不平衡工艺参数数据集的高温透平叶片铸件质量预测方法

朱铜[1]，艾松[2]，陈琨[1]，高建民[1]
（1. 西安交通大学机械制造系统工程国家重点实验室，710049，西安；
2. 清洁高效透平动力装备全国重点实验室，618000，四川德阳）

摘要：针对熔模精密铸造工艺参数数据集射线检测(RT)结果存在合格与不合格数量严重不平衡问题，提出一种基于概率分布的合成少数类集成学习(SyMProD-Stacking)的铸件质量预测方法。该方法首先对原始数据集进行预处理以保证数据质量，然后利用 Z 分数去除噪声数据，为每个少数类实例(不合格铸件)分配一个概率并基于此概率分布生成样本数据以获取平衡数据集，利用极端梯度提升模型(XGBoost)对所有工艺参数特征进行重要性排序并剔除部分排名靠后的工艺参数，最后将轻量级梯度提升机(LightGBM)、随机森林(RF)、支持向量机(SVM)和XGBoost模型进行Stacking集成并利用平衡数据集构建质量预测模型。以高温透平叶片制造过程精铸工艺为例，对所提出的质量预测方法进行验证，结果表明：相比于原始数据集构建的预测模型，利用了SyMProD过采样方法构建的预测模型不合格铸件的预测准确率提升了75.4%；相比于单一算法模型，所提质量预测方法的曲线下面积(A_{AUCROC})、几何均值(G_{m})以及 F_1 分数(F_1)这3项性能指标分别提升了5.48%～11.59%、3.78%～8.92%、5.72%～11.39%，所提出的方法能够很好地预测高温透平叶片精铸过程在不平衡问题下的铸件质量。

关键词：高温透平叶片；不平衡问题；过采样方法；集成学习；质量预测
中图分类号：TP182 **文献标志码**：A
DOI：10.7652/xjtuxb202409010 **文章编号**：0253-987X(2024)09-0094-11

Quality Prediction Method of High Temperature Turbine Blade Castings Based on Unbalanced Process Parameter Data Set

ZHU Tong[1], AI Song[2], CHEN Kun[1], GAO Jianmin[1]
(1. State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. State Key Laboratory of Clean and Efficient Turbomachinery Power Equipment, Deyang, Sichuan 618000, China)

Abstract: In response to the significant imbalance between the quantities of qualified and non-qualified results in the radiographic testing (RT) of the investment precision casting process parameters dataset, this paper proposes a casting quality prediction method using SyMProD-Stacking ensemble learning. The method begins by preprocessing the original dataset to ensure data quality. It then employs Z-scores to eliminate noisy data, assigns a probability to each minority class instance (non-qualified castings), and generates sample data based on this probability distribution to obtain a balanced dataset. XGBoost is used to rank the importance of

收稿日期：2024-02-19。 作者简介：朱铜(1998—)，男，硕士生；陈琨(通信作者)，男，教授，博士生导师。 基金项目：国家重点研发计划资助项目(2021YFF0602300)。
网络出版时间：2024-04-18 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20240417.1040.003

all process parameter features and removes some of the lower-ranking parameters. Finally, the LightGBM, RF, SVM, and XGBoost models are stacked together through ensemble learning, and a quality prediction model is constructed using the balanced dataset. Taking the precision casting process in the manufacturing of high-temperature turbine blades as an example, the proposed quality prediction method is validated. The results indicate that the predictive model constructed using the SyMProD oversampling method significantly outperforms the model built from the original dataset, with a 75.4% improvement in the accuracy of predicting non-qualified castings. Stacking ensemble learning, compared to individual algorithm models, achieves improvements of 5.48%—11.59%, 3.78%—8.92%, and 5.72%—11.39% in terms of area under curve, geometric mean, and F_1, respectively.

Keywords: high-temperature turbine blades of heavy-duty gas turbine; imbalanced dataset; oversampling methods; ensemble learning; quality prediction

传统的工业产品在生产完成后才会进行检测，若发现质量问题，对生产工艺和加工过程进行调整和优化[1]，难以提前制定响应措施。随着“工业互联网”和“中国制造 2025”等概念的提出，推动了制造生产领域的数字化和自动化，以提高竞争力、持续性和创新能力。越来越多的制造型企业建立了生产过程的数据采集系统、质量控制与分析系统[2]。如何利用制造过程的大数据进行产品质量预测是众多企业急需解决的难题。

高温透平叶片是重型燃气轮机（简称重燃）的核心关键部件，其制造质量决定了重燃的技术水平。熔模精密铸造（简称精铸）作为高温透平叶片制造过程的关键工艺，其铸件的质量很大程度上决定了高温透平叶片的质量状态。企业积累了大量的工艺参数数据，这些数据存在特征维度高和类别不平衡等问题，现有的产品质量预测方法具有一定的局限性，难以直接用于精铸不平衡工艺参数数据集。

不合格品的产生会导致企业额外的成本，通过准确地预测不合格品，企业可以采取相应的预防措施，及时调整生产过程，因此企业通常更加关注不合格品的预测准确率。在精铸工艺参数数据集中，合格铸件数远大于不合格铸件数，存在严重的类别不平衡现象，若利用该数据集直接构建预测模型，存在只有多数类实例（合格铸件）能够获得较高的预测精度，少数类实例（不合格铸件）预测准确率很低的问题，难以应用于实际生产的质量预测[3-4]。

解决类别不平衡问题的方法主要有 3 种：基于算法层面的方法、基于代价敏感层面的方法以及基于数据层面的方法[5]。基于算法层面的方法主要考虑预测模型的选择以及针对算法本身进行一定的改进，这种方法对不平衡数据集处理效果不够显著且时间开销较大。基于代价敏感层面的方法考虑了错误分类的较高成本，这种方法需要数据集的领域知识，很难估计出合适的值。基于数据层面的方法平衡了多数类和少数类之间的比例，任何算法均可以利用平衡后的数据构建预测模型，该方法更通用[6]。基于数据层面的方法包括过采样方法和欠采样方法，如果少数类数量很少，采用欠采样方法将删除大量多数类数据导致模型欠拟合[7]。一般来说，过采样方法具有更好的性能。合成少数类过采样技术（SMOTE）选择若干个最近邻的样本，然后根据这些邻居样本的特征，生成新的合成样本[8]。学者们针对 SMOTE 方法可能会生成噪声数据、区域重叠、过度泛化等问题提出了一系列改进算法。Han 等[9]提出了 Borderline-SMOTE 技术，根据邻居类的比例在决策区域的边界上创建合成实例。He 等[10]提出了自适应样本合成方法，应用基于多数类邻居比率对少数类实例进行加权的概念。基于聚类的过采样 K-means-SMOTE 算法[11]将少数类划分为几个不同的子集，然后分别在子集进行过采样。李敏波等[12]提出了基于密度聚类的过采样方法 MCDC-MF-SMOTE，生成汽车零部件质检的平衡数据集。Kunakorntum 等[13]提出了一种基于概率分布的合成少数类过采样技术，为每个少数类实例分配一个概率并基于概率分布生成样本，然后用 14 个公开数据集和 3 个分类算法与其他 7 种传统过采样方法进行性能比较，结果表明该方法取得了更好的性能。

