数据挖掘之频繁规则挖掘

-结果与分析

曹文强
2120150977
北京理工大学计算机学院

一、 数据预处理结果

首先将数据转换成 csv 格式方便读入，如下图

	A	B	C	D	E	F	G	H
1	a1	a2	a3	a4	a5	a6	d1	d2
2	35.5	no	yes	no	no	no	no	no
3	35.9	no	no	yes	yes	yes	yes	no
4	35.9	no	yes	no	no	no	no	no
5	36	no	no	yes	yes	yes	yes	no
6	36	no	yes	no	no	no	no	no
7	36	no	yes	no	no	no	no	no
8	36.2	no	no	yes	yes	yes	yes	no
9	36.2	no	yes	no	no	no	no	no
10	36.3	no	no	yes	yes	yes	yes	no
11	36.6	no	no	yes	yes	yes	yes	no
12	36.6	no	no	yes	yes	yes	yes	no
13	36.6	no	yes	no	no	no	no	no
14	36.6	no	yes	no	no	no	no	no
15	36.7	no	no	yes	yes	yes	yes	no

图 1 csv 数据

第一个属性是数值属性，我将其简化为二值属性，分别表示发烧和不发烧。

```
In [171]: data
Out[171]:
      a1   a2   a3   a4   a5   a6   d1   d2
0     no   no  yes   no   no   no   no   no
1     no   no   no  yes  yes  yes  yes   no
2     no   no  yes   no   no   no   no   no
3     no   no   no  yes  yes  yes  yes   no
4     no   no  yes   no   no   no   no   no
5     no   no  yes   no   no   no   no   no
6     no   no   no  yes  yes  yes  yes   no
7     no   no  yes   no   no   no   no   no
8     no   no   no  yes  yes  yes  yes   no
9     no   no   no  yes  yes  yes  yes   no
10    no   no   no  yes  yes  yes  yes   no
```

图 2 温度二值化后的数据

然后，将矩阵式的数据转换成事务型的列表数据，如下：

```
['a1', 'a4', 'a5', 'd1']
['a1', 'a4', 'a5', 'd1']
['a1', 'a4', 'a5', 'a6', 'd1']
['a1', 'a4', 'a5', 'd1']
['a1', 'a4', 'a5', 'd1']
['a1', 'a3']
['a1', 'a4', 'd1']
['a1', 'a3']
['a1', 'a4', 'a5', 'a6', 'd1']
['a1', 'a4', 'd1']
['a1', 'a4', 'a5', 'd1']
['a1', 'a4', 'a5', 'd1']
['a1', 'a3']
['a1', 'a4', 'a5', 'a6', 'd1']
['a1', 'a4', 'd1']
['a1', 'a3', 'a4', 'a6', 'd2']
['a1', 'a3', 'a4', 'a6', 'd2']
```

图 3 事务化的数据

二、 频繁规则统计与计算

我只计算了最大到 3 频繁项集，计算过程不赘述，结果如下。

```
frequent item 1
key: a1, value: 75
key: a3, value: 70
key: a2, value: 29
key: a5, value: 59
key: a4, value: 80
key: a6, value: 50
key: d2, value: 50
key: d1, value: 59
```

