分类模型的实现和评价

沈天奥，侯晨杨，刘韦

# 摘要

# 1.引言

# 2. 算法流程

**2.1 数据预处理**

将本课题给出的两个原始数据文件合并为一个DataFrame类型的数据集，并检测数据集中的缺失值和离散值。

对于数据集的缺失值可采取前向填充，后向填充，平均值填充以及直接剔除等方法，对于是否存在离散值可采取3σ准则，局部离群值因子，箱型图等方法进行判断，并根据实际情况判断是否需要处理离散值。

处理完成缺失值和离散值后将数据集中的27个特征属性和7个分类属性拆分为输入集和输出集，并将输入集和输出集以9:1的比例分别划出训练集和测试集。同时，考虑到本课题为分类问题，因此测试集和训练集中各个种类的数据的比例应与原数据集保持一致，这样才能保证后续模型的训练准确率。

**2.2 传统机器学习算法介绍**

传统机器学习在处理分类问题上包含了多元Logistic回归、支持向量机、决策树、朴素贝叶斯、随机森林、多层感知机等算法，考虑到本课题是一个多分类问题且数据量并不庞大，因此我们拟采用决策树和支持向量机

**2.2.1 决策树**

决策树分类算法是利用数据集生成一棵在特征空间上得条件概率分布树，是一系列关联规则的集合，侧重于数据类别之间的差异，常用于对数据集进行分类。在生成决策树时，决策树分支节点所包含的样本应尽可能属于统一类别，通常需要一定指标来表示特征的分类能力，信息熵就是最常用的一种指标，用来表示随机变量不确定性的度量。典型决策树分类算法包括ID3、C4.5、CART等，分别采取不同的标准划分属性，其中ID3使用信息增益对数据集进行划分，C4.5在前者的基础上采取增益率，而CART算法则采取基尼系数作为度量标准，本课题采取CART算法，在构建决策树时计算Gini系数划分属性。[1]

决策树在进行训练时常会出现过拟合现象，即过度依赖于训练集数据特征而导致拟合测试数据时泛化效果很差，因此通常需要对决策树进行剪枝操作。决策树的剪枝策略可分为预剪枝和后剪枝，预剪枝常采用限定深度、设定阈值与设置指标比较泛化能力等方法，而后剪枝则主要分为降低错误剪枝(REP)、悲观错误剪枝(PEP)、最小误差剪枝(MEP)、基于错误的剪枝(EBP)和代价-复杂度剪枝(CCP)五个方法。由于本课题的决策树采用CART算法，因此后剪枝方法采用相对于CART算法效果最好的代价-复杂度剪枝，预剪枝采用限定深度的方法。