产品质量预测利用历史制造数据挖掘各个质量影响因素与产品性能之间的映射关系[14-15]以评估生产过程的质量水平，提前制定响应措施，预防质量问题。目前质量预测模型的构建主要有机理建模与

数据驱动建模两种方式。机理建模通过深入理解生产过程中的物理原理和内在规律来预测产品质量，构建特定而精确的物理或数学模型，表达生产过程中工艺参数与产品质量之间的机理特性[16-17]。李传维[18]采用加权平均方法建立了大型锻件 SA508Gr.3 钢的材料强度预测模型。机理模型普遍性、可靠性高，但通常结构复杂，多数物理量在工业过程中难检测，往往要做多种假设和简化处理，使得模型计算结果和实际情况存在差异[19]，因此难以用于构建精铸工艺中铸件的质量预测模型。

利用数据驱动建模本质上是一个二分类问题，即利用历史制造数据集，并选择合适的机器学习算法训练出有效的分类模型。董海等[20]针对汽车车身装配尺寸精度控制问题提出了一种基于 XGBoost 算法的质量预测方法，解决了传统机器学习算法在处理复杂产品制造过程的质量大数据时存在的预测模型准确率低且效率差的问题。Duan 等[21]利用制造质量数据探究了产品实时质量状态与加工任务过程之间的关系并开发了实时质量预测系统。赵双凤[22]针对轴类零件车削加工测量值的预测问题提出了结合 BP 神经网络和灰色模型的质量预测模型。于文靖[23]提出了基于 PSO-SVR 的汽轮机模锻叶片的质量预测方法，通过实例验证了该方法的有效性。钟武昌等[24]以注塑成型加工过程为例，利用 Stacking 集成学习算法构建了质量预测模型，提高了产品质量预测模型性能，降低了过拟合风险。向峰等[25]为了基于 Stacking 集成学习算法对整体和分段预测模型进行融合，并以烟丝生产过程进行对比实验，验证了该方法的有效性和优越性。基于数据驱动的质量预测方法不需要先验知识或深入理解质量问题的机理，只需要具有足够的数据来进行建模和预测，减少建模的成本和复杂度，具有自适应、自学习的特点，可以保证良好的预测准确率。

针对上述问题，结合精铸工艺参数数据集特点和研究现状分析，本文提出一种基于概率分布的合成少数类（SyMProD-Stacking）集成学习的铸件质量预测方法。对精铸工艺参数数据集进行预处理后利用 SyMProD 过采样方法为每个少数类实例分配概率并基于概率分布生成样本以平衡数据集；将 LightGBM、RF、SVM 和 XGBoost 进行 Stacking 集成，构建铸件的质量预测模型。以某型号高温透平叶片的精铸工艺参数数据集为例对本文方法进行验证，帮助企业评估生产过程的质量水平，提前制定响应措施并降低生产成本。

1 SyMProD 过采样方法

本文采用文献[13]中基于概率分布的合成少数类方法（SyMProD）解决机器学习中类别不平衡问题。首先，应用 Z 分数来去除噪声数据；然后，为每个少数类实例分配一个概率并根据概率分布在合成样本过程中选择数据点；最后，在少数类组由几个少数类实例合成样本。该方法基于概率分布选择少数类实例，避免了重叠区域问题，消除了过拟合和过度泛化问题，提高预测模型的性能。

1.1 噪声去除

需要从数据中去除异常值，因为如果合成样本是从有噪声的数据中生成的，则会降低模型性能。将 Z 分数应用于原始数据集的少数类实例和多数类实例，如果绝对值高于噪声阈值 N_{T}，则去除该异常实例，移除噪声数据示意图如图 1 所示。

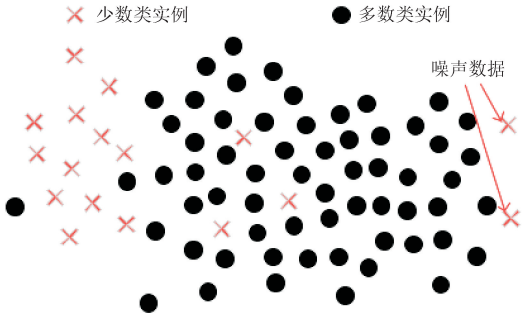

图 1 移除噪声数据示意图

Fig. 1 Remove noise samples

1.2 概率分配

需要为数据集中每个少数类实例分配概率便于后续基于此概率分布合成新样本实例。将经过噪声过滤后的数据集划分为少数类组 $\boldsymbol{X}_{\min}$ 和多数类组 $\boldsymbol{X}_{\mathrm{maj}}$，少数类组表达式为

$$\boldsymbol{X}_{\min}=\{\boldsymbol{X}_{\min}(i),\boldsymbol{Y}_{\min}(i)\},i=1,2,3,\cdots,n_{\min} \tag{1}$$

式中：$n_{\min}$ 为少数类实例数；$\boldsymbol{X}_{\min}(i)$ 为特征空间；$\boldsymbol{Y}_{\min}(i)$ 为目标值，多数类组与少数类组表示方法相同。假设 p 和 q 是具有 N 个维度的点，通过应用欧几里得距离测量两点之间的距离来确定少数类分布

$$d(p,q)=\sqrt{\sum_{i=1}^{N}(p(i)-q(i))^2} \tag{2}$$

对于每一个少数类实例样本 $\boldsymbol{X}_{\min}(i)$，计算每个点到同一类别中其他点的距离并返回少数类实例之间的总距离

$$D(\boldsymbol{X}_{\min}(i))=\sum_{j=1}^{n_{\min}}d(\boldsymbol{X}_{\min}(i),\boldsymbol{X}_{\min}(j)) \tag{3}$$

每个少数类实例中的接近因子 C 表示一个实例与同一组中其他所有实例的接近程度，表达式如下

$$C(\boldsymbol{X}_{\min}(i)) = \frac{1}{D(\boldsymbol{X}_{\min}(i))} \tag{4}$$

因此，同一组中样本点与其他点越接近则接近因子 C 越高。对于多数类组，同样也定义了多数类实例之间的距离和接近因子以确定多数类分布，计算每个实例的接近因子示意图如图 2 所示。

