关联规则挖掘

姓名：宁小东

学号：2120151024

1. 问题描述

选取以下数据集：

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

要求为：对数据集进行处理，转换成合适关联规则挖掘的形式；找出频繁项集；导出关联规则，计算其支持度和置信度；去除冗余的规则；对规则进行评价，可使用Lift，也可使用教材中所提及的其他指标；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。

1. 数据分析
   1. 数据集说明

有待关联规则挖掘的数据源是UCI“急性炎症”数据集，该数据集共有120条记录，每条记录有8个数据项：

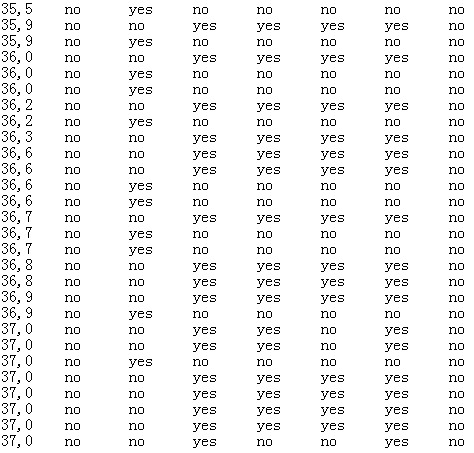
|  |  |
| --- | --- |
| 数据项 | 取值范围 |
| Temperature of patient（体温） | 35C– 42C |
| Occurrence of nausea（恶心） | yes; no; |
| Lumbar pain（腰疼） | yes; no; |
| Urine pushing（尿频） | yes; no; |
| Micturition pains（排尿疼痛） | yes; no; |
| Burning of urethra（尿道灼烧） | yes; no; |
| decision: Inflammation of urinary bladder（膀胱炎症） | yes; no; |
| decision: Nephritis of renal pelvis origin（肾盂源性肾炎） | yes; no; |

* 1. 数据集处理

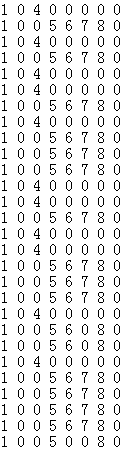
为了便于计算，将如上取值范围（{yes, no}）转化为int型整数，转化规则为：

* Temperature of patient：根据医学上对膀胱炎和肾炎的描述可知，膀胱炎患者体温一般不超过38C，而肾炎患者体温多在40C以上。因此，对于体温数据项，可以采用如下处理，当体温未超过38C时，将该项记为1，当体温超过38C时，将该项记为2；
* Occurrence of nausea：这是一种临床症状，它的取值只有yes或者no，代表患者是否有该症状，当取值为yes是，将该项记为3，当取值为no时，将该项记为0；
* Lumbar pain：这是一种临床症状，它的取值只有yes或者no，代表患者是否有该症状，当取值为yes是，将该项记为4，当取值为no时，将该项记为0；
* Urine pushing：这是一种临床症状，它的取值只有yes或者no，代表患者是否有该症状，当取值为yes是，将该项记为5，当取值为no时，将该项记为0；
* Micturition pains：这是一种临床症状，它的取值只有yes或者no，代表患者是否有该症状，当取值为yes是，将该项记为6，当取值为no时，将该项记为0；
* Burning of urethra：这是一种临床症状，它的取值只有yes或者no，代表患者是否有该症状，当取值为yes是，将该项记为7，当取值为no时，将该项记为0；
* Decision：Inflammation of urinary bladder：该项代表疾病诊断结果，它的取值只有yes或者no，代表患者是否患有该病症，当取值为yes是，将该项记为8，当取值为no时，将该项记为0；
* Decision：Nephritis of renal pelvis origin：该项代表疾病诊断结果，它的取值只有yes或者no，代表患者是否患有该病症，当取值为yes是，将该项记为9，当取值为no时，将该项记为0。

使用如上规则处理数据集，则处理前后对应的部分数据集分别为：



处理后的数据集：



* 1. 计算频繁项集

使用Apriori算法计算频繁项集，算法的步骤为：

Step1：将处理后的数据库读入到一个二维数组中，保存下来，该二维数组是一个120\*9的数组，共包含120条记录，每条记录有9个数据项，分别代表上述的数据项。

Step2：设置最小支持度，然后针对每一个数据项，扫描一遍数组，算出每个数据项的频数，将频数小于最小支持度的数据项去除，将频数大于等于最小支持度的数据项记录下来。

Step3：将第二步得到的数据项进行拼接，即将不同的数据项组合起来，得到若干个2-项集，针对每一个2-项集，扫描一遍数组，算出每个2-项集的频数，频数大于等于最小支持度的2-项集为频繁2-项集，并记录下来，同时将频数小于最小支持度的2-项集舍弃。设k=2。

