

基于 XGBoost 的冷水机组不平衡数据故障诊断*

FAULT DIAGNOSIS OF UNBALANCED DATA OF CHILLERS BASED ON XGBOOST

潘进**¹ 丁强¹ 江爱朋¹ 陈越增***² 夏宇栋¹

(1. 杭州电子科技大学自动化学院, 杭州 310018)

(2. 宁波惠康实业有限公司, 宁波 315470)

PAN Jin¹ DING Qiang¹ JIANG AiPeng¹ CHEN YueZen² XIA YuDong¹

(1. College of Automation, Hangzhou Dianzi University, Hangzhou 310018, China)

(2. Ningbo Hicon Industry Co., Ltd., Ningbo 315470, China)

摘要 冷水机组运行数据分布有不平衡、非高斯、非线性、含噪声的特点,这给基于数据的冷水机组故障诊断带来了挑战。针对这些特点,提出了一种将局部密度过采样算法(Minority Oversampling under Local Area Density, MOLAD)和极限梯度提升算法(eXtreme Gradient Boosting, XGBoost)相结合的复合算法应用于冷水机组故障诊断中,以克服样本分布不平衡问题,引入代价敏感学习理论来提升重要故障的召回率。基于离心式冷水机组常见的七个故障监测数据进行仿真,结果表明:XGBoost 相比于对照组能够更好的对冷水机组状态监测数据进行分类;MOLAD-XGBoost 复合模型能够有效处理数据不平衡问题;代价敏感权重可以有效提高重要故障的召回率。

关键词 故障诊断 冷水机组 极限梯度提升 不平衡数据 过采样

中图分类号 TP277

Abstract The chiller operating data has unbalanced, non-Gaussian, non-linear, noise-containing characteristics, which poses a challenge for data-based chiller fault diagnosis. Aiming at these characteristics, a chiller fault diagnosis method based on Minority Oversampling under Local Area Density and eXtreme Gradient Boosting is proposed to chiller fault diagnosis to overcome sample distribution imbalance. Introduce cost-sensitive learning theory to increase the recall rate of important faults. The simulations of seven fault monitoring data commonly used in centrifugal chillers show that XGBoost can better classify chiller status monitoring data compared to the control group. The MOLAD-XGBoost composite model can effectively deal with data imbalance problems; Cost sensitive weights can effectively increase the recall rate for critical failures.

Key words Fault diagnosis; Chiller; Extreme gradient boosting; Unbalanced data; Oversampling

Corresponding author: CHEN YueZen, E-mail: hicon2008@163.com, Tel: +86-574-62131866, Fax: +86-574-62132901

The project supported by the National Natural Science Foundation of China(No.61374142).

Manuscript received 20191018, in revised form 20191121.

引言

随着暖通空调(HVAC)技术的发展,HVAC 设备的系统结构愈发复杂,这给 HVAC 系统运行监测和故障诊断工作带来了挑战。冷水机组是 HVAC 系统中能耗最大的设备,相关研究表明^[1],及时排除冷水机组故障可以有效降低 20%~50%的能源消耗,因此对冷水机组运行状态进行快速准确的判断是保障冷水机组安全稳定运行和节约能源的重要基础。

对于冷水机组故障诊断的方法通常可分为基于经验、模型和数据三种^[2]。目前,基于数据驱动的故障诊断方法是研究主流^[3]。其中,机器学习方法将故障诊断问题视为一个多分类问题,通过离线训练的方法建立起冷水机组运行时的测量参数与运行状态之间的映射关系,目前的研究方法主要有神经网络类^[4-5]^{[6]5155-5163}、支持向量机类^{[7]2524-2532}^{[8]989-998}^{[9]540-547}等。

