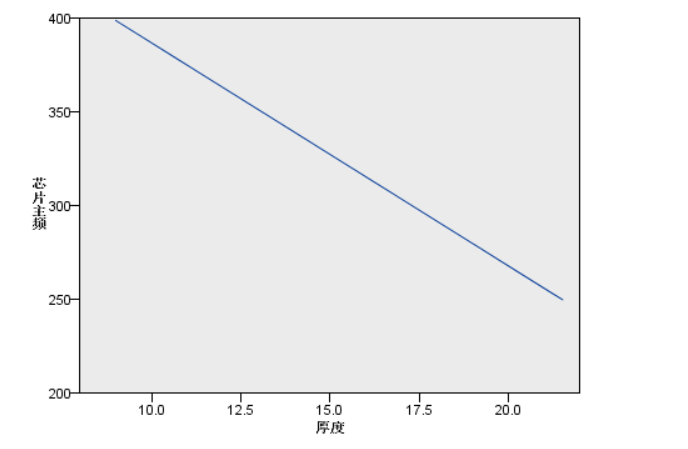
数据挖掘实验报告

## 实验一

1. 任务说明：
2. 熟悉SPSS软件和EXCEL软件的使用；
3. 基于你的理解，给出给定数据文件的描述报告；
4. 识别给定数据文件中的连续属性，并采用你认为合适的方法完成连续属性的离散化。
5. 实验结果
6. 此组数据展示的是2018年到2012年的各品牌手机上市情况，经过几个相关属性的分析，这短时间之内低端机占大多数（以芯片的主频作为评估变量），并且作为主频高的高端机来说，手机的厚度也是越薄的。

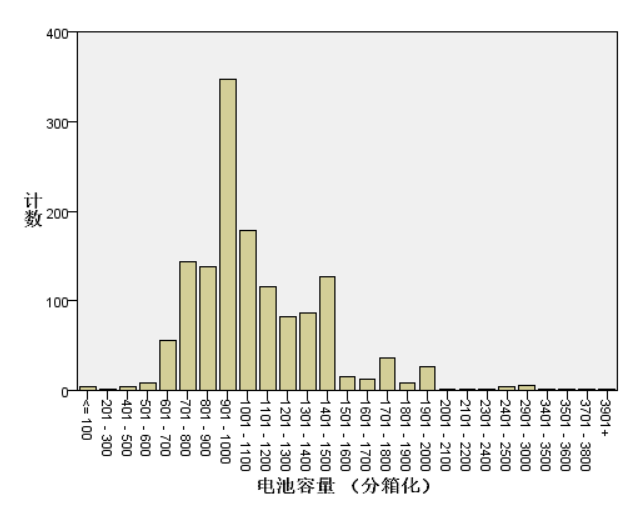


从某种意义上说，贵的手机越好看，手感越好。

1. 关于连续变量

所谓连续变量，就是取值可能为很多多情况并且还会连续，比如一个月之内的每天降水量，手机电池容量等。

在这次实验中我们选取手机电池容量来进行离散化。并用分箱操作将其离散化。最后创建名为“电池容量分类”属性来将其储存。



## 实验二

1. 任务说明
2. 采用K均值算法对给定数据集进行聚类，给出聚类结果。改变K的不同取值，研究K值改变给聚类结果所带来的变化。改变初始簇心，研究簇心变化给聚类结果所带来的变化。(SPSS和weka中都有K均值算法的实现，大家可以都试一下，对比下同一种算法在两种软件中的实现有何不同，你认为哪种更好)
3. 采用另一种你所熟悉的聚类算法，如DBSCAN，完成第一步的聚类过程，并分析与K均值算法相比，哪种聚类算法的效果最好。
4. 实验结果
5. 快速聚类(k=2, 迭代10次)

