**《数据挖掘》第二次作业报告**

**数据**：UCI的“急性炎症”数据集（diagnosis.data、diagnosis.names.txt）

**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。
7. 数据预处理

对数据进行预处理代码是date\_pre\_process.py，将数据的特征数据化。

1. 我们先规定每条记录有某一个特征，该特征位上的数据即为相对的编号（0-7），没有该特征，相应的位置上为“T”。
2. 第一维特征温度，我们以38.0℃为临界值，将数据划分为未发烧（“T”）和发烧（“0”），其他特征分为有该特征（“1-7”）和没有特征（“T”）。

数据结果保存在dataresults.txt中。

1. 使用Apriori算法进行频繁项集的选取

算法代码Apriori.py，算法中，支持度min\_support = 0.3，置信度min\_confidence = 0.5。

结果保存在frequent\_item.txt中，其中每个loop为相应级别的频繁项集以及相应的支持度（除去The 1 loop）。

1. 导出关联规则

由于有很多冗余的频繁项集的存在，我们需要实现删除冗余项。

冗余规则的的定义是：如果rule2的lhs和rhs是分别包含于rules1的lhs和rhs中，且rules2的lift小于或者等于rules1，则称rules2是rules1的冗余规则。

Lift又称为提升度，其定义为：lift(x→y)= support(x,y)/[support(x)\*support(y)]。

process.py中，我们首先根据冗余规则删除冗余的频繁项集，得到的关联规则保存在results.txt中，总共22条，规则如下：

symbol\_0-->>symbol\_1 min\_support: 0.5% min\_confidence:100.0% lift: 1.098901098901099

symbol\_7-->>symbol\_0 min\_support: 0.5% min\_confidence:85.71428571428571% lift: 1.4285714285714284

symbol\_0-->>symbol\_7 min\_support: 0.5% min\_confidence:100.0% lift: 1.4285714285714286

symbol\_2-->>symbol\_1 min\_support: 0.4166666666666667% min\_confidence:100.0% lift: 1.098901098901099

symbol\_4-->>symbol\_1 min\_support: 0.5083333333333333% min\_confidence:100.0% lift: 1.098901098901099

symbol\_7-->>symbol\_1 min\_support: 0.5833333333333334% min\_confidence:100.0% lift: 1.098901098901099

symbol\_7-->>symbol\_2 min\_support: 0.4166666666666667% min\_confidence:71.42857142857143% lift: 1.4285714285714286

symbol\_2-->>symbol\_7 min\_support: 0.4166666666666667% min\_confidence:100.0% lift: 1.4285714285714286

symbol\_4-->>symbol\_6 min\_support: 0.425% min\_confidence:83.60655737704919% lift: 1.3705993012631015

symbol\_7-->>symbol\_5 min\_support: 0.4166666666666667% min\_confidence:71.42857142857143% lift: 1.0204081632653061

symbol\_1,symbol\_7-->>symbol\_0 min\_support: 0.5% min\_confidence:85.71428571428571% lift: 1.4285714285714284

symbol\_0,symbol\_7-->>symbol\_1 min\_support: 0.5% min\_confidence:100.0% lift: 1.098901098901099

symbol\_0,symbol\_1-->>symbol\_7 min\_support: 0.5% min\_confidence:100.0% lift: 1.4285714285714286

symbol\_2,symbol\_7-->>symbol\_1 min\_support: 0.4166666666666667% min\_confidence:100.0% lift: 1.098901098901099

symbol\_1,symbol\_7-->>symbol\_2 min\_support: 0.4166666666666667% min\_confidence:71.42857142857143% lift: 1.4285714285714286

symbol\_1,symbol\_2-->>symbol\_7 min\_support: 0.4166666666666667% min\_confidence:100.0% lift: 1.4285714285714286

symbol\_4,symbol\_6-->>symbol\_1 min\_support: 0.425% min\_confidence:100.0% lift: 1.098901098901099

symbol\_1,symbol\_6-->>symbol\_4 min\_support: 0.425% min\_confidence:100.0% lift: 1.639344262295082

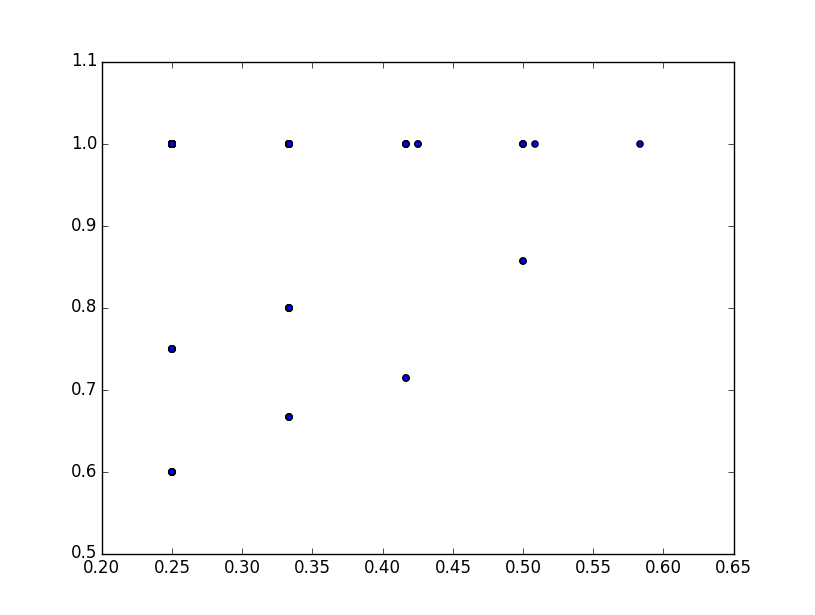
symbol\_1,symbol\_4-->>symbol\_6 min\_support: 0.425% min\_confidence:83.60655737704919% lift: 1.3705993012631015

symbol\_5,symbol\_7-->>symbol\_1 min\_support: 0.4166666666666667% min\_confidence:100.0% lift: 1.098901098901099

symbol\_1,symbol\_7-->>symbol\_5 min\_support: 0.4166666666666667% min\_confidence:71.42857142857143% lift: 1.0204081632653061

symbol\_1,symbol\_5-->>symbol\_7 min\_support: 0.4166666666666667% min\_confidence:100.0% lift: 1.4285714285714286

然后根据删除了冗余项后的频繁项集的support和confidence绘制相应的散点图，结果如下：



**Figure 1:scatter, min\_support = 0.3， min\_confidence = 0.5**