**2.2.2 支持向量机**

# 3 课题结果与分析

**3.1 数据预处理**

将给定的两个原始数据集进行合并后得到的数据集如下：

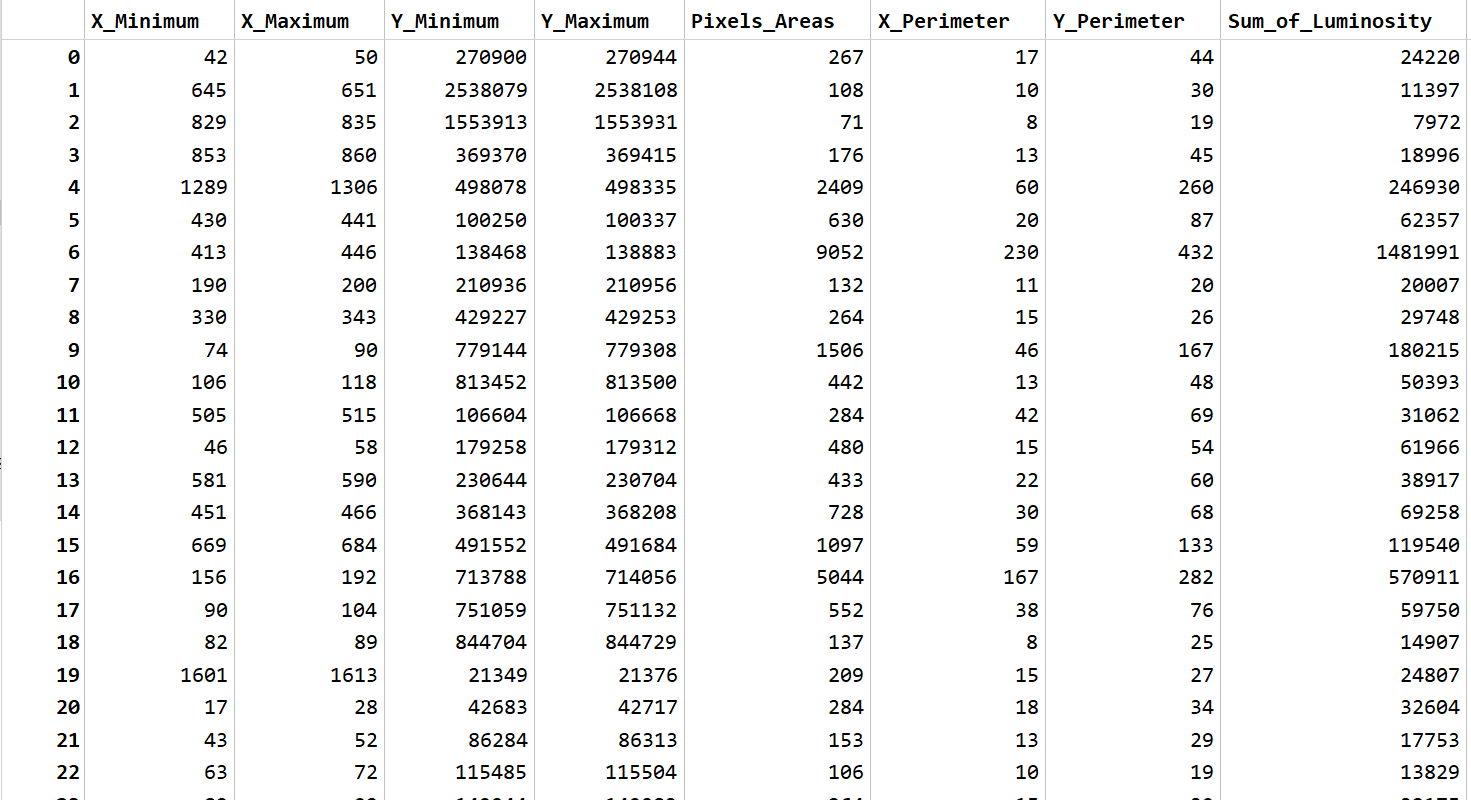


图1 合并后的数据集(部分)

在对数据集进行分析后可知，该数据集中不存在缺失值。

通过局部离群值因子法和箱型图法可知，该数据集中的数据存在含有离群值的现象，其中局部离群值因子法计算得出该数据集中共有128个样本点数据存在离群值，但分析题目可知，上述离群值均是实时测得，且部分离散值较为严重的数据可能代表了缺陷类别分类的关键信息，因此不应被修改，也不应删除。同时考虑到该课题实际情况，基本考虑官方给出的数据可信，不考虑异常。

