面向磁浮轨道异常检测的大数据分析框架研究

结题报告　刘震

一、项目总述

1.1 项目背景

磁悬浮列车（Maglev）作为一种先进的轨道交通方式，具有高速、平稳、低噪声等优点。上海磁浮示范运营线作为全球首条商业化高速磁浮线路，自 2002 年投入运营以来，在提升城市交通效率方面发挥了重要作用。然而，随着运营时间的增长，轨道系统的结构性老化和外部环境因素的影响，使得轨道的稳定性和安全性面临挑战。轨道异常（如轨道沉降、线性波动）不仅影响列车的运行稳定性，还可能对乘客的安全造成隐患。因此，研究一种基于大数据分析的轨道异常检测框架，对提高磁浮列车的安全性、可靠性和运维效率具有重要意义。

1.2 旨在解决的问题

本研究的核心目标是开发一套面向磁浮轨道异常检测的智能检测框架，利用历史运营数据，结合大数据挖掘和机器学习算法，实现对轨道健康状况的监测与预测。主要问题包括：

如何基于过往数据检测轨道异常点？

如何通过机器学习和深度学习的方法提升异常检测的准确性？

如何结合已有数据标签改进、优化数据分析流程和模型结构？

1.3 面临的挑战

本项目面临以下主要挑战：

数据复杂性：磁浮轨道运行数据具有高维度、强时序性和非线性等特点，使得异常检测难度较大。

异常检测精度要求高：微小的轨道变形可能导致严重的运行故障，因此模型需要具备高灵敏度和低误报率。

算法的适应性：磁浮轨道受温度、湿度等环境因素影响，现有检测方法难以准确适应不同的运行环境。

二、数据介绍

2.1 数据来源介绍

本项目使用的主要数据来源于上海磁浮示范运营线的历史运行数据，数据包括但不限于：轨道波形数据（左、右轨悬浮面与导向面波长数据）

数据包含了 2017-2024 年间 15 次检测的数据，每次检测可以得到轨道上 1127 个样本点 lev_left, lev_right, gui_left, gui_right 四个维度的数据，其中，由于 2017 年某次检测数据中出现了数据大量缺失的情况，我们不将该文件考虑在本次研究的数据范围中。

2.2 数据划分与预处理

为了扩充矩阵输入的维度，我们以 7 次检测为一个单位，将数据集划分为两部分，其中一部分作为训练集，另一部分作为验证集。同时，为了丰富本次异常检测任务，我们还对数

据标签进行了进一步划分：在 7 次检测中，若出现超过 4 次显示为异常数据，则将其判定为“严重异常（Sereve）”；出现 1-3 次异常标签判定为“普通异常（Abnormal）”；剩余未出现的样本点视作正常点（Normal）。

在所有数据中，有些数据文件存在单一值缺失的情况，因此我们在对数据进行预处理时，一律用该列的平均值填充该单一值，以保证数据是维度对齐的

三、模型介绍与相应结果

3.1 数据降维

在本次研究中，我们先对原先产生的 4*7=28 维的数据进行降维工作，主要涉及到的算法有：PCA（Principal Component Analysis），UMAP（Uniform Manifold Approximation and Projection）两种经典的降维方式。

