

DOI:10.19344/j.cnki.issn1671-5276.2024.06.025

基于 MSSMOTE-CNN 模型的空调冷水机组故障诊断

曹冉冉^{a,b}, 田禾^{a,b}, 樊怀聪^{a,b}, 冯明文^{a,b}

(天津理工大学 a. 机电工程国家级实验教学示范中心; b. 天津市先进机电系统设计与智能控制重点实验室, 天津 300384)

摘要:针对冷水机组运行过程中数据类别不平衡问题,提出基于马氏距离进行“三角”区域插值的 MSSMOTE 方法对故障数据进行扩充,将得到的数据输入 CNN 模型进行训练,实现对冷水机组中 7 种故障的诊断。在不同扩充比例下和同一种数据类型下分别进行仿真测试,结果显示:在扩充比例为 4 时,MSSMOTE-CNN 模型对于正常样本测试的准确率和 $F1$ -score 分别达到 0.961 和 0.971,能够较准确识别出冷水机组的故障类型。

关键词:MSSMOTE-CNN 模型;数据不平衡;故障诊断;冷水机组

中图分类号:TP277 **文献标志码:**A **文章编号:**1671-5276(2024)06-0128-05

Fault Diagnosis of Water Chiller Based on MSSMOTE-CNN Model

CAO Ranran^{a,b}, TIAN He^{a,b}, FAN Huacong^{a,b}, FENG Mingwen^{a,b}

(a. National Demonstration Center for Experimental Mechanical and Electrical Engineering Education; b. Tianjin Key Laboratory for Advanced Mechatronic System Design and Intelligent Control, Tianjin University of Technology, Tianjin 300384, China)

Abstract:To deal with the unbalanced data types during the operation of water chillers, this paper proposes the MSSMOTE method based on Mahalanobis distance and "triangle" area interpolation to expand the fault data, and input the obtained data into the CNN model for training, so as to realize the diagnosis of seven kinds of faults in water chillers. Simulation tests were conducted under different expansion ratios and the same data type. The results showed that when the expansion ratio was 4, the MSSMOTE-CNN model achieved an accuracy of 0.961 and a $F1$ -score of 0.971 respectively for normal sample testing, which was capable of accurately identifying the fault type of the chiller.

Keywords:MSSMOTE-CNN model; data imbalance; fault diagnosis; water chilling unit

0 引言

随着国内建筑全寿命周期能源消耗总量占据全国能源消耗总量比例的日益增高^[1],如果暖通空调 (heating, ventilating and air conditioning, HVAC) 系统发生故障,不仅影响室内环境的舒适度,而且会造成能源的大量浪费。所以设备的故障诊断研究是非常有必要的^[2]。基于数据的方法在异常检测^[3]和语音识别^[4-5]等领域都展现出了一定的优势。可以说工业互联网和智能化技术的不断发展为故障诊断技术带来了新的机遇^[6],但是在大数据背景下往往存在着极度的类别不平衡特性,即正常样本数量远远多于故障样本。这种情况严重影响传统的基于数据驱动故障诊断模型的诊断效果^[7-9]。本文主要针对数据采样方法中的过采样方法进行改进,研究基于数据的冷水机

组故障诊断方法,有助于及时识别并解决冷水机组故障。

为了获得更具竞争力的平衡效果,基于过采样的方法主要关注少数样本,采用各种采集少数样本的思想来扩展其样本集的大小。随机复制几种类型的样本是一种相对简单的采样思想,但单个复制的样本只重复原始样本信息,往往会导致分类模型的过拟合^[10]。面向不平衡数据集的分类方法,一般基于 SMOTE 改进的过采样方法是将 SMOTE 方法与密度思想或者 k-means 聚类思想相结合,或者通过赋予选择权重,以达到较好的样本平衡化效果。有通过构建稳定空间和改变插值方式来降低样本重叠性,进而对故障数据进行有效扩充的^[11],还有通过改变距离的度量方式来优化数据扩充方法的^[12]。除此之外,还有很多基于支持向量机的平衡化样本方法,其实验结果较传

