**报告**

**学号: 姓名: 班级:**

1. **目标**

练习如何构建决策树。认识归一化和离散化对构建决策树的影响。

1. **Data**
2. Bank-all.arff是银行的所有数据。当我们不拆分数据的时候，我们可以用10-crossvalidation来测试分类器的准确性。数据的最后一个属性是类标签。
3. Bank-tain.arff is used for constructing the model.

Bank-test.arff is used for testing the model.

The last attribute is the class label.

3）**weather-nominal.arff**

1. **Contents**

两个实验：

1. **Bank-all.arff**
2. 预处理，删除无用属性，保存到新的数据文件。
3. 选择两种方法对数据进行归一化，保存到新的数据文件中。并列出归一化的结果。
4. 选择两种方法对数据进行离散，保存到新的数据文件中。并列出离散化的结果。
5. 利用银行原始数据，用J48构建决策树。选择10-crossvalidation。比较J48与binary split或multiple split的结果。分析 "minNumObj "参数的影响（选择minNumObj=2或1）。
6. 利用归一化数据用J48构建决策树。选择10-crossvalidation。比较J48与binary split或multiple split的结果。分析 "minNumObj "参数的影响（选择minNumObj=2或1）。
7. 利用离散化数据，用ID3构建决策树。比较ID3与binary split或multiple split的结果。分析 "minNumObj "参数的影响（选择minNumObj=2或1）。
8. 对比J48 和ID3的结果。

**2．用归一化数据和离散化数据生成训练（400个对象）和测试（200个对象）文件。使用训练数据来训练模型，使用测试数据来测试模型。**

1. 对于归一化数据，比较J48中binary split或multiple split的结果。分析 "minNumObj "参数的影响（选择minNumObj=2或1）。
2. 对于离散数据，比较ID3与binary split或multiple split的结果。分析 "minNumObj "参数的影响（选择minNumObj=2或1）。

利用混淆矩阵计算准确率、错误率、精确率和召回率。写下计算的过程。用一些可视化的结果来展示你的结果。

3. **Data: weather-nominal.arff, which is included in the path of weka.**

1) use weka with ID3 to construct a tree.

2) construct a tree manually

3) compare the upper two methods.