冷水机组一般处于较为稳定的运行状态,所采集

* 20191018 收到初稿,20191121 收到修改稿。国家自然科学基金项目(61374142)资助。

** 潘进,男,1995年生,安徽宣城人,汉族,杭州电子科技大学自动化学院硕士研究生,研究方向为冷水机组故障诊断与模型优化。

*** 陈越增(通信作者),男,1965年10月生,浙江慈溪人,宁波惠康实业有限公司高级工程师,研究方向为制冷系统设计与优化。

的故障数据数量远少于正常数据,数据分布有不平衡性,使得模型分类结果偏向于正常,而对于故障诊断而言,不同类之间错分代价不同,使模型分类结果偏向于故障状态有较高的实际意义^{[10]673-688}。文献[11]指出冷水机组的部分参数不服从高斯分布且无法转换为高斯分布,不能预先假设数据的分布情况。数据在采集、传输、存储过程中可能发生的异常以及传感器自身存在迟滞、漂移、频率不一致、测量延迟等原因^[12],使冷水机组原始数据中含有离群点、缺失点和噪声,这会导致训练出来的模型过拟合,泛化能力差。冷水机组的特征参数之间存在非线性、高耦合的特点,这使得数据线性不可分。

针对上述问题提出一种基于 XGBoost 的复合故障诊断方法。XGBoost 以决策树为基础模型,无需假设数据的分布结构,并能有效处理非线性分类问题,用近似学习理论对样本数据进行划分,加快学习速度的同时提高了鲁棒性。在原有过采样技术的基础上加以改进,提出基于局部密度的过采样技术,在改善类间数量不平衡的基础上避免因离群点和噪声造成的过拟合;设置敏感权重,以改善类间价值不平衡的问题。采用冷水机组实测数据进行故障模拟实验并设置对照组,结果表明,相比于其它方法,基于 XGBoost 的故障分类方法在冷水机组故障数据分类中具有更高的精确率和召回率,本文提出的 MOLAD-XGBoost 复合模型能够在不平衡数据分类中取得较好结果。

1 基本原理

1.1 XGBoost 分类原理

Boosting 框架的思想是将多个弱分类器集成在一起形成一个强分类器,XGBoost 所选择的弱学习器为分类回归树(Classification and Regression Trees, CART)^{[13]785-794}(图1)。

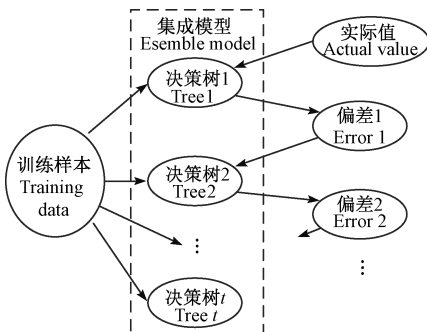


图1 XGBoost 提升原理

Fig.1 XGBoost lifting principle

如图1所示,XGBoost 是一种提升树模型,即初始一棵树去预测一个值,然后得到该值与实际值的偏差,再添加一棵树去学习这个偏差。当我们添加 t 棵树后

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad 0 < \eta < 1 \quad (1)$$

其中, $f_t(x_i)$ 为第 t 棵树对第 i 个数据的判别函数, $\hat{y}_i^{(t)}$ 为 t 棵决策树模型集成强模型的判别结果, η 为学习速率。定义 XGBoost 目标函数为

$$O_{bj}(t) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^t \Omega(f_i) \quad (2)$$

其中,第一部分是计算预测值和真实值的差距, n 为训练样本数量;另一部分是正则化项

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (3)$$

式中, T 是叶子节点的个数, ω 是叶子节点的权值。 γ 表示节点切分阈值, λ 为 L2 正则化系数,两者结合控制模型的复杂程度,防止模型过拟合。根据目标函数,解得最优输出为

$$O_{bj}^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (4)$$

$$G_j = \sum_{i \in I_j} \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (5)$$

$$H_j = \sum_{i \in I_j} \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)}) \quad (6)$$

其中,式(4)是打分函数,用来衡量树结构好坏的标准,值越小代表树的结构越好。XGBoost 采用贪心策略,会遍历每个特征,按照式(7)计算所有特征的分裂 G_{ain} , 选取最佳分裂 G_{ain} 表示分裂前后 O_{bj}^* 的差值,用 G_{ain} 最大的特征作为分裂点构建 CART

$$G_{ain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (7)$$

其中, G_L, G_R 分别为分裂前左子节点和右子节点误差函数的一阶导数之和, H_L, H_R 分别为分裂前左子节点和右子节点误差函数的二阶导数之和。

贪心策略会遍历每一个数据点分裂的增益,加重模型训练的时间复杂度。如图2所示,通过特征的核密度估计(Kernel Density Estimation, KDE)分布,将数据进行聚合统计,存放到固定 KDE 核宽的区间中,形成若干个存放数据的“桶”。在遍历分裂点时,只计算“桶”边界特征值的分裂增益,提高了模型的训练速度。