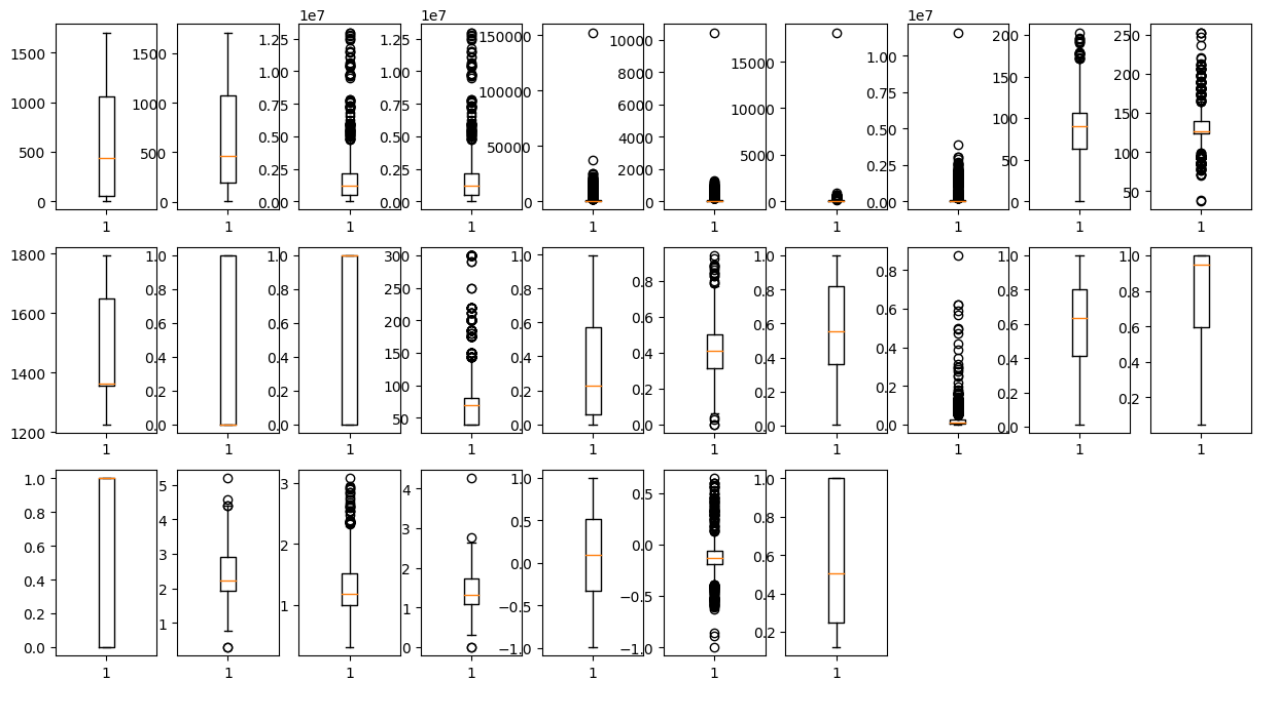


图2 各特征属性的箱型图

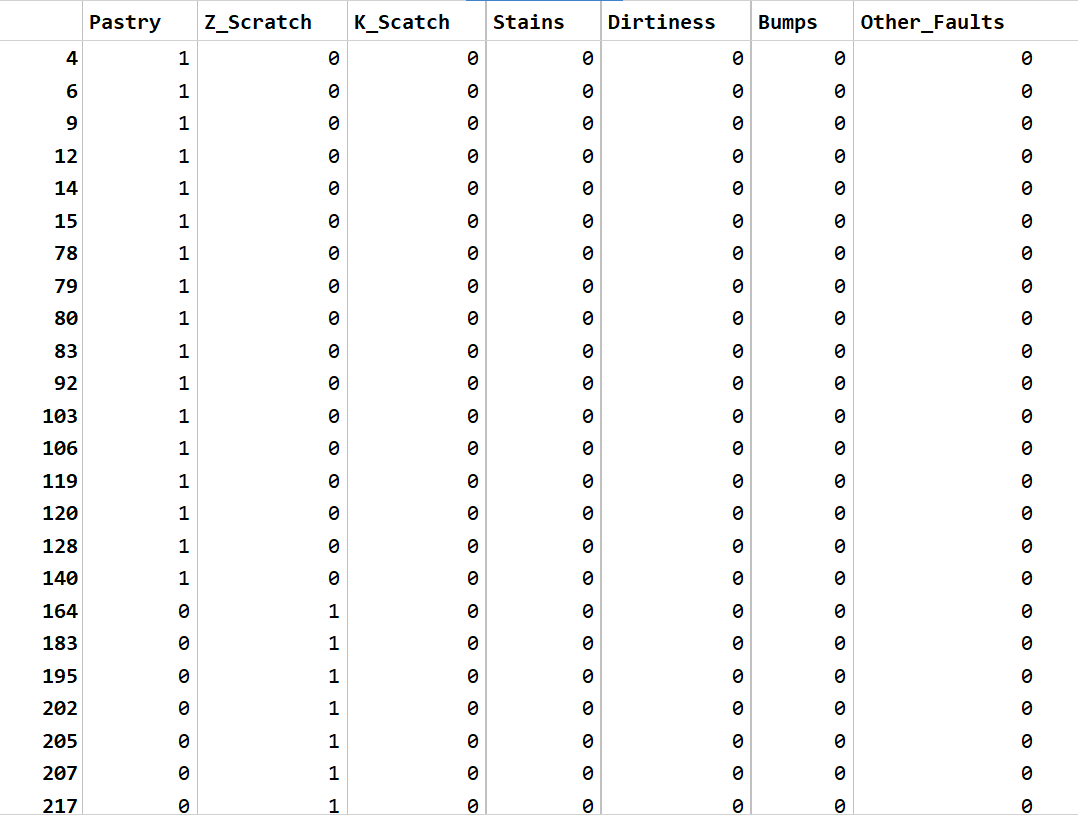


图3 存在离群值数据的样本点

**3.2 决策树模型拟合结果分析**

在不进行任何剪枝操作时，决策树的构建CPU用时为31.2ms，系统时间的时间戳为32.9ms，可以看出整体模型的构建速度较快。决策树的可视化图形和数据测试结果如图4和图5所示：