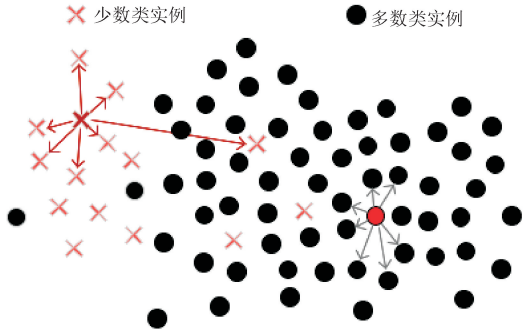

图 2　计算每个实例的接近因子示意图

Fig. 2　Calculate closeness factor of each instance

通过根据少数类和多数类的分布来决定是否选择少数类实例，避免了区域重叠问题。对于每个少数类实例 $\boldsymbol{X}_{\min}(i)$，找到每个组的 K 个最近邻，将每个少数类实例的 K 个多数类最近邻和少数类最近邻分别表示为

$$\boldsymbol{S}_{\text{maj}}(i) = \{\boldsymbol{S}_{\text{maj}}(i,1), \boldsymbol{S}_{\text{maj}}(i,2), \cdots, \boldsymbol{S}_{\text{maj}}(i,K)\} \tag{5}$$

$$\boldsymbol{S}_{\min}(i) = \{\boldsymbol{S}_{\min}(i,1), \boldsymbol{S}_{\min}(i,2), \cdots, \boldsymbol{S}_{\min}(i,K)\} \tag{6}$$

为了减少区域重叠的可能性，定义了少数类实例的少数类组接近因子和多数类组接近因子，分别表示为

$$\tau_{\min}(i) = \sum_{j=1}^{K} \frac{C(\boldsymbol{S}_{\min}(i,j))}{D(\boldsymbol{X}_{\min}(i), \boldsymbol{S}_{\min}(i,j))} \tag{7}$$

$$\tau_{\text{maj}}(i) = \sum_{j=1}^{K} \frac{C(\boldsymbol{S}_{\text{maj}}(i,j))}{D(\boldsymbol{X}_{\min}(i), \boldsymbol{S}_{\text{maj}}(i,j))} \tag{8}$$

截止阈值 C_{T} 用于选择位于少数类分布区域的少数类实例来生成新的样本，而过滤那些位于多数类区域的少数类实例以避免区域重叠问题，排除重叠实例示意图如图 3 所示。满足下列条件则保留该少数类样本实例

$$\tau_{\min}(i) > \tau_{\text{maj}}(i) C_{\text{T}} \tag{9}$$

计算每个少数类实例 $\boldsymbol{X}_{\min}(i)$ 接近因子比率 $\varphi(i)$，并转换为概率分布 $P(i)$

$$\varphi(i) = \frac{\tau_{\min}(i)+1}{\tau_{\text{maj}}(i)+1} \tag{10}$$

$$P(i) = \varphi(i) \Big/ \sum_{j=1}^{n_{\min}} \varphi(j) \tag{11}$$

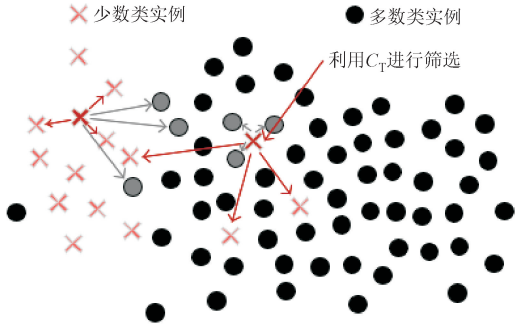

图 3　排除重叠实例示意图

Fig. 3　Exclude overlapping instance

1.3　生成样本实例

基于概率分布选择少数类实例，确定 M 个最近邻以生成样本实例，而不是像 SMOTE 技术那样沿着两点之间的线生成。集合 R 表示 M 个最近邻的实际值，生成的新样本实例在此范围内。集合 R 中每个实例的概率分布用集合 P_{r} 表示，则生成的样本实例可表示为

$$\boldsymbol{X}_{\text{new}} = \sum_{j=1}^{M+1} \beta(j) P_{\text{r}}(j) R(j) \tag{12}$$

式中：$\beta(j)$ 是 0 ～ 1 之间的随机数；$P_{\text{r}}(j)$ 表示每个被选择的少数类实例和其 M 个最近邻样本实例的第 j 个概率分布；$R(j)$ 是第 j 个样本实例的实值。$\beta(j)$ 和 $P_{\text{r}}(j)$ 均归一化为 1 并作为系数因子，满足下式

$$\beta(1)P_{\text{r}}(1) + \cdots + \beta(M+1)P_{\text{r}}(M+1) = 1 \tag{13}$$

图 4 所示为利用该方法在少数类区域生成的样本实例。

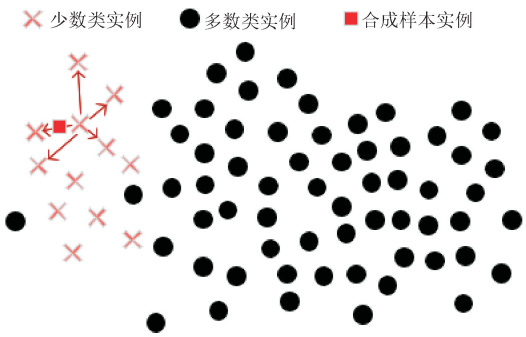

图 4　在少数类区域生成的样本实例

Fig. 4　Generate synthetic instance in minority class region

1.4 SyMProD 方法流程

SyMProD 过采样方法的输入是原始数据集 I、用于检测和去除噪声数据的噪声阈值 N_T、用于过滤位于多数类区域的少数类实例的截止阈值 C_T、用于比较组间接近因子的最近邻居数 K 以及用于生成样本实例的最近邻居数 M。输出是重新平衡后的过采样数据集。方法具体实现步骤如下。