图 4 1-频繁项集

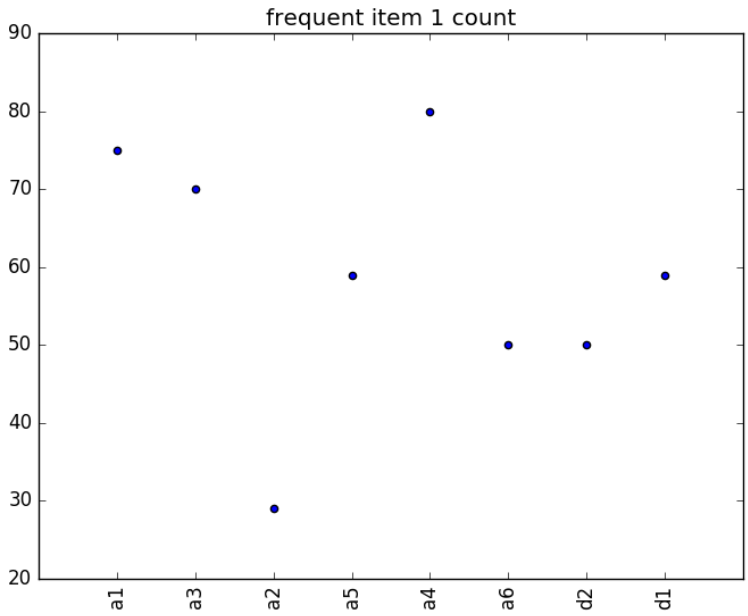
```
frequent item 2
('key: ', ('d2', 'a1'), ' value: 50')
('key: ', ('d2', 'a5'), ' value: 29')
('key: ', ('a3', 'a2'), ' value: 29')
('key: ', ('a4', 'a1'), ' value: 52')
('key: ', ('d1', 'a4'), ' value: 59')
('key: ', ('a5', 'a4'), ' value: 49')
('key: ', ('d1', 'a1'), ' value: 31')
('key: ', ('d2', 'a4'), ' value: 40')
('key: ', ('a3', 'a1'), ' value: 53')
```

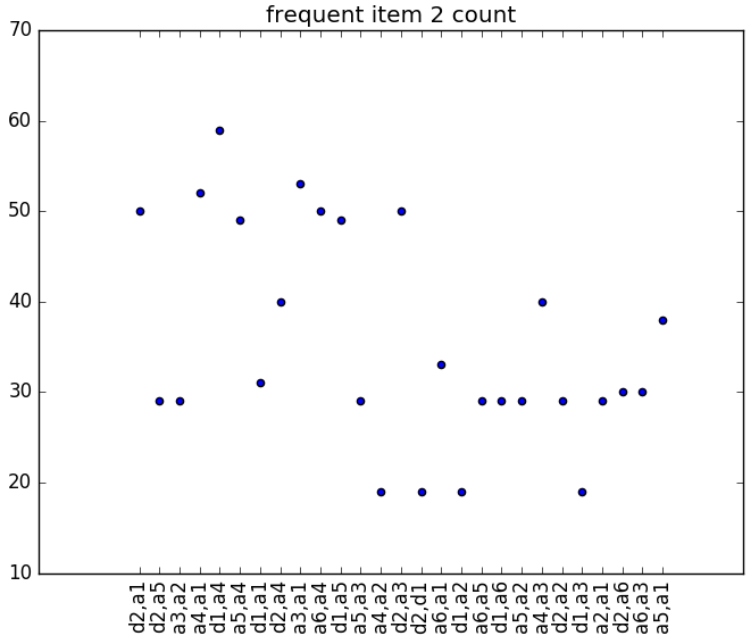
图 5 2-频繁项集

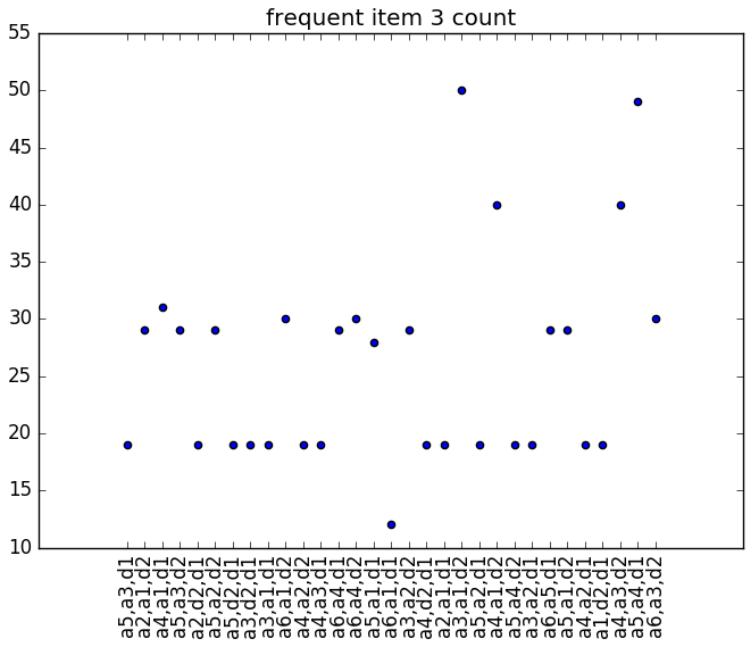
```
frequent item 3
('key: ', ('a5', 'a3', 'd1'), ' value: 19')
('key: ', ('a2', 'a1', 'd2'), ' value: 29')
('key: ', ('a4', 'a1', 'd1'), ' value: 31')
('key: ', ('a5', 'a3', 'd2'), ' value: 29')
('key: ', ('a2', 'd2', 'd1'), ' value: 19')
('key: ', ('a5', 'a2', 'd2'), ' value: 29')
('key: ', ('a5', 'd2', 'd1'), ' value: 19')
('key: ', ('a3', 'd2', 'd1'), ' value: 19')
('key: ', ('a3', 'a1', 'd1'), ' value: 19')
('key: ', ('a6', 'a1', 'd2'), ' value: 30')
```

