

融合领域知识和集成模型的挤压铸造工艺数据 正确性检测方法

邓建新^{1,2} 吴秀松^{1,2} 李修明³ 尹 政^{1,2}

(1. 广西大学广西制造系统与先进制造技术重点实验室, 南宁 530003; 2. 广西大学机械工程学院, 南宁 530003;
3. 广西机械工业研究院有限责任公司研发中心, 南宁 537000)

摘要 为构建高质量的挤压铸造工艺数据集, 保证数据驱动模型的质量, 针对多源挤压铸造工艺数据存在的异常数据问题, 提出了融合挤压铸造领域工程知识, 集成多个检测模型的数据正确性检测方法。首先结合挤压铸造工程知识构建挤压铸造工艺数据属性取值规则, 建立基于属性取值规则的异常判定模型; 然后基于集成学习思想, 使用投票法建立基于局部离群因子、箱型图和孤立森林模型的挤压铸造工艺数据的异常集成检测模型, 并与属性值取值规则结合形成两阶段检测方法, 实现对挤压铸造工艺数据集从工程知识和数据特征两个维度的数据正确性检测。结果表明, 所提出的方法提高了泛化能力, 能对挤压铸造工艺数据集中的异常数据进行有效识别, 平均召回率达95%。

关键词 挤压铸造; 数据正确性; 异常检测; 领域知识; 集成模型

中图分类号 TG249.2; TH166; TP391

文献标志码 A

DOI: 10.15980/j.tzzz.T20240071

Correctness Detection Method of Squeeze Casting Process Data Based on Domain Knowledge and Integrated Model

DENG Jianxin^{1,2}, WU Xiusong^{1,2}, LI Xiuming³, YIN Zheng^{1,2}

(1. Guangxi Key Laboratory of Manufacturing System & Advanced Manufacturing Technology, Guangxi University, Nanning 530003; 2. School of Mechanical Engineering, Guangxi University, Nanning 530003;
3. Guangxi Research Institute of Mechanical Industry Co., Ltd., Nanning 537000)

Abstract: To construct a high-quality squeeze casting process dataset and ensure the quality of data-driven model, a data correctness detection method was proposed to address the problem of anomaly data in multi-source squeeze casting process data, which integrates engineering knowledge in the field of squeeze casting and integrates multiple detection models. Firstly, the rule of attribute value of squeeze casting process data was constructed based on the engineering knowledge of squeeze casting, and the anomaly judgment model was established. Then, based on the idea of ensemble learning and voting method, an integrated detection model of abnormal squeeze casting process data based on local outlier factor, boxplot, and isolation forest was established, and combined with the attribute value rules to form a two-stage detection method, which realizes the data correctness detection of squeeze casting process dataset from two dimensions of engineering knowledge and data characteristics. The results reveal that the proposed method can improve the generalization ability and effectively identify anomaly data in the squeeze casting process dataset, and the average recall rate can reach up to 95%.

Key Words: Squeeze Casting, Data Correctness, Anomaly Detection, Domain Knowledge, Integrated Model

当前, 数据驱动科学正成为继试验科学、理论推导和模拟仿真之后的第四范式^[1], 数据驱动方法成为各领域新的发展基础和方向。材料领域研究者正通

过利用以往的试验数据实现新材料的快速设计, 有效解决传统试验研究模式存在的周期长、成本高等问题^[2], 在制造领域数据驱动制造也成为实现智能制造

收稿日期: 2024-02-29; 修订日期: 2024-11-04

基金项目: 国家自然科学基金资助项目(51965006); 广西自然科学基金资助项目(2018GXNSFAA050111); 国家金属材料近净成形工程技术研究中心开放基金课题资助项目(2019001)。

第一作者简介: 邓建新, 男, 1979年出生, 教授, E-mail: dengjxin@163.com;

引用格式: 邓建新, 吴秀松, 李修明, 等. 融合领域知识和集成模型的挤压铸造工艺数据正确性检测方法[J]. 特种铸造及有色合金, 2025, 45(9): 1281-1290.

DENG J X, WU X S, LI X M, et al. Correctness detection method of squeeze casting process data based on domain knowledge and integrated model[J]. Special Casting & Nonferrous Alloys, 2025, 45(9): 1281-1290.

的一条路径^[3]。但数据驱动方法的性能严重依赖于其驱动数据样本的质量^[4](正确性、一致性、低缺失等),较差的数据质量不仅导致错误的知识获取,影响数据驱动模型、方法的正确性、学习效果和有效性^[5-6],甚至可能造成严重后果和巨大损失^[7],如基于有错误数据的数据样本训练得到的机器学习预测模型,预测结果与实际差异过大,无法应用^[8];据统计,较差的数据质量甚至造成了美国零售企业平均每年超过 1 500 万美元的损失^[9]。

挤压铸造是一种融合了铸造和锻造优势的零件近净成形制造工艺。为了适应智能制造的发展趋势,需要基于已有工艺数据建立数据驱动的工艺参数设计方法,解决当前工艺参数依赖“试错法”设计成本高、效率低等问题,从而提高挤压铸造工艺的生产效率和制件质量。数据驱动方法依赖的工艺数据可来源于物理试验或数值模拟试验记录、理论计算、工业生产和已有文献等^[10-12],但因生产环境、试验方案、测量手段等存在差异,以及数据采集、记录和发布过程中的误操作等因素,不可避免会导致从多源渠道收集到的挤压铸造工艺数据彼此间存在冲突,或与真实数据有一定误差,即某些数据可能存在异常或不正确,如有关研究^[13-14]均采用物理试验的方法开展 AZ31 镁合金挤压铸造工艺参数优化研究,但得出挤压压力和保压时间的最优值分别相差 50 MPa 和 20 s。为保证构建的挤压铸造数据驱动工艺参数设计模型的质量,对其依赖数据进行有效的正确性检测,进而清洗或提升至关重要。

为了实现数据驱动应用,通常从不同渠道收集数据或采用不同的数据增强算法来产生新的数据^[15-16],但这些数据的质量良莠不齐。当前,不同领域研究者逐渐意识到数据质量的重要性,并采用基于统计分析、聚类、基于密度、最近邻等异常检测技术来评价数据正确性,以识别和剔除(清洗)异常数据,从而获取对象的正确认识、知识和模型等。主要出现了基于密度的局部离群因子(Local Outlier Factor, LOF)^[17]、基于隔离思想的孤立森林算法(Isolation Forest, IForest)^[18]和基于统计分析的箱型图(Boxplot)^[19]等典型的数据异常(即不正确的数据)检测方法。LOF 由 BREUNIG M M 等^[17]首次提出,其核心思想是数据点异常与否取决于局部环境,即一个样本点所在位置的密度与该样本点周围的样本所处位置的平均密度相比,差距越大则越可能是异常点。LOF 因其简单有效被广泛应用于现实场景下的异常数据检测,如工业过程故障检测^[20-21]、网络入侵检测^[22-23]等。何九虎等^[20]采用 LOF 算法对工业过程异常数据进行检测,并利用 CDC(Closest Distance