XGBoost 预先计算了每个特征分裂的增益,然后排序并保存为有序链表结构,在训练时重复地使用此结构,进一步提高了模型的训练速度。

噪声会导致测量数据在真实值附近波动。基于近似学习的分裂点查找方法,会将每个“桶”中的数据近似为“桶”的边界值,降低了对噪声的敏感性^{[13]785-794}。

1.2 局部密度过采样算法

过采样算法通过增加少数类样本的数量达到类别的平衡。简单的随机过采样只是随机复制少数样本使

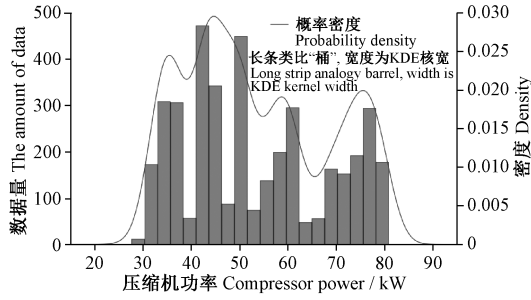


图2 离心式压缩机运行功率的核密度分布估计

Fig.2 Estimation of nuclear density distribution of operating power of centrifugal compressor

样本数量保持平衡,但这会导致模型重复学习特定样本,易产生模型过拟合问题。合成少数类过采样技术^[14] (Synthetic Minority Oversampling Technique, SMOTE)对少数类样本进行分析并合成新的样本添加到数据集中,具体步骤如下:

(1)对少数类中的每个样本 x_i , 按计算其到该类下其他样本的欧式距离,得到样本 x_i 的 k 近邻。

(2)根据样本失衡比例(少数类与多数类的比例)设置一个采样比例以确定采样倍率 N ,对于每个少数类样本 x_i ,从其 k 近邻中随机选择 N 个样本得到 $X^k(x_1^k, x_2^k, \dots, x_n^k)$ 。

(3)将原样本 x_i 与随机选出来的 N 个新样本按照公式(8)生成新样本 x_{new}

$$x_{new} = x_i + \text{rand}(0,1)(x^k - x_i) \quad (8)$$

文献[8]⁹⁸⁹⁻⁹⁹⁸[9]⁵⁴⁰⁻⁵⁴⁷指出不平衡数据集往往伴随着类别重叠和小析取项问题。冷水机组状态监测数据中正常状态和故障状态的特征表达高度耦合,两种状态的数据在空间分布上存在一定重叠;其数据分布概率密度如图2所示呈多个峰值,表示数据存在多个类簇,存在一定的小析取项问题。SMOTE 算法相比于简单的随机采样有较大的优势,但其未考虑到少数类样本的空间分布情况,导致合成的样本质量不稳定。

假设有一类小样本数据,分布如图3所示的三个类簇中,A点为1号类簇内的样本,B点为类簇外的孤立样本。

从图3中可知,B点处的样本分布密度要小于A点。按式(9)计算两个样本点的 k 阶近邻均值 $d(x_i, x_i^k)$ (前 k 个欧氏距离的平均值)。

$$d(x_i, x_i^k) = \frac{1}{k} \sum \sqrt{(x_i, x_i^k)(x_i, x_i^k)^T} \quad (9)$$

设 $k=7$,A点与其最近七个点的欧式距离的均值要小于B点,所以将所有小样本的 k 阶近邻均值按照从小到大的顺序排列,就等同于按样本分布密度从大到小排序。考虑样本分布的局部密度^[15],参照 SMOTE 提出一种基于局部密度的少数类过采样技术

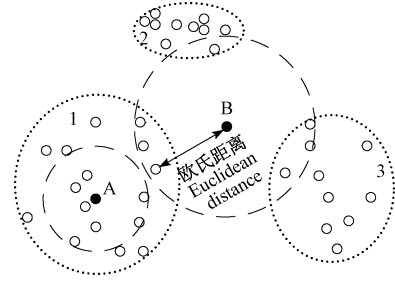


图3 单类样本下局部密度概念图

Fig.3 Conceptual map of local density under a single class of samples

(MOLAD),其具体步骤如下:

(1)按式(9)计算少数类中的每个样本 x_i 的 k 阶近邻均值 $d(x_i, x_i^k)$,并将所有样本按 k 阶近邻均值从小到大顺序排列成升数组 Array。

(2)确定采样倍率 N ,并根据其计算需要生成新样本数量 M 。

(3)顺序遍历 Array 中的样本,按式(8)计算样本和该样本 k 个近邻生成的新样本 x_{new} ;同时统计新样本的个数,当个数达到 M 时,停止算法。

MOLAD 算法可以有效考虑样本的空间分布信息,使局部密度偏小的一些样本不参与新样本的生成,样本的生成范围偏向于密度较大的类簇中,使新生成样本集的质量有所提高,避免模型过度学习异常数据和高噪声数据。

在 MOLAD 算法中,近邻阶数 k 需要大于采样倍率 N ,若 $k=N$,则效果与 SMOTE 算法相同;若 k 小于 N ,则生成的样本数小于 M ,样本数量仍然不平衡。若 k 远大于 N 的话,那么新生成的样本都是以局部密度最大的几个样本和其 k 阶近邻样本所生成,导致模型重复学习高局部密度样本,易发生过拟合,从而降低泛化性能。综上所述, k 的取值应在采样倍率 N 的 1~2 倍之间。

1.3 代价敏感学习

模型在诊断时会假设各类错分代价相等,而在冷水机组故障诊断问题中,故障状态错分的代价比正常状态错分的代价要高,在不同的需求下,不同故障的错分代价也不一致。根据代价敏感学习原理^{[10]673-688},给需要调整的故障类型样本设定敏感权重 b ,以使得模型对该种故障类型更为敏感,降低漏检概率。

XGBoost 在训练时默认所有样本的权重系数都为 1,设有 c 类故障数据和正常数据组成的数据集,每类数量均为 i ,设正常数据的敏感权重为 b_{nor} , c 种故障数据的敏感权重分别为 $b_{e1}, b_{e2}, \dots, b_{ec}$,那么调整敏感权重后的样本权重系数矩阵 w_b 为

$$w_b = [\underbrace{b_n, b_n, \dots, b_n}_{\text{正常样本数}}, \underbrace{b_{e1}, b_{e1}, \dots, b_{e1}}_{\text{故障类型1样本数}}, \underbrace{b_{e2}, b_{e2}, \dots, b_{e2}}_{\text{故障类型2样本数}}, \dots, \underbrace{b_{ec}, b_{ec}, \dots, b_{ec}}_{\text{故障类型c样本数}}] \quad (10)$$

将式(10)代入式(2),得到改进后的 XGBoost 损失函数为

$$O_{bj}(t) = w_b L + \sum_{i=1}^l \Omega(f_i) \quad (11)$$

$$L = [l(y_1, \hat{y}_1), l(y_2, \hat{y}_2), \dots, l(y_n, \hat{y}_n)]^T \quad (12)$$

其中, L 表示 n 个样本的预测损失组成向量的转置, $w_b L$ 表示将样本预测的损失乘上该样本的敏感权重, 让模型能够学习样本的敏感权重信息。

所有类型的敏感权重都默认为 1, 增大某一故障类型失衡权重的敏感权重, 可以提高该故障类型的错分代价, 增大对该故障的检出概率。

2 冷水机组故障诊断实例

2.1 案例介绍

本文选用数据源自 ASHRAE RP-1043 冷水机组故障模拟实验^{[16]54-120}。实验对象为一台 90 t 的离心式冷水机组, 在 27 个工况下进行了测试, 共模拟了 8 种状态类型: 正常状态和 7 种故障状态, 详见表 1, 偏差表示与正常值的偏离程度。实验中采集了冷冻水进出口温度, 冷却水进出口温度, 压缩机功率等 48 个测量参数; 并记录了制冷量、过冷度和能效比等 16 个计算参数。该实验对 7 个故障类型做了四个程度的模拟, 根据文献^{[6]5155-5163} ^{[7]2524-2532}, 故障的正确识别概率和发生程度成正比, 而及时性和早期故障诊断是故障诊断的要求之一, 为了尽早识别出故障的发生, 本文选择故障最小发生程度的监测数据作为仿真对象。

