Comprehensive Experiment 实验报告

PB20010429 侯相龙

2023 年 1 月 20 日

1 实验内容

通过数据预处理、数据划分、模型训练、模型验证等步骤，完整地处理机器学习实验任务

2 实验原理

决策树模型

3 实验步骤

3.1 数据预处理

3.1.1 数据读取

从文件"train_feature.csv" 和"train_label.csv" 读取带标签的训练集

3.1.2 清除离群数据

注意到本实验包含了大量的离群数据，且数据的分布无明显规律性，异常值偏离较明显。下面类似于四分位法处理异常值。即：首先取数据集各特征第 10% 的点 M 和第 90% 的点 L，只保留范围在 $M-5(L-M)\sim L+5(L-M)$ 的值，其余的点赋空值。（每个特征处理的异常值点约占 0.5%）

3.1.3 单位化

$x=\frac{x-min}{max-min}$。其中 max，min 分别为每个特征的最大、最小值。这样即可将各特征约化到 [0,1] 区间。

3.1.4 空值填充

利用 knn 近邻法，将空值填充为其近邻的属性值。

3.1.5 特征提取

本实验的特征提取会极大影响最终的实验结果，故采取了两种特征选择方式：其一 ReliefF 方法过滤式选择，其二，通过 XGBoost 树模型选择。事实上，此两种方法似乎都并不能显著提升分类准确率。在综合效果、成本等方面考虑后，采取了如下基于 ReliefF 方法的特征选择方法：因为数据量大（10000 组），我们多次少量随机采样计算最近邻以及属性距离等相关参数，最后选取在多次采样评估均较优的一组属性。

下表表明了特征选择和分类正确率的关系。

实验方法	有特征选择	无特征选择
SVM	0.256	0.261
XGBoost	0.253	0.245
决策树	0.268	0.244
对率回归	0.260	0.253

3.2 数据划分

按 7：3 的比例随机分为训练集、测试集。在选择参数时使用了 k-折交叉验证法（k=5）。

3.3 模型训练

在选择合适参数后，测试线性回归模型、对率回归、决策树模型、神经网络模型、支持向量机以及 XGBoost，利用 k-折交叉验证的平均正确率均能达到 25% 以上。但是，正确率都并不能达到 30% 以上。下面是部分实验结果汇总。

实验方法	SVM	XGBoost	决策树	对率回归
正确率	0.256	0.253	0.268	0.260

我们选择了其中准确率最高的模型：决策树。下面将基于此模型介绍调参

3.4 参数调试

我们采用 sklearn 库中网格搜索和交叉验证的方法，进行了参数选择。

```
'max_depth': np.arange(1, 10, 1),
'random_state': np.arange(1, 50, 10),
'min_samples_leaf': np.arange(1, 100, 10),
'min_samples_split': np.arange(1, 50, 10)
```

最终选择出最优的参数组合，使得正确率为 26.8%

3.5 模型评估

模型在验证集上的表现不尽如人意，准确率（Acc）为 26.8%，利用卡方检验可知，模型预测的准确率也并不高。究其本因，特征选择未能选择合适的方法。

4 实验结果

见附件 test_label.csv

5 实验分析与总结

本次实验体现了机器学习中完整的数据处理、模型训练与预测的过程。从这次实验也可以看出，训练出良好的预测模型并不是那么简单。其中特征选取，调参都是极为关键的步骤，它们也会极大影响实验的结果。而本次实验，也正是由于没有选择合适的方法进行特征选取，导致了准确率并不高。