to Center)采样点预处理方法,数据仿真表明其效率有所提高,可满足该领域的实时性要求。针对多工况数据,冯立伟等^[21]采用 LOF 算法结合统计模量对故障样本进行检测,发现其检测率比主元分析(PCA)、核主元分析(k PCA)和基于 k 近邻等检测方法均更高。LUO Y 等^[22]设计了一种基于 LOF 的检测方法用于识别网络攻击,并通过实例验证了该方法的可用性。吴俊^[23]基于 LOF 和 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)构建了一种网络入侵检测模型,经入侵试验证明其具有低误报率、高检测率的特性。此外,为提高 LOF 算法的异常检测效率, TU B 等^[24]提出了融合谱角和 LOF 的 SALOF(Spectral Angle and the Local Outlier Factor)算法,依据事先设置好的分割阈值筛选计算得到数据点的异常概率。刘芳等^[25]通过设计融合索引结构和多层 LOF 上界的多粒度剪枝方法,提出了 MTLOF(Multi-Granularity upper Bound Pruning Based Top-n LOF Detection)算法。IForest 是由 LIU F T 等^[18]提出的一种无监督异常点检测集成方法,其利用异常数据普遍存在的“少而不同”的特点,将异常数据定义为容易被孤立的离群点进行识别,因其时间复杂度低、算法简单高效而广受关注,广泛应用于金融^[26-27]、医疗^[28]、网络安全^[29-30]和工业^[31]等领域。汤淳等^[26]利用 IForest 算法同时对 4 种测度指标进行异常值识别,并根据多数投票原则,将 2 种以上测度同时出现异常值的样本视为极端系统性金融风险,实例说明该算法能够很好地适用于系统性金融风险分类。熊文滔等^[28]利用 IForest 算法对老年人的日常行为数据进行建模,以检测不符合正常行为模式的跌倒数据,通过公开数据集验证表明该模型的平均识别准确率可达 96.76%。针对网格攻击中的拒绝服务攻击,陈佳等^[29]将 IForest 算法应用于网络数据异常检测,具有较高的异常检测性能。WANG H 等^[31]将 IForest 算法应用于船用核电汽轮机系统级、设备级和参数级的异常数据检测与定位,其对每个异常状态的查准率超过 94%。STAERMAN G 等^[32]将 IForest 算法拓展到功能数据的异常检测,并通过数值模拟验证了算法的有效性和准确性。箱型图是一种用作显示一组数据分散情况的统计图,用于反映数据分布特征,由 TUKEY J W^[33]首次提出。因其简单易用和良好的抗掩蔽效应性能,箱型图被广泛应用^[34-35]。顾国庆等^[35]将箱型图应用于构建数据预测模型的数据预处理阶段,过滤传感节点采集的异常数据点,发现经异常点过滤处理的预测模型预测精度提高了 8.4%。

综上,现有数据异常检测通常只采用单一的异常

检测方法对数据样本进行检测,但单一异常检测方法只能获得数据集上单角度的测度,泛化能力不足,适用场景有限。如IForest方法虽然无需数据先验知识,计算复杂度低,但其设计原理决定了该方法只适用于低异常程度的数据集,对异常数据比例较高的数据集则效果不佳,同时其结果不够直观和易解释;基于密度的LOF方法虽然结果可解释性好,但其性能严重依赖参数选择,且计算复杂度较高,处理高维数据集效果不佳。Boxplot方法具有非参数性,适用于任何类型的数据分布,且结果直观易比较,但其受数据量限制,当数据量很小时,结果往往不够精确。其次,单一的异常检测方法对所有数据均采用同一种异常评价标准,无法综合考虑数据的总体和局部信息,可能导致检测精度不足和效率不高等问题。如IForest方法异常检测过程均采用全局异常标准,对局部异常点检测效果不佳,而LOF方法则是通过计算给定数据点的局部偏差来识别异常点,拥有充分检测局部异常点的优势。再次,现有检测方法主要从数据分布特性出发对数据进行正确性检测,很少融入数据涉及的工程知识来帮助检测,导致难以发现有悖于领域知识的异常数据,如在挤压铸造工艺数据中,对液态挤压铸造,其浇注温度要高于金属的液相线温度,而对半固态挤压铸造,其浇注温度应处于固液相线温度之间,两者虽存在明确分界,但其数据分布可能相近,仅从数据分布特性出发则难以发现异常。

基于此,本研究面向挤压铸造工艺的数据驱动研发需求,为保证其依赖的多源、异构的工艺数据的质量,基于挤压铸造工艺特点,提出集成LOF、Boxplot和IForest算法,并融合挤压铸造领域知识的工艺数据正确性检测方法(LBI-Based Data Correctness Detection Method Incorporating Squeeze Casting Domain Knowledge, LBISCDK),该方法综合考虑不同异常检测模型的优缺点,并融合挤压铸造领域知识,从工艺数据本身和领域知识两个角度对挤压铸造工艺数据进行正确性检测,提高对异常数据的检查能力和检测效率,可为后续的数据驱动应用提供高质量的数据样本,保证数据驱动应用的质量。

1 挤压铸造工艺数据正确性检测方法

1.1 挤压铸造工艺数据

挤压铸造工艺数据主要包含工艺参数数据、工艺影响因素数据和铸件性能数据,其中工艺参数数据主要包括浇注温度、挤压压力、模具预热温度等,工艺影

响因素数据主要由影响工艺参数的因素组成,包括材料成分和铸件形状特征参数等,铸件性能数据则由不同铸件的力学性能等组成,如硬度、抗拉强度等,对应类别在数据集都可称为属性。设建立的挤压铸造工艺数据集为 DT ,有 n 个样本, m 个属性; $S_i(i=1,2,\dots,n)$ (表示第 i 个数据样本); A_j 为第 j 个属性; a_{ij} 为第 i 个数据样本对应的第 j 个属性数据,则 $DT=\{a_{ij}|i=1,2,\dots,n,j=1,2,\dots,m\}$,其结构见表1。

表1 挤压铸造工艺数据集
Tab. 1 Squeeze casting process dataset

编号	A_1	A_2	$\dots$	A_m
S_1	a_{11}	a_{12}	$\dots$	a_{1m}
S_2	a_{21}	a_{22}	$\dots$	a_{2m}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
S_n	a_{n1}	a_{n2}	$\dots$	a_{nm}

1.2 基于领域知识的正确性检测

挤压铸造工艺数据中的每个属性都具有特定的含义、数据类型、取值范围和量化计算方法,如浇注温度必须高于材料的液相温度。对某一特定样本,需要判断其属性值的数据类型、取值范围是否与领域工艺知识一致。因此定义了挤压铸造工艺数据属性值的取值规则来检测数据的正确性。