表 1 RP-1043 故障类型
Tab.1 RP-1043 fault types

编号 Number	故障类别 Fault type	偏差 Deviation/%
1	冷凝器结垢 Condenser fouling (CF)	-6
2	润滑油过量 Excessive over (EO)	-14
3	冷凝器水流量减少 Flow water of condenser Insufficient (FWC)	-10
4	蒸发器水流量减少 Flow water of evaporator Insufficient (FWE)	+10
5	含不凝性气体 Non-condensable gas contained (NC)	+1
6	制冷剂泄漏 Refrigerant leak (RL)	-10
7	制冷剂过量 Refrigerant over (RO)	+10

2.2 特征构建

水冷式冷水机组主要结构由制冷剂循环回路、冷却水回路、冷冻水回路构成, 常用测点有蒸发器进出口温度 (TEI、TEO) 和冷凝器进出口温度 (TCI、TCO) 等^{[16]54-120}。考虑到冷水机组实际运行成本和部分特征冗余的问题, 结合文献^[17]和文献^[18], 设定如表 2 所示的特征集合。

表 2 特征信息

Tab.2 Characteristic information

编号 Number	特征变量 Characteristic variable
1	蒸发器侧进口温度 Temperature of evaporator in
2	蒸发器侧出口温度 Temperature of evaporator out
3	冷凝器侧进口温度 Temperature of condenser in
4	冷凝器侧出口温度 Temperature of condenser out
5	压缩机功率 Compressor power
6	蒸发温度 Evaporation temperature
7	冷凝温度 Condensing temperature
8	过冷度 Subcooling
9	压缩机吸气温度 Compressor suction temperature
10	压缩机吸气过热度 Compressor suction superheat
11	压缩机排气温度 Compressor exhaust temperature
12	压缩机排气过热度 Compressor exhaust superheat
13	压缩机油温 Compressor oil temperature
14	压缩机油压 Compressor oil pressure

2.3 故障诊断模型构建

基于 XGBoost 的故障诊断模型主要由离线训练和在线诊断两部分构成, 整体流程如图 4 所示。离线训练是将冷水机组运行的历史数据进行预处理和特征选择, 并将预先计算好的样本权重信息带入 XGBoost 模型进行训练, 然后将测试集带入训练好的 XGBoost 分类模型得到分类结果并与标准标签进行对比, 计算出模型的评价指标, 进而评估模型的性能; 在线诊断为采集冷水机组实时运行数据, 判断其运行状态。

