**北京邮电大学软件学院**

**2016－2017学年第一学期实验报告**

**课程名称： 数据挖掘**

**项目名称： 分类任务**

**项目完成人：**

**姓名：\_肖逸敏\_\_\_\_\_\_\_学号：\_\_2014211990\_\_\_\_\_\_**

**姓名：\_\_\_\_\_\_\_\_学号：\_\_\_\_\_\_\_\_**

**姓名：\_\_\_\_\_\_\_\_学号：\_\_\_\_\_\_\_\_**

**姓名：\_\_\_\_\_\_\_\_学号：\_\_\_\_\_\_\_\_**

**姓名：\_\_\_\_\_\_\_\_学号：\_\_\_\_\_\_\_\_**

**指导教师：\_\_\_\_牛琨\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**日 期： 2016 年 12 月28 日**

1. **实验目的**

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。

1. **实验内容**

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。

任务二：

1. 使用C4.5算法，生成决策树；
2. 分析模型，考察分类准确率等指标。

任务三：

综合评定以上各模型，生成符合该数据集的最佳模型。

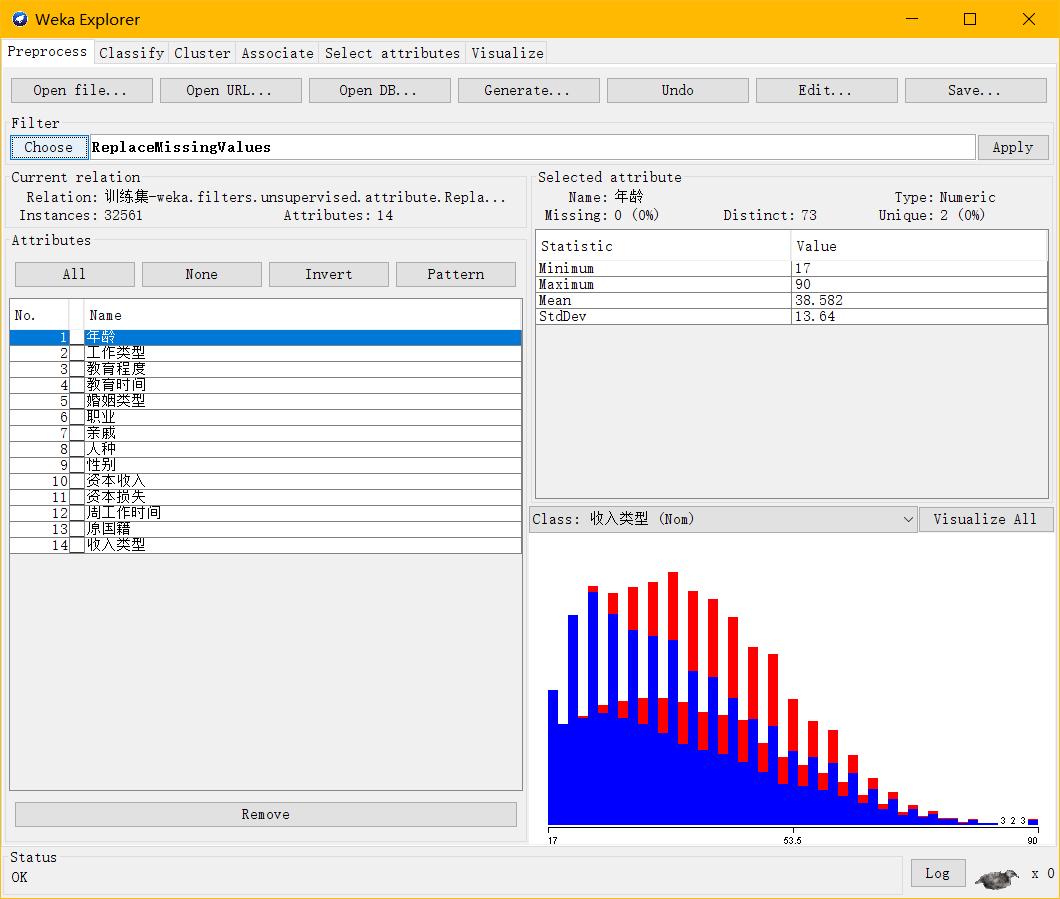
1. **实验环境**

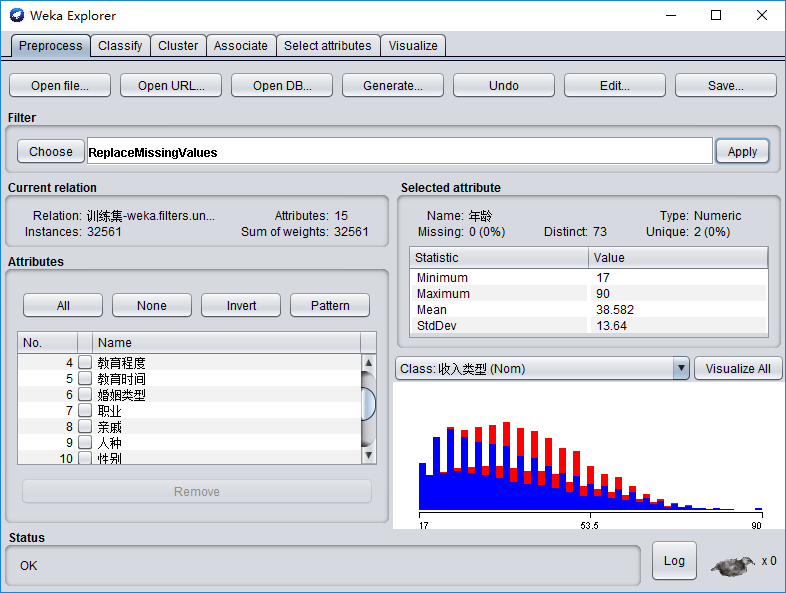
Windows10，weka

1. **实验结果**
2. C4.5算法：

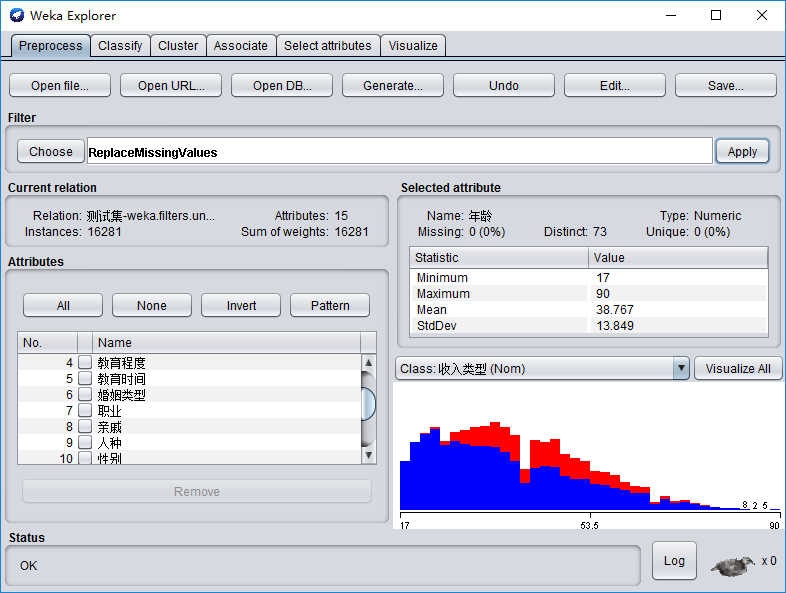
进行数据预处理：

训练集（缺失值处理：weka.filters.unsupervised.attribute.ReplaceMissingValues。 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。）：

