**UCI的”急性炎症”数据集关联规则挖掘过程报告**

**姓 名：王旗林**

**学 号：2120151041**

1. **对数据集进行处理，转换成适合关联规则挖掘的形式**

原始数据集格式为：

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

……

第一列表示体温，后七列分别表示不同的症状。

通过prepare\_data.py文件进行处理。

将第一列按37.5为分界点，大于等于37.5度的设置为fever。

其余列对应不同的症状，当症状为yes时，symptom\_\*\_y，其中\*分别代表症状1到7。

转换后的数据集的格式为：

symptom\_2

symptom\_3,symptom\_4,symptom\_5,symptom\_6

symptom\_2

symptom\_3,symptom\_4,symptom\_5,symptom\_6

symptom\_2

symptom\_2

symptom\_3,symptom\_4,symptom\_5,symptom\_6

symptom\_2

symptom\_3,symptom\_4,symptom\_5,symptom\_6

……

fever,symptom\_3,symptom\_4,symptom\_6

fever,symptom\_2

fever,symptom\_2

fever,symptom\_3,symptom\_4,symptom\_5,symptom\_6

fever,symptom\_3,symptom\_6

fever,symptom\_3,symptom\_6

……

每一行分别代表一个事务。

项目集合 = {fever, symptom\_1, symptom\_2, symptom\_3, symptom\_4, symptom\_5, symptom\_6, symptom\_7}

1. **频繁项集**

main.py中，将最小支持度设置为0.3，通过数据分析，将得到频繁项集：

1-频繁项目集：

(['symptom\_6'])

(['symptom\_7'])

(['symptom\_4'])

(['symptom\_3'])

(['symptom\_5'])

(['fever'])

(['symptom\_2'])

2-频繁项目集：

(['symptom\_6', 'fever'])

(['symptom\_7', 'symptom\_2'])

(['symptom\_4', 'symptom\_6'])

(['symptom\_4', 'symptom\_3'])

(['symptom\_7', 'fever'])

(['symptom\_4', 'fever'])

(['fever', 'symptom\_3'])

(['symptom\_6', 'symptom\_3'])

(['fever', 'symptom\_2'])

(['symptom\_5', 'symptom\_3'])

3-频繁项目集：

(['symptom\_4', 'fever', 'symptom\_3'])

(['symptom\_6', 'symptom\_3', 'fever'])

(['symptom\_4', 'symptom\_6', 'symptom\_3'])

(['symptom\_7', 'fever', 'symptom\_2'])

(['symptom\_4', 'symptom\_6', 'fever'])

4-频繁项目集：

(['symptom\_4', 'symptom\_6', 'symptom\_3', 'fever'])

1. **导出关联规则，计算其支持度和置信度**

项集的支持度就是该项集出现的次数除以总的记录数。

置信度的意义在于项集{X，Y}同时出现的次数占项集{X}出现次数的比例。

http://img.blog.csdn.net/20150420154115423?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvZ2p3YW5nMTk4Mw==/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

通过apriori.py文件对数据的处理，得到关联规则及支持度和置信度：

(['fever']) --> (['symptom\_6']) support: 0.384615384615 conf: 0.555555555556 lift: 0.902777777778

(['symptom\_6']) --> (['fever']) support: 0.384615384615 conf: 0.625 lift: 0.902777777778

(['symptom\_2']) --> (['symptom\_7']) support: 0.307692307692 conf: 0.666666666667 lift: 2.16666666667

(['symptom\_7']) --> (['symptom\_2']) support: 0.307692307692 conf: 1.0 lift: 2.16666666667

(['symptom\_6']) --> (['symptom\_4']) support: 0.461538461538 conf: 0.75 lift: 1.39285714286

(['symptom\_4']) --> (['symptom\_6']) support: 0.461538461538 conf: 0.857142857143 lift: 1.39285714286

(['symptom\_3']) --> (['symptom\_4']) support: 0.461538461538 conf: 0.666666666667 lift: 1.2380952381

(['symptom\_4']) --> (['symptom\_3']) support: 0.461538461538 conf: 0.857142857143 lift: 1.2380952381

(['fever']) --> (['symptom\_7']) support: 0.307692307692 conf: 0.444444444444 lift: 1.44444444444

(['symptom\_7']) --> (['fever']) support: 0.307692307692 conf: 1.0 lift: 1.44444444444

……

以上X-->Y，其中X称为前件，Y称为后件。其后的值分别表示支持度、置信度以及lift值。

1. **去除冗余的规则**

根据冗余的规则的定义，其代码如下：

rules[i][0] < rules[j][0] and rules[i][1] < rules[j][1] and rules[i][4] <= rules[j][4]

即

1. 规则i的LHS（Left Hand Side)的项集是规则j的LHS（Left Hand Side)的项集的子集
2. 规则i的RHS（Right Hand Side)的项集是规则j的HS（RightHand Side)的项集的子集
3. 规则i的lift值小于等于规则j的lift值

main.py中，当最小支持度设置为0.3时，根据去除冗余的规则，规则由初始的45条变为33条。

main.py文件中，rules用来存储原始的规则，rules\_remove\_redundancy用来存储去除了冗余的规则。

1. **对规则进行评价，可使用Lift**

规则的提升度的意义在于度量项集{X}和项集{Y}的独立性。

http://img.blog.csdn.net/20150420154204408?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvZ2p3YW5nMTk4Mw==/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

通过apriori.py文件对数据的处理，得到关联规则及支持度和置信度以及lift值：

(['fever']) --> (['symptom\_6']) support: 0.384615384615 conf: 0.555555555556 lift: 0.902777777778

(['symptom\_6']) --> (['fever']) support: 0.384615384615 conf: 0.625 lift: 0.902777777778

(['symptom\_2']) --> (['symptom\_7']) support: 0.307692307692 conf: 0.666666666667 lift: 2.16666666667

(['symptom\_7']) --> (['symptom\_2']) support: 0.307692307692 conf: 1.0 lift: 2.16666666667

(['symptom\_6']) --> (['symptom\_4']) support: 0.461538461538 conf: 0.75 lift: 1.39285714286

(['symptom\_4']) --> (['symptom\_6']) support: 0.461538461538 conf: 0.857142857143 lift: 1.39285714286

(['symptom\_3']) --> (['symptom\_4']) support: 0.461538461538 conf: 0.666666666667 lift: 1.2380952381

(['symptom\_4']) --> (['symptom\_3']) support: 0.461538461538 conf: 0.857142857143 lift: 1.2380952381

(['fever']) --> (['symptom\_7']) support: 0.307692307692 conf: 0.444444444444 lift: 1.44444444444

(['symptom\_7']) --> (['fever']) support: 0.307692307692 conf: 1.0 lift: 1.44444444444

……

其中最后一列为lift值。

1. **散点图对规则进行展示**

main.py文件中，对去除了冗余的规则绘制散点图如下：