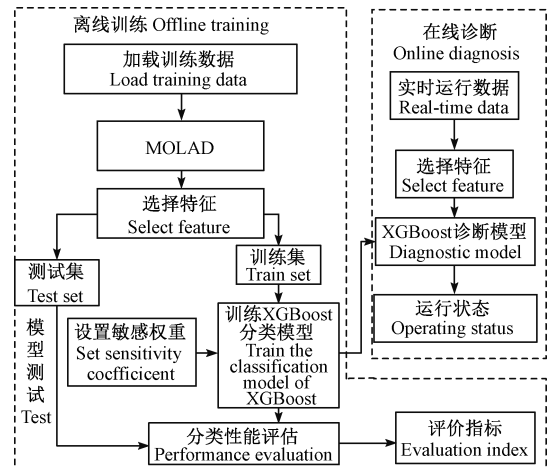


图 4 冷水机组故障诊断流程

Fig.4 Chiller fault diagnosis process

2.4 分类性能评价指标

参考文献^[19]定义如下指标评价分类器性能。

精确率 (Precision, P): 用来衡量模型对某一类故障诊断精确性的指标。定义如下

$$P = \frac{T_p}{T_p + F_p} \quad (10)$$

召回率 (Recall, R): 用来衡量模型对某一类故障检出能力的指标

$$R = \frac{T_p}{T_p + F_N} \quad (11)$$

对于某一类, T_p 是正确识别的数量, F_p 是其它类被错分到该类的样本数量, F_N 是该类被错分到其它类的数量。

P 和 R 是评价多分类模型对某一类的识别能力, 两者往往相互制约, 引入 F_1 , 对精确率和召回率进行折中, F_1 越高, 分类器的分类性能越好

$$F_1 = \frac{2PR}{(P + R)} \quad (12)$$

3 结果与分析

共选取 68 000 条数据作为样本集, 正常数据和每种故障数据的比例为 10 : 1, 以模拟冷水机组数据不平衡性的特性, 采样倍率 $N = 10$ 。为了验证所提出方法的有效性, 设置三个可非线性分类的对照组: 核函数为径向基的支持向量机 (Support Vector Machine, SVM)、多层感知器 (Multi-layer Perceptron, MLP)、标

准 GBDT, 对 SVM 和 MLP 模型的训练样本进行标准化, 以提升模型的性能。仿真基于 Python3.5.2 平台, 并使用了 Numpy1.11.1、Scikit-learn 0.17.1^[20] 和 XGBoost 0.6 等多个科学计算包, 算法参数使用 Scikit-learn 默认值。

3.1 近邻阶数 k 的确定

从前文可知 k 值在 N 的 1~2 倍之间, 取 k 值为 10~20, 将 MOLAD 算法分别与 XGBoost 和三个对照组组合, 对近邻阶数 k 进行验证, 以多个状态的平均 F_1 值为评价指标确定 k 的最佳值, 详见表 3。当 $k = 10$ 时, MOLAD 等同于 SMOTE 算法, 从表 3 可知, MOLAD 算法合成的样本质量要高于 SMOTE 算法。MOLAD 算法中最优近邻阶数 $k = 12$, MOLAD-XGBoost 的诊断表现要优于其余复合算法。

3.2 不平衡数据集下不同模型分类性能

使用多个模型对上文所述样本集进行诊断测试, 其中 MOLAD-XGBoost 中近邻阶数 $k = 12$ 。其分类结果如表 4 所示。

表 3 不同 k 值下的 MOLAD 算法与各分类模型组合的平均 F_1 值

Tab.3 The average F_1 of the MOLAD algorithm combined with each classification model under different k

模型 Model	$k = 10$	$k = 11$	$k = 12$	$k = 13$	$k = 14$	$k = 15$	$k = 16$	$k = 17$	$k = 18$	$k = 19$
MOLAD-XGBoost	0.948 1	0.950 3	0.951 1	0.951 1	0.949 7	0.946 5	0.948 4	0.943 5	0.947 3	0.949 1
MOLAD-SVM	0.869 9	0.870 9	0.870 9	0.870 6	0.870 4	0.870 4	0.870 6	0.868 9	0.868 7	0.869 1
MOLAD-MLP	0.897 5	0.897 5	0.899 1	0.901 5	0.905 1	0.898 5	0.898 6	0.898 5	0.897 7	0.897 5
MOLAD-GBDT	0.911 1	0.912 8	0.913 7	0.913 5	0.913 5	0.913 1	0.913 1	0.913 7	0.913 5	0.912 8

表 4 不平衡样本不同分类方法下的精确率和召回率

Tab.4 Precision and recall under different classification methods for unbalanced samples

故障 Fault	SVM		MLP		GBDT		XGBoost		MOLAD-XGBoost	
	$P/\%$	$R/\%$	$P/\%$	$R/\%$	$P/\%$	$R/\%$	$P/\%$	$R/\%$	$P/\%$	$R/\%$
Normal	83.98	99.55	80.88	95.36	76.01	98.71	91.48	99.52	94.69	96.20
CF	95.74	52.38	73.14	22.40	92.55	38.27	97.98	72.72	96.88	85.37
EO	98.48	83.38	74.10	67.00	91.68	61.84	99.59	93.57	98.66	95.69
FWC	98.74	71.64	97.35	96.97	93.74	48.20	98.15	90.63	97.46	97.06
FWE	99.18	73.99	95.70	94.64	97.26	45.83	98.77	87.48	94.07	93.08
NC	99.89	99.49	99.12	99.31	97.70	97.79	100	99.43	99.91	99.44
RL	98.71	40.10	55.86	6.13	96.63	26.54	96.15	74.59	96.52	87.10
RO	96.31	78.98	81.66	76.30	90.27	63.92	98.95	91.46	96.37	94.67