步骤 1 计算需要生成的样本实例数 n_{gen}，$n_{gen} = n_{maj} - n_{min}$，其中 n_{maj} 为多数类实例数。

步骤 2 将 Z 分数应用于原始数据集的少数类实例和多数类实例，如果绝对值高于 N_T，则去除该异常的少数类实例。

步骤 3 根据式(3)计算每个实例的欧几里得距离。

步骤 4 根据式(4)计算接近因子。

步骤 5 对于每一个少数类实例，根据式(7)、(8)计算少数类组接近因子 τ_{min} 和多数类组接近因子 τ_{maj}。

步骤 6 根据式(9)判定是否需要去除少数类实例。

步骤 7 根据式(10)计算接近因子比率并根据式(11)计算每个少数类实例的概率分布 $P(i)$。

步骤 8 基于 $P(i)$ 选择 n_{gen} 个少数类实例(可重复)。

步骤 9 对于每一个少数类实例，找到 M 个少数类最近邻并将所选择的少数类实例及其最近邻收集在 R 中。

步骤 10 生成值介于 0～1 之间的 $M+1$ 维随机正向量。

步骤 11 根据式(12)生成样本实例，返回重新平衡后的过采样数据集。

SyMPro 过采样方法需要设定 4 个超参数即噪声阈值(N_T)、截止阈值(C_T)、概率分配时 K 最近邻以及生成样本时 M 最近邻来调节分类算法性能。N_T 用于噪声去除，其适当范围是 3～5，因为较低的 N_T 值可能会消除有效点并降低分类算法性能。另一方面，较高的 N_T 可能没有去除异常实例而增加噪声产生的可能性。截止阈值 C_T 与 K 最近邻用于降低少数类与多数类区域重叠问题的可能性。对于每个少数类实例，搜索 K 个多数类最近邻和少数类最近邻，分别用于计算 τ_{maj} 和 τ_{min}，若满足式(9)，则去除该少数类实例。C_T 的合理范围是 0.8～1.2，因为较低的 C_T 值可能不能有效处理区域重叠问题，而较高的 C_T 值可能会删除有效样本实例降低分类算法性能。M 最近邻用于确定少数类邻居数量以生成样本实例。K 和 M 的默认值均设定为 5[13]。

2 Stacking 集成的质量预测模型构建

通过上述 SyMProD 过采样方法能够产生用于质量预测模型构建的平衡数据集，避免重叠区域问题，解决过拟合和过度泛化问题，提高了分类算法的性能。考虑到高温透平叶片精铸流程及工艺参数数据集的复杂性，单一的分类算法难以适应复杂多变的生产制造场景[24]。提出了将 LightGBM、RF、SVM 和 XGBoost 模型进行 Stacking 集成的预测模型构建方法。整个铸件质量预测模型构建流程如图 5 所示。对数据集进行预处理并划分数据集后利用 SyMProD 过采样方法进行数据集平衡，然后基于 XGBoost 的特征重要度进行工艺参数的维度缩减，最后构建基于 Stacking 集成学习的质量预测模型。

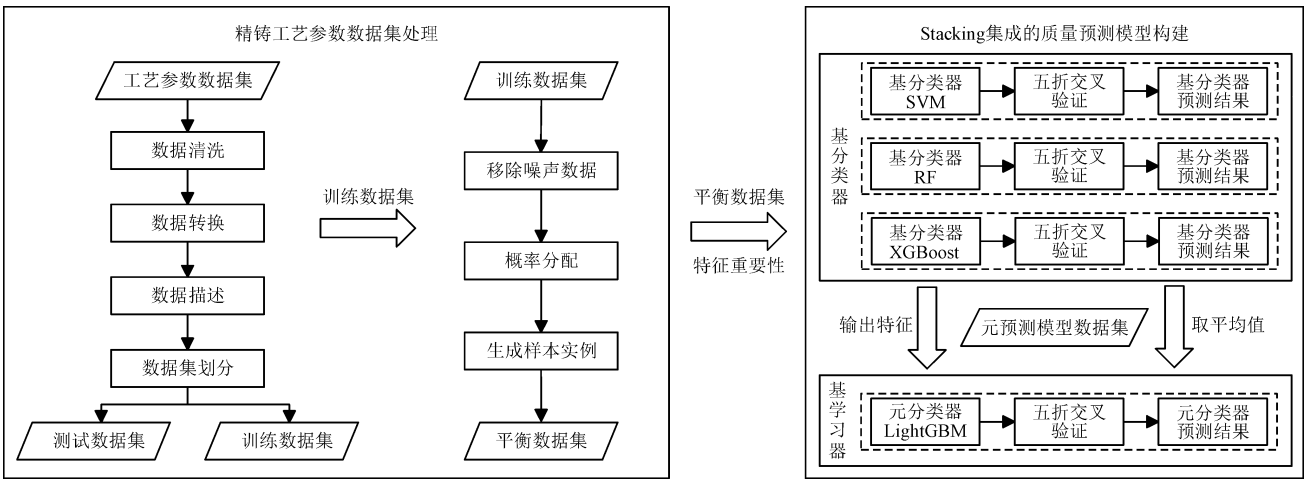

图 5 精铸工艺质量预测模型构建整体流程

Fig. 5 Overall flow of constructing the quality prediction model of precision casting process

2.1 数据预处理

数据的质量直接决定了模型的预测精度和泛化能力，在重燃叶片精铸数据采集过程中不可避免地出现数据缺失、数据噪声、各维数据特征分布范围差

异大等问题，因此，本文通过数据清洗和数据转换确保数据集质量。

例如，某生产设备故障或工人记录遗漏导致部分工艺参数数据缺失，需要利用列变量均值进行填充。例如某些工艺参数属于非数值型，需要将其转换为数值型以便学习算法后续处理。为了消除数据变量之间的量纲影响，比较不同变量之间的相互作用，本文采用 Min-Max 归一化处理，能够在不改变数据分布的前提下消除变量量纲和变异范围影响

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

式中：x 为某个参数的样本数据；$x_{\min}$ 与 $x_{\max}$ 分别表示 x 中的最小值和最大值；x' 为经过归一化处理后的样本值。

经过上述处理后，需要进行数据描述，计算集中趋势指标（平均数）和离散趋势指标（最小值、最大值和变异系数）等统计量，进而观察出数据的整体分布情况、波动情况和数据异常情况。

为保证训练出的质量预测模型符合真实制造生产情况并具有一定泛化性，需要在过采样之前划分数据集，将训练数据集进行过采样而测试数据集保持不变，这样处理才能反映出预测模型性能的真实性和可靠性。

2.2　不平衡数据集处理

在精铸工艺参数数据集中，合格铸件数远大于不合格铸件数，存在严重的类别不平衡问题，若利用该数据集直接构建预测模型，则只有多数类实例（合格铸件）能够获得较高的预测精度，不能对少数类实例（不合格铸件）进行准确分类，难以应用于实际生产的质量预测模型构建。SyMProD 过采样方法充分考虑了数据分布情况并基于分布情况分配概率，能够生成高质量的平衡数据集，可以有效解决上述问题。

2.3　工艺参数特征降维

精铸过程铸件的质量特性受到众多工艺参数取值的影响，如果将所有工艺参数作为特征构建质量预测模型会使得模型复杂度较高、训练时间较长、参数调整难以进行。因此，利用 XGBoost 模型对所有特征的重要度排序，根据特征重要性的排名，将特征按照重要性从高到低进行排序，选择保留排名靠前的特征，而剔除排名靠后的特征，减少特征的数量，从而降低模型的复杂度和训练时间，同时保留对模型性能影响较大的特征。

