基于数据驱动的镁合金压铸件质量智能预测

汪星辰[1] 王鑫[1] 付彭怀[1] 童胜坤[2] 陈滨[2] 彭立明[1]

（1.上海交通大学材料科学与工程学院，上海 200240；2.万丰镁瑞丁新材料科技有限公司，绍兴 312500）

摘要：为实现镁合金压铸件质量的智能预测，降低人工下线检测成本，提升镁合金压铸产业智能化水平，通过收集镁合金大型薄壁压铸件“工艺参数-质量参数”大数据，采用随机森林模型建立工艺参数与铸件产生的缺陷种类间的关系，分析了工业数据中的标签长尾分布现象对机器学习模型预测性能的影响，通过“随机降采样+SMOTE过采样”算法对数据集分布进行均衡化，最终获得了准确率为89.54%、受试者工作特征曲线（ROC）下面积为0.983 8、平均真正率为87.65%的准确预测模型，实现了极少数含缺陷样本的精准检出，并获得了镁合金压铸关键工艺参数重要性排序。

关键词：高压铸造 镁合金 机器学习 质量智能预测

中图分类号：TG292；TG249.2 **文献标志码：**B **DOI:** 10.19710/J.cnki.1003-8817.20240363

Intelligent Quality Prediction of Magnesium Alloy Die-Casting Parts Based on Data-Driven Method

Wang Xingchen[1], Wang Xin[2], Fu Penghuai[1], Tong Shengkun[2], Chen Bin[2], Peng Liming[1]

(1. School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240; 2. Meridian Lightweight Technologies, Shaoxing 312500)

Abstract: In order to achieve intelligent predication of magnesium alloy die-casting parts, reduce offline labor inspection cost, and improve intelligent level of magnesium alloy die-casting industry, this paper collects big data on “process parameters-quality parameters” of large thin-walled magnesium alloy castings, and uses random forest model to establish the relationship between process parameters and the types of defects in castings, and analyzes the effect of long-tailed distribution of labels in the industrial data on the predictive performance of machine learning models. Then the “Random Downsampling + SMOTE Over-sampling” algorithm is emptoyed to balance the distribution of the data set. Finally, an accurate prediction model with an accuracy of 89.54%, an area under ROC curve of 0.983 8, and an average true rate of 87.65% are obtained, which achieves a precise detection of a small number of defective samples, and obtains the ranking of the importance of key process parameters for magnesium alloy casting.

Key words: High pressure die-casting, Magnesium alloy, Machine learning, Intelligent quality prediction

1 前言

采用压铸铝、镁合金生产车身大型薄壁一体化结构件可大幅减少零件设计、装配与焊接工序数量，从而减少生产能源消耗，并获得显著的车身轻量化效果。镁合金密度比铝合金更低，具有更广阔的应用前景。对于大型薄壁一体化结构件而言，气孔、缩孔、裂纹以及冷隔等压铸缺陷均会对

作者简介：汪星辰（1991—），男，助理研究员，博士学位，研究方向为镁/铝合金智能研发与制造。

通信作者：付彭怀（1980—），男，副研究员，博士学位，研究方向为镁/铝合金材料与铸造、增材制造技术。

基金项目：国家重点研发计划项目（2021YFB3701000）；宁波科技项目(20241ZDYF020400)；自然科学基金项目（U21A2048、51821001）；广东省重点领域研发计划（2020B010186001）；上海市科委重点项目(21DZ1208200)；广东省基础与应用基础研究基金(2022B1515120046)。

参考文献引用格式：

汪星辰, 王鑫, 付彭怀, 等. 基于数据驱动的镁合金压铸件质量智能预测[J]. 汽车工艺与材料, 2025(4): 40-45.

WANG X C, WANG X, FU P H, et al. Intelligent Quality Prediction of Magnesium Alloy Die-Casting Parts Based on Data-Driven Method[J]. Automobile Technology & Material, 2025(4): 40-45.

最终服役性能产生巨大的负面影响。与铝合金压铸过程相比,镁合金压铸过程中合金熔体流动规律、凝固规律、缺陷形成机制与影响因素的研究较少,铸件品质控制更困难,导致镁合金压铸件次品率较高。

为更好地研究镁合金压铸过程缺陷形成规律,并准确预测缺陷的形成,需建立完善的“工艺参数-压铸缺陷”内在关系模型。传统研究多通过试验试错建立工艺、组织、性能之间的映射关系模型,从而指导压铸工艺参数控制与优化。例如,Ma等[1]通过压铸试验与电子计算机断层扫描(Computed Tomography,CT)三维重建技术,研究了不同慢压射速率对AE44镁合金压铸缩孔形成的影响。Hou等[2]对AE44镁合金中缺陷带的形貌、组成以及形成规律进行了系统性研究,初步揭示了缺陷带的形成机制与影响因素。由于此类研究计算与试验成本高、周期长,尽管能科学地揭示部分缺陷的产生机制,但仅能聚焦于少量样本中的少量影响因素的定性分析。大型一体化压铸工艺参数众多,可监测的工艺参数高达3 000余个,其中,上百个工艺参数对气孔、缩孔、冷隔等不同铸造缺陷均有复杂的影响。采用传统的“经验+试错”模式无法准确预测镁合金压铸生产过程中多物理场耦合、多参数共同作用情况下铸造缺陷的产生情况与铸件的最终质量。