|  |  |  |
| --- | --- | --- |
| **备注** | | |
| 已创建输出 | | 05-NOV-2018 10:24:53 |
| 注释 | |  |
| 输入 | 活动数据集 | 数据集1 |
| 过滤器 | <无> |
| 权重 | <无> |
| 拆分文件 | <无> |
| 工作数据文件中的行数 | 2800 |
| 缺失值处理 | 对缺失的定义 | 将用户定义的缺失值视为缺失。 |
| 使用的个案数 | 统计基于那些对任何所用聚类变量都没有缺失值的个案。 |
| 语法 | | QUICK CLUSTER 语音时长分钟 平均带宽Kbps 流量4G占比 上网时长分钟 上网流量MB  /MISSING=LISTWISE  /CRITERIA=CLUSTER(2) MXITER(10) CONVERGE(0)  /METHOD=KMEANS(NOUPDATE)  /PRINT INITIAL. |
| 资源 | 处理程序时间 | 00:00:00.05 |
| 耗用时间 | 00:00:00.07 |
| 所需工作空间 | 816 字节 |

[数据集1]

|  |  |  |
| --- | --- | --- |
| **初始聚类中心** | | |
|  | 聚类 | |
| 1 | 2 |
| 语音时长分钟 | 162.766666666666680 | 86.122222216666660 |
| 平均带宽Kbps | 903.449855 | 27626.745615 |
| 流量4G占比 | .872515000000000 | .999332000000000 |
| 上网时长分钟 | 178683.097222216700000 | 70.850000000000000 |
| 上网流量MB | 2471.613517284393300 | 3.206974983215332 |

|  |  |  |
| --- | --- | --- |
| **迭代历史记录a** | | |
| 迭代 | 聚类中心中的变动 | |
| 1 | 2 |
| 1 | .000 | 26848.013 |
| 2 | .000 | .000 |

|  |
| --- |
| a. 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 2。初始中心之间的最小距离为 180617.187。 |
|  |

|  |  |  |
| --- | --- | --- |
| **最终聚类中心** | | |
|  | 聚类 | |
| 1 | 2 |
| 语音时长分钟 | 162.766666666666680 | 28.442169307511335 |
| 平均带宽Kbps | 903.449855 | 919.795741 |
| 流量4G占比 | .872515000000000 | .925246329878484 |
| 上网时长分钟 | 178683.097222216700000 | 2816.264641311280000 |
| 上网流量MB | 2471.613517284393300 | 121.524376504114950 |

|  |  |  |
| --- | --- | --- |
| **每个聚类中的个案数目** | | |
| 聚类 | 1 | 1.000 |
| 2 | 2798.000 |
| 有效 | | 2799.000 |
| 缺失 | | 1.000 |

1. 快速聚类(k=3, 迭代10次)

|  |  |  |
| --- | --- | --- |
| r | | |
| 已创建输出 | | 05-NOV-2018 10:26:52 |
| 注释 | |  |
| 输入 | 活动数据集 | 数据集1 |
| 过滤器 | <无> |
| 权重 | <无> |
| 拆分文件 | <无> |
| 工作数据文件中的行数 | 2800 |
| 缺失值处理 | 对缺失的定义 | 将用户定义的缺失值视为缺失。 |
| 使用的个案数 | 统计基于那些对任何所用聚类变量都没有缺失值的个案。 |
| 语法 | | QUICK CLUSTER 语音时长分钟 平均带宽Kbps 流量4G占比 上网时长分钟 上网流量MB  /MISSING=LISTWISE  /CRITERIA=CLUSTER(3) MXITER(10) CONVERGE(0)  /METHOD=KMEANS(NOUPDATE)  /PRINT INITIAL. |
| 资源 | 处理程序时间 | 00:00:00.05 |
| 耗用时间 | 00:00:00.05 |
| 所需工作空间 | 1024 字节 |

