关联规则挖掘

学院：计算机学院

专业：计算机科学与技术

班级：2015级硕士五班

学号：2120150992

姓名：侯军

一、处理数据集

1. 数据集

采用的是Titanic数据集（下载地址：<https://www.kaggle.com/c/titanic/data>）。数据集共有891条记录，每条记录有12个变量：PassengerId（乘客ID）、 Survived（0代表遇难，1代表幸存）、Pclass（乘客的社会阶层）、Name（乘客姓名）、Sex（乘客性别）、Age（乘客年龄）、SibSp（乘客的兄弟姐妹/配偶登船人数）、Parch（乘客的父母/孩子登船人数）、Ticket（船票号）、Fare（费用）、Cabin（客舱号）、embarked（登船的港口，C=Cherbourg；Q=Queenstown；S=Southampton）。

2. 处理数据集

（1）对于数据挖掘任务来说，PassengerId、Name、Ticket、Fare、Cabin这样的信息是无用的，所以将之删除。

（2）把Age属性归类，0~10（包含10），10~20为“teenager”，20~40为“youth”，40~60为“middle-ager”，60以上的为“elder”；SibSp属性值为0的共608条，标记为“no”，其他标记为“has”；Parch属性值为0的共678条，标记为“no”，其他标记为“has”；处理后的数据如下图：

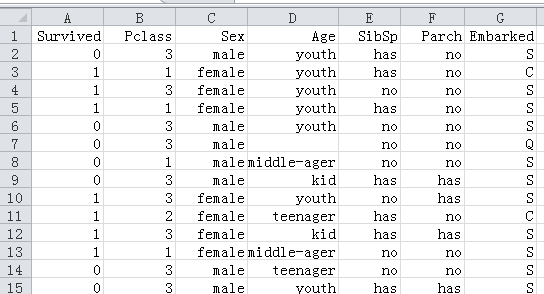
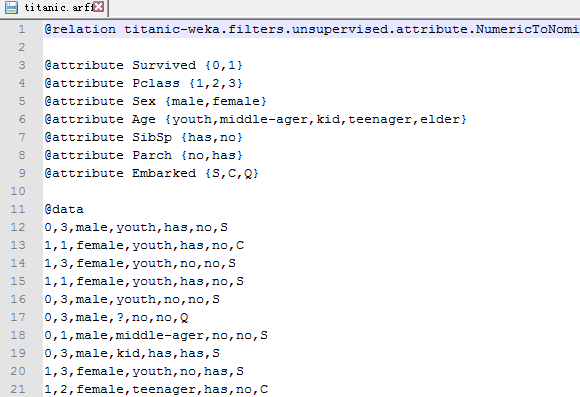


图1

（3）运用数据挖掘工具weka将CSV文件转化为ARFF文件：进入“Explore”模块，点击“Openfile”按钮打开CSV文件然后另存为ARFF文件即可。处理后数据如图2所示：



关系声明：

关系名称在ARFF文件的第一个有效行来定义，格式为@relation<relation-name>

属性声明：

属性声明用一列以“@attribute”开头的语句表示。数据集中的每一个属性都有它对应的“@attribute”语句，来定义它的属性名称和数据类型。属性声明的格式为@attribute <attribute-name> <datatype>