Step4：将k-项集进行拼接，对于项数大于等于2的k-项集拼接时，拼接满足如下规则：

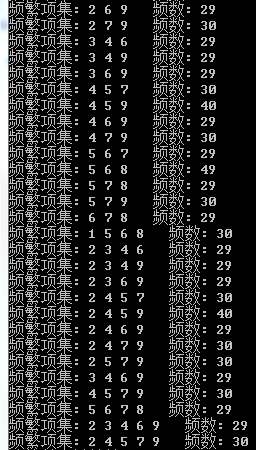
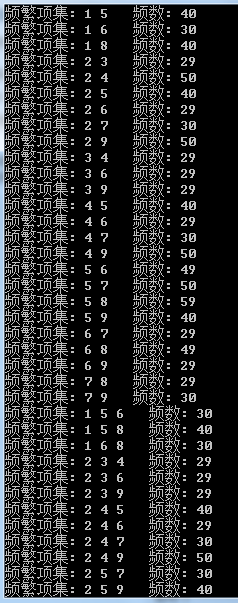
* 扫描并比较两个k-项集的每一项，若这两个k-项集有且仅有一项不同，则这个k-项集可以拼接成一个k+1-项集，否则不能拼接成功。拼接后的k+1-项由两个k项集的所有项组成；
* 将所有不同的k-项集拼接后，得到若干个k+1-项集，针对每一个k+1-项集，扫描一遍数组，算出每个k+1-项集的频数，频数大于等于最小支持度的k+1-项集为频繁k+1-项集，并记录下来，同时将频数小于最小支持度的k+1-项集舍弃。若频繁k+1-项集的数目小于等于1，执行第五步，否则，令k=k+1，然后执行Step4；

Step5：输出所有的频繁i-项集和它们的频数。（i>=2）

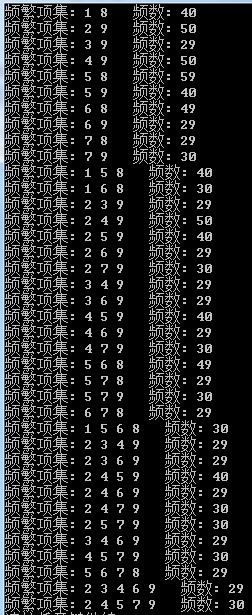
1. 计算关联规则
   1. 去除冗余规则

频繁集计算模块可得到所有的频繁集。而在本例中，研究目标是根据患者症状推导患者病情，因此，在所有的频繁集中，只需要研究包含有患病情况8或者9的频繁集，因此，不含有8或者9属性的频繁集将被舍弃。则去除冗余频繁集前后的频繁集分别为：

处理前：



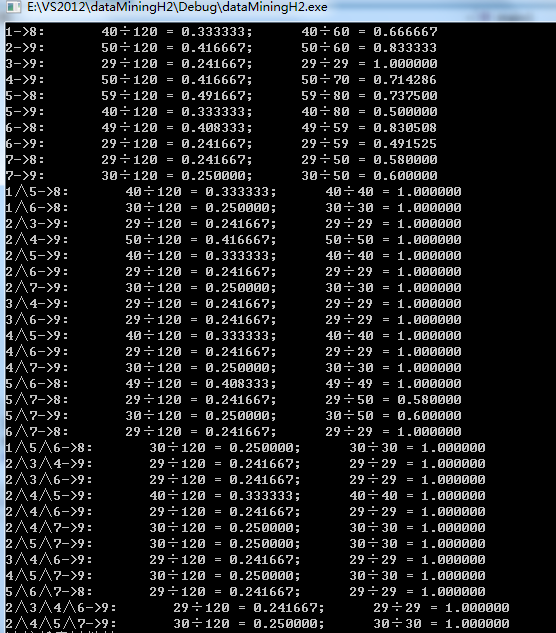
处理后：



* 1. 计算关联规则

在本例中，令最小置信度为0.8。对处理后的所有频繁集计算其关联规则：

计算关联规则支持度置信度：



图中从左至右分别表示关联规则、支持度和置信度。当置信度设为0.8时，右图中数据可以得出所有的强关联规则，如下：

2->9; 3->9; 6->8; 1∧5->8; 1∧6->8; 2∧3->9; 2∧4->9;

2∧5->9; 2∧6->9; 2∧7->9; 3∧4->9; 3∧6->9; 4∧5->9;

4∧6->9; 4∧7->9; 5∧6->8; 6∧7->8; 1∧5∧6->8; 2∧3∧4->9;

2∧3∧6->9; 2∧4∧5->9; 2∧4∧6->9; 2∧4∧7->9; 2∧5∧6->9;

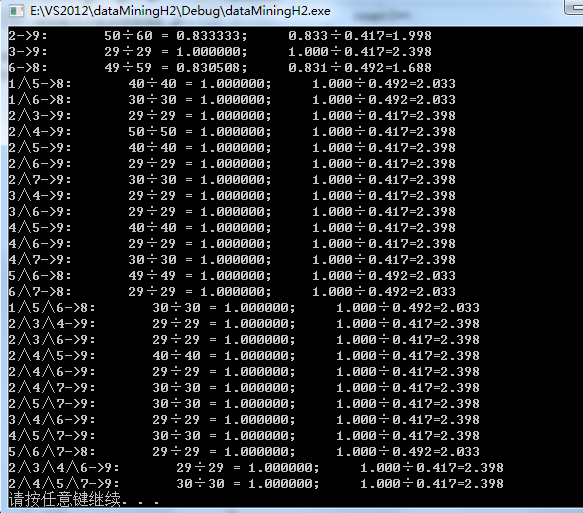
3∧4∧6->9; 4∧5∧7->9; 5∧6∧7->8; 2∧3∧4∧6->9; 2∧4∧5∧7->9;