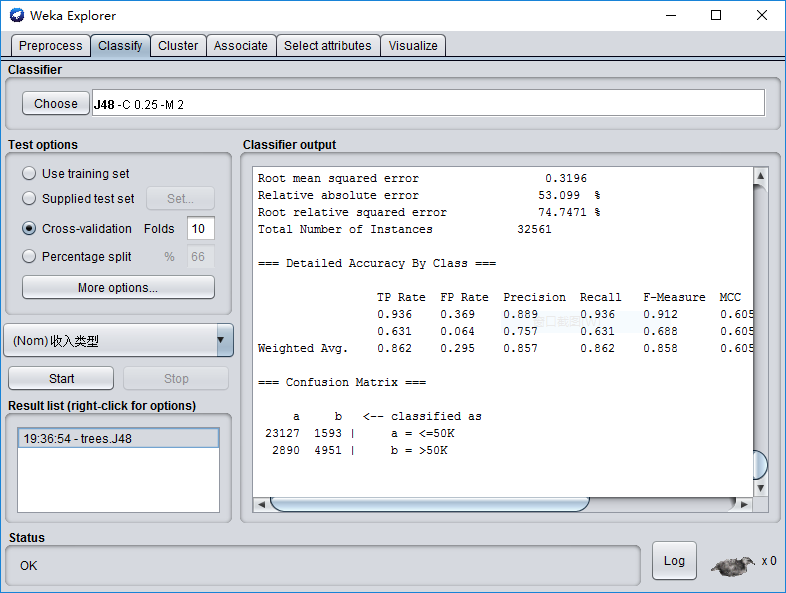




测试集：



用“Explorer”打开数据“训练集.arff”，然后切换到“Classify”。点击“Choose”，选择算法“trees-J48”，再在“Test options”选择“Cross-validation（Flods=10）”，点击“Start”，开始运行。



系统默认trees-J48决策树算法中minNumObj=2，得到如下结果：

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 28078 86.232 %

Incorrectly Classified Instances 4483 13.768 %

Kappa statistic 0.6009

Mean absolute error 0.1942

Root mean squared error 0.3196

Relative absolute error 53.099 %

Root relative squared error 74.7471 %

Total Number of Instances 32561

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.936 0.369 0.889 0.936 0.912 0.605 0.891 0.951 <=50K

0.631 0.064 0.757 0.631 0.688 0.605 0.891 0.761 >50K

Weighted Avg. 0.862 0.295 0.857 0.862 0.858 0.605 0.891 0.905

=== Confusion Matrix ===

a b <-- classified as

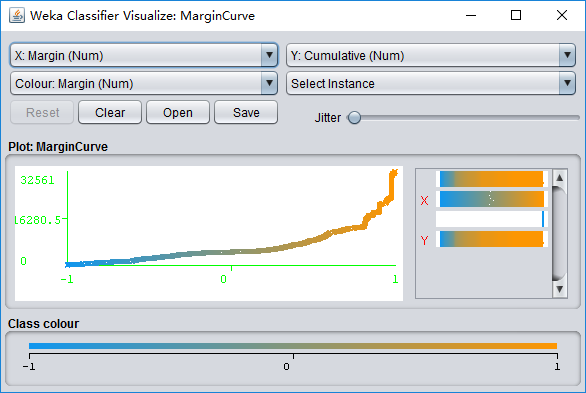
23127 1593 | a = <=50K

2890 4951 | b = >50K

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MinNumObj | 2 | 3 | 4 | 5 |
| Correctly Classified Instances | 28078  86.232% | 28102  86.3057% | 28091  86.2719% | 28091  86.2719% |

