

# 燃气轮机气路故障诊断-项目报告

## 数据处理过程

手工去除两行 `BAD` 的数据后，经过一定的思考，我认为这次这些数据除了去除 `BAD` 数据，其余不需要什么操作，并且也是一个较为简单的多次二分类过程。最后我选择了逻辑回归（Logistic Regression）来编写这一个程序。为了方便每一步的检测与查看，我使用了 `colab` 来编写程序，最后完成了一个 `ipynb` 的文件。图标在 `Plot` 中有体现。

## 模型设计思路

- 

```
# drop 'type'
X_train = training_data.iloc[:, :-1]
X_validate = validating_data.iloc[:, :-1]
X_test = testing_data.iloc[:, :-1]
# use 'type'
y_train = training_data.iloc[:, -1]
y_validate = validating_data.iloc[:, -1]
print(X_train)
print(y_train)
X_train.shape, y_train.shape

X_train = np.array(X_train)
X_validate = np.array(X_validate)
X_test = np.array(X_test)
y_train = np.array(y_train)
y_validate = np.array(y_validate)
smo = SMOTE(random_state=43)
X_train, y_train = smo.fit_sample(X_train, y_train)
```

使用这两段代码将 `train` 和 `validate` 里的数据提取出来。并使用 `SMOTE` 函数将过拟合部分去除。

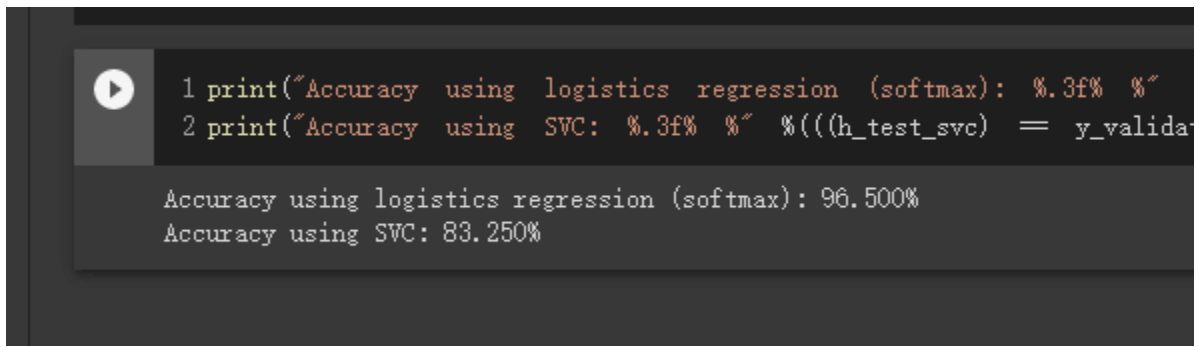
`SMOTE` 全称是 Synthetic Minority Oversampling Technique 即合成少数类过采样技术，它是基于随机过采样算法的一种改进方案，由于随机过采样采取简单复制样本的策略来增加少数类样本，这样容易产生模型过拟合的问题，即使得模型学习到的信息过于特别 (Specific) 而不够泛化 (General)，`SMOTE` 算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中。虽然该算法可以较好的处理过拟合的问题，但是该算法无法克服非平衡数据集的数据分布问题，容易产生分布边缘化问题。

- 

```
model = LogisticRegression(multi_class="multinomial", solver="lbfgs", C=10,
max_iter=500)
model.fit(X_train, y_train)
```

使用简单多元 `Logistic` 回归将数据进行拟合。有六种类别，那么使用多次回归进行迭代即可。

## 模型预测结果



```
1 print("Accuracy using logistics regression (softmax): %.3f%%" % accuracy_score(y_test, y_pred))
2 print("Accuracy using SVC: %.3f%%" % accuracy_score(y_test, y_pred))

Accuracy using logistics regression (softmax): 96.500%
Accuracy using SVC: 83.250%
```

可以看出，精确度较好。

最后将 `array[]` 写入 `Dataframe` 再导入 `.xlsx` 中即可。