# 数据挖掘第二次作业

敖权

2120150974

## 实验环境

Ubuntu14.04 + python + Rstudio

## 数据集

[UCI的”急性炎症”数据集](http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations)

数据描述

|  |
| --- |
| 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  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 |

## 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 |

1. 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)) |

1. 频繁项集存放在freqencySet.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)) |

1. 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：

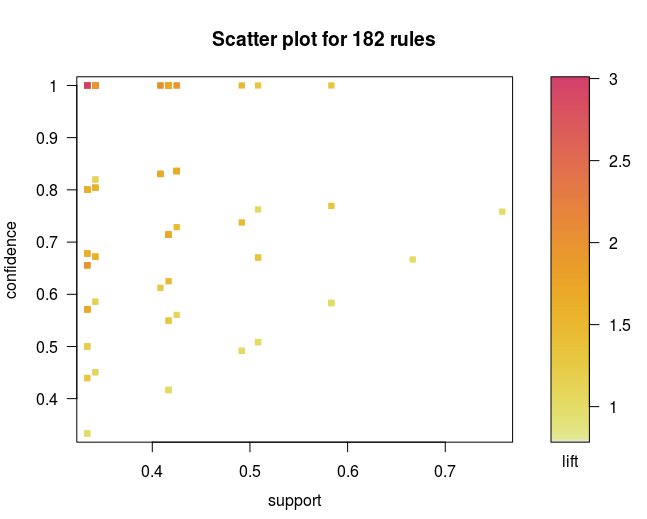


图1

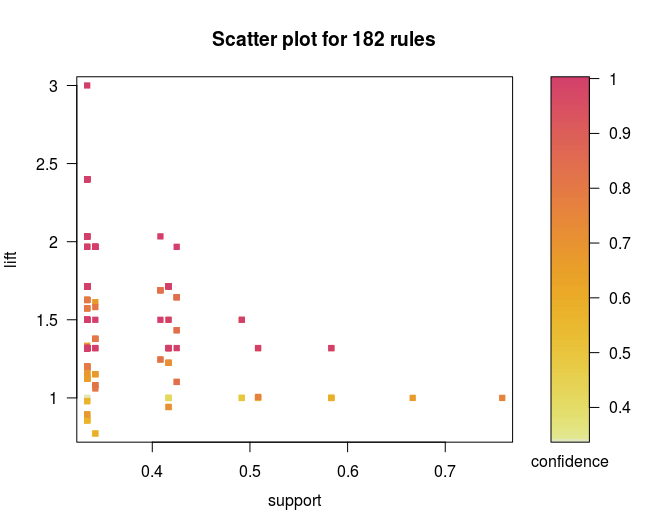


图2

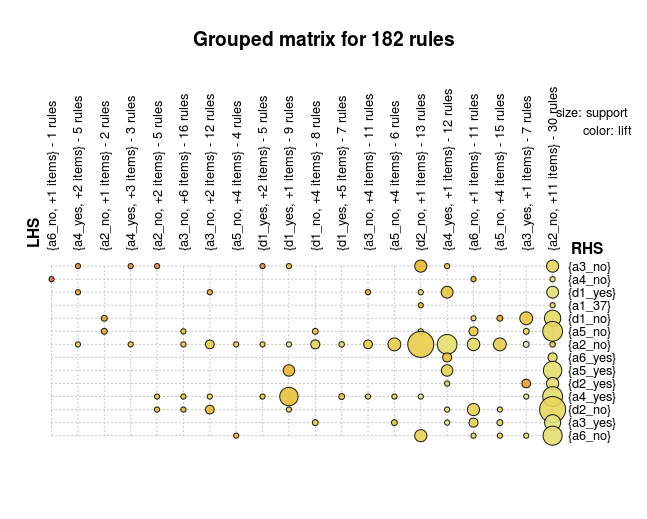


图3

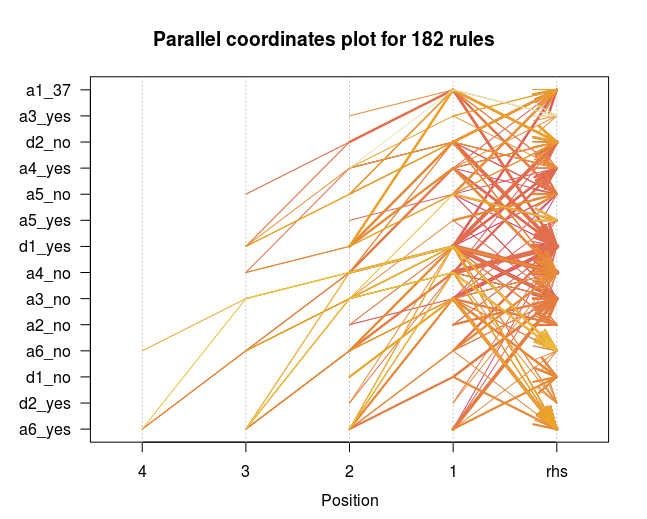


图4