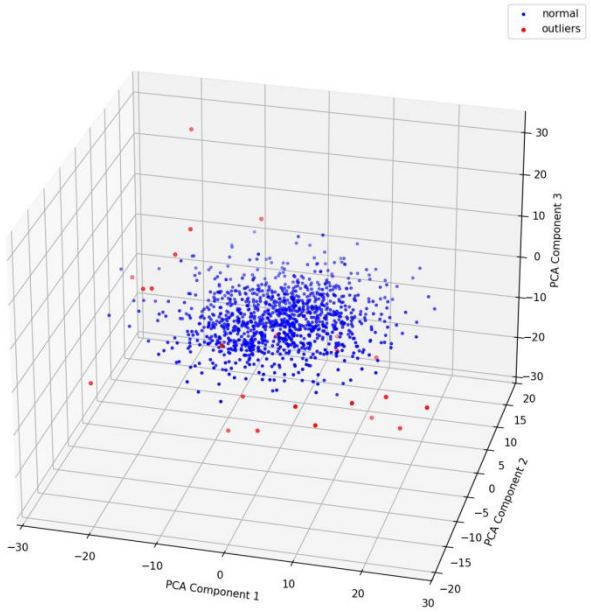


图 1 PCA 降维效果可视化

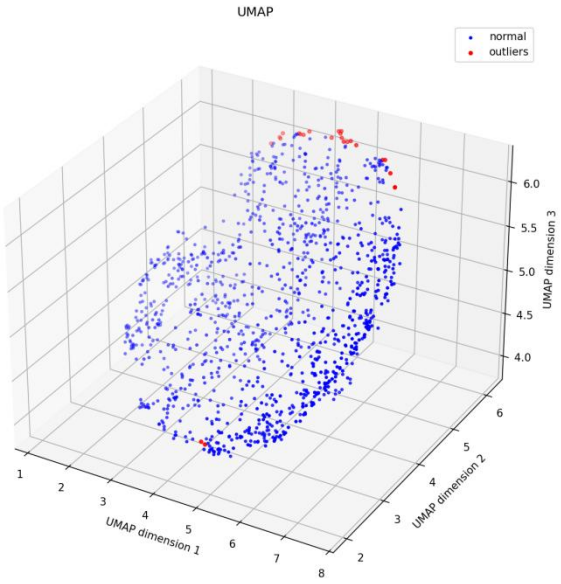


图 2 UMAP 降维效果可视化

3.2 模型训练

在本次任务中，我们选取了 KNN（K-Nearest Neighbors）、RF（Random Forest）、MLP（Multilayer Perceptron）、 Decision Tree（决策树）、SVM（Support Vector Machine）、Naïve Bayes（朴素贝叶斯）这些有监督算法作为主要检测模型。

3.2.1 模型介绍

(1) K 近邻（KNN）

KNN 通过计算测试样本与训练样本之间的欧几里得距离，选取最近的 K 个邻居进行分类 。

$$\hat{y} = \arg\max_{c} \sum_{i \in \mathcal{N}_k} \mathbb{1}(y_i = c)$$

其中， $\mathcal{N}_k$表示测试样本的 K 个最近邻, $\mathbb{1}(y_i = c)$表示该邻居是否属于类别。

(2) 支持向量机（SVM）

SVM 通过构造最优超平面实现分类，目标是最大化数据间的间隔。

$$\min_{w,b} \frac{1}{2}|w|^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1, \quad \forall i$$

