一种基于不平衡数据的矿用脱介筛故障监测方法

杨　军,栗轩华,张雷云
(国能神东煤炭集团公司保德煤矿,山西 忻州　036600)

摘　要:矿用设备故障监测对提高煤矿生产的连续性和安全性具有重要意义,但是由于实际工业过程中故障数据稀少且难以采集,造成了不平衡问题,对基于数据的故障监测算法实际应用不利。本文针对实际设备故障监测中少数类样本集的类内不平衡问题,提出改进的加权过采样算法。算法以 Borderline-SMOTE 为基础,基于 K 近邻样本的分布情况,以边界样本作为根样本进行加权过采样,利用 LOF 实现异常新样本的识别,提高生成样本的准确率。在实际矿用精煤脱介筛上进行了故障监测实验表明,与传统过采样方法新疆比,本文方法能获得更好的精度和分类效果。

关键词:LOF;Borderline-SMOTE;不平衡数据分类;过采样;脱介筛故障监测
中图分类号:TD94　　**文献标志码**:A　　**文章编号**:1006-6772(2024)S2-0078-04

Fault monitoring method for mining demineralization screens based on imbalanced data

YANG Jun,LI Xuanhua,ZHANG Leiyun
(*Baode Coal Mine of CHN Energy Shendong Coal Group Co.,Ltd.,Xinzhou*　036600,*China*)

Abstract:Fault monitoring of mining equipment is crucial for enhancing the continuity and safety of coal mining operations. However, the scarcity and difficulty in collecting fault data in actual industrial processes create imbalance issues, which hinder the practical application of data-driven fault monitoring algorithms. This paper addresses the intra-class imbalance problem in real equipment fault monitoring by proposing an enhanced weighted oversampling algorithm. The algorithm is based on Borderline-SMOTE, which uses the distribution of K-nearest neighbor samples to perform weighted oversampling with borderline samples as the root samples. It utilizes LOF for the identification of new outlier samples, thereby improving the accuracy of the generated samples. Experimental fault monitoring on an actual mining demineralization screen demonstrates that the proposed method outperforms traditional oversampling approaches in terms of accuracy and classification effectiveness.

Key words:LOF;borderline-SMOTE;imbalanced data classification;oversampling;fault monitoring of mining demineralization

0　引　　言

脱介筛是一种典型工业设备,主要用于不同工业场所进行脱水、脱泥或脱介。在砂石料厂洗沙选煤厂煤泥回收、选矿厂矿干排中都有应用。在选煤厂应用中,脱介筛作为选煤工艺的重要组成部分,若出现故障不能正常工作则整个选煤系统都无法生产,造成经济损失。因此对脱介筛的故障监测是十分必要的[1]。

随着机器学习理论的不断进步,数据集的质量对机器学习算法的准确性产生了越来越重要的影响。在实际应用中,数据集常常面临多种挑战,如数据量不足、标签缺失或错误、噪声以及数据不平衡等。本文以数据不平衡问题为研究对象,即数据集中某一类别的样本数量明显少于其他类别[2-3]。在故障检测中针对故障数据的判断问题上,如果算法对少数类样本识别准确率低,将造成严重的后果[4]。在机器学习领域,传统分类算法在不平衡数据集上的效果并不理想[5]。这是由于传统分类算法以整体分类准确率为优化目标,因此当样本数量不均衡时,分类器往往会偏向于预测多数类。此外,少数类样本的不足还可能导致特征表示不充分,使得模型容易过拟合且对噪声敏感。因此,为提高算法对少数类样本的识别能力,需要对不平衡数据进

收稿日期:2024-08-22;**责任编辑**:戴春雷　　**DOI**:10.13226/j.issn.1006-6772.24082204
作者简介:杨 军(1983—),男,陕西神木人,工程师。E-mail:junyang66@126.com
引用格式:杨军,栗轩华,张雷云.一种基于不平衡数据的矿用脱介筛故障监测方法[J].洁净煤技术,2024,30(S2):78-81.
YANG Jun,LI Xuanhua,ZHANG Leiyun.Fault monitoring method for mining demineralization screens based on imbalanced data[J].Clean Coal Technology,2024,31(S2):78-81.