从表 4 可以得到四点结论:

(1) 前四个模型对于 Normal 状态召回率都达到了 97% 以上, 精确率明显小于召回率, 表示有大量故障类型数据被错分为正常类型; 其余 7 个故障类型由于训练样本数量少, 在精确率和召回率上的表现与正常类型相反, 证明了类间样本数量不平衡对模型分类性能影响很大。

(2) XGBoost 模型对于 8 种状态分类的精确率和召回率表现都明显优于前三种模型, 说明 XGBoost 算法更适用于冷水机组故障数据分类任务。

(3) 所有模型对 NC 故障诊断的精确率和召回率

均接近 100%, 表明 NC 故障状态的特征表达和其它几种状态相差较大, 是更容易识别出的故障; 相反, CF 故障和 RL 故障相对来说就难以识别。

(4) MOLAD-XGBoost 相比于 XGBoost 模型, 大部分状态类型的 F_1 都有提高, 尤其是 CF 状态和 RL 状态; NC 的 F_1 有所下降, 可能是在对少样本过采样后, 增大了故障特征的比重, 导致与 NC 状态故障特征表达类似的故障模糊了 NC 故障数据集的边界。从整体 F_1 的变化来说, 模型在增加失衡权重后对不平衡数据的分类能力有所提高。

3.3 敏感权重验证

在实际应用中,往往会要求提高某些损失大的故障的召回率,以 CF 故障为例,验证敏感权重对 MOLAD-XGBoost 模型的作用。设敏感权重为 0、5、10、15、20 和 25,CF 故障诊断的精确率、召回率和 F_1 的变化如图 5 所示。

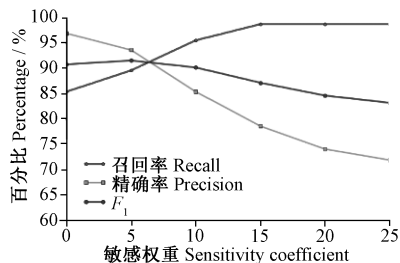


图 5 敏感权重对分类器性能影响

Fig.5 Effect of sensitive weights on classifier performance

从图 5 可以看出,在敏感权重小于 15 时,CF 故障的召回率 R 随着敏感权重增加,当敏感权重大于 15 时,召回率不再增加;精确率随着敏感权重的增加持续减小; F_1 在敏感权重较小时有一定增长,敏感权重大于 5 时 F_1 开始下降。同时考虑到增加 CF 故障的敏感权重会使得分类器对其它状态的分类能力下降,所以在设定敏感权重时应当取一个较小值,以分类器性能小幅下降的代价换取对错判代价高的故障类型更高的召回率。

4 结论

本文通过分析冷水机组运行数据的特点,将一种基于决策树的提升算法-XGBoost 应用到冷水机组故障诊断任务中;基于 SMOTE 算法,结合样本的分布密度,提出 MOLAD 算法,并与 XGBoost 相结合。利用离心式冷水机组的实测数据对所提出的方法进行仿真,得到如下结论。

1) XGBoost 算法不考虑数据的分布情况且能处理线性不可分问题,相比于其他算法,在冷水机组故障数据分类任务中拥有更高的精确率、召回率,是一种适用于冷水机组故障诊断的方法。

2) 针对数据分布不平衡问题,提出的 MOLAD 算法合成的新样本质量更高,MOLAD-XGBoost 复合模型相对于 SMOTE-XGBoost,将平均 F_1 从 0.948 1 提高到 0.951 1。

3) 基于代价敏感学习,提出了敏感权重并进行验证,结果表明设定适当的敏感权重能够在小幅降低分类器准确度的代价下提高对指定故障的召回率,有一定的实用价值。

参考文献 (References)