将挤压铸造工艺数据表示为一个三元组 $\langle A_s, T, V \rangle$,其中 A_s 为属性, T 为其数据类型, V 为数据范围, $A_s=\{A_1, A_2, \dots, A_j, \dots, A_m\}$; $T=(t_1, t_2, \dots, t_j, \dots, t_m)$, t_j 为第 j 个属性的数据类型; $V=\{v_1, v_2, \dots, v_j, \dots, v_m\}$, $v_j=\langle v_j^{\min}, v_j^{\max} \rangle$ 为第 j 个属性的经验取值范围, $v_j^{\min}$ 和 $v_j^{\max}$ 分别为经验取值的最小值和最大值。给定 DT 中 a_{ij} ,可根据下式判断其是否为异常数据点:

$$F(a_{ij}) = \begin{cases} 0, & \text{type} = t_j \text{ 且 } a_{ij} \in [v_j^{\min}, v_j^{\max}] \\ 1, & \text{其他} \end{cases} \quad (1)$$

式中, $\text{type}(a_{ij})$ 为获取 a_{ij} 的数据类型,如果 a_{ij} 的数据类型 $\text{type}(a_{ij})$ 与其属性的经验取值类型 t_j 相匹配且 $a_{ij} \in [v_j^{\min}, v_j^{\max}]$,则将其判定为正常数据,结果记为0,否则判定为异常数据,记为1。

挤压铸造工艺数据属性值的数据类型和经验取值范围等通过挤压铸造工艺相关手册、文献等获取,如通常结合铸件形状,依据近液相线法选择比合金液相线温度 T_l 高50~100℃的温度作为挤压铸件的浇注温度 T_p ,且控制在较低值,以减少能源浪费^[12],因此一般设定挤压铸造浇注温度数据类型为浮点型,其经验取值

范围为高于合金液相线温度 50~100 ℃,即确定其取值规则如下式:

$$T_p \in [T_l, T_l + 100] \quad (2)$$

1.3 基于集成模型的正确性检测

1.3.1 基础模型构建

(1)LOF 模型 LOF 主要聚焦于数据点的邻域密度,将数据视为空间中的点,具有低密度的数据点视为异常点。如在以两个数据属性为坐标轴形成的二维空间中(见图1), C_1 和 C_2 为两个正常的数据点簇,点 O_1 、 O_2 为两个离群的异常数据,其数据密度明显不同。因此如果一个样本点的密度比较小,而它周围临近样本点的密度都比较高,那么其可能是异常点(见图1中 O_1)。在本研究中,将挤压铸造工艺数据的属性抽象为坐标轴,形成LOF中多维空间的工艺数据样本点。为检测其中的异常样本,则需要计算数据密度。

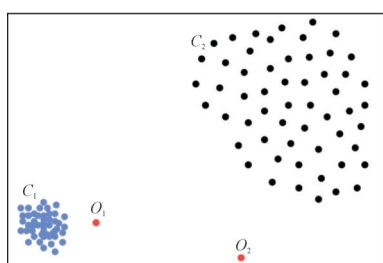


图1 含异常数据的二维数据集

Fig.1 Two-dimensional data set with abnormal data

为此,对转化为多维空间的挤压铸造工艺数据点簇,计算样本 S_i 到其他样本的欧式距离,并按从小到大排序,记序列中第 k 个距离为 $k_dist(S_i)$;然后得到小于 $k_dist(S_i)$ 的所有样本集合,记 S_i 的 k 距离邻域为 $N_k(S_i)$;取 S_i 到其他任意样本 $S_{i'}$ 的距离和 $k_dist(S_i)$ 中的较大值为样本 S_i 的可达距离 $reachdist_k(S_i, S_{i'})$,记 $N_k(S_i)$ 中的所有样本到样本 S_i 的平均可达距离的倒数为 S_i 的局部可达密度 $lrd(S_i)$ 。

在此基础上,通过计算样本点 S_i 的邻域点 $N_k(S_i)$ 的局部可达密度的平均值与点 S_i 的局部可达密度之比来判断 S_i 的异常程度,该值越大表明该样本点越可能是异常,见图2。

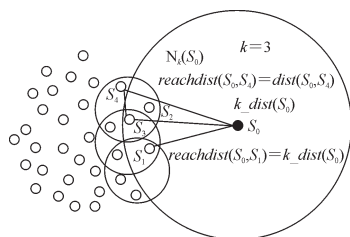


图2 LOF 示意图

Fig.2 Schematic diagram of LOF

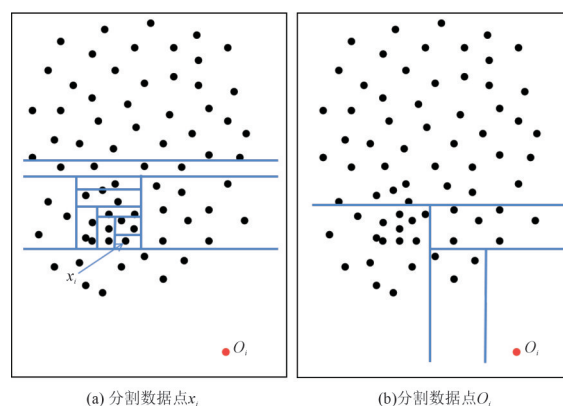
构建的检测挤压铸造工艺数据正确性的LOF模型的伪代码见表2。

表2 检测挤压铸造工艺数据正确性的LOF模型伪代码

Tab.2 Pseudo-code for LOF model to check the correctness of squeeze casting process data

算法名称:局部离群因子(LOF)
输入:挤压铸造工艺数据集 DT ,邻居数 k ,异常样本阈值 e
输出:每个数据样本的LOF得分
1.对工艺数据进行归一化处理,将各属性的取值转化至 $[0,1]$ 区间
2.对于每个数据样本 $S_i \in DT$,计算 $k_dist(S_i)$
3.找到样本 S_i 的 $N_k(S_i)$ 及计算 $reachdist_k(S_i, S_{i'})$
4.计算样本 S_i 的 $lrd(S_i)$
5.计算样本 S_i 的 $N_k(S_i)$ 中所有样本的 $lrd(S_{i'})$ 之和记为 $sumlrd_k(S_i)$
6.计算样本 S_i 的LOF得分 $lof(S_i) = \frac{sumlrd_k(S_i)}{ N_k(S_i) } \div lrd_k(S_i)$
7.根据异常样本阈值 t 标记异常数据样本
8.If $lof(S_i) > e$ then
9.将样本 S_i 标记为异常数据样本,标签记为1
10.else
11.将样本 S_i 标记为正常数据样本,标签记为0
12.end if

(2)IForest模型 IForest是一种基于树结构的异常检测方法,由预设的多棵隔离二叉树iTree组成,其利用数据点被二叉树分割的次数来度量一个点是聚集(正常)还是孤立的(异常)。图3为随机分隔数据点。从图3a可知,对于数据集簇中的正常点 x_i ,需要对其进行多次随机分割才能隔离出来;而对于异常点 O_i ,见图3b,仅需要对其进行几次分割就可以隔离出来。



(a) 分割数据点 x_i (b) 分割数据点 O_i

图3 随机分割数据点

Fig.3 Randomly segment data points

基于IForest的挤压铸造工艺数据正确性检测过程主要分为训练和评估两个阶段^[18],见图4。

训练阶段生成指定数目的孤立树,其伪代码见表3。

评估阶段给出样本的异常分数 s 。

首先计算从iTree的根节点到外部节点所经历边的数目即路径长度,记为 $h(S_i)$,对于给定样本量为 ϕ 的样本子空间和一条样本数据 S_i ,其异常分数 s 定义为:

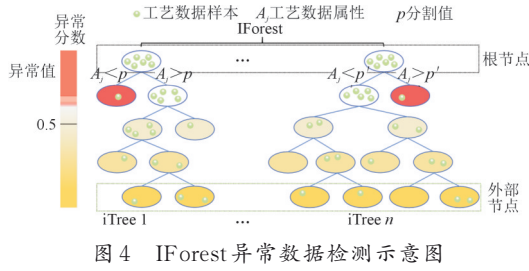


图4 IForest异常数据检测示意图
Fig.4 Schematic diagram of IForest method for detecting abnormal data

表3 IForest训练生成指定数目的孤立树

Tab.3 A specified number of isolated trees generated by IForest training

算法名称: $iTrees(DT, h, h_{lim}, n_{estimators}, \phi)$
输入: $DT, h, h_{lim}, n_{estimators}, \phi$
输出: 孤立森林 $iTrees$
1. 对工艺数据进行归一化处理, 将各属性的取值转化至 $[0, 1]$ 区间
2. if $h > h_{lim}$ or $ DT \leq 1$ then
3. return
4. else
5. for $i = 1$ to $n_{estimators}$
6. for $h = 1$ to h_{lim}
7. 从数据集 DT 随机选择一个属性 A_j 和一个分割点 p
8. 将数据集 DT 分为两个子集 DT_{left} (属性 A_j 的值小于 p)
9. 和 DT_{right} (属性 A_j 的值大于 p)
10. end for
11. end for
12. end if

注: DT 为工艺数据集; h 为当前树的高度; h_{lim} 为孤立树的限制高度; $n_{estimators}$ 为孤立树数量; ϕ 为子采样样本量。

$$s(S_i, n) = 2^{-\frac{E[h(S_i)]}{c(\phi)}} \in [0, 1] \quad (3)$$

式中, $E[h(S_i)]$ 为样本数据 S_i 在 n 棵 $iTree$ 中路径长度的平均值; $c(\phi)$ 为在二叉搜索树 (Binary Search Tree, BST) 中搜索失败的平均路径长度^[20], 在此处主要起标准化 $h(S_i)$ 的作用, 其定义为:

$$c(\phi) = 2H(\phi - 1) - 2(\phi - 1)/\phi \quad (4)$$

式中, $H(\phi) = \ln(\phi) + \xi$ 为调和级数; ξ 为欧拉常数。

数据样本的异常分数 s 越接近 1, 越有可能是异常值, 样本标签记为 1。如果大部分样本异常得分 s 都接近 0.5, 表示整个数据集中无明显的异常值, s 远小于 0.5 时, 则说明该数据样本有很大可能为正常值, 标签记为 0。

(3) Boxplot 模型 通过箱型图 (Boxplot) 可以直观了解数据的集中趋势、离散程度和异常值情况, 将异常大或异常小的极端值识别为异常值。图 5 为箱型图示意图。下四分位数 Q_1 、上四分位数 Q_3 分别指一组属性值按从小到大顺序排列后, 处于 25%、75% 位置的数值, 四分位距定义为 $IQR = Q_3 - Q_1$, 各属性上边缘为 $Q_3 + 1.5IQR$, 下边缘为 $Q_1 - 1.5IQR$, 将超过上下边缘

区间的数据识别为异常值, 标记为 1, 其余正确数据标记为 0。

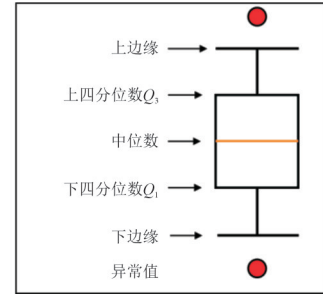


图5 箱型图示意图

Fig.5 Schematic diagram of Boxplot

基于以上原理, 构建的检测挤压铸造工艺数据正确性的 Boxplot 模型的伪代码见表 4。

表4 Boxplot 算法伪代码

Tab.4 Pseudo-code for Boxplot algorithm

算法名称: 箱型图模型 (Boxplot)
输入: 工艺数据集 DT
输出: 每个属性的 5 个统计量和数据样本标签
1. 对每个属性的数据进行排序
2. 计算每个属性的最小值 (Min)、下分位数 (Q_1)、中位数 (Median)、上分位数 (Q_3) 和最大值 (Max)
3. 计算各属性的四分位距 $IQR = Q_3 - Q_1$
4. 计算各属性的上限 $Up = Q_3 + 1.5IQR$, 下限 $Low = Q_1 - 1.5IQR$
5. if $a_{ij} > Up$ or $a_{ij} < Low$ then
6. 将数据点 a_{ij} 所属样本标记为异常, 标签记为 1
7. else
8. 将数据点 a_{ij} 所属样本标记为正常, 标签记为 0
9. end if

1.3.2 检测的集成模型

如前所述, 以上 3 种模型各有优势, 为了提高模型的泛化检测能力, 采用集成学习策略, 将以上 3 种模型视为单元模型进行集成。由于以上 3 种单元模型的输出均为数据样本的分类标签, 故采用投票法进行模型集成, 综合得到一个分类结果, 即当两个及以上基础模型均判定某个挤压铸造工艺数据样本为异常时, 则判定该样本最终为异常。设 LOF、IForest 和 Boxplot 模型的输出标记分别为 y_1 、 y_2 和 y_3 , 其输出都是 1 或 0, 集成输出结果 y 可表示为:

$$y = \begin{cases} 1, & \text{当 } y_1 + y_2 + y_3 \geq 2 \\ 0, & \text{其他} \end{cases} \quad (5)$$

1.4 融合领域知识和集成模型的正确性检测

基于以上方法原理, 提出融合领域知识和集成模型的挤压铸造工艺数据正确性检测方法 LBISCDK, 该方法流程见图 6。检测过程大致分为 3 步:

(1) 校验材料成分取值及判断样本材料系列 首先运用材料成分属性值取值规则对样本的材料成分

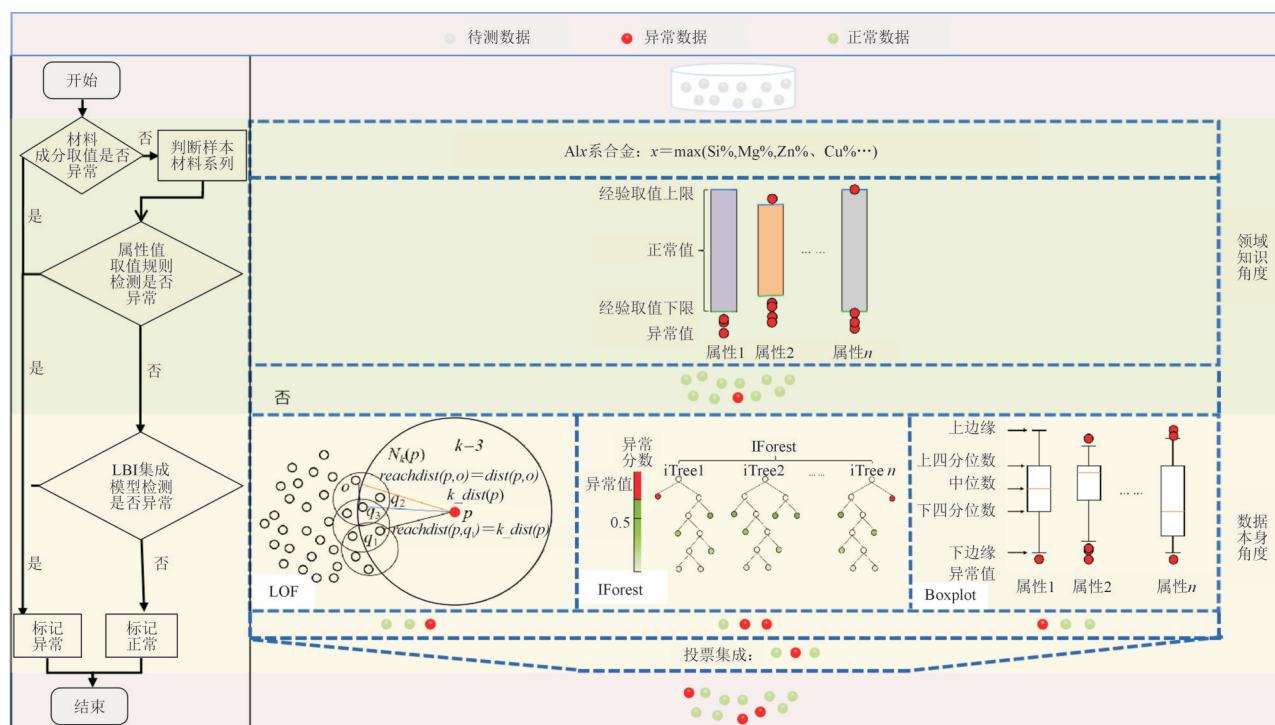


图6 LBISCDK方法检测流程
Fig.6 Detection process of LBISCDK method

数据进行检验,确保材料成分数据的正确性。对于材料成分数据取值合理正确的样本,通过比较合金成分中除主元素(如Al)外的其他元素的质量分数,取最大值判定数据样本所属材料系列,如一条数据样本的Si元素质量分数为除Al元素外的最大值,则判定该样本为Al-Si合金的挤压铸造工艺数据,并自动选择Al-Si合金系列的属性值取值规则和集成模型。

(2)基于领域知识的正确性检测 基于对应合金系列的工艺数据属性值类型、经验取值范围对样本数据进行检测,如果样本数据属性值类型相符且数值在经验取值范围内,那么将其标记为正常样本并进入下一步检测;否则将其标记为异常,结束检测。

(3)LBI集成模型检测 通过第二步检测的样本再应用集成模型LBI进行检测,输出相应结果标签。

2 运行试验

为了验证所提数据正确性检测方法的有效性,通过试验来检测数据集中的异常值,以召回率($R_{\text{召回}}$)、精确率($P_{\text{精确}}$)和正确率($A_{\text{正确}}$)作为评价指标。其中召回率表示的是数据集所有异常样本被识别为异常的比例;精确率是被识别为异常的样本中实际为异常的比例;正确率是被正确识别的样本数占总样本数的比例:

$$R_{\text{召回}} = T_p / (T_p + F_N) \quad (6)$$

$$P_{\text{精确}} = T_p / (T_p + F_p) \quad (7)$$

$$A_{\text{正确}} = (T_p + T_N) / (T_p + T_N + F_p + F_N) \quad (8)$$

式中, T_p 为检测异常且实际也为异常的样本数; F_p 为检测异常实际却为正常的样本数; T_N 为检测正常实际也为正常的样本数; F_N 为检测正常实际却为异常的样本数。

首先基于真值数据集进行异常模拟添加后得到4个异常数据集,然后采用融合挤压铸造领域知识的LBISCDK方法和未融合领域知识的方法(即LOF、IForest、Boxplot和集成检测模型LBI)分别开展异常检测,并对比几种方法的检测效果。

2.1 试验数据

试验真值数据为本研究从CNKI、Web of Science等数据库检索的挤压铸造相关文献中提取整理的4个不同系列铝合金(Al-Si系、Al-Mg系、Al-Cu系和Al-Mg系合金)的近液相线温度挤压铸造工艺数据集,共332个样本,其中每个样本包含4个工艺参数属性(浇注温度、挤压压力、保压时间和模具预热温度)、5个铸件几何特征属性(挤压方式、形状复杂度、最大壁厚、最小壁厚和材料成分)和3个质量指标属性(抗拉强度、屈服强度和缩松缩孔率),见表5。

挤压铸造工艺数据异常可能由多种因素引起,包括铸造设备不良状态或故障、生产环境条件的不稳

表 5 挤压铸造工艺真值数据

Tab.5 True value data of squeeze casting process

数据集	样本编号	浇注温度/℃	比压/MPa	保压时间/s	模具预热温度/℃	缩松缩孔率/%	...	w(Al)/%	...
Al-Si 系	1	660	60	25	220	5.06	...	92.19	...
	2	660	80	30	240	4.038	...	92.19	...
	...	...	...	...	...	...	...	...	...
	164	690	120	40	280	2.527	...	91.28	...
Al-Cu 系	1	725	90	15	225	6.04	...	93.11	...
	2	730	50	35	250	7.21	...	92.92	...
	...	...	...	...	...	...	...	...	...
	76	730	80	30	200	6.44	...	91.95	...
Al-Zn 系	1	600	100	60	250	2.99	...	90.34	...
	2	640	110	40	300	3.63	...	89.96	...
	...	...	...	...	...	...	...	...	...
	48	610	90	50	250	3.45	...	87.20	...
Al-Mg 系	1	650	75	15	250	6.32	...	93.05	...
	2	670	100	20	250	5.73	...	92.58	...
	...	...	...	...	...	...	...	...	...
	44	690	110	16	270	3.66	...	97.20	...