|  |  |  |  |
| --- | --- | --- | --- |
| **初始聚类中心** | | | |
|  | 聚类 | | |
| 1 | 2 | 3 |
| 语音时长分钟 | .233333333333333 | 86.122222216666660 | 162.766666666666680 |
| 平均带宽Kbps | 757.938855 | 27626.745615 | 903.449855 |
| 流量4G占比 | 1.000000000000000 | .999332000000000 | .872515000000000 |
| 上网时长分钟 | 39372.833333333336000 | 70.850000000000000 | 178683.097222216700000 |
| 上网流量MB | 1707.584392547607400 | 3.206974983215332 | 2471.613517284393300 |

|  |  |  |  |
| --- | --- | --- | --- |
| **迭代历史记录a** | | | |
| 迭代 | 聚类中心中的变动 | | |
| 1 | 2 | 3 |
| 1 | 19978.526 | 26736.261 | .000 |
| 2 | 100.532 | 6.622 | .000 |
| 3 | 96.967 | 6.865 | .000 |
| 4 | 198.962 | 13.748 | .000 |
| 5 | 50.321 | 3.477 | .000 |
| 6 | 50.693 | 3.483 | .000 |
| 7 | .000 | .000 | .000 |

|  |
| --- |
| a. 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 7。初始中心之间的最小距离为 47639.175。 |

|  |  |  |  |
| --- | --- | --- | --- |
| **最终聚类中心** | | | |
|  | 聚类 | | |
| 1 | 2 | 3 |
| 语音时长分钟 | 129.217526234907400 | 21.513382352992085 | 162.766666666666680 |
| 平均带宽Kbps | 745.336342 | 931.790658 | 903.449855 |
| 流量4G占比 | .897178038888889 | .927176158899924 | .872515000000000 |
| 上网时长分钟 | 19910.763472222123000 | 1640.936226657358700 | 178683.097222216700000 |
| 上网流量MB | 820.404415527184500 | 73.473036922696950 | 2471.613517284393300 |

|  |  |  |
| --- | --- | --- |
| **每个聚类中的个案数目** | | |
| 聚类 | 1 | 180.000 |
| 2 | 2618.000 |
| 3 | 1.000 |
| 有效 | | 2799.000 |
| 缺失 | | 1.000 |

1. 快速聚类(k=4, 迭代10次)

|  |  |  |
| --- | --- | --- |
| **备注** | | |
| 已创建输出 | | 05-NOV-2018 10:30:15 |
| 注释 | |  |
| 输入 | 活动数据集 | 数据集1 |
| 过滤器 | <无> |
| 权重 | <无> |
| 拆分文件 | <无> |
| 工作数据文件中的行数 | 2800 |
| 缺失值处理 | 对缺失的定义 | 将用户定义的缺失值视为缺失。 |
| 使用的个案数 | 统计基于那些对任何所用聚类变量都没有缺失值的个案。 |
| 语法 | | QUICK CLUSTER 语音时长分钟 平均带宽Kbps 流量4G占比 上网时长分钟 上网流量MB  /MISSING=LISTWISE  /CRITERIA=CLUSTER(4) MXITER(10) CONVERGE(0)  /METHOD=KMEANS(NOUPDATE)  /PRINT INITIAL. |
| 资源 | 处理程序时间 | 00:00:00.06 |
| 耗用时间 | 00:00:00.06 |
| 所需工作空间 | 1240 字节 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **初始聚类中心** | | | | |
|  | 聚类 | | | |
| 1 | 2 | 3 | 4 |
| 语音时长分钟 | 86.122222216666660 | 284.400000000000000 | 440.783333333333360 | 162.766666666666680 |
| 平均带宽Kbps | 27626.745615 | 274.605670 | 61.032033 | 903.449855 |
| 流量4G占比 | .999332000000000 | 1.000000000000000 | .258989000000000 | .872515000000000 |
| 上网时长分钟 | 70.850000000000000 | 12792.450000000000000 | 43113.988888883340000 | 178683.097222216700000 |
| 上网流量MB | 3.206974983215332 | 201.008565902709960 | 135.633234659831060 | 2471.613517284393300 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **迭代历史记录a** | | | | |
| 迭代 | 聚类中心中的变动 | | | |
| 1 | 2 | 3 | 4 |
| 1 | 10297.690 | 10333.638 | 9551.118 | .000 |
| 2 | 3372.998 | 397.143 | 6871.337 | .000 |
| 3 | 1879.804 | 193.472 | 3178.246 | .000 |
| 4 | 1791.993 | 114.741 | 1678.094 | .000 |
| 5 | 673.219 | 70.399 | 1018.845 | .000 |
| 6 | 311.015 | 36.237 | 494.823 | .000 |
| 7 | 60.228 | 18.145 | 265.867 | .000 |
| 8 | .000 | 3.673 | 52.443 | .000 |
| 9 | .000 | .000 | .000 | .000 |

