多变量时间序列异常检测

实现方法

时间序列异常检测是基于论文 Outlier detection for multidimensional time series using deep neural networks 实现的，将 LSTM 和自编码器结合，可以有效地利用 LSTM 捕捉时间序列数据中的复杂模式和长期依赖关系，而自编码器学习正常时间序列的特征，通过重构误差来区分正常样本和异常样本。

实验步骤

1. 数据预处理

- 缺失值处理：检测数据是否含有 NaN 或 Inf，并将检测到为 NaN 的值用 0 填充
- 标准化： 使用 StandardScaler 对训练集和测试集的多变量时间序列数据进行标准化处理，确保各特征在相似的数值范围内。

2. 模型搭建

- 参考论文相应代码的 LSTMAE 模块，采用 LSTM 网络构建编码器-解码器结构。编码器将输入序列编码为一个低维度的隐藏状态向量，解码器部分根据该向量重构输入序列。
- 使用 MSE 均方误差作为损失函数。
- 使用 ReLU 作为激活函数（原文使用 Sigmoid）：可以缓解梯度消失问题。

3. 模型训练

- 基于 keras 框架，batchsize=256，epochs=15，取训练集中 10%的数据作为验证集。由于是重构模型，同时使用处理好的训练集数据作为输入输出。

4. 检测异常

- 使用训练好的模型对正常数据进行重构，并计算原始正常数据和重构数据之间的 MSE 均方误差，接着对测试数据进行重构，计算 MSE。
- 设定阈值：根据正态分布的规则 95%的数据点落在均值的两个标准差范围内，设定阈值为 $Threshold = \mu + 2\sigma$ 。若样本的重构误差 MSE 大于 Threshold，标记为 anomaly

可能的改进方向

- LSTM 自编码器的网络结构简单，可以增加模型的复杂度，增加网络层数实现深度自编码器。
- 在 LSTM 中引入注意力机制，使模型能够更好地关注重要时间步。
- 根据数据动态调整阈值，避免固定阈值带来的误差。
- 将无监督变为半监督，引入少量已标记的数据，提高异常检测结果。