基金项目:国网天津市电力公司科技项目 (KJ22-2-02);国网天津市电力公司科技项目 (KJ21-1-21);天津理工大学 2022 年天津市研究生科研创新项目 (2022SKYZ070);天津理工大学 2022 年校级研究生科研创新实践项目 (YJ2209);企业科技特派员项目 (20YDTPJC01670)

第一作者简介:曹冉冉 (1996—),女,河北邯郸人,硕士研究生,研究方向为机械及机器学习算法,3401517075@qq.com。

统的分类算法具有更好的性能和更高的精度。

实际冷水机组运行过程中通常存在类别不平衡以及多类型故障识别准确率低的问题。针对这两种问题,本文提出一种改进的 MSSMOTE 样本平衡化方法。最后将平衡化后得到的数据集输入 CNN 模型进行训练,实现对冷水机组运行过程中多类故障的诊断。

1 基于 MSSMOTE-CNN 的故障诊断模型

1.1 故障的类型

根据 COMSTOCK 等^[13]对美国主流冷水机组产品进行的调查,同时列出了 7 种值得研究的故障类型,如表 1 所示。

表 1 7 种故障类型的介绍

故障名称	缩写
冷凝器水流量减少	FWC
蒸发器水流量减少	FWE
制冷剂泄漏/不足	RL
制冷剂过量	RO
机油过多	EO
冷凝器结垢	CF
制冷剂中含有不凝性气体	NC

在同一条件下,由于不同故障类型和同一故障类型的不同方面都有一定的影响,这会大大增加故障诊断模型构建的难度。

1.2 数据扩充方法的对比

在传统 SMOTE 方法的基础上,SSMOTE 和 MSMOTE 分别在插值方式和距离度量方式上进行了改进。本文将利用 SSMOTE 方法的插值方式改进 MSMOTE 方法样本重叠性低的问题,其中 SMOTE、SSMOTE 和 MSMOTE 在距离度量方式和插值方式上的对比如表 2 所示。

表 2 SMOTE、SSMOTE 和 MSMOTE 方法的对比

项目	方法		
	SMOTE	SSMOTE	MSMOTE
距离度量方式	欧氏距离	欧氏距离	马氏距离
插值方式	线性插值	“三角”区域插值	线性插值

1.3 改进的 MSSMOTE 方法

由于 MSMOTE 方法的聚类效果好,但是采用

线性插值导致样本重叠性较高,而 SSMOTE 方法的聚类效果较差,而采用的“三角”区域插值能降低样本重叠性。为了得到更好的聚类效果,同时降低样本重叠性,将 SSMOTE 方法中的“三角”区域插值方式与 MSMOTE 方法基于马氏距离这两个方式相结合,以改进 MSMOTE 方法样本重叠性低的问题,以下是基于改进后的 MSSMOTE 方法的具体步骤。

1) 利用式(1)计算故障样本集 $\hat{X}_{\min}$ 中的任意一个样本 $\hat{x}_{\min,i}$ 到 $\hat{X}_{\min}$ 中其他样本的马氏距离,筛选出其中最小的 k 个样本作为近邻样本,并记为 $\hat{X}_{\min,j}$ 。将少数类样本 $\hat{x}_{\min,i}$ 与其近邻样本组合成形式为 $\{\hat{x}_{\min,i}, \hat{x}_{\min,i1}, \hat{x}_{\min,i2}, \dots, \hat{x}_{\min,ik}\}$ 的样本组。

$$D_M(\hat{x}) = \sqrt{(\hat{x} - \hat{\mu})^T S^{-1} (\hat{x} - \hat{\mu})} \quad (1)$$

式中: S^{-1} 为 S 的逆矩阵, S 为样本协方差矩阵; k 一般取值为 5。

2) 对于每个样本组 $\{\hat{x}_{\min,i}, \hat{x}_{\min,i1}, \hat{x}_{\min,i2}, \dots, \hat{x}_{\min,ik}\}$, 统计其中的样本 $\hat{x}_{\min,i1} \sim \hat{x}_{\min,ik}$ 所对应的故障类型标签,并设定比率阈值 r_s 。对于样本 $\hat{x}_{\min,i}$, 若在 k 个近邻样本的类型标签中与 $\hat{x}_{\min,i}$ 类型标签相同样本的数目所占比例超过 r_s , 则样本 $\hat{x}_{\min,i}$ 被保留,否则被去除。通过这种方式对各个类型的样本集进行样本筛选,每个类型的保留样本构成了所属类型的稳定特征空间集合 $S_t (t=1, 2, \dots)$ 。