其中，ω是超平面法向量，b 是偏置项。

(3) 朴素贝叶斯（Naive Bayes）

朴素贝叶斯基于贝叶斯定理进行分类，假设特征条件独立。

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

(4) 多层感知机（MLP）

MLP 采用前馈神经网络，使用反向传播进行训练。

$$h = \sigma(Wx + b)$$

其中，W 为权重矩阵，b 为偏置，б为激活函数（ReLU）。

(5) 随机森林（RF）与决策树（DT）

决策树基于信息增益进行特征选择 [5]，其计算公式如下：

$$IG(D, a) = H(D) - \sum_{v \in V} \frac{|D_v|}{|D|} H(D_v)$$

其中，H(D)为熵函数。随机森林是多个决策树的集成，通过多数投票进行决策。

3.2.2 训练过程

我们分别利用这六种算法对数据集进行了检测，并且基于 GridSearch 对每个模型进行了参数优化，最终可以得到模型的最优参数，以及相应的训练结果

model	best_macro_f1_score	best_params
KNN	0.542143439	{'metric': 'euclidean', 'n_neighbors': 15, 'weights': 'distance', 'smote__k_neighbors': 3}
RandomForest	0.631262333	{'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 300, 'smote__k_neighbors': 3}
MLP	0.53567211	{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.0001, 'smote__sampling_strategy': 'auto'}
SVM	0.562062487	{'C': 1, 'gamma': 'auto', 'kernel': 'rbf', 'smote__k_neighbors': 3}
DecisionTree	0.600285471	{'criterion': 'entropy', 'max_depth': 10, 'min_impurity_decrease': 0.0, 'smote__k_neighbors': 5}
GaussianNB	0.532100485	{'var_smoothing': 1e-9, 'smote__k_neighbors': 3}

Algorithm	weighted f1-score	macro f1-score
RF	0.7296	0.6438
KNN	0.6279	0.5430
MLP	0.6817	0.5660
DecisionTree	0.7014	0.5667
SVM	0.6705	0.5434
Naive Bayes	0.6496	0.5441

图 3 利用 GridSearch 得到的各个模型最优参数以及评价参数

除了利用单一模型检测之外，我们还采用了多模型联合的方式，主要有模型堆叠、模型投票两种方式。

(1) 模型堆叠（Model Stack）

模型堆叠是一种集成学习方法，通过结合多个基模型（Base Models）的预测提取特征，再通过一个元模型（Meta Model）进行最终预测。

我们遍历的所有可能作为 Base Models 的模型组合（一共有 26 种），结果如下表：

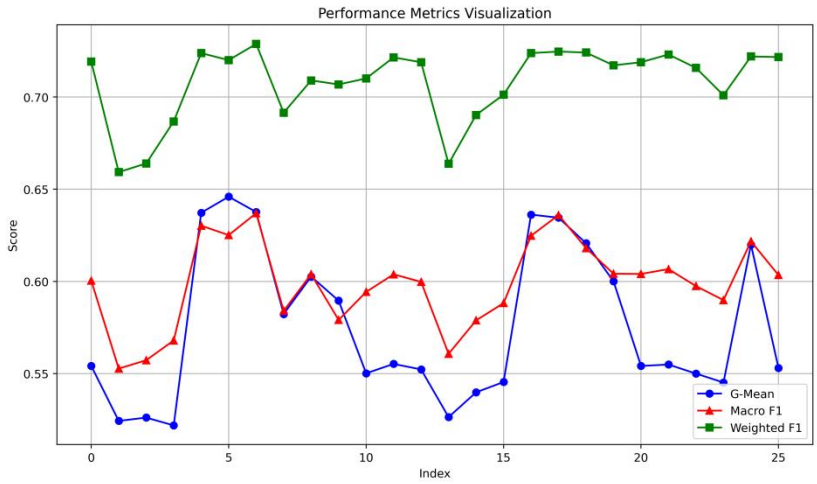


图 4 模型堆叠：不同元模型之间的表现差异

可以直接观测到具有较高 F1-score 的组合，他们有：MLP + KNN,MLP + KNN + DecisionTree,MLP + KNN + DecisionTree + SVM，MLP + DecisionTree + SVM，MLP + SVM

等。

（2）模型投票（model vote）

$$All\_Models: [MLP，RF，KNN，SVM，DecisionTree，NaiveBayes]$$
$$Votemodels \in All\_models$$

