**决策树应用说明**

**应用名称：**决策树

**应用ID：**ClassificationAndRegressionTree

**应用功能：**根据给定数据集进行构建四种决策树模型。

**应用作者：Baiyaer**

**应用输入（参数）：**

|  |  |  |
| --- | --- | --- |
| 应用输入（参数） | 参数类型 | 参数说明 |
| Author.txt | 文件上传 | 用户输入的作者姓名，用于个性化 |
| Title.txt | 文件上传 | 用户输入的绘图标题，您可以说明您的数据集名称 |
| Height.txt | 文件上传 | 生成PDF的尺寸大小，高度参数，为了保证决策树正常输出，您应该尽量输入大一点的图片尺寸，源程序已将图片默认放大两倍。 |
| Width.txt | 文件上传 | 生成PDF的尺寸大小，宽度参数，为了保证决策树正常输出，您应该尽量输入大一点的图片尺寸，源程序已将图片默认放大两倍。 |
| DataInput.txt | 文件上传 | 用户输入的需要进行构树的数据集，demo文件为糖尿病数据集，每一行是一个sample/患者（无行名，第一行顶格）每一列是一个属性/变量名称（需要分类的变量如身高体重等影响因，最后一列列名为Tag），而程序内自动以8：2划分训练集与测试集 |

**应用输出（结果）：**

|  |  |
| --- | --- |
| 结果文件 | 结果说明 |
| .\Result\CART\\* | CART算法结果 |
| .\Result\CART\TestSetMixMatrix.txt | 使用测试集通过模型得到的混淆矩阵 |
| .\Result\CART\TrainSetMixMatrix.txt | 使用训练集通过模型得到的混淆矩阵 |
| .\Result\CART\TestSetMixMatrixFixtree.txt | 使用测试集通过修剪后的模型得到的混淆矩阵 |
| .\Result\CART\TrainSetMixMatrixFixtree.txt | 使用训练集通过修剪后的模型得到的混淆矩阵 |
| .\Result\CART\PDF\Cart.pdf | 原始CART树 |
| .\Result\CART\PDF\TestRoc.pdf | 原始CART树对于测试集的ROC图 |
| .\Result\CART\PDF\CartFixtree.pdf | 修剪后的Cart树 |
| .\Result\CART\PDF\TestFixtreeRoc.pdf | 修剪后的CART树对于测试集的ROC图 |
|  |  |
|  |  |

**应用场景：**

**临床诊断或是分类人群**

**应用说明：**

主要是临床信息分析进行结果筛选，通过不同变量构建四种决策树分别是CART、ID3、C4.5、C5.0。

决策树简单说明：

决策树(decision tree)是一类常见的机器学习方法。以二分类任务为例，我们希望从给定训练数据集学得一个模型用以对新示例进行分类，这个把样本分类的任务，可看做对“当前样本属于正常吗？”这个问题的‘决策’或者‘判定’过程。顾名思义，决策树是基于树结构来进行决策的，这恰是人类在面临决策问题时一种很自然的处理机制。

常用的决策树算法：

ID3 以信息增益作为分类标准

CART 以基尼系数作为分类标准

算法的具体理论可以参考周志华的《机器学习》

剪枝（pruning）是决策树学习算法对付“过拟合”的主要手段，在决策树学习中，为了尽可能正确分类训练样本，结合划分过程将不断重复，有事会造成决策树分支过多，这时就可能因训练样本学得“太好了”，以致于把训练集自身的一些特点当做所有数据都具有的一般性质而导致过拟合。因此，可通过主动去掉一些分支来降低过拟合的风险。cp是参数复杂度（complexity parameter）作为控制树规模的惩罚因子，简而言之，就是cp越大，树分裂规模（nsplit）越小。输出参数（rel error）指示了当前分类模型树与空树之间的平均偏差比值。xerror为交叉验证误差，xstd为交叉验证误差的标准差。

**参考文献：**