分类模型的实现和评价

沈天奥，侯晨杨，刘韦

# 摘要

# 1.引言

# 2. 实验步骤

**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。

# 3 实验验证

**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：

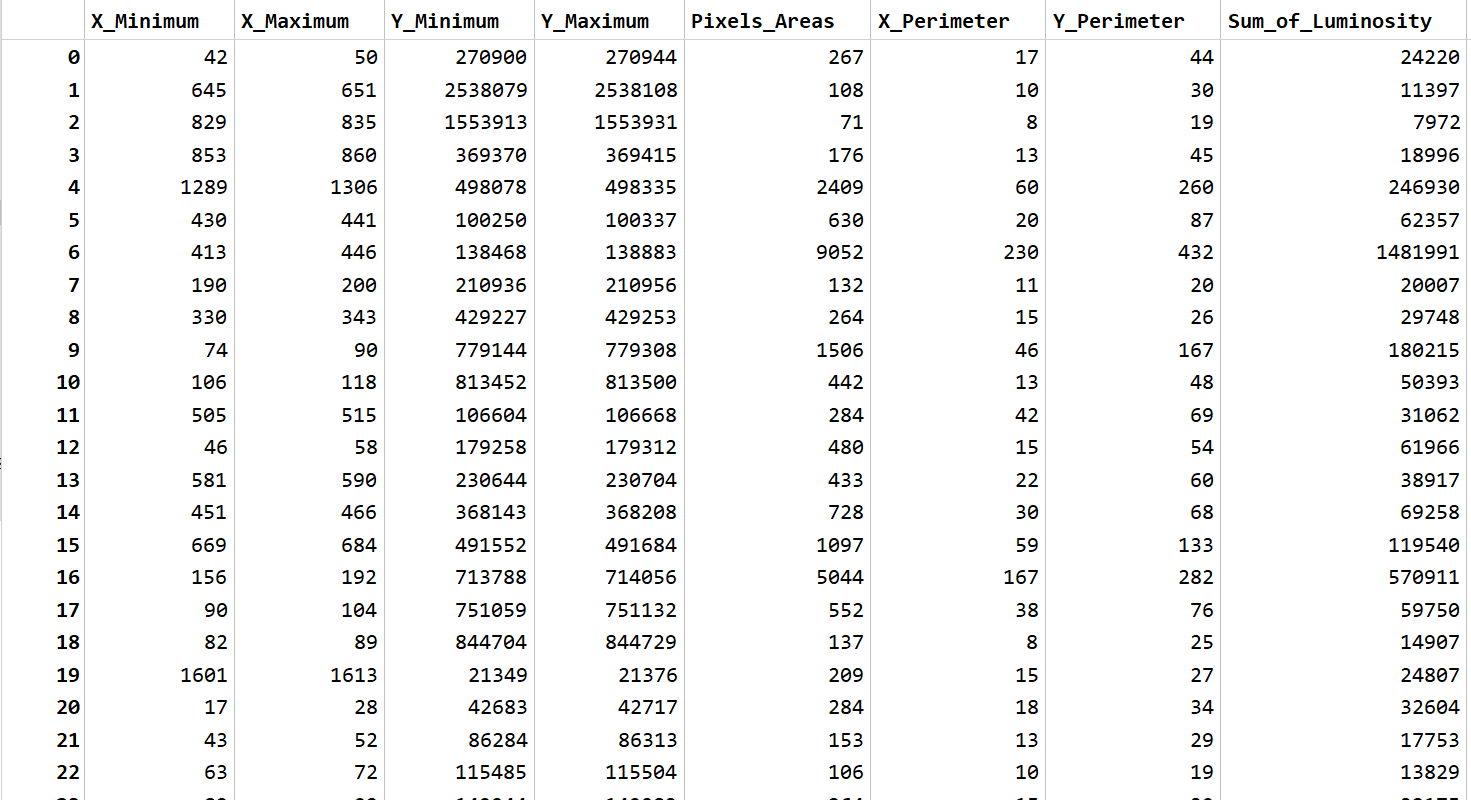


图1 合并后的数据集(部分)

在对数据集进行分析后可知，该数据集中不存在缺失值：

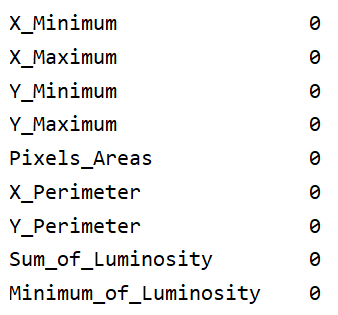


图2 各特征列的缺失值数量(部分)