其中<attribute-name>是必须以字母开头的字符串。和关系名称一样，如果这个字符串包含空格，它必须加上引号。

WEKA支持的<datatype>有四种，分别是：

numeric-------------------------数值型

<nominal-specification>-----分类（nominal）型

string----------------------------字符串型

date [<date-format>]--------日期和时间型

数据信息：

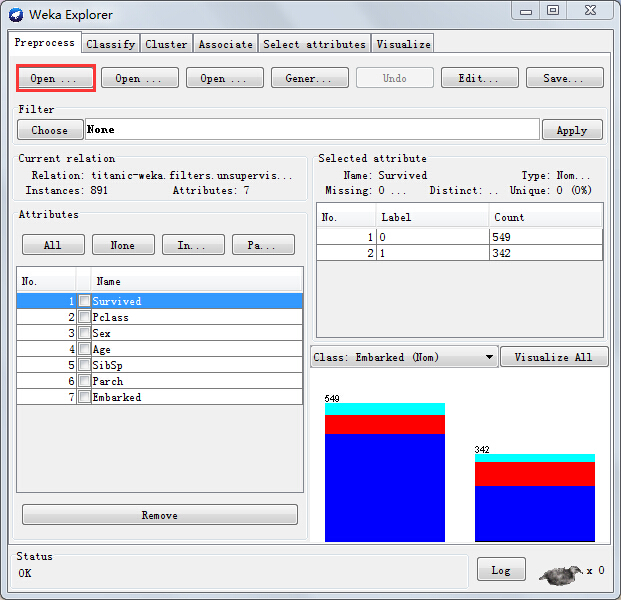
数据信息中“@data”标记独占一行，剩下的是各个实例的数据。每个实例占一行。实例的个属性值用逗号隔开。如果某个属性值是缺失值，则用“？”表示，且这个问号不能省略。

二、找出频繁项集、关联规则、对规则进行评价

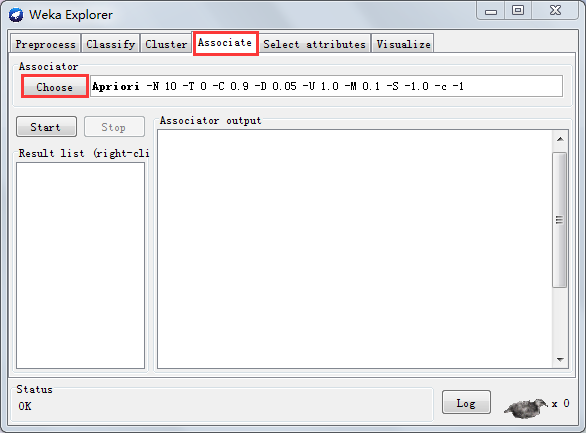
1、安装好weka后，打开Explorer



2、打开titanic.arff文件



3、选择关联规则挖掘，选择Apriori算法



4、设置参数

参数主要是选择支持度（lowerBoundMinSupport），规则评价机制metriType及对应的最小值，参数设置说明如下：

1. car 如果设为真，则会挖掘类关联规则而不是全局关联规则。

2. classindex 类属性索引。如果设置为-1，最后的属性被当做类属性。

3. delta 以此数值为迭代递减单位。不断减小支持度直至达到最小支持度或产生了满足数量要求的规则。

4. lowerBoundMinSupport 最小支持度下界。

5. metricType 度量类型。设置对规则进行排序的度量依据。可以是：置信度（类关联规则只能用置信度挖掘），提升度(lift)，杠杆率(leverage)，确信度(conviction)。

在 Weka中设置了几个类似置信度(confidence)的度量来衡量规则的关联程度，它们分别是：

a) Lift ： P(A,B)/(P(A)P(B)) Lift=1时表示A和B独立。这个数越大(>1)，越表明A和B存在于一个购物篮中不是偶然现象,有较强的关联度.

b) Leverage :P(A,B)-P(A)P(B)Leverage=0时A和B独立，Leverage越大A和B的关系越密切

c) Conviction:P(A)P(!B)/P(A,!B) （!B表示B没有发生） Conviction也是用来衡量A和B的独立性。从它和lift的关系（对B取反，代入Lift公式后求倒数）可以看出，这个值越大, A、B越关联。