由上表可知，当minNumObj=3时，准确率最高。

根据测试数集，利用准确率最高的模型得到的结果：



分析说明：

在用J48对数据集进行分类时采用了10折交叉验证（Folds=10）来选择和评估模型，其中属性值有两个Y，N。一部分结果如下：

Correctly Classified Instances 28078 86.232 %

Incorrectly Classified Instances 4483 13.768 %

=== Confusion Matrix ===

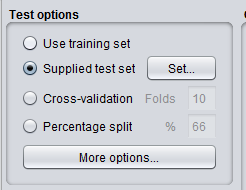
a b <-- classified as

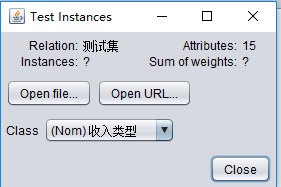
23127 1593 | a = <=50K

2890 4951 | b = >50K

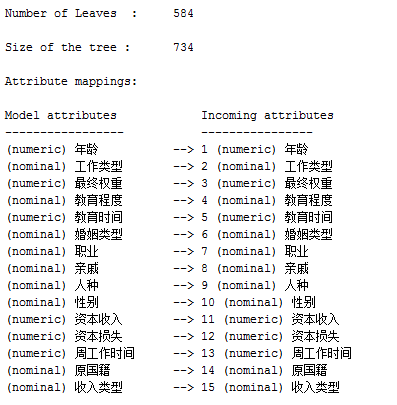
这个矩阵是说，原来是“Y”的实例，有23127个被正确的预测为“Y”,有1593个错误的预测成了“N”。原本是“NO”的实例有2890个被正确的预测成为“Y”，有4951个正确的预测成了“N”。“23127+1593+2890+4951=32561”是实例的总数，而（23127+4951）/32561=0.86231正好是正确分类的实例所占比例。这个矩阵对角线上的数字越大，说明预测得越好

将测试集插入：





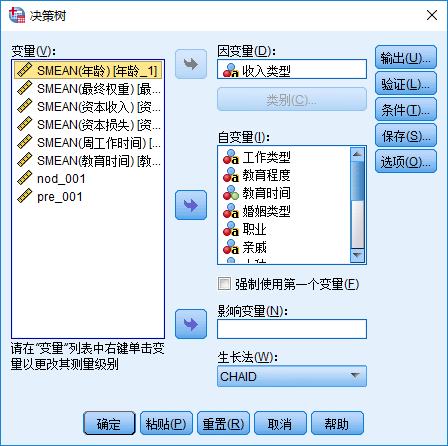
会输出：



1. ID3算法：使用Spss进行

训练集：

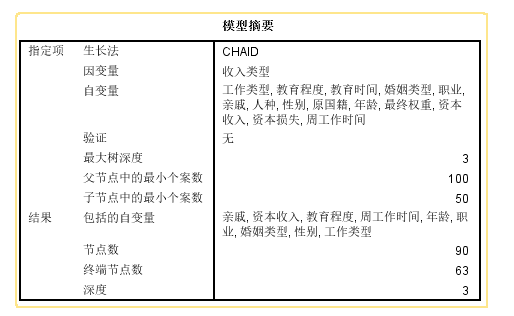
导入训练集.xls，打开数据后选择分析---分类---树，选择转化——替换缺失值，进行数据预处理。在决策树框中，将你待预测的量放到因变量中，将影响因变量的量放到自变量中。

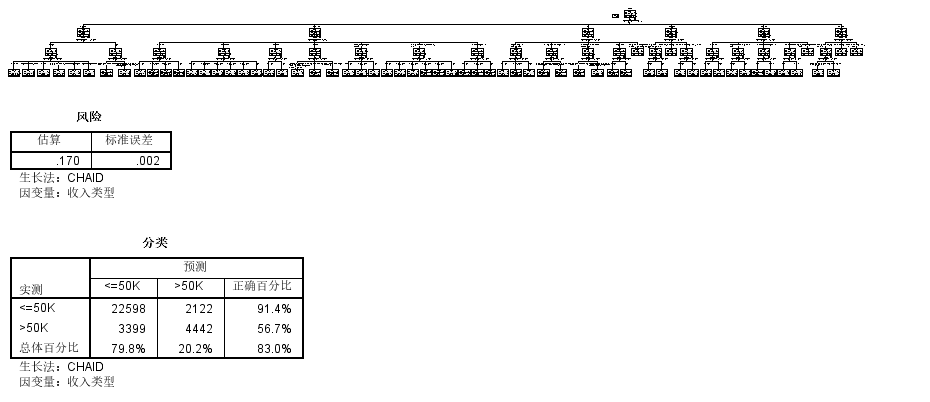


点击输出，可以出现如下图，选择规则的位置，方便以后使用：



确定以后可以出现决策树：





打开测试集，新建——语法：



运行——全部：