|  |
| --- |
| a. 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 9。初始中心之间的最小距离为 30167.153。 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **最终聚类中心** | | | | |
|  | 聚类 | | | |
| 1 | 2 | 3 | 4 |
| 语音时长分钟 | 15.444756944212950 | 21.711924727725460 | 131.271358901893960 | 162.766666666666680 |
| 平均带宽Kbps | 9350.103235 | 693.955951 | 743.144180 | 903.449855 |
| 流量4G占比 | .915507680555555 | .927276462352941 | .899816471590909 | .872515000000000 |
| 上网时长分钟 | 1302.719675926157400 | 1664.948970915052500 | 20116.436215277736000 | 178683.097222216700000 |
| 上网流量MB | 605.237290583716600 | 58.771898232110980 | 832.839659344337000 | 2471.613517284393300 |

|  |  |  |
| --- | --- | --- |
| **每个聚类中的个案数目** | | |
| 聚类 | 1 | 72.000 |
| 2 | 2550.000 |
| 3 | 176.000 |
| 4 | 1.000 |
| 有效 | | 2799.000 |
| 缺失 | | 1.000 |

1. 快速聚类(k=4, 迭代10次,并且更改了初始聚类中心)

|  |  |  |
| --- | --- | --- |
| **备注** | | |
| 已创建输出 | | 05-NOV-2018 11:45:49 |
| 注释 | |  |
| 输入 | 活动数据集 | 数据集1 |
| 过滤器 | <无> |
| 权重 | <无> |
| 拆分文件 | <无> |
| 工作数据文件中的行数 | 2800 |
| 缺失值处理 | 对缺失的定义 | 将用户定义的缺失值视为缺失。 |
| 使用的个案数 | 统计基于那些对任何所用聚类变量都没有缺失值的个案。 |
| 语法 | | QUICK CLUSTER 语音时长分钟 平均带宽Kbps 流量4G占比 上网时长分钟 上网流量MB  /MISSING=LISTWISE  /CRITERIA=CLUSTER(4) MXITER(10) CONVERGE(0)  /METHOD=KMEANS(NOUPDATE)  /PRINT INITIAL  /FILE='E:\study\大学\大三\大三上\数据挖掘\实验二\k=4聚类中心.sav'. |
| 资源 | 处理程序时间 | 00:00:00.03 |
| 耗用时间 | 00:00:00.05 |
| 所需工作空间 | 1160 字节 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **初始聚类中心** | | | | |
|  | 聚类 | | | |
| 1 | 2 | 3 | 4 |
| 语音时长分钟 | 10.444756944212950 | 20.711924727725460 | 100.271358901893960 | 150.766666666666680 |
| 平均带宽Kbps | 9000.000000 | 700.000000 | 750.000000 | 900.000000 |
| 流量4G占比 | .990000000000000 | .900000000000000 | .850000000000000 | .800000000000000 |
| 上网时长分钟 | 1200.000000000000000 | 1500.000000000000000 | 20000.000000000000000 | 150000.000000000000000 |
| 上网流量MB | 600.000000000000000 | 500.000000000000000 | 600.000000000000000 | 2000.000000000000000 |