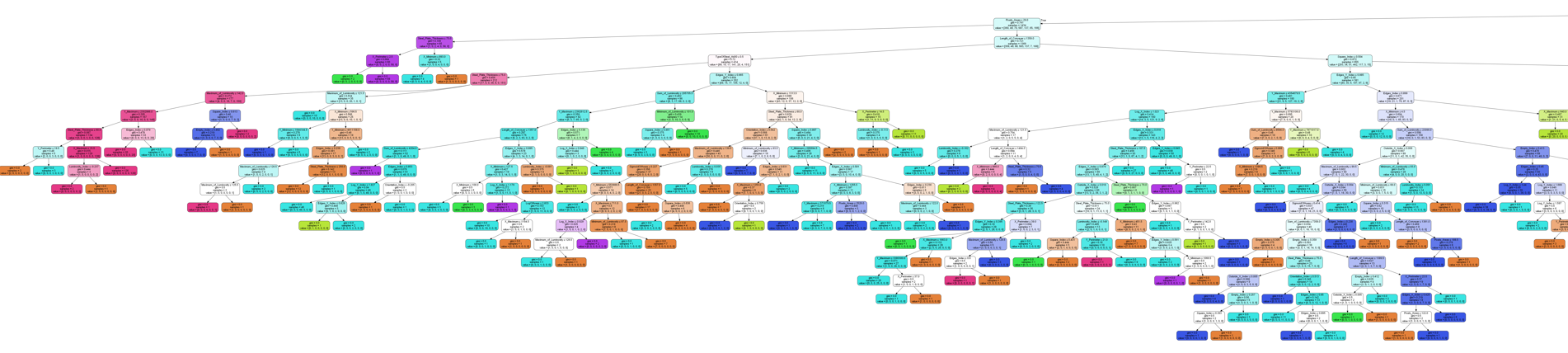


图4 决策树可视化图形(部分)



图5 决策树预测结果

可以看出，决策树在未进行剪枝时整体树状结构十分复杂，其树的深度达到了19层，并且根据图5结果可知决策树对Pastry、Bumps、Other\_Faults三个类别的数据预测结果较差，其整体数据的预测准确率为0.65。