3) 在任一稳定特征空间集合 S_t 中,随机选出一个保留样本 $\hat{x}_i^{S_t}$ 。然后,在样本 $\hat{x}_i^{S_t}$ 的 k 个近邻样本中随机选出两个近邻样本 $\hat{x}_{i1}^{S_t}$ 和 $\hat{x}_{i2}^{S_t}$ 。最后,利用上述 3 个样本合成新样本 $\hat{x}_{\text{new}}$, 计算公式如下:

$$\hat{x}_{\text{new}}^1 = \hat{x}_{\min,i}^{S_t} + r_{e1} \times (\hat{x}_{\min,i1}^{S_t} - \hat{x}_{\min,i}^{S_t}) \quad (2)$$

$$\hat{x}_{\text{new}}^2 = \hat{x}_{\min,i}^{S_t} + r_{e2} \times (\hat{x}_{\min,i2}^{S_t} - \hat{x}_{\min,i}^{S_t}) \quad (3)$$

$$\hat{x}_{\text{new}} = \hat{x}_{\text{new}}^1 + r_{e3} \times (\hat{x}_{\text{new}}^2 - \hat{x}_{\text{new}}^1) \quad (4)$$

式中 r_{e1} 、 r_{e2} 和 r_{e3} 均为取值在 0~1 之间的随机数。

MSSMOTE 方法流程如图 1 所示。首先将训练样本集分为故障样本集和正常样本集,然后分别进行归一化,得到归一化后的故障样本集和正常样本集。按照上面的步骤得到新样本 $\hat{x}_{\text{new}}$, 直到 $\hat{x}_{\text{new}}$ 和 $\hat{X}_{\min}$ 数量达到一定比例。将得到的新样本添加到故障样本集 $\hat{X}_{\min}$ 中,此时 $\hat{X}_{\max}$ 和 $\hat{X}_{\min}$ 构成新的训练数据集。

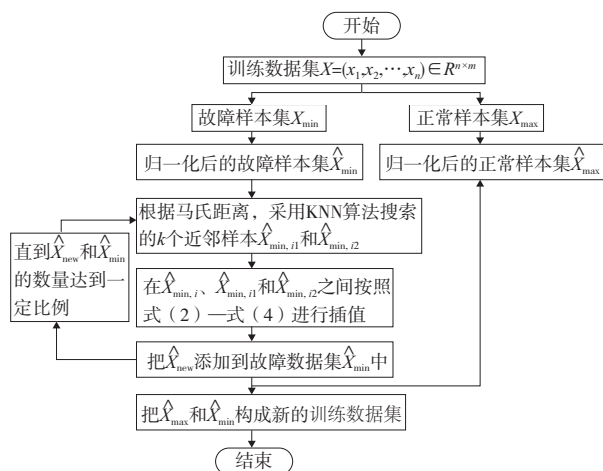


图1 MSSMOTE 算法流程图

2 实验方案及仿真分析

2.1 MSSMOTE 方法可行性分析

在本节仿真过程中,选用了 RP-1043 公开数据集^[14]中的 LEVEL2 数据集来训练 MSSMOTE-CNN、MSSMOTE-LSTM 和 MSSMOTE-TCN 模型,并测试各模型在正常和 7 种故障数据中的故障诊断性能。

故障诊断结果如图 2 所示,混淆矩阵的横向代表样本预测类别,纵向为样本的真实类别。主对角线方块中的数字代表正确分类的样本数量,除主对角线以外的深色方块中的数字表示错误分类的样本数量。可以看出,图 2(b)和图 2(c)中的第一行错误分类的样本数字和分别为 20 和 9,大于图 2(a)中第一行错误分类的样本数字和 3;同样图 2(b)和图 2(c)中其余几行错误分类的样本数字和都大于图 2(a)中相应行的错误分类的样本数字和,这说明 MSSMOTE-LSTM 和 MSSMOTE-TCN 错误分类的样本更多。从图 2(a)来看, MSSMOTE 方法产生的数据在 CNN 模型上的适用性更好。