6. minMtric 度量的最小值。

7. numRules 要发现的规则数。

8. outputItemSets 如果设置为真，会在结果中输出项集。

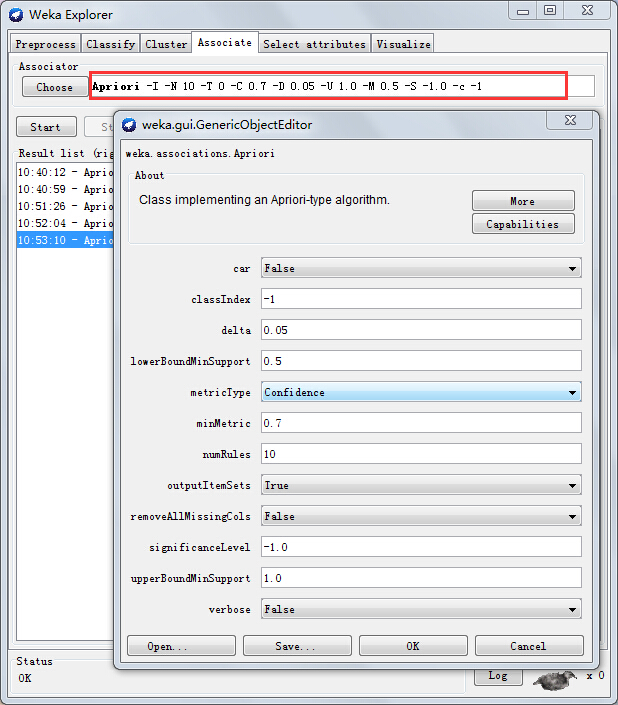
9. removeAllMissingCols 移除全部为缺省值的列。

10. significanceLevel 重要程度。重要性测试（仅用于置信度）。

11. upperBoundMinSupport 最小支持度上界。 从这个值开始迭代减小最小支持度。

12. verbose 如果设置为真，则算法会以冗余模式运行。

1)用置信度评价



设置好参数后点击start运行，结果如下：

=== Run information ===

Scheme: weka.associations.Apriori -I -N 10 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1

Relation: titanic-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 891

Attributes: 7

Survived

Pclass

Sex

Age

SibSp

Parch

Embarked

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.5 (446 instances) //最小支持度0.5.即最小需要446个实例

Minimum metric <confidence>: 0.7 //最小度量<置信度>：0.7

Number of cycles performed: 10 //进行了10轮搜索

Generated sets of large itemsets: //生成的频繁项集

Size of set of large itemsets L(1): 6 //频繁1项集：6个

Large Itemsets L(1):

Survived=0 549

Pclass=3 491

Sex=male 577

SibSp=no 608

Parch=no 678

Embarked=S 644

Size of set of large itemsets L(2): 4 //频繁2项集：4个

Large Itemsets L(2):

