**数据挖掘作业二**

学院：计算机学院

专业：计算机科学与技术

学号：2120150994

姓名：黄建峰

## 一、源数据的处理：

1. 数据集

关联规则挖掘的数据源是UCI的“急性炎症”数据集，下载地址是（<http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations>）。

该数据集共有120条记录，每条记录有8个数据项，具体如下：

|  |  |
| --- | --- |
| 数据项 | 取值情况 |
| Temperature of patient（体温） | 35 C– 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 | yes ; no; |
| decision: Nephritis of renal pelvis origin | yes ; no; |

其中，第一个数据项表示患者体温，是一个具体的数值，取值范围是35C到42C；第二到五个数据项表示各种症征，取值为yes或者no，yes代表有这种症状，no代表没有；第八和第九个数据项是诊断结果，分别对应膀胱炎和肾炎两种疾病，yes代表患有该种疾病，no代表未患有该种疾病。

1. 处理数据集

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。

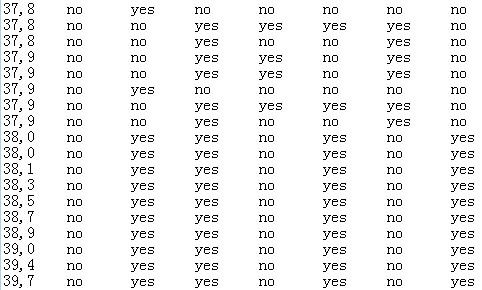
Burning of urethra：这是一种临床症状，它的取值只有yes或者no，代表患者是否有该症状，当取值为yes是，将该项记为7，当取值为no时，将该项记为0。

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。

下面分别给出处理前的数据集和用上述规则处理后的数据集：

处理前（部分）：



处理后（部分）：



## 二、计算频繁集

对于频繁集的计算，采用Apriori算法的思想实现，其具体过程如下:

第一步：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。

第二步：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。

第三步：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。

第四步：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,然后执行第四步。

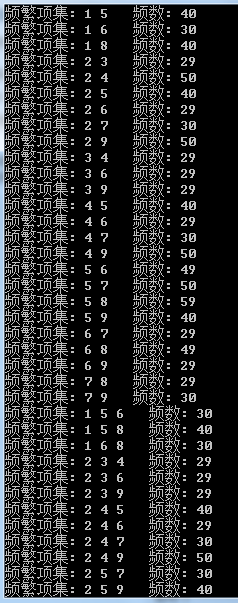
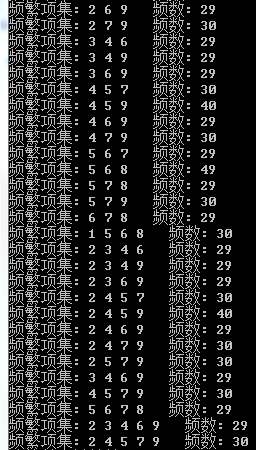
第五步：输出所有的频繁i-项集和它们的频数。（i>=2）

## 三、计算关联规则

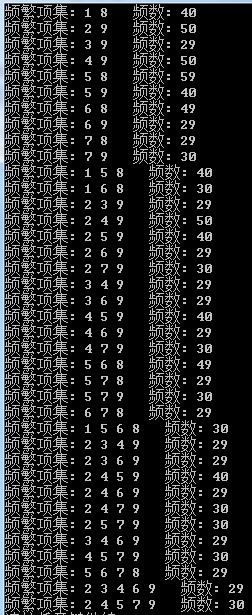
1.去除冗余规则

频繁集计算模块得到的是所有的频繁集，但是在本例中，研究目标是根据患者症状推导患者的病情，因此，在所有的频繁集中，我们只需要研究包含有患病情况8或者9的频繁集，因此，对于不含有8或者9的频繁集，都将被舍弃，不去计算，下面分别是去除冗余频繁集前后的频繁集：

处理前：

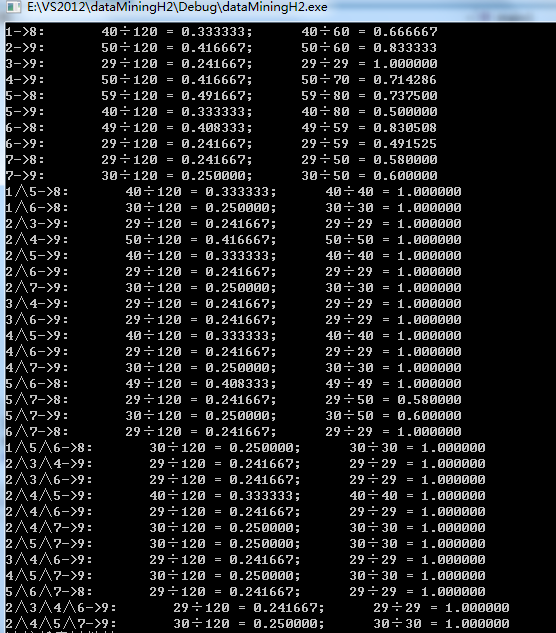
处理后：



2．计算关联规则

在本例中，设置最小置信度为0.8，下面对处理后的所有频繁集，计算其关联规则。

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;