|  |
| --- |
| 从 FILE 子命令输入 |

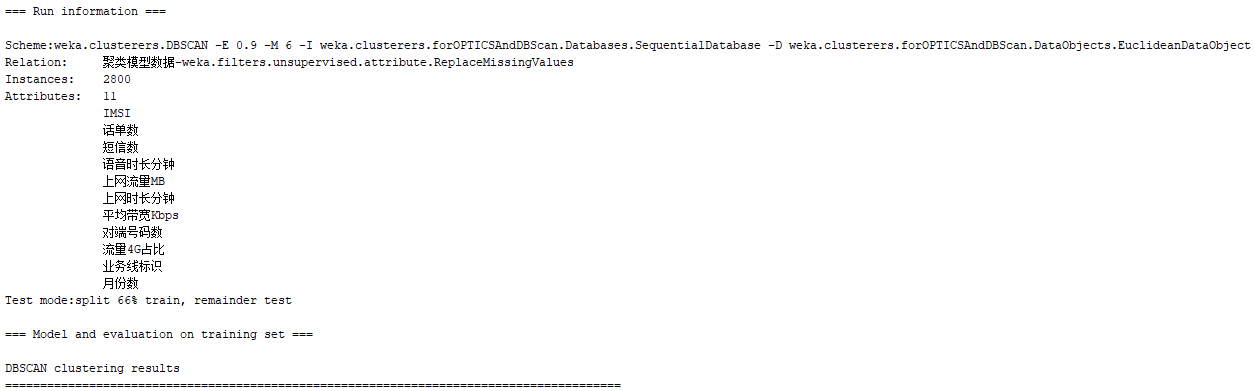
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **迭代历史记录a** | | | | |
| 迭代 | 聚类中心中的变动 | | | |
| 1 | 2 | 3 | 4 |
| 1 | 157.613 | 467.838 | 239.585 | 28686.977 |
| 2 | 52.521 | 1.689 | .000 | .000 |
| 3 | .000 | .000 | .000 | .000 |

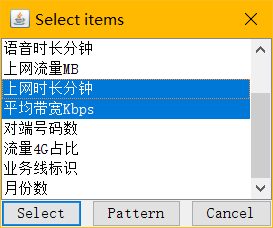
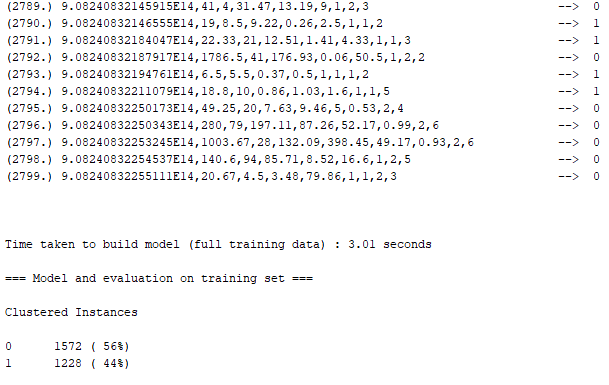
|  |
| --- |
| a. 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 3。初始中心之间的最小距离为 8306.028。 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **最终聚类中心** | | | | |
|  | 聚类 | | | |
| 1 | 2 | 3 | 4 |
| 语音时长分钟 | 16.358048696707804 | 21.677979131139107 | 129.217526234907400 | 162.766666666666680 |
| 平均带宽Kbps | 8860.078297 | 678.660465 | 745.336342 | 903.449855 |
| 流量4G占比 | .914414000000000 | .927583622388648 | .897178038888889 | .872515000000000 |
| 上网时长分钟 | 1316.885301783333000 | 1651.282353939501700 | 19910.763472222123000 | 178683.097222216700000 |
| 上网流量MB | 571.961230342182100 | 57.557568390186795 | 820.404415527184500 | 2471.613517284393300 |

|  |  |  |
| --- | --- | --- |
| **每个聚类中的个案数目** | | |
| 聚类 | 1 | 81.000 |
| 2 | 2537.000 |
| 3 | 180.000 |
| 4 | 1.000 |
| 有效 | | 2799.000 |
| 缺失 | | 1.000 |