下面对决策树进行剪枝操作，图6为从预剪枝和后剪枝两个维度对决策树剪枝的结果：

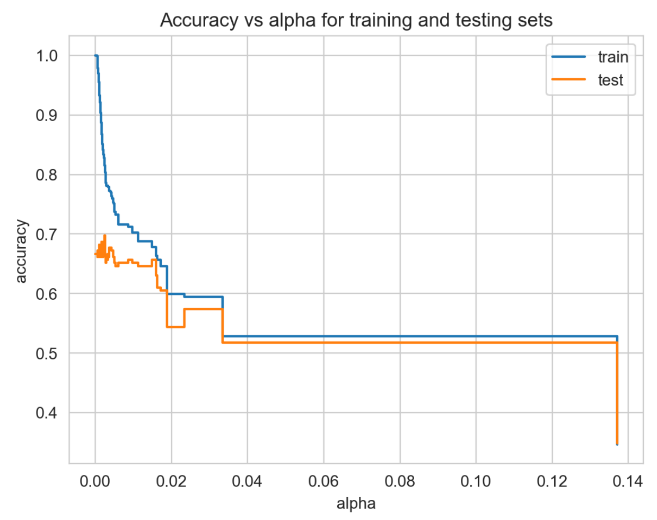
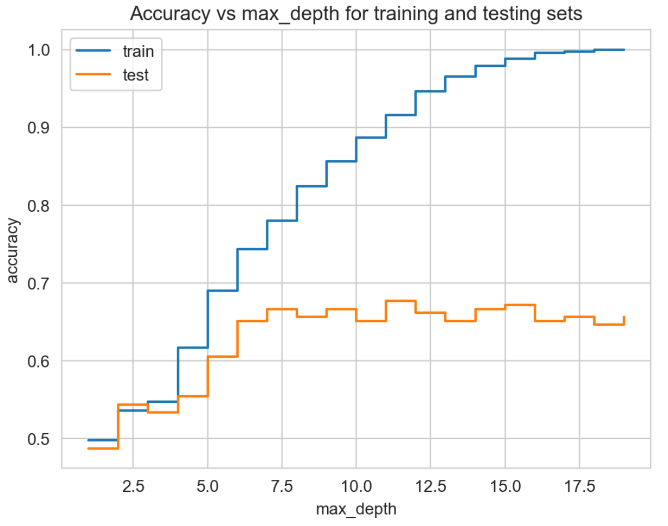


图6 决策树剪枝后的准确率比较

根据上图可知，预剪枝通过限定深度的方式和后剪枝的CCP策略得出的拟合准确率在训练集中均呈现下降趋势，在测试集中也没有得到明显的提升，即该课题中的数据利用决策树拟合并未出现过拟合现象，剪枝操作只能在一定程度上解决决策树模型结构过于复杂的问题，但无法提升模型的精度。

# 4 实验总结

# 5 实验体会

# 6 致谢

内容包括：借鉴其他组的小组报告内容、获得的帮助、参考的非文献资料。

# 7 参考文献

[1]范劭博,张中杰,黄健.决策树剪枝加强的关联规则分类方法[J/OL].计算机工程与应用:1-10[2022-11-19]. http://kns.cnki.net/kcms/detail/11.2127.TP.20220916.1612.014.html

这是报告模板和提交要求间的分割线

**1.实验报告书写规范。**

1) 全文（包括所有的章节题目）的汉字字体为宋体，章节序号、所有字母与数字的字体为Times New Roman。一级标题字号为三号加粗；二级标题四号加粗；三级标题小四号加粗；正文小四号。

2) 页面设置格式：行距：1.25倍；页码：居中；边距：上下左右各空2cm。

3) 公式应居中书写，公式的编号用圆括号括起放在公式右边行末，字号为五号。

4) 所有表格要求三线表，每个表格应有表序和表题，表序和表题应写在表格上方正中，表序后空一格书写表题，表题、内容的字号均为五号。

5) 凡有引用他人成果之处，均应按实验中所出现的先后次序列于参考文献中，并且只应列出引言中以标注形式引用或参考的有关著作和论文。引文的标注应在一段引文后的右上角，用小方括号中填写数字表示，如：“Buck变换器是单管不隔离型DC-DC变换器中的一种基本结构[8]”，并与参考文献中的序列号相对应。

**2.实验报告提交要求。**

1) 每个组只要提交一份实验报告和一份实验代码。

2) 实验报告和实验代码打包成一个压缩文件，主文件名设置为：实验2实验报告\_组号\_小组所有成员姓名，将该压缩包用为邮件附件发送到dareports\_jnuiot@163.com，邮件的主题设置为实验2实验报告\_组号。

3) 实验报告提交截止时间：2022-10-17 23:59。