定、相关人员的操作不当或故意欺诈等。此外,数据提取和整理过程中出现的失误也可能导致异常数据的产生,如铝合金近液相线温度挤压铸造工艺数据集中错误地收集了铜合金的半固态挤压铸造工艺数据等,这都会影响后续数据驱动应用的有效性和准确性。基于此,通过随机数据生成和人为构造两种方法来模拟异常数据的生成,以构造出不同异常程度的待测数据集。

(1)随机数据生成 对表 5 的工艺数据真值数据集通过生成随机异常偏移量来进行异常数据模拟添加,并通过设定异常数据比例来实现生成不同异常程度的待测数据集,算法的伪代码见表 6。

(2)人为构造 通过在收集的铝合金近液相线挤压铸造工艺数据真值数据集中随机加入其他材料成分(如 Cu 合金、Mg 合金等)或其他工艺方式(如半固态挤压铸造等)的工艺数据样本,以模拟在数据收集和整理过程中因操作失误而导致数据异常的情况。

经过异常数据模拟添加后得到 4 个不同异常程度的异常数据集,见表 7。

真值数据集和异常数据集的 4 个关键属性的数据

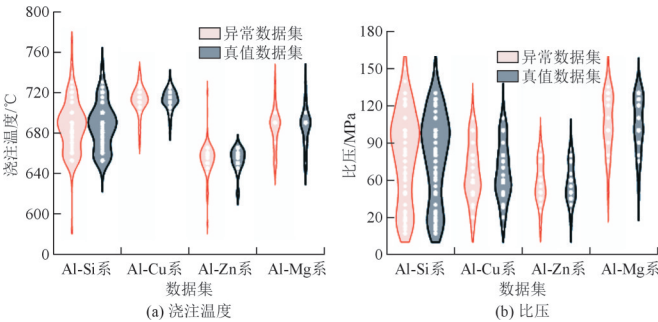


表 6 随机异常数据模拟添加算法伪代码

Tab.6 Pseudo-code for random abnormal data simulation addition algorithm

算法名称:随机异常数据模拟添加算法
输入: 真值数据集
输出: 存在点异常数据的异常数据集
1: 导入真值数据集,在数据表中添加数据标签列,并将标签初始化为 0
2: 设定异常数据比例阈值及点异常数据的添加数量
3: for $i = 1$ to 点异常数量限制
4: 随机选取属性、随机选取单点位置
5: 计算所选属性的均值 μ 和标准差 σ
6: 产生服从 $N(\mu, \sigma^2)$ 的随机数作为点异常偏移量进行单点异常模拟
7: 将此数据样本的数据标签设置为 1
8: end for
9: 保存异常数据添加后的异常数据集

表 7 异常数据集详细信息

Tab.7 Details of the abnormal data set

数据集	样本数/个	异常样本数/个	异常占比/%
Al-Si 系合金	166	6	3.6
Al-Cu 系合金	77	6	7.8
Al-Zn 系合金	49	4	8.2
Al-Mg 系合金	44	4	9.1

分布见图 7。可以看出,相较于真值数据集,由于异常

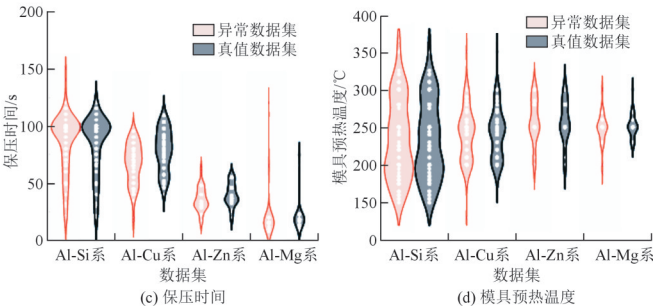


图 7 工艺数据集关键属性的数据分布

Fig.7 Data distribution of key attributes in the process dataset

数据的存在,异常数据集中各属性的数据分布范围变大,即有异常数据超出原正常数值范围。

2.2 数据正确性检测

校验材料成分数值步骤,发现 Al-Zn 系合金数据集中样本 37 的材料组分数值之和为 102.46,其违反材料成分属性取值规则且明显不符合工程实际。基于挤压铸造工艺数据属性值的取值范围、数值类型等信息对数据样本进行正确性检测,发现 4 条异常样本(见表 8,其中标签 0 代表正常,1 代表异常),其中 Al-Si 系合金数据集的样本 67 和样本 90 的浇注温度数据偏离 Al-Si 合金的经验取值范围,核实数据源文献发现其为半固态挤压铸造工艺而非近液相线温度挤压铸造。Al-Cu 系合金数据集中样本 46 的浇注温度数据偏离经验取值范围,Al 材料成分数值也明显异常,核实数据源文献发现其为 AZ91D 镁合金的近液相线温度挤压铸造工艺数据。Al-Mg 系合金数据集中样本 37 的浇注温度数值偏离经验取值范围。

删除违反属性值取值规则的数据样本,然后将其余数据样本采用 LBI 集成模型进行二次检测,得 LBISCDK 方法的检测结果见表 9。

表 8 违反属性值取值规则的异常样本

Tab.8 Abnormal samples violating the rule of attribute value selection

数据集	样本 编号	浇注温 度/℃	比压/ MPa	保压 时间/s	模具预热 温度/℃	...	$w(\text{Al})/\%$	...	样本 标签
Al-Si 系	67	600	100	30	215	...	91.36	...	1
	90	595	45	16	250	...	92.72	...	1
Al-Cu 系	46	770	80	40	200	...	9.06	...	1
Al-Mg 系	24	645	60	45	280	...	97.46	...	1

表 9 LBISCDK 方法检测异常样本

Tab.9 Abnormal samples detected by LBISCDK method

数据集	样本 编号	浇注温 度/℃	比压/ MPa	保压时 间/s	模具预热 温度/℃	...	$w(\text{Al})/\%$	...	样本 标签
Al-Si 系	46	700	21.5	30	350	...	92.32	...	1
	67	600	100	30	215	...	91.36	...	1
	90	595	45	16	250	...	92.72	...	1
	112	720	80	26	175	...	91.76	...	1
	119	685	150	135	250	...	92.34	...	1
	125	617	85	45	250	...	91.36	...	1
Al-Cu 系	139	680	100	120	240	...	92.47	...	0
	21	680	80	40	200	...	92.19	...	1
	34	687	100	60	190	...	93.04	...	1
	36	720	23	50	225	...	92.19	...	1
	46	770	80	40	200	...	9.06	...	1
	65	730	100	20	250	...	92.66	...	1
Al-Zn 系	13	600	22.6	45	300	...	90.04	...	1
	37	640	35	60	200	...	94.72	...	1
	39	590	45	60	280	...	89.97	...	1
	43	680	70	120	200	...	90.20	...	1
Al-Mg 系	7	700	80	123	250	...	93.05	...	1
	18	690	100	30	230	...	92.58	...	0
	24	645	60	45	280	...	97.46	...	1
	29	690	80	60	270	...	94.27	...	1
	43	700	75	16	178	...	97.54	...	1

对相同数据集,只使用 LBI 集成模型的检测结果见表 10。

表 10 LBI 方法检测异常样本

Tab.10 Abnormal samples detected by LBI method

数据集	样本 编号	浇注温 度/℃	比压/ MPa	保压时 间/s	模具预热 温度/℃	...	$w(\text{Al})/\%$	...	样本 标签
Al-Si 系	46	700	21.5	30	350	...	92.32	...	1
	90	595	45	16	250	...	92.72	...	1
	112	720	80	26	175	...	91.76	...	1
	119	685	150	135	250	...	92.34	...	1
	125	617	85	45	250	...	91.36	...	1
	139	680	100	120	240	...	92.47	...	0
Al-Cu 系	21	680	80	40	200	...	92.19	...	1
	34	687	100	60	190	...	93.04	...	1
	36	720	23	50	225	...	92.19	...	1
	46	770	80	40	200	...	9.06	...	1
	65	730	100	20	250	...	92.66	...	1
Al-Zn 系	13	600	22.6	45	300	...	90.04	...	1
	39	590	45	60	280	...	89.97	...	1
	43	680	70	120	200	...	90.20	...	1
Al-Mg 系	7	700	80	123	250	...	93.05	...	1
	18	690	100	30	230	...	92.58	...	0
	29	690	80	60	270	...	94.27	...	1
	43	700	75	16	178	...	97.54	...	1