* 1. 关联规则评价及可视化

首先采用相关性系数Lift评价：

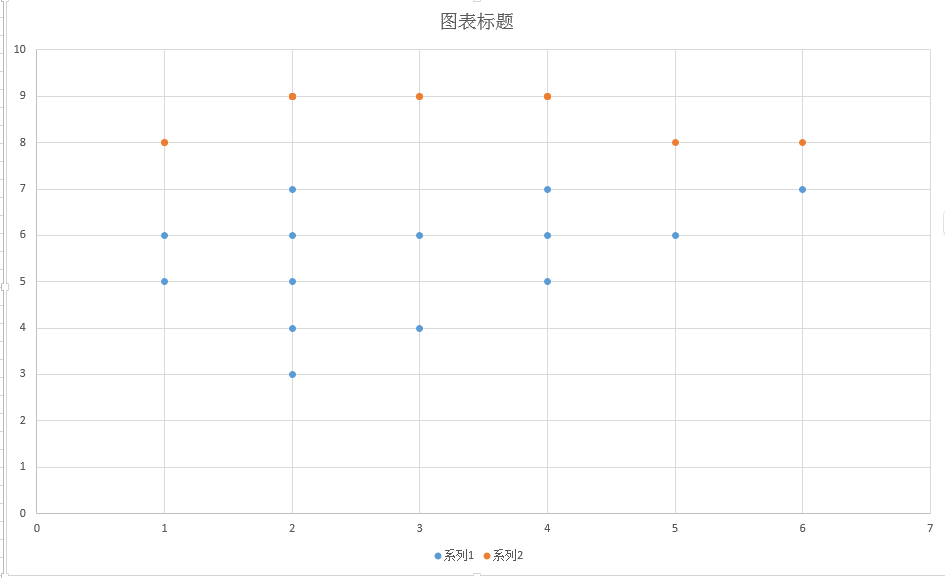
Lift（A->B）=confidence(A->B)/support(B),根据计算公式，可得各关联规则的相关性系数如下：

关联规则置信度相关性系数



对于规则A—>B或者B—>A，lift(A,B)=P(A∩B)/(P(A)\*P(B))，如果lift(A,B)>1表示A、B呈正相关，lift(A,B)<1表示A、B呈负相关，lift(A,B)=1表示A、B不相关（独立）。正相关和负相关都是我们需要关注的，而独立往往是我们不需要的，两个商品都没有相互影响也就是不是强规则。根据图中的计算结果可知，除了关联规则2->9和6->8的相关性系数小于2外，其他的都在2以上，远离于1说明大部分关联规则有较强的正相关。

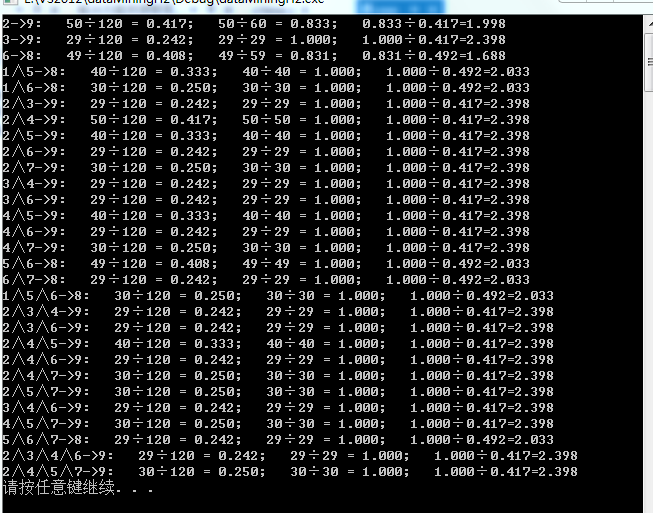
使用散点图对关联规则进行可视化，可得3-频繁集对应的关联规则散点图如下：



1. 挖掘结果及分析

挖掘结果如下：

关联规则支持度置信度相关性系数



数字1到9分别对应如下：

1：体温低于38C

2：体温高于38C

3：恶心现象

4：腰疼

5：尿频

6：排尿疼痛

7：尿道灼烧

8：膀胱炎

9：肾炎

根据挖掘的关联规则，可以得出如下结果：

膀胱炎患者具有以下临床症状：体温低于38C；尿频；排尿疼痛；尿道灼烧。

肾炎患者具有以下临床症状：体温高于38C；恶心；腰疼；尿频；排尿疼痛；尿道灼烧。