1. DBCSCAN的处理结果(需要忽略两个属性才可以聚类，如图所示)





1. 分析

**K值越大，聚类越精细，每个簇的聚合度会更高，就从给定的例子看，就能明显发现从2到3到4的过程中，每一类的聚集程度也就是聚合度越来越高。**

## 实验三

1. 任务说明

实验内容如下：对给定数据集进行分类任务，并建立相应的决策树。分析分类结果指标，比较不同的实验结果，以生成最佳模型。

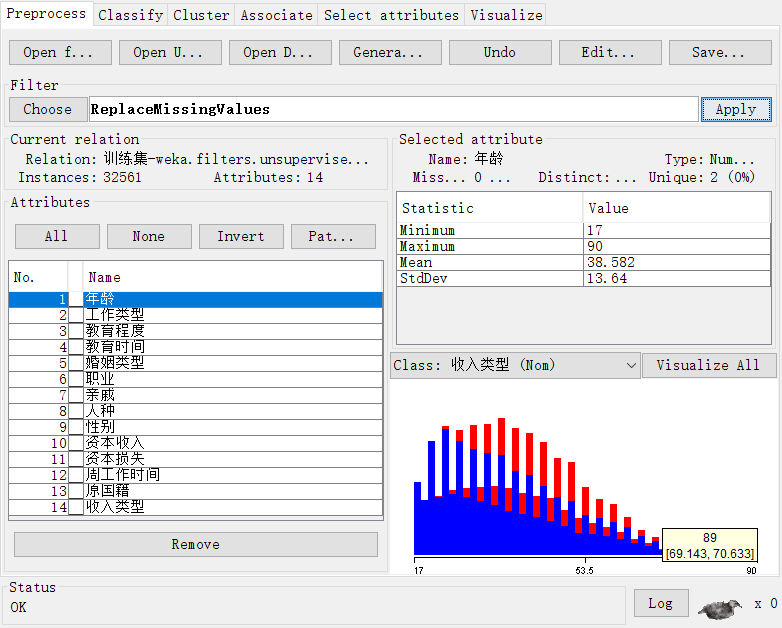
实验任务如下：

1. 任务一
   1. 对训练集和测试集分别进行预处理，使之能够符合ID3算法的要求
   2. 使用ID3算法生成决策树；
   3. 分析模型，考察分类准确率等指标。
2. 任务二
3. 使用C4.5算法，生成决策树；
4. 分析模型，考察分类准确率等指标。
5. 任务三
6. 综合评定以上各模型，生成符合该数据集的最佳模型。
7. 实验结果