3 分析结果与讨论

融合领域工程知识的 LBISCDK 方法和纯数据驱动的 LBI 方法各检测指标结果见表 11。

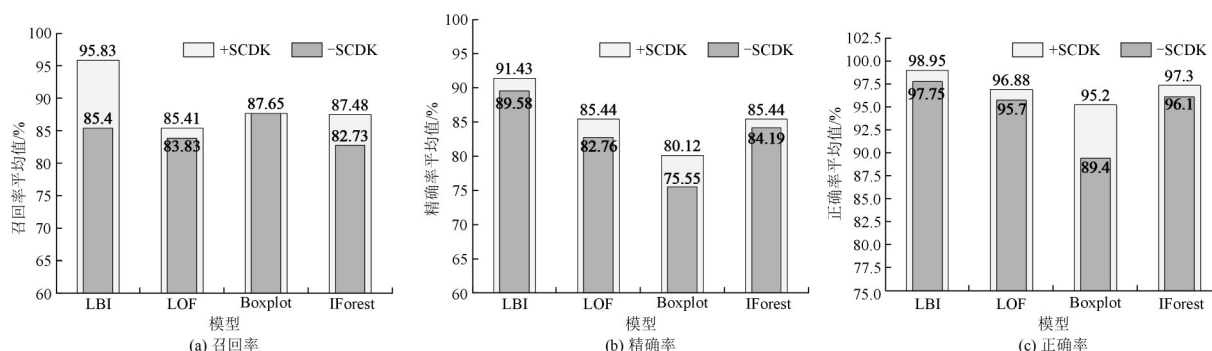
分析表 11 可得, LBISCDK 方法在 4 个不同异常程度的异常数据集上的检测召回率、精确率和正确率的平均值分别为 94.43%、95.23% 和 99.37%, 其中 Al-Zn 系合金数据集的 3 个检测评价指标均为 100%, 说明该方法能够有效检测挤压铸造工艺数据集中的异常数据, 检测效果可满足数据驱动相关应用需求。

表 11 异常检测结果

Tab.11 Abnormal detection result

数据集	LBISCDK			LBI		
	召回率 /%	精确率 /%	正确率 /%	召回率 /%	精确率 /%	正确率 /%
Al-Si 系	100	85.7	99.4	83.3	83.3	98.8
Al-Cu 系	83.3	100	98.7	83.3	100	98.7
Al-Zn 系	100	100	100	75.0	100	98.0
Al-Mg 系	100	80	97.7	75.0	75.0	95.5

图 8 为几种方法融合领域知识前后在 4 个异常数据集上的平均检测召回率、准确率和正确率。可见相较于纯数据驱动的检测方法, 融合领域知识后的方法检测效果均有一定提升, 异常数据召回表现尤其优秀; 其中 LBI、LOF 和 IForest 方法的平均召回率分别提升了 10.43%、1.58% 和 4.75%, LOF 和 Boxplot 方法的平



+SCDK表示融合领域知识的两阶段检测方法,-SCDK表示未融合领域知识的纯数据驱动检测方法。

图8 几种方法融合领域知识前后的检测效果对比

Fig.8 Comparison of detection effects of several methods before and after integrating domain knowledge

均精确率分别提升了2.68%和4.57%,Boxplot方法的平均正确率提升了5.6%,表明领域知识和数据驱动相结合能够对挤压铸造工艺数据集中的异常数据进行更加有效的识别,可为数据驱动的挤压铸造工艺参数设计等数据驱动应用提供高质量的数据样本。

分析图8可得,在不同异常程度的数据集上,LBI集成模型相较于3个基础模型,融合挤压铸造领域知识前后的检测效果表现均更优异,其平均召回率、准确率和正确率提升最大值分别为10.42%、14.03%和8.35%,说明采用集成学习的策略进行挤压铸造工艺数据的正确性检测是可行且有效的,可以实现不同类型基础模型的性能互补,提升模型的检测效率和泛化能力。此外,在3个基础模型中,IForest模型和LOF模型的3个指标表现相对均衡,而Boxplot模型对单点异常数据十分敏感,检测召回率指标表现最好,但其检测精确率指标表现不佳,易将正常的数据样本误判为异常,其在4个数据集上融合领域知识前后的检测平均精确率仅为75.55%和80.12%。

模型参数直接影响模型的复杂度、鲁棒性和泛化能力,合理选择和调整模型参数至关重要。如在构建IForest基模型过程中,选择不同的子采样量 ϕ 和孤立树数量 $n_{estimators}$ 组合,其检测结果存在很大差异,所得召回率、准确率和正确率指标的最大差值均可达40%。为提升模型的异常检测性能,在基础模型构建过程中对于经验取值的每一个参数组合,采用 k 折交叉验证技术计算其在交叉验证过程中的平均性能指标,选择使得性能指标最优的参数组合作为最终的参数设置,以达到最优的异常检测效果,同时也避免模型对特定数据集的过拟合,提高模型的泛化能力。

集成模型的计算量分析需要考虑单元模型的独立计算量及其在集成过程中的相互作用。分析可知,对 n 个数据样本,LOF模型在最坏情况下的时间复杂度为 $O(n^2)$,IForest模型的时间复杂度为 $O(n_{estimators} \times \phi \times$

$\log \phi)$,Boxplot模型的时间复杂度为 $O(n \times \log n)$;建立的LBI模型基于投票法进行集成,集成过程只涉及计数和比较操作,其计算复杂度为 $O(n)$,则LBISCDK方法的总体计算时间复杂度为 $O(n \times \log n)$ 到 $O(n^2)$ 。针对当前的挤压铸造工艺数据量级(百千级),该方法有效可行。后续如果涉及到更大规模或更高维度的数据集时,可探索算法优化和并行处理等方法提升检测效率,以确保方法的检测性能。

4 结 论

(1)数据驱动方法的准确性和有效性严重依赖于正确、高质量可靠的数据样本,在数据驱动应用前对数据样本进行正确性检测和提升至关重要。现有数据正确性检测方法往往仅采用单一模型从数据特性检测,忽视了工程知识,无法有效检测出有悖于工程知识的异常数据。

(2)面向数据驱动的挤压铸造工艺参数设计方法对数据质量的需求,提出了融合挤压铸造工程知识和多个模型的数据正确性检测方法。该方法将多源挤压铸造工艺数据的正确性检测划分为两个阶段,第一阶段使用挤压铸造工程知识构建工艺数据属性取值规则,识别背离工程知识的异常数据,第二阶段从数据维度出发,根据数据特征使用LOF、Boxplot和IForest 3种模型来优化检测,实现了3种方法优势互补,提高了3种方法的泛化能力。