图 6 3-频繁项集

这些频繁项集的可视化目标主要是频数，如下：


frequent item 1 count
90
80
70
60
50
40
30
20
a1
a3
a2
a5
a4
a6
d2
d1


frequent item 2 count
70
60
50
40
30
20
10
d2,a1
d2,a5
a3,a2
a4,a1
d1,a4
a5,a4
d1,a1
d2,a4
a3,a1
a6,a4
d1,a5
a5,a3
a4,a2
d2,a3
d2,d1
a6,a1
d1,a2
a6,a5
d1,a6
a5,a2
a4,a3
d2,a2
d1,a3
a2,a1
d2,a6
a6,a3
a5,a1



三、 规则、可信度、支持度

规则从 3-频繁项集中选取，数据中，最后两个属性是结果属性，前 6 个属性是特征属性，所以这里规则的选取应该是从特征到结果的映射，如下所示：

```
-----------------------export association rule----------------------
('a5', 'a3') -> ('d1',), support: 0.158333  confidence: 0.655172  lift: 1.332554
('a2', 'a1') -> ('d2',), support: 0.241667  confidence: 1.000000  lift: 2.400000
('a4', 'a1') -> ('d1',), support: 0.258333  confidence: 0.596154  lift: 1.212516
('a5', 'a3') -> ('d2',), support: 0.241667  confidence: 1.000000  lift: 2.400000
('a2',) -> ('d2', 'd1'), support: 0.158333  confidence: 0.655172  lift: 4.137931
('a5', 'a2') -> ('d2',), support: 0.241667  confidence: 1.000000  lift: 2.400000
('a5',) -> ('d2', 'd1'), support: 0.158333  confidence: 0.322034  lift: 2.033898
('a3',) -> ('d2', 'd1'), support: 0.158333  confidence: 0.271429  lift: 1.714286
('a3', 'a1') -> ('d1',), support: 0.158333  confidence: 0.358491  lift: 0.729133
('a6', 'a1') -> ('d2',), support: 0.250000  confidence: 0.909091  lift: 2.181818
```