行有针对性的处理。

目前,针对数据不平衡问题主要存在两类解决方法:数据层面方法和算法层面方法。作为数据层面方法的代表,数据预处理方法独立于分类器,通过改变原始数据的分布来调整不同类别样本的比例,进而提高少数类样本的识别准确率。数据预处理手段包括欠采样、过采样和混合采样算法。欠采样算法通过减少多数类样本的数量来实现类别均衡,过采样算法通过合成新的少数类样本来达到类别均衡,而混合采样算法则结合了上述两种算法的优点。然而,考虑到欠采样算法可能导致原始数据的特征丢失,目前过采样算法的研究相对更为广泛和深入。

SMOTE 是目前最受欢迎的过采样技术之一,它通过在少数类样本及其邻近样本之间进行线性插值来生成新样本,以达到平衡不同类别数据量的目的[6]。然而 SMOTE 在生成新的少数类样本时,仅考虑了类别不平衡率和少数类样本间的距离关系,而忽略了正负类样本的分布等其他重要因素。这种局限性可能会导致噪声样本的影响被放大,同时弱化样本分类边界[7]。因此,减少 SMOTE 生成的噪声样本量以及增强决策边界成为了当前热点研究方向[8]。为减少噪声样本生成,许多研究采用了与噪声过滤相结合的策略。

LOF 是一种常用异常点检测算法,本文利用 LOF 对 Borderline-SMOTE 生成的新样本进行筛选,减低过采样算法生成异常样本的风险。

本文针对实际工业设备故障检测中广泛存在的故障样本稀少、难以采集,最终形成不平衡数据集的问题,提出一种基于 LOF 的 B-SMOTE,利用 LOF 实现新异常样本的筛选。并根据 K 近邻设计权重系数对少数类边界样本进行过采样,生成均衡有效的训练数据集。通过在实际脱介筛的实验证明,该方法能够在不平衡数据集的基础上,有效提高设备故障监测的准确性。

1　算法基础

1.1　LOF 算法

LOF[10](Local Outlier Factor)是一种基于密度的离群点检测算法,属于无监督方法。LOF 通过计算每个样本相对于其近邻的局部密度来发现局部异常点,是一种无监督学习方法。

局部可达距离的定义如图 1 所示,如果 p_2 距离 o 较远,那么两者之间的可达距离就是它们的实际距离。如果距离足够近,如点 p_1,实际距离将被 o 的第 k 距离代替。

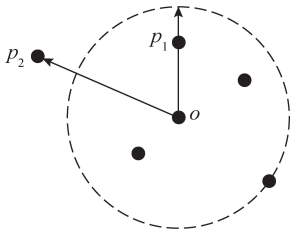

图 1　局部可达距离

该处理可以显著减少所有接近 o 的 p 点统计波动 $d(p,o)$,并通过参数 k 来控制,k 值越高,同一邻域内的点的局部可达距离越相似。点 o 到 x 的第 k 局部可达距离计算方式如下:

$$reach_dist_{k(x,o)} = Max\{d_k(x), dist(x,o)\}, \quad (1)$$

式中,$d_k(x)$ 为 x 的第 k 近邻到 x 的欧式距离;$dist(x,o)$ 为点 o 到 x 的欧氏距离。

数据点 x 的第 k 局部可达密度 $lr\ d_k(x)$ 计算为点 x 的第 k 距离邻域内的所有点到点 x 的平均第 k 可达距离的倒数。它表征了点 x 的局部密度情况,点 x 与周围点密集度越高,各点的可达距离越可能是较小的各自的第 k 距离,lrd 值越大;点 x 与周围点的密集度越低,各点的可达距离越可能是较大的两点间的实际距离,lrd 值越小。$lr\ d_k(x)$ 的计算方式如下:

$$lr\ d_k(x) = 1/\left(\frac{\sum_{y\in N_k(x)} reachdist_k(x,y)}{|N_k(x)|}\right), \quad y \in N_k(x), \quad (2)$$

点 x 的 LOF 值为 x 的局部可达密度与 x 的 k 个邻居的平均距离之比。LOF 值越接近 1 表示该点越可能为正常样本,LOF 值越大于 1 表示该点越可能为离群点。$LOF_k(x)$ 的计算公式如下:

$$LOF_k(x) = \frac{\sum_{o\in N_k(x)} lr\ d_k(y)}{|N_k(x)|} lr\ d_k(x), \quad y \in N_k(x), \quad (3)$$

