##### QG工作室数据挖掘小组实验报告

|  |  |  |
| --- | --- | --- |
| 实习生： 李炜乐 | 导师：张润鹏 | 日期：2019 年 7 月 20日 |

|  |
| --- |
| **实验名称：C4.5决策树算法** |
| **已完成内容：**   1. **C4.5决策树算法的编写** 2. **C4.5决策树的绘图** |
| **未完成内容：暂无** |
| **未完成原因：暂无** |
| **需要帮助：暂无** |

|  |  |
| --- | --- |
| **实验总结** | |
| **知识点总结：**   1. **C4.5决策树包括计算信息增益率、选取最佳分支点，递归构建决策树三个步骤。其中停止分裂的条件是：（1）当前节点包含的样本完全属于同一个类别，无需划分（2）当前属性集为空，或是所有样本在所有属性上相同，无法划分。（3）当前节点包含的样本集合为空，不能划分。** 2. **C4.5决策树的工作原理是通过信息增益公式计算每一个信息增益，选取信息增益率最大的特征作为分支点，通过递归来实现决策树的构建。** 3. **数据集的环境为lenses数据集，因为数据集太少，在选取训练集和测试集的时候，有可能会导致测试集的数据预测不到结果。但是相比较于ID3决策树总的正确率大概在75%-100%之间，运行时间为0.000784964s** | |
| **遇到问题**：跟ID3一样，由于数据集比较少，测试集的数目不能过多，过多的测试集会导致训练集训练出来的数目不一样 | **解决过程：全部数据作为训练集或者去少量数据作为测试集** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **导师评价** | | | | |
| **实验分数** | **知识掌握情况** | **代码编写能力** | **建议** | **评价日期** |
|  |  |  |  |  |