- [1] Kim W, Katipamula S. A review of fault detection and diagnostics methods for building systems[J]. Science and Technology for the Built Environment, 2018, 24(1): 3-21.
- [2] Beghi A. Data-driven fault detection and diagnosis for HVAC water chillers[J]. Control Engineering Practice, 2016(53): 79-91.
- [3] 文成林,吕菲亚,包哲静,等.基于数据驱动的微小故障诊断方法综述[J].自动化学报,2016,42(9):1285-1299.
WEN ChengLin, LV FeiYa, BAO ZheJing, et al. A review of data driven-based incipient fault diagnosis [J]. Acta Automatica Sinica, 2016,42(9):1285-1299(In Chinese).
- [4] Shi S, Li G, Chen H, et al. Refrigerant charge fault diagnosis in the VRF system using bayesian artificial neural network combined with relieff filter[J]. Applied Thermal Engineering, 2017(112): 698-706.
- [5] 梁晴晴,韩 华,崔晓钰,等.基于主元分析-概率神经网络的制冷系统故障诊断[J].化工学报,2016,67(3):1022-1031.
LIANG QingQing, HAN Hua, CUI XiaoYu, et al. Fault diagnosis for refrigeration system based on PCA-PNN[J]. CIESC Journal, 2016, 67(3):1022-1031(In Chinese).
- [6] 刘旭婷,李益国,孙栓柱,等.基于稀疏局部嵌入深度卷积网络的冷水机组故障诊断方法[J].化工学报,2018,69(12):5155-5163.
LIU XuTing, LI YiGuo, SUN ShuanZhu, et al. Fault diagnosis of chillers using sparsely local embedding deep convolutional neural network [J]. CIESC Journal, 2018, 69(12): 5155-5163 (In Chinese).
- [7] Han H, Gu B, Hong Y, et al. Automated FDD of multiple-simultaneous faults (MSF) and the application to building chillers [J]. Energy & Buildings, 2011, 43(9):2524-2532.
- [8] Sun K, Li G, Chen H, et al. A novel efficient SVM-based fault diagnosis method for multi-split air conditioning system's refrigerant charge fault amount [J]. Applied Thermal Engineering, 2016(108): 989-998.
- [9] Han H, Cui X, Fan Y, et al. Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features[J]. Applied Thermal Engineering, 2019(154): 540-547.
- [10] 李艳霞,柴 毅,胡友强,等.不平衡数据分类方法综述[J].控制与决策,2019,34(4):673-688.
LI YanXia, CHAI Yi, HU YouQian, et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019,34(4): 673-688(In Chinese).
- [11] Chakraborty D, Elzarka H. Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold [J]. Energy and Buildings, 2019(185): 326-344.
- [12] 刘 强,柴天佑,秦泗钊,等.基于数据和知识的工业过程监视及故障诊断综述[J].控制与决策,2010,25(6):801-807.
LIU Qiang, CHAI TianYou, Qin SiZhao, et al. Progress of data-driven and knowledge-driven process monitoring and fault diagnosis for industry process[J]. Control and Decision, 2010,25(6):801-807(In Chinese).

- [13] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.
- [14] 董振波. 面向不平衡数据集分类的层次引力模型研究[D]. 重庆:重庆大学,2016:8-9.
DONG ZhenBo. Research on hierarchical gravity model for classification of imbalanced datasets[D]. Chongqing: Chongqing University,2016:8-9(In Chinese).
- [15] Yang J, Deng T, Sui R. An adaptive weighted one-class svm for robust outlier detection[C]//Proceedings of the 2015 Chinese Intelligent Systems Conference. Springer, Berlin, Heidelberg, 2016: 475-484.
- [16] COMSTOCK M C, BRAUN J E. Development of analysis tools for the evaluation of fault detection and diagnostics for chillers[R]. ASHRAE, 1999: 54-120.
- [17] Zhao Y, Wang S, Xiao F. Pattern recognition-based chillers fault detection method using support vector data description (SVDD)[J]. Applied Energy, 2013,112(C): 1041-1048.
- [18] 郑超瑜,肖 赋,陈 武.冷水机组故障诊断新模型[J].建筑科学,2012,28(12): 104-107.
ZHENG ChaoYu, XIAO Fu, CHEN Wu. New reference models for fault detection and diagnose in water chillers[J]. Buliding Science, 2012,28(12):104-107(In Chinese).
- [19] Chakraborty D, Elzarka H. Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold[J]. Energy and Buildings, 2019(185): 326-344.
- [20] Swami A, Jain R. Scikit-learn: Machine learning in python[J]. Journal of Machine Learning Research, 2012, 12(10):2825-2830.