为了找到最佳的模型投票组合，我们遍历了 All_models 所有可能的子集，得到的结果如下：

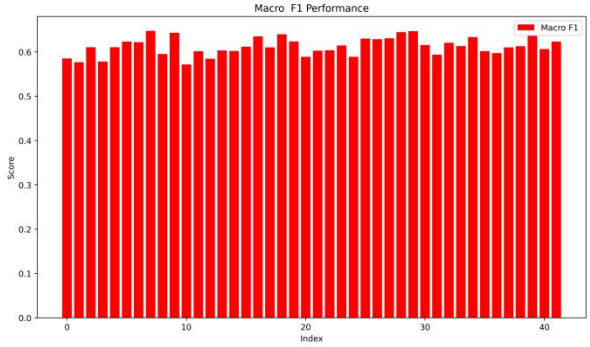


图 5 模型投票：各种组合之间的差异

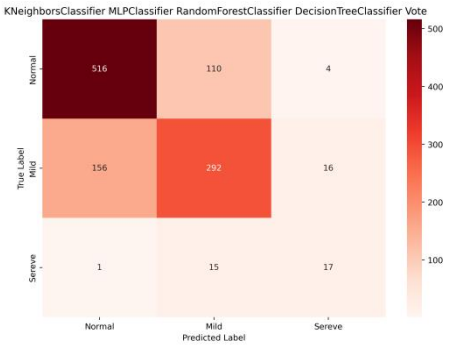


图 6 模型投票：最佳方法对应的混淆矩阵

其中，最佳组合为：KNN + MLP + RF + DecisionTree，其 Confusion Matrix 见图 6。

3.3 检测过程中的 Tricks

（1） SMOTE 处理对部分模型的必要性

SMOTE（Synthetic Minority Over-sampling Technique）是一种用于解决分类问题中类不平衡问题的算法。该算法通过在特征空间中生成合成的少数类样本来增加少数类样本的数量，从而平衡类别之间的比例。SMOTE 的核心思想是通过插值生成新的样本，而不是简单地复制少数类样本，这样有助于提高模型的泛化能力。

$$x_{new} = x_i + \lambda(x_j - x_i), \lambda \sim U(0,1)$$

下图展示了在有 SMOTE 和无 SMOTE 的情况下，算法 F1-score 的对比，可以看见的是，对于 KNN，SVC，MLP 算法来说，SMOTE 扩充数据可以对检测准确率有明显提升。

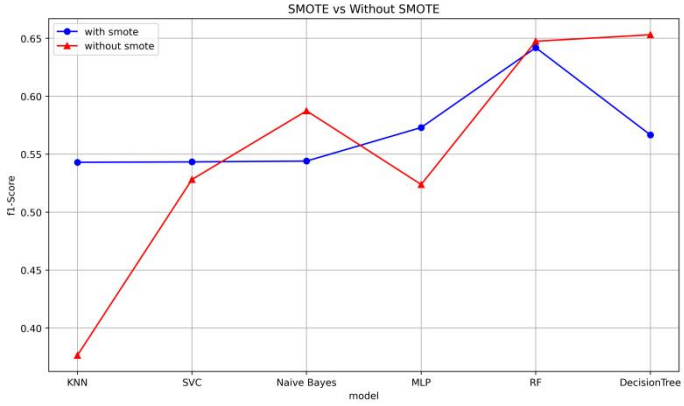


图 7 SMOTE vs without SMOTE

（2）阈值优化

$$\text{Loss} = \frac{1}{\text{TotalSamples}} \sum_{i=1}^{\text{TotalSamples}} \text{ClassWeight}[y_i] \cdot \text{SampleLoss}(h(x_i), y_i)$$