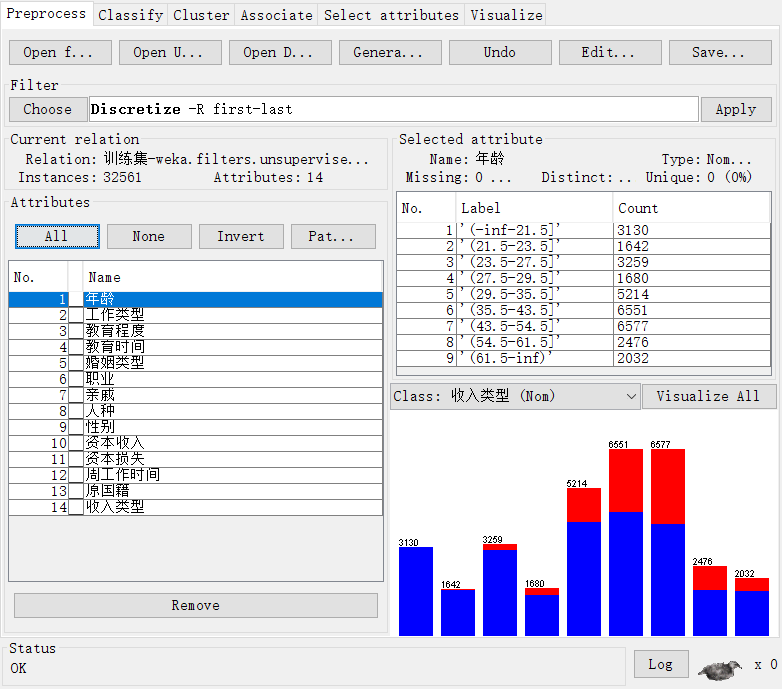
训练集

缺失值的处理 weka ->filters ->unsupervised->attribute->Replace Missing Values

对于数值属性，用平均值代替缺失值，对于nominal属性，用它的mode(出现最多的值)来代替缺失值。

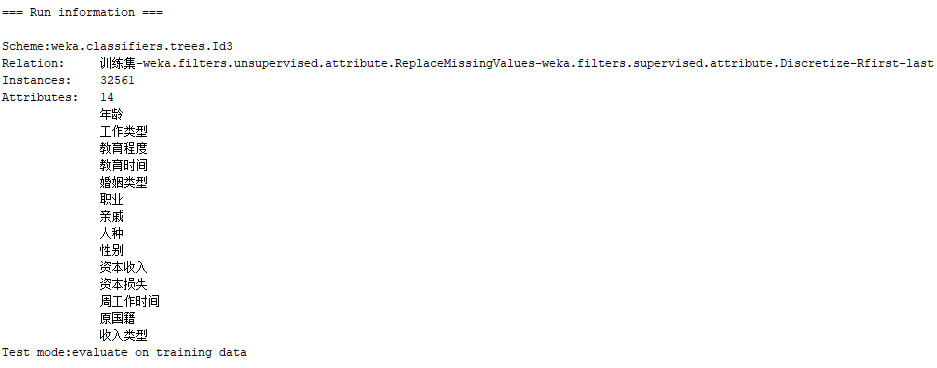


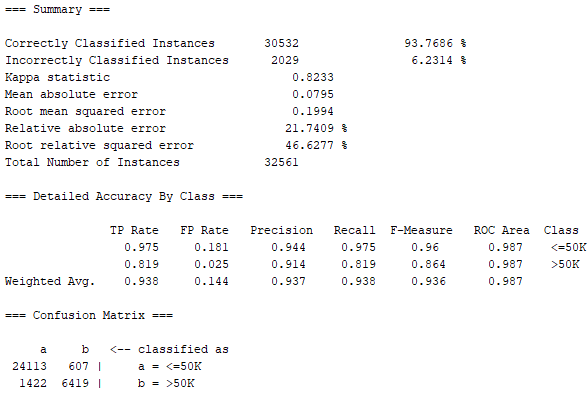
然后进行数据的离散化



然后用训练集进行ID3算法的训练

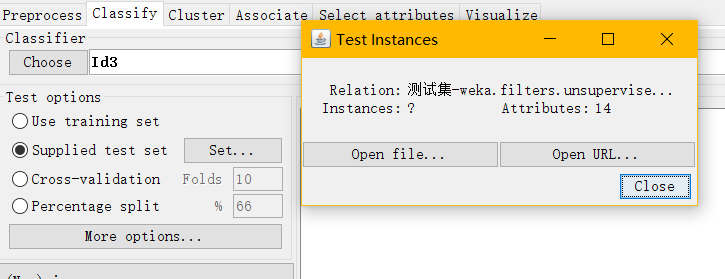
训练日志：

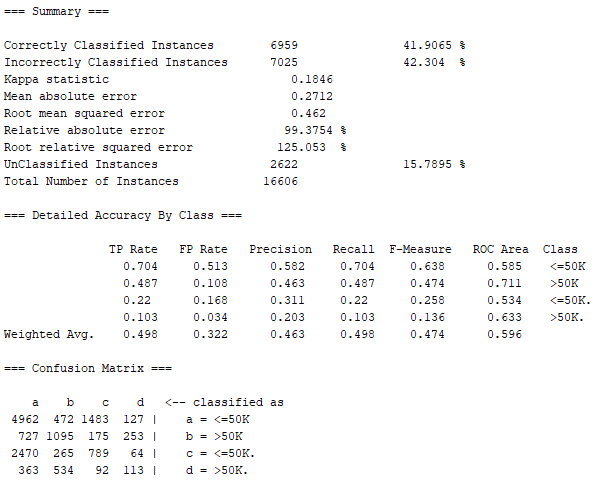




接着队训练集进行填补空缺值以及数据离散化