根据局部可达密度的定义,如果某数据点跟其他数据点越疏远,它的局部可达密度就越小。该处理使得 LOF 算法衡量某个数据点的异常程度时,主要关注周围邻近数据点的相对密度,并能在数据分布不均匀,密度差异较大的情况下,有效地发现局部离群点。

1.2　Borderline-SMOTE 算法

Borderline-SMOTE[9]是一种针对边界处的少数类样本进行过采样的方法,主要解决 SMOTE 处理类别不平衡问题时生成样本过于泛化,难以保证样本生成质量的问题。B-SMOTE 使学习算法更关注分类边界,提高对正类的识别能力,同时根据样本的

危险程度有选择性地进行过采样,减少由过采样带来的过拟合风险。

B-SMOTE 通过统计每个少数类样本周围的多数类数量,将少数类样本区分为 3 类,"危险"样本 D,"安全"样本 S 和"噪声"样本 N。并将"危险"样本 D 视为边界样本,作为生成新样本的起点,在 D 和其最近的少数类邻居样本之间线性插值合成新样本。

本文利用 LOF 算法去除新生成的噪声样本。设计了一种针对边界样本的加权采样策略,达到更好地利用边界样本的分布信息,合成更有效的少数类样本的目的。

2　基于 LOF 去噪的加权 B-SMOTE 方法

B-SMOTE 基于样本 K 近邻分布情况确定边界样本,对边界样本采用均匀权重进行新样本生成。本文算法针对少数类边界样本广泛存在的类内不均衡问题,利用 LOF 估计少数类边界样本的密度分布特征,与 B-SMOTE 结合,为边界样本分配权重并进而合成少数类样本。

算法基本原理如下所示:

X 为两类数据的不平衡数据集,一类是多数类数据X_{maj},其中包括 p 个多数类样本;另一类是少数类数据X_{min},其中包括 q 个少数类样本,显然有 $p \gg q$。需要利用过采样算法生成的样本数量为 n。

1) 按照 B-SMOTE 方法,找到少数类样本x_i($x_i \in X_{min}$)的 m 个近邻,记作$X_{i;m}$。对每个x_i,根据其近邻中多数类的样本个数,按照$\frac{m}{2} \leqslant \{X_{i;m} \cap X_{maj}\} < m$,$0 \leqslant \{X_{i;m} \cap X_{maj}\} < \frac{m}{2}$,$\{X_{i;m} \cap X_{maj}\} = m$ 的规则,分别将x_i划分三类:"DANGER","SAFE"以及"NOISE"。并将 DANGER 类样本组成的集合记作X_{dan}。

2) 遍历 X_{dan},以线性插值的方式生成新的少数类实例。对于每个 x_d,找到样本在总数据集中 X 的 m 个近邻,记作 $X_{d;m}$。

3) 计算每个新样本的 LOF 值,并进行排序。

4) 删除排名靠前的 α%新样本。

5) 回到步骤 2)再次进行样本生成,一共循环 n 次。

6) 按照采样率设定,选择所有剩余新样本中 LOF 值最低的样本,与原样本组成新样本集。完成过采样。

3　脱介筛故障监测试验

3.1　脱介筛故障监测问题

由于故障数据的稀缺性,导致故障诊断算法在实际应用领域常面临数据集不平衡的问题,矿用设备故障诊断亦如此。为验证本文所提方法的有效性,本文利用某煤矿的精煤脱介筛为实验对象。精煤脱介筛上共安装 6 个传感器监测电机和激振器,分别在电机驱动端、自由端、1 号激振器主动轴、1 号激振器从动轴、2 号激振器主动轴和 2 号激振器从动轴上。其中,1 号激振器主动轴上的传感器检测到了故障。对这期间的 6 个传感器数据进行采集,测量的数据量包括速度均方根、电压、温度与振动加速度。由于这四个特征数据的采样时间不完全同步,以振动加速度的时间间隔为基准,对间隔期间的其他测量数据加和求均处理,得到样本大小为 476 的数据集,共两种样本类型(故障与正常),其中正常样本 379 个,故障样本 97 个。正常与故障数据的不平衡比为 3.91。为验证故障数据缺少条件下,本文算法有效性,选择正常样本 125 个,并 3 次选择不同故障数据 20 个,组成不平衡的训练数据集,验证本文算法的有效性。