2.4　Stacking 集成的质量预测模型构建

Stacking 集成学习模型通过结合多个基础模型的预测结果来生成最终的预测结果。在选择基分类器时需要考虑其组合方式的差异性，尽可能选择差异较大的学习算法，发挥不同模型的优势，避免相关性较高的模型重复学习导致过拟合。因此，第一层基分类器主要选用实用效果较好的 RF、SVM 以及 XGBoost。为了减少过拟合的风险，通常会采用五折交叉验证来对基分类器进行训练，将训练数据分成五个子集，然后进行五折交叉验证。在每一轮交叉验证中，其中 4 个子集用于训练基分类器，而剩下的 1 个子集则用于验证。这样，每个基分类器都会被训练 5 次，并在不同的验证集上进行验证。对于每个基分类器，将其在验证集上的预测结果整合起来，作为该分类器在整个训练数据上的预测结果，这样就可以得到一个在训练数据上经过交叉验证调优过的基分类器集合。在得到这些基分类器后，再利用这些基分类器的预测结果来训练元模型，从而得到最终的 Stacking 集成模型。LightGBM 在训练速度、内存占用、准确性和灵活性等方面都具有明显的优势，因此本文选择 LightGBM 模型作为元分类器。

2.5　模型性能评估

在处理不平衡数据时，常规的评价指标可能会受到数据不平衡的影响而产生偏见。从而无法提供准确的评估。例如，对于准确率，如果一个分类算法只能预测多数类，那么即使总体准确率很高，也可能会出现对少数类的预测准确率几乎为 0 的情况。因此，本文选用 A_{AUCROC}、G_{m} 以及 F_1 作为质量预测模型性能评估指标，在计算这几个模型性能评估指标时，以下几个变量常常被用到：真正例的数量（T_{P}）、真反例的数量（T_{N}）、假正例的数量（F_{P}）和假反例的数量（F_{N}），各项评价指标描述如下。

A_{AUCROC} 是受试者工作特征（ROC）即真正例率（T_{PR}）对假正例率（F_{PR}）曲线下的面积，A_{AUCROC} 越接近 1，表示模型分类性能越好

$$T_{\mathrm{PR}} = \frac{T_{\mathrm{P}}}{T_{\mathrm{P}} + F_{\mathrm{N}}} \tag{15}$$

$$F_{\mathrm{PR}} = \frac{F_{\mathrm{P}}}{F_{\mathrm{P}} + T_{\mathrm{N}}} \tag{16}$$

G_{m} 用来表示多数类和少数类精度的几何平均，可以同时反映出少数类和多数类的分类情况

$$G_{\mathrm{m}} = \sqrt{\frac{T_{\mathrm{P}}}{T_{\mathrm{P}} + F_{\mathrm{N}}}\left(\frac{T_{\mathrm{N}}}{T_{\mathrm{N}} + F_{\mathrm{P}}}\right)} \tag{17}$$

模型的精确度（P）和召回率（R）表示如下

$$P = \frac{T_{\mathrm{P}}}{T_{\mathrm{P}} + F_{\mathrm{P}}} \tag{18}$$

$$R = \frac{T_{\mathrm{P}}}{T_{\mathrm{P}} + F_{\mathrm{N}}} \tag{19}$$

F_1 综合考虑了模型的精确度和召回率，能够平衡这两个指标

$$F_1 = \frac{2PR}{P+R} \tag{20}$$

3 应用案例

3.1 数据描述

为了验证上述提出的面向不平衡数据集的质量预测方法，本文以高温透平叶片制造过程精铸工艺为例，构建质量预测模型，探究精铸工艺参数与质检参数之间的映射关系。实验数据来源于某汽轮机有限公司高温透平叶片精铸工艺制造生产数据。精铸制造流程较为复杂，图 6 所示为精铸工艺流程。对铸件质量产生影响的关键制造工序主要是制芯、制蜡、制壳、浇注 4 个部分，包含的参数称为工艺参数，铸件成品质量检测涉及的参数称为质检参数。由这两部分参数组成的数据集称为精铸工艺参数数据集。

对原始实验数据进行数据预处理以保证数据质量，利用均值、最小值、最大值以及变异系数进行数据描述，观察数据集整体分布情况与波动情况。变异系数在标准差的基础上引入均值，可以消除度量数据离散程度时受不同水平均值影响。通常，变异系数越小，说明数据的离散程度越小；反之，变异系数越大，说明数据的离散程度越大。精铸工艺参数集的数据描述情况见表 1。V 为浆料的黏度；pH 为浆料的酸碱度，下标 1、2、3、f 分别表示第 1、2、3 及封浆层；w(Si) 为浆料中硅元素的质量分数；T 为环境温度；$w(H_2O)$ 为浆料中水的质量分数，下标 2、3、f 分别表示第 2 层、第 3 层和封浆层。由表 1 可以看出，注蜡温度、型壳层数、T_2、浇注时长等工艺参数的变异系数为 0，表 1 中已经加粗处理，表明这些参数的样本数据没有变化，对于铸件后续的质量预测模型构建没有意义，需要将这些工艺参数剔除。经过预处理后，最终的精铸工艺参数数据共计 10 000 条，其中合格铸件为8 502 条，不合格铸件为 1 498 条。