图 7 规则、可信度、支持度

相关标准的计算方法在另一个文档中已经说明。可视化如下(横坐标是规则，纵坐标是 lift 值)：

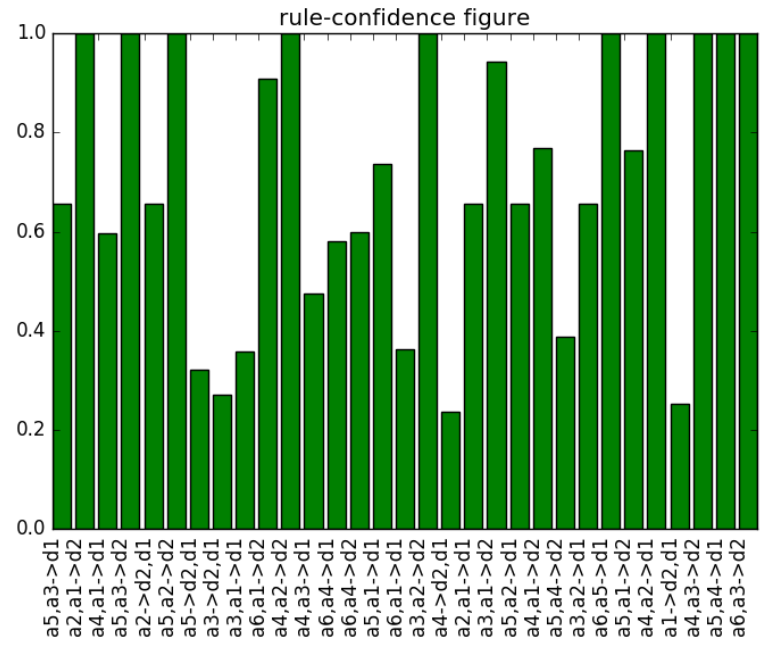


图 8 规则-可信度图示

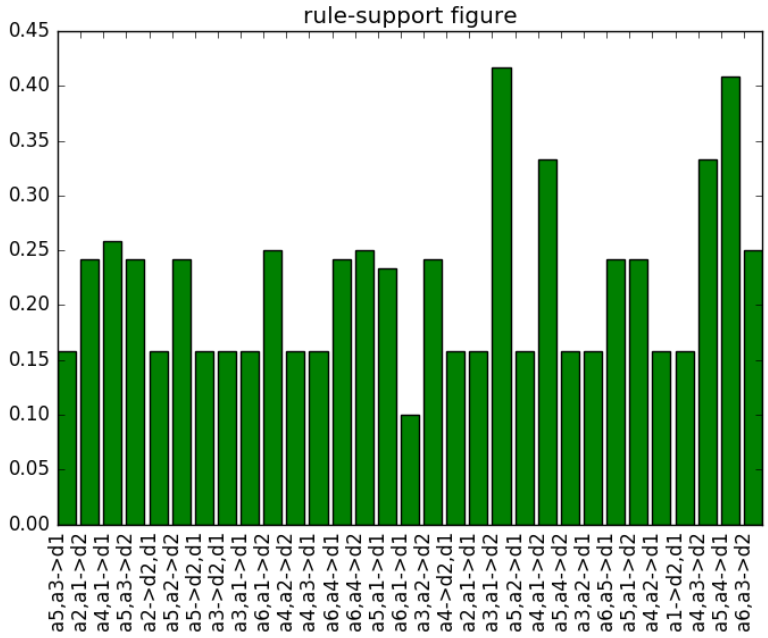


图 9 规则-支持度图示

四、 Lift 评价因素

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：

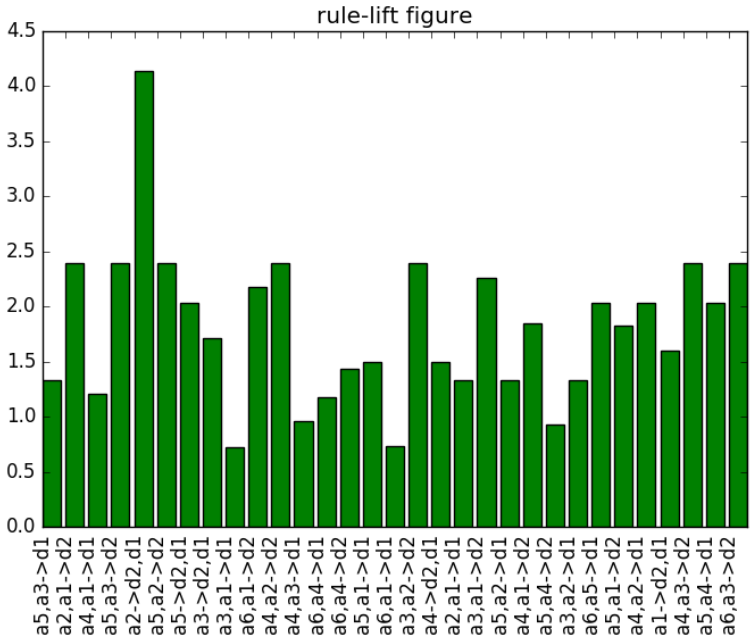


图 10 规则-提升度图示

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。