Survived=0 Sex=male 468

Sex=male Parch=no 484

SibSp=no Parch=no 537

Parch=no Embarked=S 484

Best rules found: //最佳关联规则

1. SibSp=no 608 ==> Parch=no 537 <conf:(0.88)> lift:(1.16) lev:(0.08) [74] conv:(2.02)

2. Survived=0 549 ==> Sex=male 468 <conf:(0.85)> lift:(1.32) lev:(0.13) [112] conv:(2.36)

3. Sex=male 577 ==> Parch=no 484 <conf:(0.84)> lift:(1.1) lev:(0.05) [44] conv:(1.47)

4. Sex=male 577 ==> Survived=0 468 <conf:(0.81)> lift:(1.32) lev:(0.13) [112] conv:(2.01)

5. Parch=no 678 ==> SibSp=no 537 <conf:(0.79)> lift:(1.16) lev:(0.08) [74] conv:(1.52)

6. Embarked=S 644 ==> Parch=no 484 <conf:(0.75)> lift:(0.99) lev:(-0.01) [-6] conv:(0.96)

7. Parch=no 678 ==> Sex=male 484 <conf:(0.71)> lift:(1.1) lev:(0.05) [44] conv:(1.23)

8. Parch=no 678 ==> Embarked=S 484 <conf:(0.71)> lift:(0.99) lev:(-0.01) [-6] conv:(0.96)

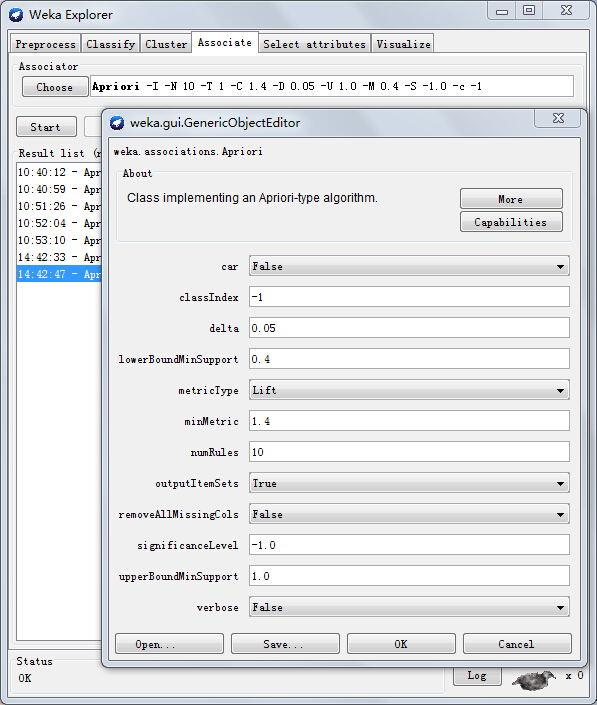
结果分析：

当最小支持度设为0.5时，有10项频繁项集，计算可得出76.1%的乘客没有带父母/孩子登船，72.3%的乘客是从Southampton 港口登船的，68.2%的乘客的兄弟姐妹/配偶没有登船，64.5%的乘客是男性，61.7%的乘客最后遇难，第三阶层的乘客占总人数的55.1%；另外男性乘客遇难人数占总人数的52.5%。

通过最佳关联规则可以看出，遇难的乘客中有85%都是男性，而男性乘客中有81%遇难。

2)用Lift评价

参数配置：



运行结果如下：

Apriori

=======

Minimum support: 0.4 (356 instances) //最小支持度0.4.即最小需要356个实例

Minimum metric <lift>: 1.4 //最小度量<Lift>：1.4

Number of cycles performed: 12 //进行了12轮搜索

Generated sets of large itemsets: //生成的频繁项集

Size of set of large itemsets L(1): 7 //频繁1项集：7个

Large Itemsets L(1):

Survived=0 549

Pclass=3 491

Sex=male 577

Age=youth 385

SibSp=no 608

Parch=no 678

Embarked=S 644

Size of set of large itemsets L(2): 12 //频繁2项集：12个

Large Itemsets L(2):

Survived=0 Pclass=3 372

Survived=0 Sex=male 468

Survived=0 SibSp=no 398

Survived=0 Parch=no 445

Survived=0 Embarked=S 427

Pclass=3 Parch=no 381

Sex=male SibSp=no 434

Sex=male Parch=no 484

Sex=male Embarked=S 441

SibSp=no Parch=no 537

SibSp=no Embarked=S 438

Parch=no Embarked=S 484

Size of set of large itemsets L(3): 7 //频繁3项集：10个

Large Itemsets L(3):

Survived=0 Sex=male SibSp=no 361

Survived=0 Sex=male Parch=no 404

Survived=0 Sex=male Embarked=S 364

Survived=0 SibSp=no Parch=no 374

Sex=male SibSp=no Parch=no 411

Sex=male Parch=no Embarked=S 373

SibSp=no Parch=no Embarked=S 393

Best rules found: //最佳关联规则

1. Sex=male 577 ==> Survived=0 Parch=no 404 conf:(0.7) < lift:(1.4)> lev:(0.13) [115] conv:(1.66)

2. Survived=0 Parch=no 445 ==> Sex=male 404 conf:(0.91) < lift:(1.4)> lev:(0.13) [115] conv:(3.73)

3. Sex=male 577 ==> Survived=0 SibSp=no 361 conf:(0.63) < lift:(1.4)> lev:(0.12) [103] conv:(1.47)

4. Survived=0 SibSp=no 398 ==> Sex=male 361 conf:(0.91) < lift:(1.4)> lev:(0.12) [103] conv:(3.69)

分析结果：

频繁项集的前十项与用置信度评价的结果一样，不再赘述。

通过最佳关联规则可以看出，男性与是否遇难有呈正相关。