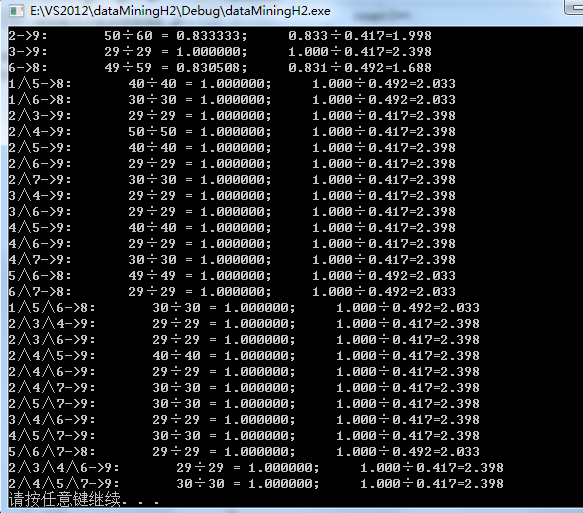
3∧4∧6->9; 4∧5∧7->9; 5∧6∧7->8; 2∧3∧4∧6->9; 2∧4∧5∧7->9;

## 四、关联规则评价及可视化

1.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：

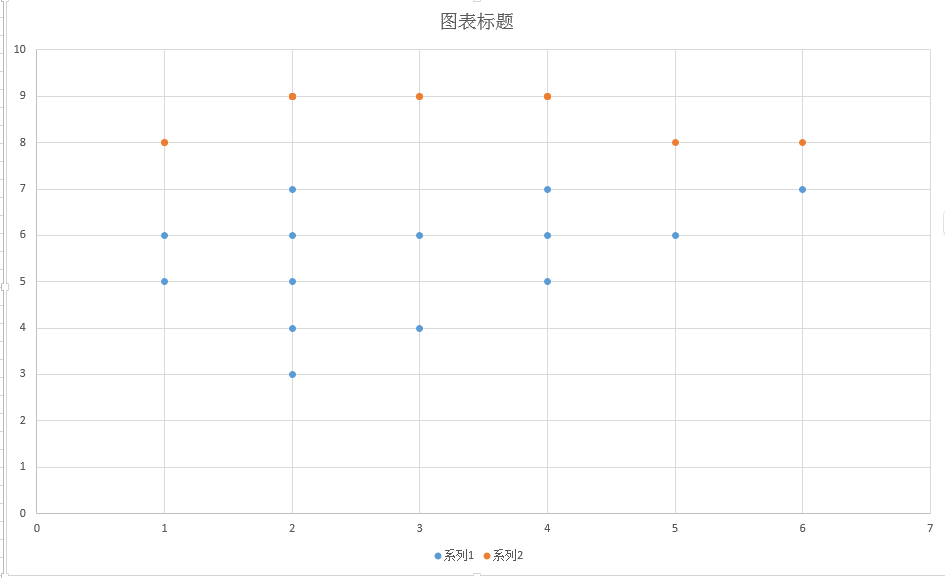
关联规则 置信度 相关性系数



对于规则A—>B或者B—>A，lift(A,B)=P(A交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。

2.用散点图实现关联规则的可视化

针对3-频繁集对应的关联规则，作出其对应的散点图如下所示：

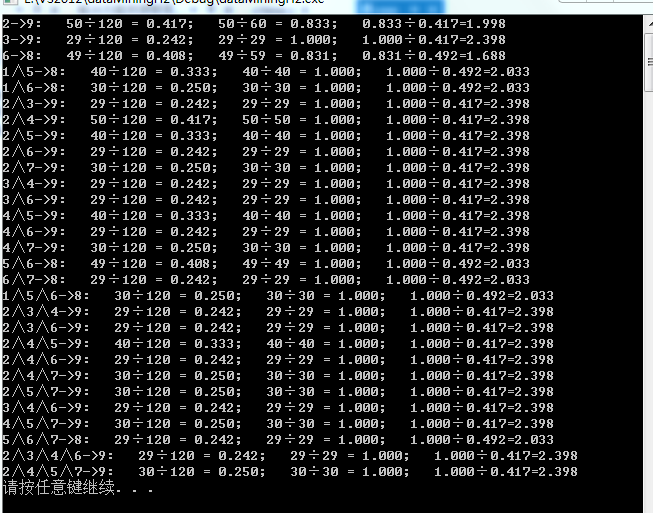


蓝点表示对应的临床症状，橙色点表示患病类型。

## 五、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。