	Normal	CF	EO	FWC	FWE	NC	RL	RO
Normal	413				3			
CF		415				1		
EO			415	1				
FWC				416				
FWE					415		1	
NC						414		2
RL	2					1	412	1
RO	1				2			413
	Normal	CF	EO	FWC	FWE	NC	RL	RO

(a) MSSMOTE-CNN

	Normal	CF	EO	FWC	FWE	NC	RL	RO
Normal	396	3			8		2	7
CF		413				1	2	
EO		2	412					2
FWC	2		1	413				
FWE					413		3	
NC						413		3
RL	11	3			5		391	6
RO	13					2	9	392
	Normal	CF	EO	FWC	FWE	NC	RL	RO

(b) MSSMOTE-LSTM

	Normal	CF	EO	FWC	FWE	NC	RL	RO
Normal	407			2			1	6
CF		413					3	
EO		2	413		1			
FWC				414				2
FWE			2		413		1	
NC		1				412	3	
RL	3	5		1	9	7	391	
RO	10				4		9	393
	Normal	CF	EO	FWC	FWE	NC	RL	RO

(c) MSSMOTE-TCN

图2 3种模型的故障诊断结果

2.2 故障诊断方案

基于 MSSMOTE 方法的冷却系统故障诊断整体方案如图 3 所示,具体流程如下。

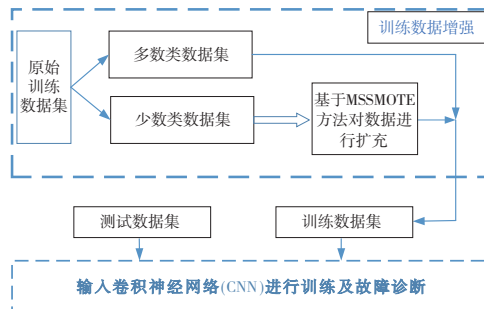


图3 基于 MSSMOTE-CNN 方法的冷却系统故障诊断整体方案

1) 仿真数据设定

本节所使用的数据集是从原始数据集中随机选取了 6 668 个正常样本,而每一种故障类型都仅选取了 800 个样本,由此组成不平衡数据集。测试集则是由与不平衡数据集不重复的 13 344 个样本组成。为了验证 MSSMOTE 方法的数据扩充能力,在仿真过程中,选择了 SSMOTE 和 MSMOTE 这两种作为数据扩充对比方法。分类器选用了在上节表现较好的 CNN 模型进行训练。本文选择准确率 (Accuracy) 和 F1 分数 (F1-score) 作为评价指标,其中 F1 分数定义为精确率 (Precision) 和召回率 (Recall) 的调和平均值,是一

个衡量分类模型综合能力的指标, $F1$ -score 值越高说明模型的综合识别能力越好。评价指标的混沌矩阵如表 3 所示,计算过程如下:

$$A_{\text{accuracy}} = \frac{T_p + T_n}{T_p + F_p + T_n + F_n} \quad (5)$$

$$F1\text{-score} = 2 \cdot \frac{P_{\text{recision}} \cdot R_{\text{ecall}}}{P_{\text{recision}} + R_{\text{ecall}}} \quad (6)$$

$$P_{\text{recision}} = \frac{T_p}{T_p + F_p} \quad (7)$$

$$R_{\text{ecall}} = \frac{T_p}{T_p + F_n} \quad (8)$$

表 3 评价指标的混沌矩阵

项目	预测正值	预测负值
真正值	T_p	F_n
真实负值	F_p	T_n

2) 按照扩充比例对 MSSMOTE 方法的性能测试

为了验证 MSSMOTE 方法的数据扩充能力,在实验分析中,选择了 SSMOTE 和 MSMOTE 作为数据扩充对比方法,并将扩充之后的数据输入 CNN 模型中训练, MSSMOTE、SSMOTE 和 MSMOTE 与 CNN 模型的组合分别简记为 MSSMOTE-CNN、SSMOTE-CNN 和 MSMOTE-CNN。最后,使用测试集获取 3 种模型各自的 Accuracy 和 $F1$ -score 指标。