在该数据集上,采用了多种过采样算法,包括 SMOTE、B-SMOTE,ADASYN 并与本文所提出的方法进行了对比。选择支持向量机(SVM)作为基础分类器。

为了平衡数据集中的类别分布,将过采样环节的平衡比设置为 50:50。同时,将 SMOTE 等过采样的 k 近邻参数设置为 5。将选取多数类邻居时随机系数 α 的最大值设定为 20,循环采样次数 n 设为 3。本文采用 Gmean 值作为在不平衡数据条件下故障监测的准确性。

3.2　脱介筛故障监测问题

在该数据集上进行实验的 F1 值与 G-mean 结果对比如图 2 所示。

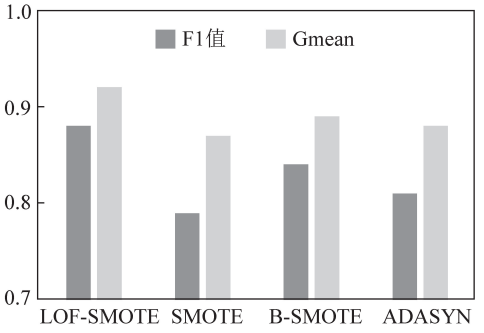

图 2　不同过采样算法的故障监测准确率对比

通过对数据进行分析,可以观察到本文算法可以有效提高分类准确率,在 F1 值和 Gmean 两个指标上,本文算法都获得了显著提高。

4 结　论

1)针对实际工业设备的故障监测中数据不平衡问题,本文提出了一种基于局部离群因子(LOF)的边界样本加权过采样方法。本文首先分析了当前 LOF 在过采样中的应用状态,对其在现有方法中的使用 合理性进行了评估。然后,我们提出利用 LOF 作为密度估计方法的独特视角,阐述了其合理性。并基于该视角将 LOF 算法引入到少数类边界样本的密度分布评估中。

2)在传统 B-SMOTE 方法以最近邻方法确定边界 样本的基础上,进一步利用 LOF 值获取少数类边界样本的密度分布信息,设计了相应的加权过采样算法。脱介筛故障监测的实验结果表明,本文方法在生成新样本时,能够充分利用局部密度信息。与传统 BSMOTE 方法相比,本文方法在故障分类性能上取得了较大的提升。同时,与其他常见的过采样方法相比, 本文方法总体上展现出优势,能够生成更加有效的合成样本,且运算时间成本在可接受范围内,且比传统 SMOTE 与 LOF 结合的算法有大幅降低。

参考文献(References):

[1] MOHAMAD M, SELAMAT A, SUBROTO I M, et al. Improving the classification performance on imbalanced data sets via new hybrid parameterisation model [J]. Journal of King Saud University - Computer and Information Sciences, 2021, 33(7): 787-797.

[2] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[3] LIU N, LI X, QI E, et al. A novel ensemble learning paradigm for medical diagnosis with imbalanced data [J]. IEEE Access, 2020, 8: 171263-171280.

[4] LI J, LIU Y, LI Q. Generative adversarial network and transfer-learning-based fault detection for rotating machinery with imbalanced data condition [J]. Measurement Science and Technology, 2022, 33(4): 045103.

[5] SPELMEN V S, PORKODI R. A review on handling imbalanced data [C]//2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). Coimbatore, IEEE, 2018: 1-11.

[6] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[7] GYOTEN D, OHKUBO M, NAGATA Y. Imbalanced data classification procedure based on SMOTE [J]. Total Quality Science, 2020, 5(2): 64-71.

[8] FERNANDEZ A, GARCIA S, HERRERA F, et al. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary [J]. Journal of Artificial Intelligence Research, 2018, 61: 863-905.

[9] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]//International conference on intelligent computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 878-887.

[10] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers [C]//Proceedings of the 2000 ACM SIGMOD international conference on Management of data. Dallas Texas USA. ACM, 2000: 93-104.