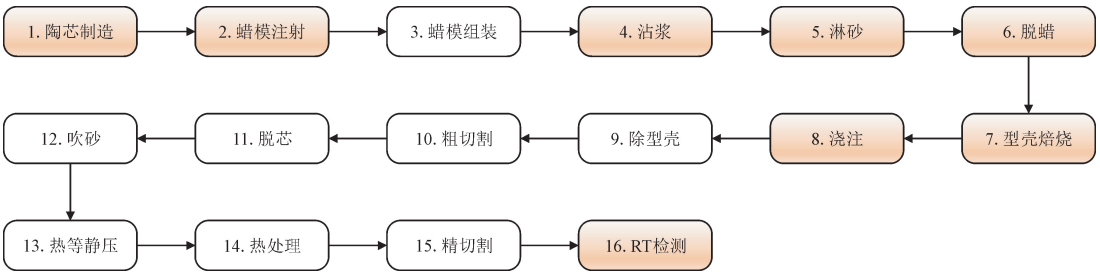

图 6 精铸工艺流程

Fig. 6 Manufacturing process of investment precision casting

表 1 精铸工艺参数数据描述

Table 1 Data description of process parameters in investment precision casting

工艺参数	数值				工艺参数	数值			
	最小值	最大值	平均值	变异系数		最小值	最大值	平均值	变异系数
低温强度/MPa	8.51	11.00	10.037	0.059 0	V_f	17.70	18.51	18.211	0.005 5
高温强度/MPa	20.01	25.00	22.660	0.0579	pH_f	8.08	8.84	8.502	0.029 2
材料变形量/mm	0.10	0.30	0.199	0.282 8	w_f(Si)/%	27.81	30.16	29.110	0.018 0
陶芯材料中方石英的质量分数	0.15	0.30	0.199	0.150 0	T_f/℃	**22.00**	**22.00**	**22.000**	**0**
					$w_f(H_2O)$/%	**42.00**	**42.00**	**42.000**	**0**
注蜡温度/℃	**68.00**	**68.00**	**68.000**	**0**	升压时长/s	9.00	11.00	10.007	0.081 5
型壳层数	**8.00**	**8.00**	**8.000**	**0**	脱蜡压力/MPa	0.53	0.58	0.563	0.019 8
V_2/(MPa·s)	42.55	43.95	43.281	0.004 7	脱蜡时长/s	294.00	305.00	300.290	0.009 6
pH_2	8.62	9.42	9.200	0.017 3	排气时长/s	248.00	259.00	253.012	0.011 9
w(Si)/%	25.00	27.17	26.437	0.013 7	清洗次数	6.00	9.00	7.508	0.147 8
T_2/℃	**22.00**	**22.00**	**22.000**	**0**	模壳转运时长/s	189.00	210.00	200.062	0.028 6
$w_2(H_2O)$/%	**62.00**	**62.00**	**62.000**	**0**	浇注温度/℃	1 481.00	1 484.00	1 482.299	0.000 7
V_3/(MPa·s)	18.04	18.70	18.356	0.005 7	浇注时长/s	**3.00**	**3.00**	**3.000**	**0**
pH_3	16.82	17.98	17.414	0.015 6	浇注质量/kg	28.80	29.60	29.184	0.006 8
w(Si)/%	28.25	30.41	29.324	0.018 3	重熔浮渣比/%	0.01	0.06	0.043	0.251 8
T_3/℃	**23.00**	**23.00**	**23.000**	**0**	熔炼真空度	0.04	0.39	0.219	0.469 5
$w_3(H_2O)$/%	**42.00**	**42.00**	**42.000**	**0**	坩埚使用次数	1.00	6.00	3.326	0.446 1

注：加粗表示这些参数的样本数据没有变化，应剔除这些参数。

一般地，当多数类和少数类之间的比例超过 4∶1，即类不平衡率超过 4，就可以认为数据集是不平衡的，需要应用少数类组处理，这是由于当类不平衡率超过 4 时，在构建产品质量预测模型前若不进行少数类组处理，容易出现模型性能偏差，即模型更倾向于预测多数类而忽略少数类。

铸件质量由射线检测（RT）结果决定，结果有合格与不合格两种质量状态，精铸工艺参数数据集中合格铸件数远大于不合格铸件数，存在严重的类别不平衡问题，类不平衡率为 5.67，铸件质量类不平衡度情况如图 7 所示，图中比例表示合格铸件数或不合格铸件数占总铸件数的比值。

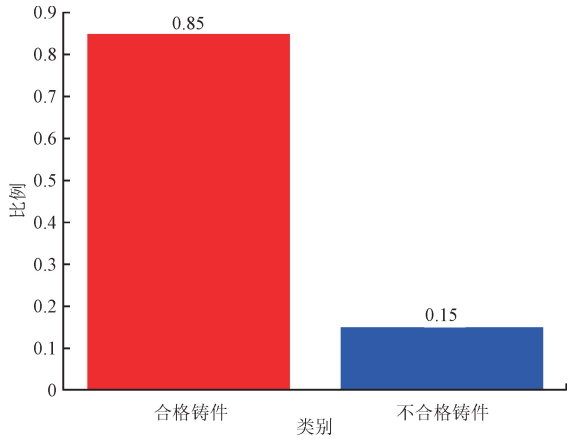

图 7　铸件质量类别的不平衡度

Fig. 7　Category balance ratio of quality state of castings

3.2　不平衡数据集处理前后模型性能对比实验

本实验探讨利用 SyMProD 过采样方法对精铸工艺参数不平衡数据集处理后模型的性能提升情况，对比实验运行环境见表 2。

表 2　对比实验运行环境

Table 2　Operating environment

名称	型号	名称	型号
处理器	i7-12700H	Python	3.9.7
显卡	3060	Pycharm	2021.2.2
运行内存	16 GB	Sklearn	1.2.2
操作系统	Win11	Imblearn	0.10.1

在过采样与特征降维之前，将原始数据的 80% 作为训练集，20% 作为测试集。利用 XGBoost 特征重要性进行特征降维，剔除部分排名靠后的特征，特征选择结果为转运时长、浇注温度、浇注质量、重熔浮渣比、熔炼真空度、坩埚使用次数、脱蜡时长、脱蜡压力、升压时长以及排气时长共 10 个关键工艺参数特征，作为后续预测模型构建的输入特征。针对原始数据集与过采样后的数据集，分别使用 3 种分类算法 LightGBM、RF 以及 SVM 构建预测模型对比过采样前后模型性能。其中，SyMProD 过采样方法 K 最近邻和 M 最近邻均设置为 5，噪声阈值（N_T）设置为 3，尽可能保留数据分布情况，过采样率设为 1。截止阈值（C_T）使用管道方式与后续分类算法组合进行网格搜索（GridSearchCV）获取最优值。为了更好地得到预测结果，实验结果取五折交叉验证的平均值，对比结果如图 8 所示。由图 8 可以看出，经过过采样后，虽然整体上铸件的预测准确率有轻微下降，而对于少数类别即铸件不合格品的预测准确率提升了 75.4%，更加满足实际生产情况。无论采用哪种机器学习分类算法，经过 SyMProD 过采样方法后的数据集构建的预测模型，其 A_{AUCROC}、G_{m}、F_1 相比于原始不平衡数据集构建的预测模型都有一定的提升，A_{AUCROC} 增大了 24%～28%，G_{m} 增大了 30%～35%，F_1 增大了 32%～50%。说明了 SyMProD 过采样方法能够有效解决精铸工艺参数不平衡数据集严重的类别不平衡问题，利用过采样后的数据集构建质量预测模型能够显著提高预测少数类的准确度和整体模型预测效果。