(3)实现了知识和数据特征双维度检测,拓展了数据异常(正确性)检测的维度,为数据清洗等应用提供了新的参考视角,其中的集成模型可以单独用于其他数据集异常检测。试验表明,所提方法提高了异常数据的检测召回率、正确率,对异常数据的召回率平均可达95%。

参 考 文 献

[1] TOLLE K M, TANSLEY D S W, HEY A J G. The fourth para-

- digm: Data-intensive scientific discovery [J]. Proceedings of the IEEE, 2011, 99(8): 1 334-1 337.
- [2] AGRAWAL A, CHOUDHARY A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science[J]. APL Materials, 2016, 4(5): 53 208.
- [3] 任磊, 贾子翟, 赖李媛君, 等. 数据驱动的工业智能: 现状与展望[J]. 计算机集成制造系统, 2022, 28(7): 1 913-1 939.
- [4] 任明仑, 宋明丽. 大数据: 数据驱动的过程质量控制与改进新视角[J]. 计算机集成制造系统, 2019, 25(11): 2 731-2 742.
- [5] BERBERICH J, KOHLER J, MULLER M A, et al. Data-driven model predictive control with stability and robustness guarantees[J]. IEEE Transactions on Automatic Control, 2020, 66(4): 1 702-1 717.
- [6] BRINAL T P, ARROYO A. Empirical study on distance based data reduction for truth discovery using conflicted data [J]. Mukti Shabd Journal, 2020, 7(9): 2 204-2 208.
- [7] TIAN H, SHENG W W, SHEN H, et al. Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets[J]. Knowledge-Based Systems, 2020, 187: 104 828.
- [8] LIU Y, ZHAO T L, JU W W, et al. Materials discovery and design using machine learning [J]. Journal of Materiomics, 2017, 3(3): 159-177.
- [9] EHRLINGER L, WÖB W. A survey of data quality measurement and monitoring tools[J]. Frontiers in Big Data, 2022, 5: 850 611.
- [10] KARTHIK A, KARUNANITHI R, SRINIVASAN S A, et al. The optimization of squeeze casting process parameter for AA2219 alloy by using the Taguchi method [J]. Materials Today-Proceedings, 2020, 27: 2 556-2 561.
- [11] JIANG J F, YAN J, LIU Y Z, et al. Numerical simulation and experimental validation of squeeze casting of AlSi9Mg aluminum alloy component with a large size [J]. Materials, 2022, 15(12): 4 334.
- [12] DENG J X, XIE B, YOU D D, et al. Process parameters design of squeeze casting through an improved KNN algorithm and existing data [J]. Journal of Manufacturing Processes, 2022, 84: 1 320-1 330.
- [13] 傅松桥, 常建娥. 液态模锻 AZ31 镁合金汽车轮毂的性能分析[J]. 热加工工艺, 2021, 50(7): 92-96.
- [14] 黄志伟, 黄静, 邓天泉. AZ31 镁合金挤压铸造工艺参数优化研究 [A]//重庆市铸造年会[C]. 重庆, 2013.
- [15] 黄盟, 毕晓阳, 杨晓, 等. 基于人工数据融合的柴油机故障数据增强方法[J]. 振动与冲击, 2023, 42(13): 278-286.
- [16] MAHARANA K, MONDAL S, NEMADE B. A review: Data pre-processing and data augmentation techniques[J]. Global Transitions Proceedings, 2022, 3(1): 91-99.
- [17] BREUNIG M M, KRIEGER H, NG R T, et al. LOF identifying density-based local outliers [A]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of data, Association for Computing Machinery[C]. New York, USA, 2000.
- [18] LIU F T, TING K M, ZHOU Z. Isolation-based anomaly detection [J]. ACM Transactions on Knowledge Discovery from Data, 2012, 1(1): 1-39.
- [19] GAURA E I, BRUSEY J, ALLEN M, et al. Edge mining the internet of things[J]. IEEE Sensors Journal, 2013, 13(10): 3 816-3 825.
- [20] 何九虎, 刘飞. 工业过程数据异常检测的改进局部离群因子法[J]. 计算机与应用化学, 2013, 30(1): 53-56.
- [21] 冯立伟, 张成, 李元, 等. 基于统计模量和局部近邻标准化的局部离群因子故障检测方法[J]. 计算机应用, 2018, 38(4): 965-970.
- [22] LUO Y, LIU Y, YANG W, et al. Distributed filtering algorithm based on local outlier factor under data integrity attacks [J]. Journal of the Franklin Institute, 2023, 360(12): 9 290-9 306.
- [23] 吴俊. 基于 DBSCAN 和 LOF 的网络入侵异常检测算法优化研究 [D]. 福州: 福州大学, 2017.
- [24] TU B, ZHOU C L, KUANG W L, et al. Hyperspectral imagery noisy label detection by spectral angle local outlier factor[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(9): 1 417-1 421.
- [25] 刘芳, 齐建鹏, 于彦伟, 等. 基于密度的 Top-n 局部异常点快速检测算法[J]. 自动化学报, 2019, 45(9): 1 756-1 771.
- [26] 汤淳, 刘晓星. 基于集成分类算法的系统性金融风险预警研究[J]. 运筹与管理, 2023, 32(6): 186-191.
- [27] LIU J, TIAN J, CAI Z, et al. A hybrid semi-supervised approach for financial fraud detection [A]//2017 International Conference on Machine Learning and Cybernetics (ICMLC)[C]. Ningbo, China, IEEE, 2017.
- [28] 熊文滔, 郑建立. 基于扩展孤立森林的个性化跌倒检测研究[J]. 软件工程, 2023, 26(4): 5-9.
- [29] 陈佳, 欧阳金源, 冯安琪, 等. 边缘计算构架下基于孤立森林算法的 DoS 异常检测[J]. 计算机科学, 2020, 47(2): 287-293.
- [30] 黄家续, 曾献辉, 施陈俊. 基于实时数据流特征提取的设备能耗异常识别算法研究[J]. 信息技术与网络安全, 2021, 40(5): 45-50.
- [31] WANG H T, LI Y J, ZHANG X P, et al. Research on anomaly detection and positioning of marine nuclear power steam turbine unit based on isolated forest[J]. Nuclear Engineering and Design, 2023, 412: 112 466.
- [32] STAERMAN G, MOZHAROVSKIY P, CLÉMENÇON S, et al. Functional isolation forest [A]//Asian Conference on Machine Learning[C]. Macau, China, PMLR, 2019.
- [33] TUKEY J W. Exploratory Data Analysis [M]. MA: Addison-Wesley, 1977.
- [34] DOVOEDO Y H, CHAKRABORTI S. Boxplot-based outlier detection for the location-scale family [J]. Communications in Statistics-Simulation and Computation, 2015, 44(6): 1 492-1 513.
- [35] 顾国庆, 李晓辉. 基于箱线图异常检测的指数加权平滑预测模型 [J]. 计算机与现代化, 2021(1): 28-33.

(编辑: 张正贺)