测试结果如图 4 所示,从图中可以看出,在扩充比例为 1、2 和 3 的情况下,3 种方法扩充的数据都可以使模型的 Accuracy 和 $F1$ -score 指标得到不同程度的上升,并且在扩充比例为 4 时达到了最大值,但在 4 以上都有不同程度的下降。

表 4 MSSMOTE-CNN、SSMOTE-CNN 和 MSMOTE-CNN 详细结果

指标	模型	数据扩充比例						
		0	1	2	3	4	5	6
A_{accuracy}	MSSMOTE-CNN	0.774	0.891	0.913	0.934	0.961	0.957	0.954
	SSMOTE-CNN	0.774	0.819	0.832	0.847	0.861	0.843	0.837
	MSMOTE-CNN	0.774	0.821	0.841	0.849	0.852	0.820	0.812
$F_{1\text{-score}}$	MSSMOTE-CNN	0.781	0.913	0.927	0.946	0.971	0.967	0.952
	SSMOTE-CNN	0.781	0.838	0.845	0.867	0.882	0.866	0.858
	MSMOTE-CNN	0.781	0.818	0.857	0.875	0.874	0.845	0.822

总之,相比较于其他方法, MSSMOTE 方法可以实现更大比例的数据扩充,仿真结果表明:

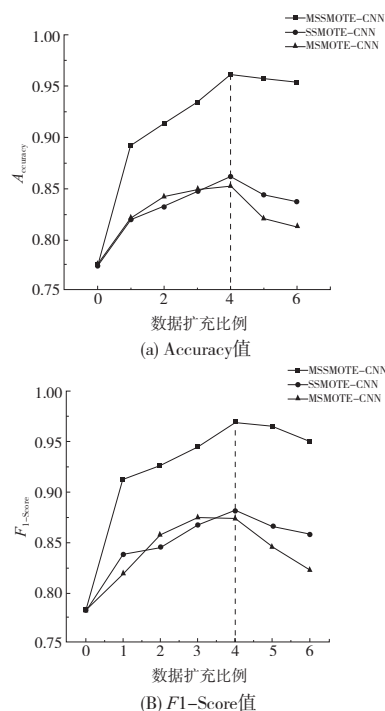


图 4 不同扩充比例下 3 种模型的对比结果

详细数据结果如表 4 所示。可以发现,在扩充比例从 0~1 的变化阶段, MSSMOTE-CNN 方法的性能提升得更多,在 Accuracy 指标中上升了 15.1%,而 SSMOTE-CNN 和 MSMOTE-CNN 方法的性能在 Accuracy 指标中分别上升了 5.8% 和 6.0%; MSSMOTE-CNN 方法的性能在 $F1$ -score 指标中提升了 16.9%,而 SSMOTE-CNN 和 MSMOTE-CNN 方法的性能在 $F1$ -score 指标中分别提升了 7.3% 和 4.7%; MSSMOTE-CNN 在扩充比例上升到 4 时, Accuracy 和 $F1$ -score 两项指标达到最高值,分别为 0.961 和 0.971。达到最大值之后 MSSMOTE-CNN 还可以保持较稳定的性能。

MSSMOTE-CNN 模型在解决冷水机组故障诊断问题上具有出色的表现。

3) 按照数据类型对 MSSMOTE 方法的性能测试。

为了进一步分析 MSSMOTE 方法的性能,对 MSSMOTE-CNN、SSMOTE-CNN 和 MSMOTE-CNN

3 种模型在数据扩充比例为 4 时的故障诊断表现进行了详细比较。使用测试集获取了 3 种模型在每一种数据类型下的 Accuracy 和 $F1$ -score 结果,3 种模型的对比情况如图 5 所示。