3.3　过采样方法性能对比实验

本实验将 SyMProD 过采样方法与 SMOTE、ADASYN、K-Means-SMOTE 算法对原始数据进行扩充处理，采用 3 种不同的分类算法（LightGBM、RF、SVM）分别利用平衡后的数据集构建铸件质量预测模型进行对比实验。SyMProD 过采样方法设置 $K=5$，$M=5$，$N_T=3$，C_T 利用管道组合的方式进行网格搜索最优值，其他过采样方法均采用 Python 算法库 imblearn 对数据进行处理，所有过采样方法的过采样率均设置为 1，各个分类算法参数均按 3.2 节进行设置与调优。各项模型评估指标均采用五折交叉验证求取平均值，对比实验结果见表 3。由表 3 可以看出，SyMProD 过采样方法在 3 种分类算法情况下，预测模型评估指标几乎都取得了最优，因此，在精铸工艺参数不平衡数据集的数据分布情况与复杂性下，采用 SyMProD 方法相比于其他过采样方法能够更好地改善铸件质量预测模型。

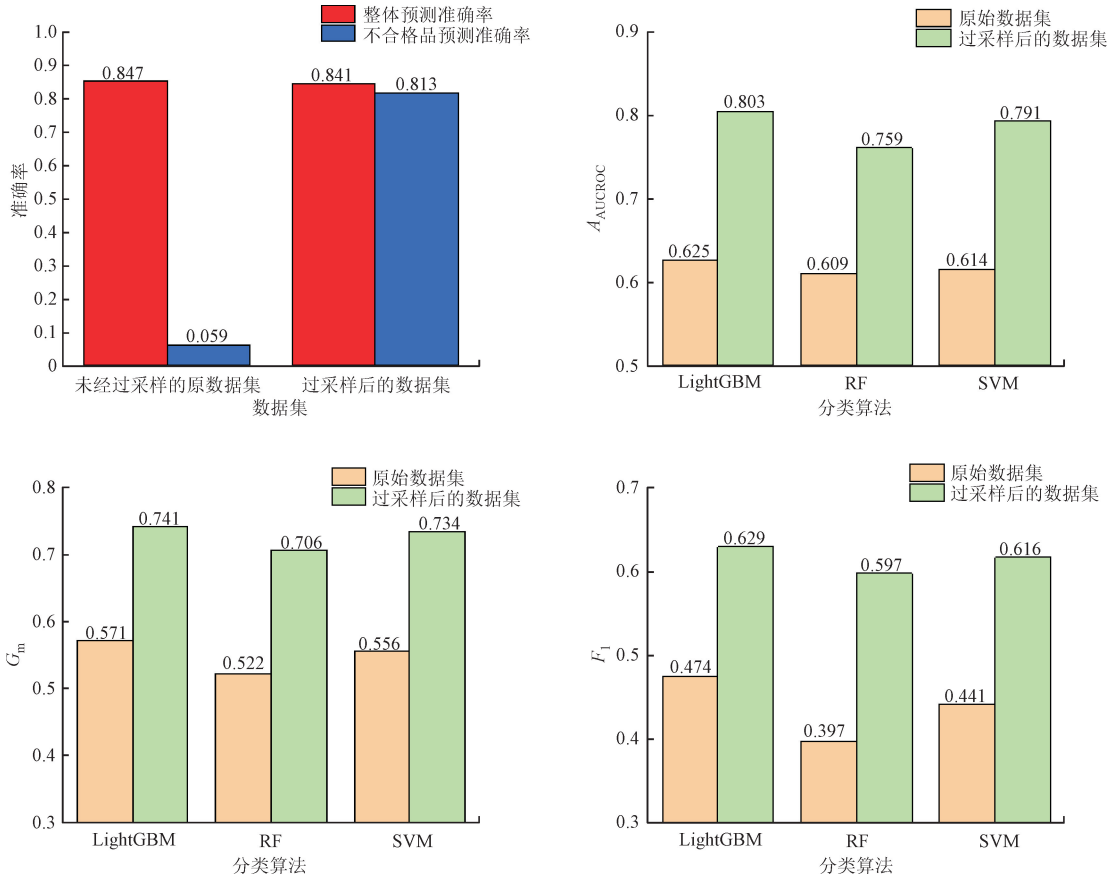

图 8 SyMProD 过采样前后不同分类算法性能对比结果

Fig. 8 Performance comparison results of different classification algorithms before and after SyMProD oversampling

表 3 过采样方法性能对比实验结果

Table 3 Experimental results of performance comparison for oversampling methods

分类方法	过采样方法	A_{AUCROC}	G_m	F_1
LightGBM	SMOTE	0.724	0.697	0.512
	ADASYN	0.737	0.682	0.504
	K-means-SMOTE	0.780	0.721	0.573
	SyMProD	**0.803**	**0.741**	**0.629**
RF	SMOTE	0.701	0.677	0.486
	ADASYN	0.696	0.654	0.507
	K-means-SMOTE	0.746	0.719	0.574
	SyMProD	**0.759**	0.706	**0.597**
SVM	SMOTE	0.712	0.681	0.497
	ADASYN	0.719	0.691	0.528
	K-means-SMOTE	0.775	0.714	0.564
	SyMProD	**0.791**	**0.734**	**0.616**

注:加粗数据表示每种学习算法对应的最高平均模型性能指标。

3.4 分类算法性能对比实验

本实验探讨利用 Stacking 集成学习分类算法相比于单一分类算法 LightGBM、XGBoost、SVM 以及 RF 构建质量预测模型性能的提升情况。每种算法均采用 SyMProD 过采样方法平衡数据。其超参数设置为 $K=5$、$M=5$、$N_T=3$、$C_T=1.2$,过采样率为 1.0。Stacking 集成模型的第一层采用五折交叉验证,其中基分类器的输入为经过特征降维后原始数据的训练集 (8 000,10)以及原始数据对应的标签 y,每个基分类器的输出为训练集对应的预测值;元分类器的输入应为 3 个基分类器输出数据的组合(24 000,10)以及基分类器的预测标签 y_1,元分类器的输出就是整个 Stacking 集成学习模型的预测值。单一分类算法的实验结果取五折交叉验证的平均值,实验结果见表 4。由表 4 可知,Stacking 集成学习算法预测模型性能最优。相比于单一模型中表现较好的 LightGBM 算法,Stacking 集成学习算法在 A_{AUCROC} 的性能上提升了 5.48%,G_m 提升了 3.78%,F_1 提升了 5.72%。相比于最差的 RF 算法,3 项性能指标分别提升了 11.59%、8.92%、11.39%。结果表明 Stacking 集成学习算法能够结合多个基

模型的预测结果，相比单一模型更适应复杂多变的高温透平叶片精铸过程生产制造场景。

表 4　分类算法性能对比实验结果

Table 4 Experimental results of performance comparison for classification algorithms

算法	A_{AUCROC}	G_{m}	F_1
Stacking	0.847	0.769	0.665
LightGBM	0.803	0.741	0.629
XGBoost	0.794	0.736	0.621
SVM	0.791	0.724	0.616
RF	0.759	0.706	0.597