针对本数据集数据标签不平衡的特点，我们对损失函数进行了进一步优化，让模型在检测异常点时能够表现更佳。

（3）降维在本数据集中失效

对于有监督方法，对部分模型可能具有有效性，但是大部分情况下会使效果更差。

右图分别展现了:不降维、pca 降维至 7 维、umap 降维至 4 维后在各个模型上的表现，明显可以看出的是，不降维的效果是要明显优于其他两种的。

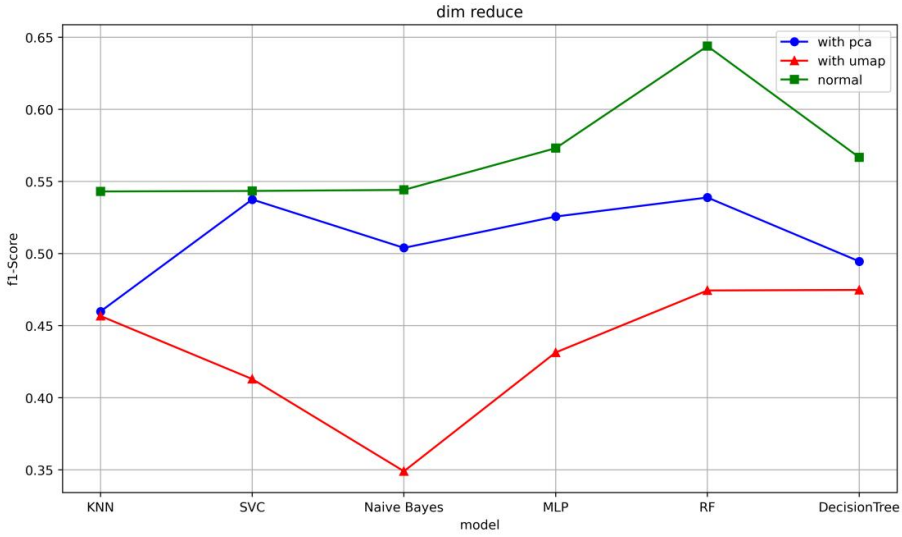


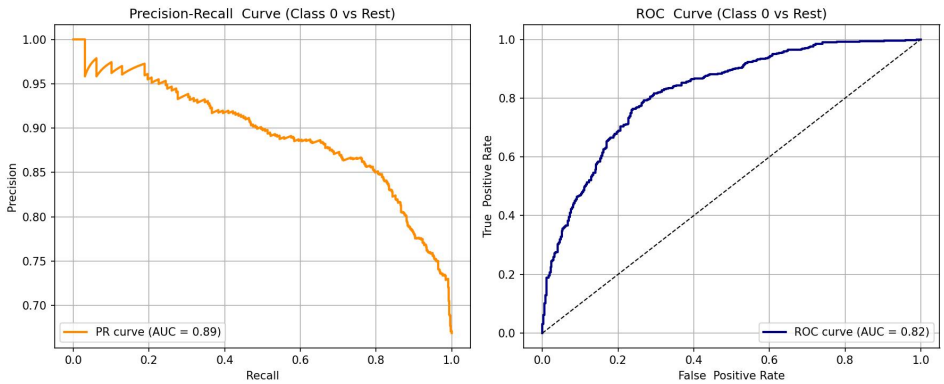
图 8　降维与不降维的效果对比

四、检测结果

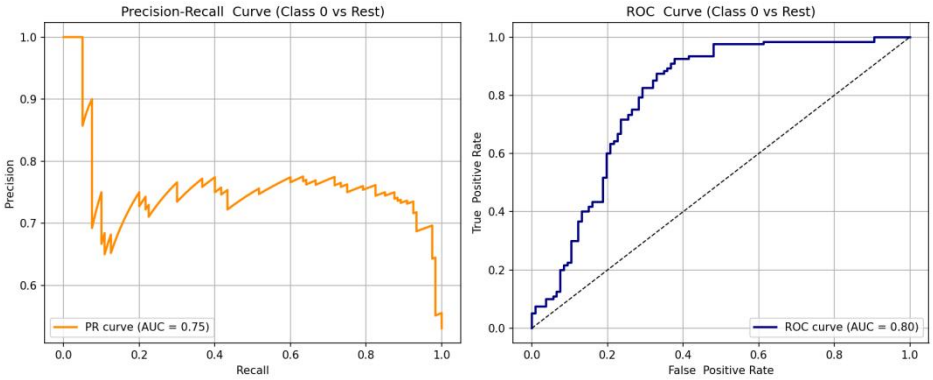
最佳参数的 RF 模型取不同 random_state （5 次）的结果

	Precision	Recall	F1-score	AUROC	AUPRC
训练集	0.892（0.002）	0.867（0.001）	0.860（0.004）	0.96	0.97
验证集	0.762（0.008）	0.752（0.006）	0.744（0.008）	0.80	0.75
测试集	0.713（0.014）	0.732（0.008）	0.711（0.005）	0.89	0.82