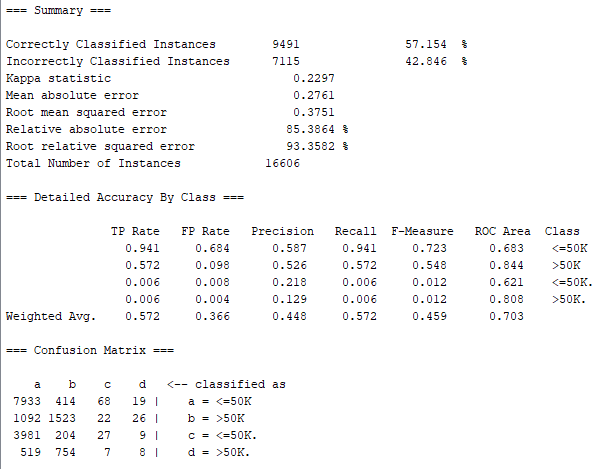
然后将训练集导入用ID3算法将测试集导入进行验证

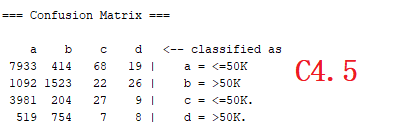
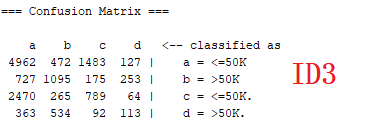




最终训练测试结果如上图

使用C4.5算法结果如下





根据Confusion 矩阵的结果看，C4.5的算法训练出来的模型更好一些，能够比较好的将测试集进行分类。

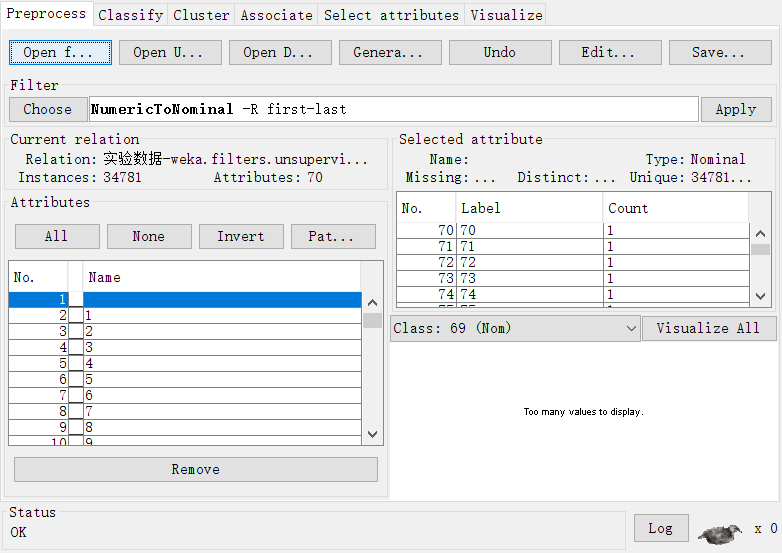
## 实验四

1. 任务说明

完成给定数据集的关联规则挖掘任务，分析所找到的关联规则结果集，改变支持度和可信度阈值，比较不同的实验结果，给出你找到的最终规则集。

1. 实验结果

先对数据进行预处理如下图所示



再使用Apriori算法对数据进行关联规则数据挖掘参数如下



算法结果规则集如下图所示