4　结　论

本文提出了 SyMProD-Stacking 集成学习的铸件质量预测方法。针对精铸工艺参数数据集的不平衡问题，将原始数据集经预处理后进行了对比实验，利用 SyMProD 过采样方法对数据进行平衡，相比于原始不平衡数据集，铸件不合格品的预测准确率提升了 75.4%，A_{AUCROC}、G_{m} 以及 F_1 性能指标分别增大了 24%～28%、30%～35%、32%～50%，同时，与其他过采样方法相比，该方法在本文实验数据集分布情况下各个性能指标几乎都取得了最优。为满足企业对高温透平叶片精铸过程铸件质量预测的迫切需求，提出了一种基于 Stacking 集成学习的重燃叶片精铸过程铸件质量预测模型构建方法，帮助企业提前制定响应措施，降低成本。利用 SyMProD 过采样方法对数据集进行处理并采用 Stacking 集成学习构建质量预测模型相比于单一算法模型 3 项性能指标分别增大 5.48%～11.59%、3.78%～8.92%、5.72%～11.39%。后续研究工作中将基于该预测模型，探究如何利用多目标优化算法和多属性决策对各个工艺参数进行事前优化调控。

参考文献：

[1] 邹昌利，张恃铭，刘银华. 面向装配质量预测控制的随机 Kriging 建模方法 [J]. 机械设计与研究，2020，36(2)：108-112.
ZOU Changli，ZHANG Shiming，LIU Yinhua. Study on the stochastic Kringing modeling for predictive quality control in assembly process [J]. Machine Design & Research，2020，36(2)：108-112.

[2] 赵京鹤，高明，梁楠. 工业互联网助推制造企业数字化转型 [J]. 中国自动识别技术，2022(1)：58-60.
ZHAO Jinghe，GAO Ming，LIANG Nan. Industrial internet promotes digital transformation of manufacturing enterprises [J]. China Auto-ID，2022(1)：58-60.

[3] GUO Haixiang，LI Yijing，SHANG J，et al. Learning from class-imbalanced data：review of methods and applications [J]. Expert Systems with Applications，2017，73：220-239.

[4] YADAV S S，BHOLE G P. Learning from imbalanced data in classification [J]. International Journal of Recent Technology and Engineering，2020，8(5)：1007-1016.

[5] DOUZAS G，BACAO F，LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE [J]. Information Sciences，2018，465：1-20.

[6] PIRI S，DELEN D，LIU Tieming. A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets [J]. Decision Support Systems，2018，106：15-29.

[7] 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 Youqiang，et al. Review of imbalanced data classification methods [J]. Control and Decision，2019，34(4)：673-688.

[8] 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(1)：321-357.

[9] HAN Hui，WANG Wenyuan，MAO Binghuan. Borderline-SMOTE：a new over-sampling method in imbalanced data sets learning [C]//Advances in Intelligent Computing. Berlin，Germany：Springer，2005：878-887.

[10] HE Haibo，BAIYang，GARCIA E A，et al. ADASYN：adaptive synthetic sampling approach for imbalanced learning [C]//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Piscataway，NJ，USA：IEEE，2008：1322-1328.

[11] DOUZAS G，BACAO F，LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE [J]. Information Sciences，2018，465：1-20.

[12] 李敏波，董伟伟. 面向不平衡数据集的汽车零部件质量预测 [J]. 中国机械工程，2022，33(1)：88-96.
LI Minbo，DONG Weiwei. Quality prediction of automotive parts for imbalanced datasets [J]. China Me-

chanical Engineering, 2022, 33(1): 88-96.

[13] KUNAKORNTUM I, HINTHONG W, PHUNCHONGHARN P. A synthetic minority based on probabilistic distribution (SyMProD) oversampling for imbalanced datasets [J]. IEEE Access, 2020, 8: 114692-114704.

[14] ZHOU Ping, JIANG Yue, WEN Chaoyao, et al. Data modeling for quality prediction using improved orthogonal incremental random vector functional-link networks [J]. Neurocomputing, 2019, 365: 1-9.

[15] ZHENG Wenjian, LIU Yi, GAO Zengliang, et al. Just-in-time semi-supervised soft sensor for quality prediction in industrial rubber mixers [J]. Chemometrics and Intelligent Laboratory Systems, 2018, 180: 36-41.

[16] 高建琛. 基于集成学习的铝合金薄板带材力学性能预测方法研究 [D]. 北京：北京工业大学，2020.

[17] LU Ningyun, GAO Furong. Stage-based online quality control for batch processes [J]. Industrial & Engineering Chemistry Research, 2006, 45(7): 2272-2280.

[18] 李传维. 核电压力容器大型锻件组织与性能研究及热处理数值模拟 [D]. 上海：上海交通大学，2016.

[19] 杨健，吴思炜. 基于机器学习的钢铁轧制过程性能预测 [J]. 钢铁，2021，56(9)：1-9.
YANG Jian, WU Siwei. Property prediction of steel rolling process based on machine learning [J]. Iron & Steel, 2021, 56(9): 1-9.

[20] 董海，冯晔. 基于 XGBoost 的车身尺寸装配质量智能预测模型 [J]. 工业工程，2021，24(3)：77-82.
DONG Hai, FENG Ye. An intelligent prediction model of body size assembly quality based on XGBoost algorithm [J]. Industrial Engineering Journal, 2021, 24(3): 77-82.

[21] DUAN Guijiang, YAN Xin. A real-time quality control system based on manufacturing process data [J]. IEEE Access, 2020, 8: 208506-208517.

[22] 赵双凤. 基于 MES 的机加车间制造过程工序质量控制方法与系统研究 [D]. 重庆：重庆大学，2016.

[23] 于文靖. 汽轮机模锻叶片加工质量预测及控制方法研究 [D]. 哈尔滨：哈尔滨工业大学，2018.

[24] 钟武昌，战洪飞，林颖俊，等. 基于机器学习多算法集成的产品质量问题预测方法 [J]. 机械设计与研究，2023，39(5)：100-107.
ZHONG Wuchang, ZHAN Hongfei, LIN Yingjun, et al. Product quality problem prediction method based on machine learning multi algorithm integration [J]. Machine Design & Research, 2023, 39(5): 100-107.

[25] 向峰，杨磊，张萌，等. 基于模型融合的复杂生产过程产品质量预测 [J]. 中国科学(技术科学)，2023，53(7)：1127-1137.
XIANG Feng, YANG Lei, ZHANG Meng, et al. Model fusion based product quality prediction for complex manufacturing process [J]. Scientia Sinica(Technologica), 2023, 53(7): 1127-1137.

（编辑 武红江）