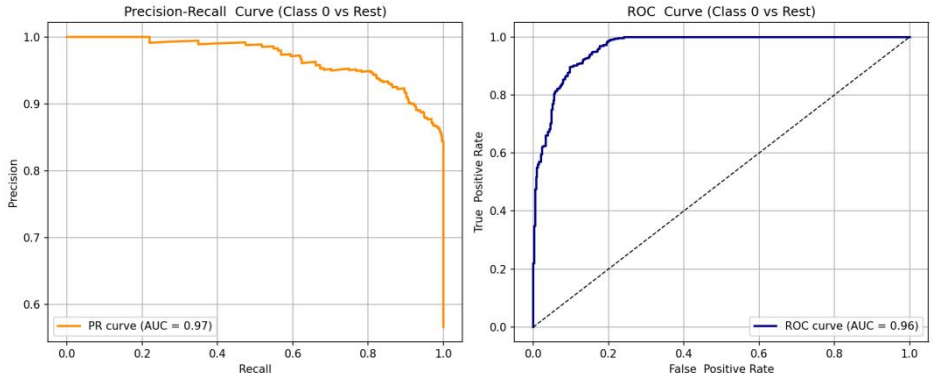
测试集 PR/ROC 曲线：



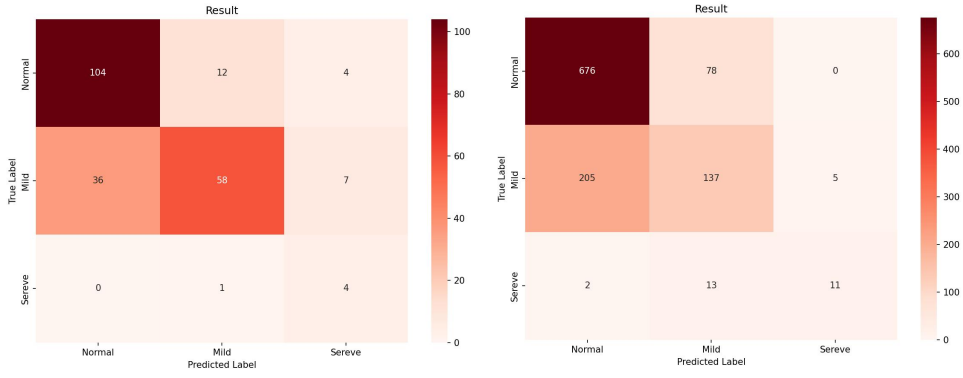
验证集 PR/ROC 曲线



训练集 PR/ROC 曲线：



训练集和测试集对应的 confusion_matrix



五、参考文献

[1]Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

[2]Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297.

[3]McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. AAAI Workshop on Learning for Text Categorization.

[4]Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

[5]Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

[6]Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[7]Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[8]He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[9] 吴峻;李洪鲁;张雨馨;张云洲;汤钧元.中低速磁浮轨道疑似不平顺的 Box-Whisker 图筛选法[J].交通运输工程学报，2023，03：68-76

[10] 贺航宇;王岁儿;吴春晓;王晟;李擎.基于 FBG 和 EMD 的悬挂式永磁磁浮轨道厢梁应力检测技术研究[J]. 铁路通信信号工程技术，2023，08：52-58

[11] 洪小波.高速磁浮轨道不平顺检测系统的研究[D].湖南：国防科技大学，2021

[12] 李梦雪;张敏;马卫华;罗世辉.中低速磁浮轨道不平顺功率谱研究[J].铁道标准设计，2023-7-21

[13] 周旭;温韬;龙志强.基于漏检率的磁浮列车悬浮系统异常检测[J].西南交通大学学报，2023，058（004）：903-912