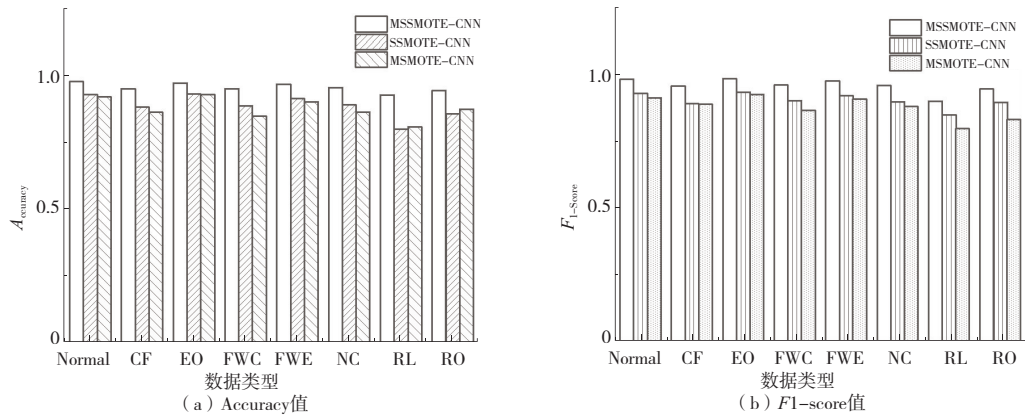


图 5 数据扩充比例为 4 时 3 种模型在不同数据类型下的比较结果

通过图 5 中 7 种故障类型的 Accuracy 和 $F1$ -score 指标比较结果可以看出,SSMOTE-CNN 和 MSMOTE-CNN 的表现相对接近;MSSMOTE-CNN 明显更高。

表 5 为在同一数据类型中 3 种方法的实验数据结果。从表 5 仿真结果数据可以看出,在 RL 故障的 Accuracy 指标中, MSSMOTE-CNN 比 SSMOTE-CNN 提高了 15.9%, MSSMOTE-CNN 比 MSMOTE-CNN 提升了 14.7%。在其他数据类型

中,相较于 SSMOTE-CNN 和 MSMOTE-CNN 两种方法, MSSMOTE-CNN 的 Accuracy 指标都有不同程度的提升。而在 CF 故障的 $F1$ -score 指标结果中, MSSMOTE-CNN 比 SSMOTE-CNN 提升了 7.3%;在 RO 故障的 $F1$ -score 结果中, MSSMOTE-CNN 比 MSMOTE-CNN 提高了 13.7%。在其他数据类型中, MSSMOTE-CNN 比 SSMOTE-CNN 和 MSMOTE-CNN 在 $F1$ -Score 上都有不同程度的提升。

表 5 在同一数据类型中 3 种方法的实验数据结果

指标	模型	数据类型							
		Normal	CF	EO	FWC	FWE	NC	RL	RO
$A_{accuracy}$	MSSMOTE-CNN	0.976	0.950	0.971	0.949	0.967	0.954	0.926	0.942
	SSMOTE-CNN	0.928	0.881	0.929	0.885	0.912	0.889	0.799	0.856
	MSMOTE-CNN	0.919	0.861	0.927	0.847	0.901	0.862	0.807	0.873
$F1$ -Score	MSSMOTE-CNN	0.983	0.957	0.985	0.961	0.976	0.959	0.901	0.946
	SSMOTE-CNN	0.929	0.892	0.935	0.902	0.921	0.899	0.849	0.897
	MSMOTE-CNN	0.914	0.889	0.927	0.867	0.908	0.881	0.798	0.832

通过上述分析可以发现, MSSMOTE-CNN 模型在每一种数据类型中都具备更高的 Accuracy 和 $F1$ -score 值。

3 结语

本文为了解决冷水机组数据类别不平衡问题,利用 MSSMOTE 方法实现数据的有效扩充。训练结果显示:在扩充比例为 4 时,对于正常样本

测试的 Accuracy 和 $F1$ -score 值分别达到 0.961 和 0.971,所构建的 MSSMOTE-CNN 模型具有更好的故障诊断性能。但是从仿真结果数据可以看出,仍有一定的提升空间,并且本文仿真分析所采用的是公开数据集,必然会受到一些相关外在因素的影响,但本文尚未考虑这些问题。

(下转第 137 页)

- 收稿日期:2023-04-20

参考文献:

- 收稿日期:2023-04-14