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。

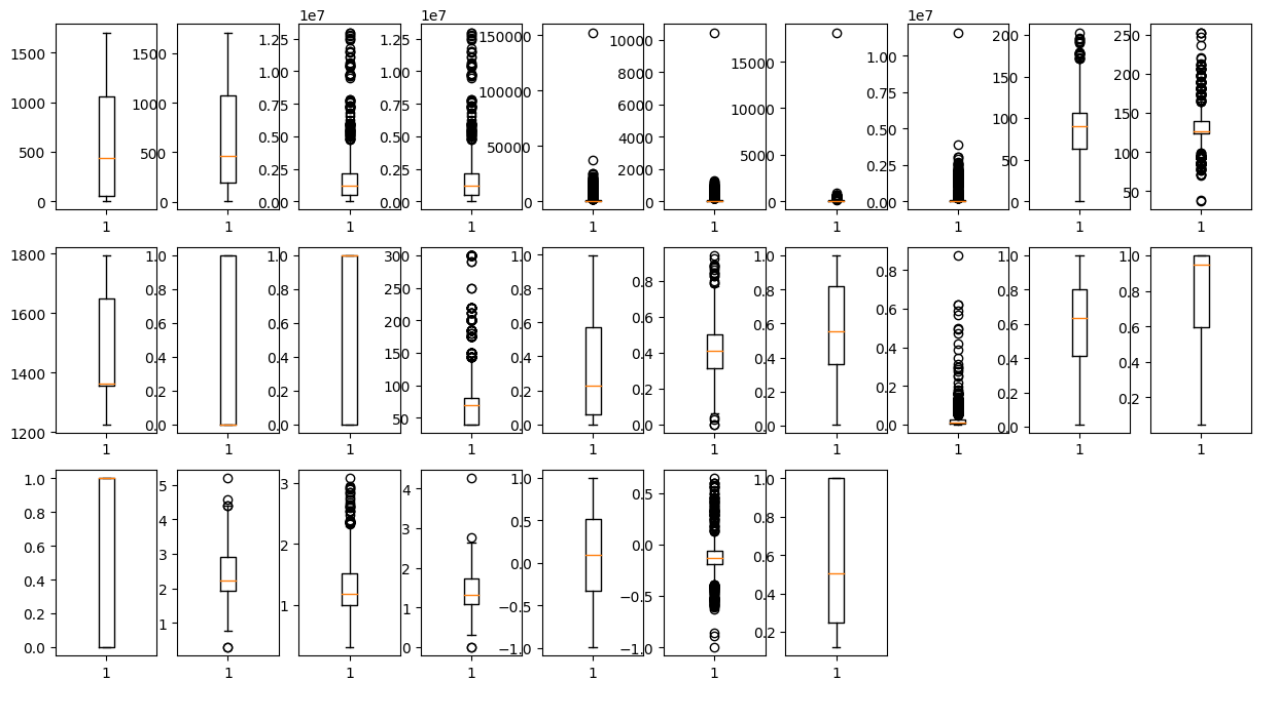


图3 各特征属性的箱型图

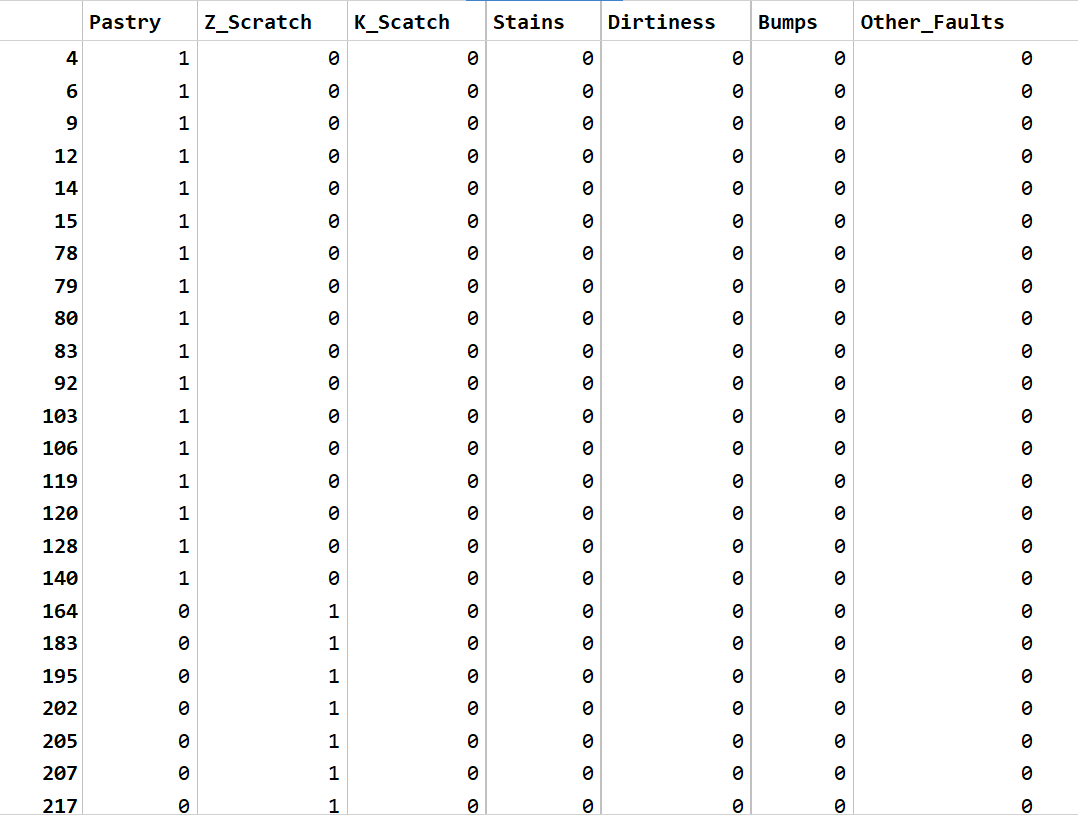


图4 存在离群值数据的样本点

# 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

-------------------------------------------------------------------------------------------

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。