日志异常检测 模型说明

1. 建模思路

在对数据集的观察中，可以注意到错误信息多带有一些特定的词语，如“异常”、“错误”、“失败”等，因此可以通过对错误信息的文本特征进行提取，然后通过机器学习的方法进行异常检测。在这里，我们使用了支持向量机进行正常与异常数据的二分类识别。

2. 特征提取

首先，观察训练集可以得知，绝大多数日志信息由`timestamp date hostname [RAS/NULL] module {log_str}`的结构构成。考虑到常识上，时间与异常日志应为弱相关、主机名与异常日志应为弱相关，因此仅提取了`RAS/NULL module`与`log_str`作为特征，使用`sklearn.feature_extraction.text.CountVectorizer`对日志信息进行特征提取。

3. 模型训练与测试

1. 将数据集中正常日志与异常日志，分别抽取`80%`作为训练集，`20%`作为测试集。
2. 使用vectorizer对训练集进行特征提取。
3. 使用支持向量机进行训练，其中正常日志为`1`，异常日志为`0`。
4. 使用测试集进行测试，计算准确率、召回率、F1值等指标。

4. 结果分析

基于随机划分的测试、训练集不同，训练得到的模型性能存在一定浮动。但在多次实验中，我们的模型准确度均在`99%`以上，召回率基本为`100%`,F1值在`99%`以上。但特异性较低，仅在`90%`到`97%`浮动。推测是由于异常日志样本量较少，特征与词汇种类也不如正常日志丰富，因此在异常日志的识别上存在一定的困难。

整体而言，我们的异常检测模型在日志异常检测上表现良好，可以作为日志异常检测的一个有效工具。