联规则，计算其支持度和置信度

采用 R 语言，设置支持度为 0.3，置信度为 0.3，利用 apriori 算法进行关联规则挖掘。
代码如下：

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

结果保存在：RulesResult.txt
按置信度排序后的前 10 条如下：

```
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
```

按支持度排序后的前 10 条如下：

```
    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
```

4. 去除冗余规则

该利用算法得到的数据比较合理，可以不用去除冗余规则

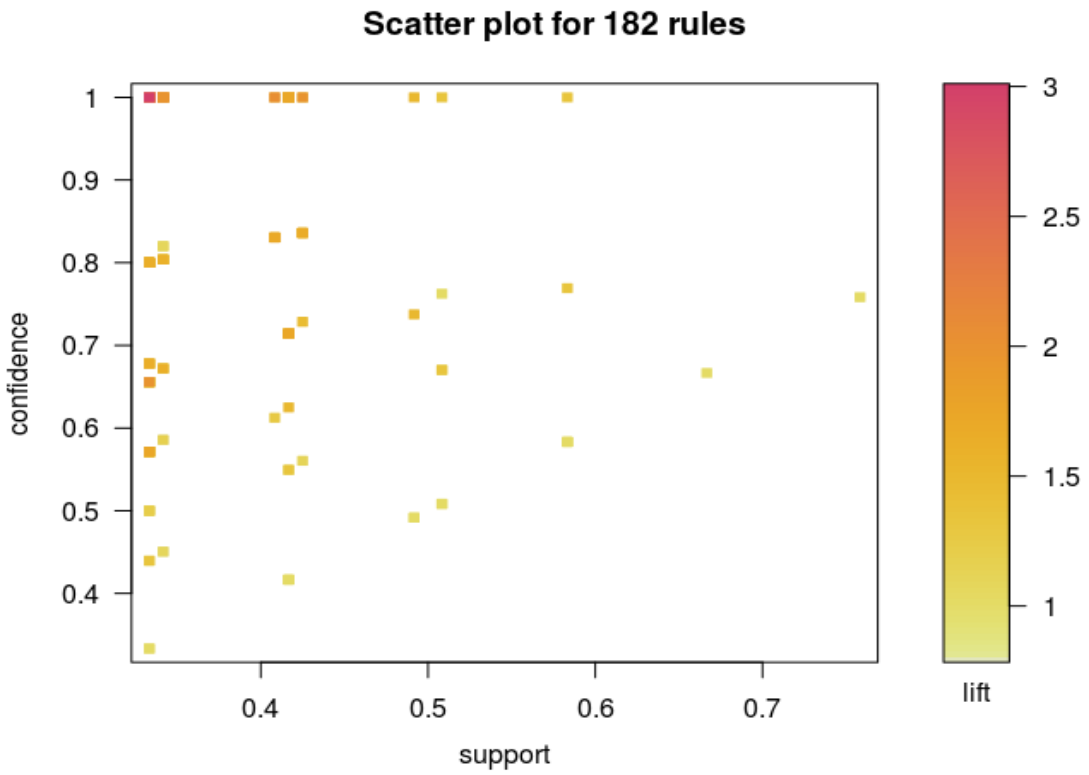
5. 利用 Lift 对规则进行评价

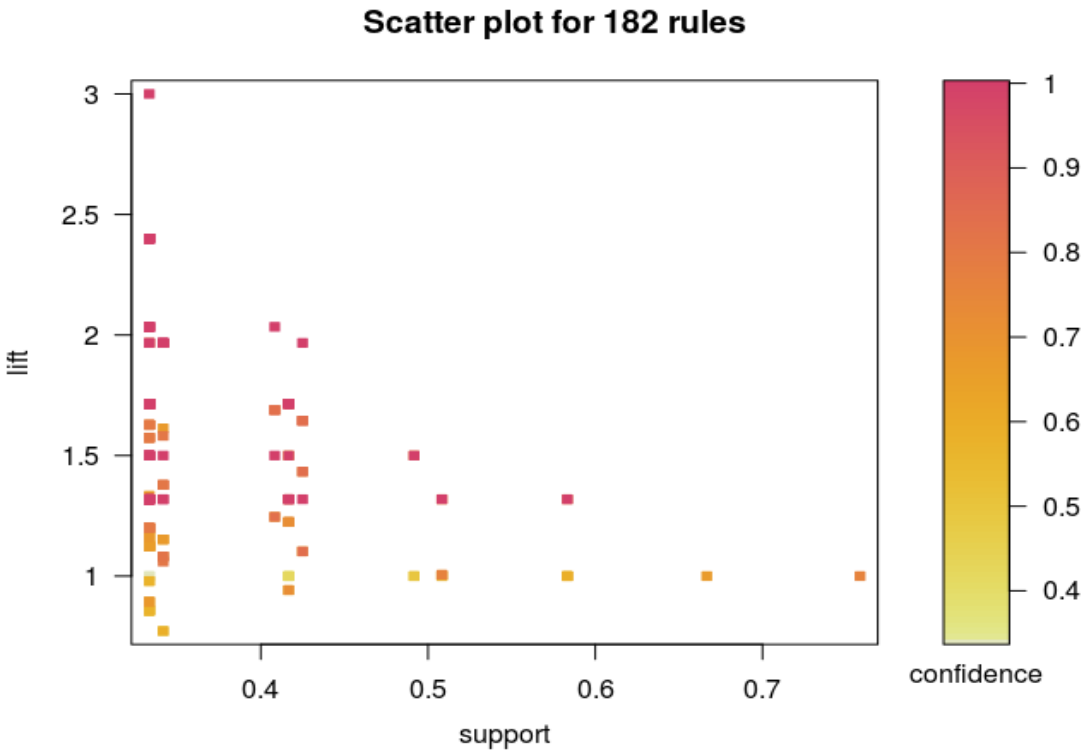
采用 Lift 排序后的关联规则的前 10 条如下：

	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

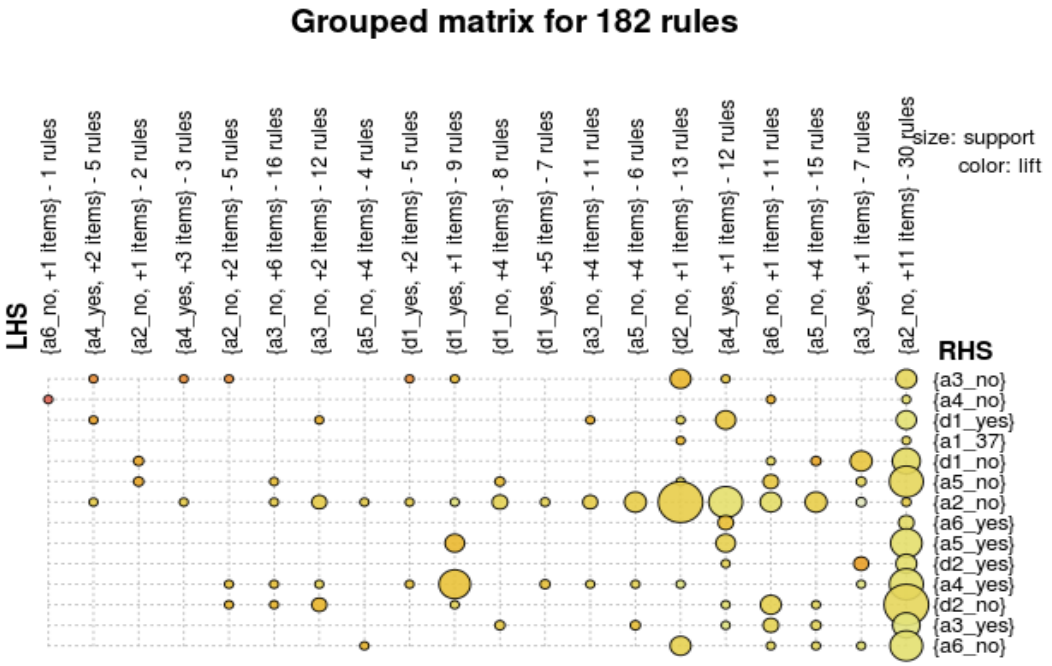
6. 使用可视化技术对规则进行展示

散点图表示关联规则如下：




Scatter plot for 182 rules
lift
support
confidence
3
2.5
2
1.5
1
0.4
0.5
0.6
0.7
1
0.9
0.8
0.7
0.6
0.5
0.4

泡泡图表示如下：


Grouped matrix for 182 rules
LHS
{a6_no, +1 items} - 1 rules
{a4_yes, +2 items} - 5 rules
{a2_no, +1 items} - 2 rules
{a4_yes, +3 items} - 3 rules
{a2_no, +2 items} - 5 rules
{a3_no, +6 items} - 16 rules
{a3_no, +2 items} - 12 rules
{a5_no, +4 items} - 4 rules
{d1_yes, +2 items} - 5 rules
{d1_yes, +1 items} - 9 rules
{d1_no, +4 items} - 8 rules
{d1_yes, +5 items} - 7 rules
{a3_no, +4 items} - 11 rules
{a5_no, +4 items} - 6 rules
{d2_no, +1 items} - 13 rules
{a4_yes, +1 items} - 12 rules
{a6_no, +1 items} - 11 rules
{a5_no, +4 items} - 15 rules
{a3_yes, +1 items} - 7 rules
{a2_no, +11 items} - 30 rules
size: support
color: lift
RHS
{a3_no}
{a4_no}
{d1_yes}
{a1_37}
{d1_no}
{a5_no}
{a2_no}
{a6_yes}
{a5_yes}
{d2_yes}
{a4_yes}
{d2_no}
{a3_yes}
{a6_no}