采用基于数据驱动的机器学习模型可有效建立高维度映射关系模型,在材料研发、制造领域已获得广泛的关注与应用。例如,Xie等[3]以热轧钢板成分和工艺参数等27个材料参数为模型输入,以屈服强度、抗拉强度、延伸率和冲击功等力学性能为输出,建立深度学习模型,准确预测了热轧钢板力学性能。Xu等[4]利用人工神经网络与支持向量机建立了AZ31镁合金成分与挤压比、轧制比等多种加工工艺对屈服强度、抗拉强度与延伸率间的内在映射关系,获得了力学性能精确预测模型。除此之外,多项研究表明,机器学习模型可有效建立各类金属材料“工艺”与“性能”间的复杂映射关系[5-10]。

目前,基于机器学习的压铸生产过程智能控制研究较少,尚处于起步阶段。本文通过收集镁合金大型薄壁压铸件生产数据,采用随机森林模型建立工艺参数与铸件质量参数间的关系,分析工业数据中的长尾分布现象对模型预测性能的影响,获得较为准确的铸件质量预测模型,并对关键工艺参数重要性进行排序,以期为揭示大型薄壁铸件压铸过程中的缺陷形成机制、推动镁合金压铸件质量智能预测与工艺参数智能优化技术发展提供算法模型基础。

2　数据集建立

2.1　压铸大数据采集

对3 200 t压铸机与其周边设备进行压铸生产数据采集,并对每一模次产品进行质量检验,获得了54 931条新能源汽车镁合金仪表盘支架(Cross Car Beam,CCB)的压铸数据样本。每一条样本包含产品模次编号、压铸始末时间节点、71个重要压铸工艺参数以及5个表征铸件质量的质量参数,如表1所示。其中,质量参数为欠铸、冷隔、裂纹、孔洞、拉模/粘模5种铸造缺陷的出现情况,值为0或1,0代表该组工艺参数下对应的压铸件没有产生上述5种铸造缺陷,1则代表产生了铸造缺陷。

表1　镁合金CCB压铸产线原始数据集的数据结构

参数类型	参数数量
模次编号	1
生产时间节点	2
压铸工艺参数	71
质量参数	5

2.2　工艺参数(特征值)降维

为实现压铸件质量的智能实时预测,须建立镁合金压铸过程工艺参数与质量参数间的映射关系模型。其中,工艺参数作为模型的输入值(特征值),质量参数作为模型的输出值(标签值)。原始数据集包含了71个压铸工艺参数,若将其全部作为输入特征值建立机器学习模型,可能会因为维度过高而增加过拟合的风险与计算成本。在此71个工艺参数中,并不是每个工艺参数均对铸件质量有独立的影响,部分工艺参数可能存在线性相关性。理想情况下,一个机器学习模型的输入特征应该由一组N个互相独立(线性不相关或弱相关)的特征组成,因此,需要通过相关性计算分析

不同参数间的相关性系数，并删除其中线性强相关的特征，实现模型输入参数的降维与进一步筛选。本文通过计算71个工艺参数两两间的皮尔逊系数，并形成相关性矩阵进行工艺参数的降维。工艺参数A与工艺参数B间的皮尔逊系数可由如下公式获得：

$$\rho(A,B)=\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right) \quad (1)$$

式中：μ_A和σ_A分别为参数A的平均值和标准差，μ_B和σ_B分别为参数B的平均值和标准差，i为样本序号，N为样本的总数。

图1所示为71个压铸工艺参数的相关性矩阵热图，其中，横坐标与纵坐标的数字代表工艺参数在原始数据表中的序号，颜色深浅代表皮尔逊系数大小（亮黄色表示皮尔逊系数为-1，即完全负线性相关；黑色表示皮尔逊系数为1，即完全正线性相关）。由图1可知，部分工艺参数间具有极强的线性相关性。

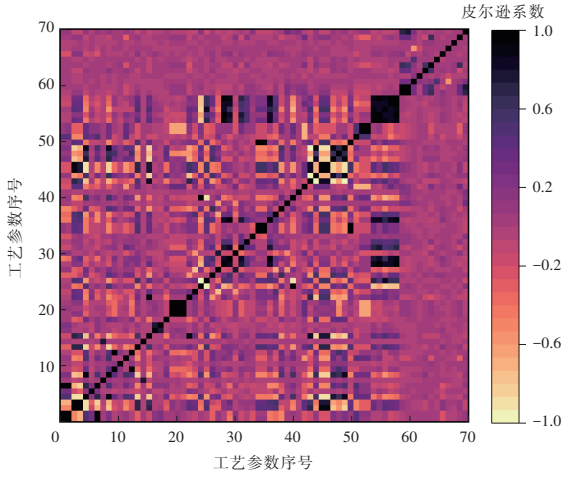

图1 71个压铸工艺参数的相关性热图

在本文中，若皮尔逊系数满足$|\rho(A,B)|>0.6$，则视为A与B具有强线性相关性，并按如下规则进行工艺参数筛选降维：若一组工艺参数$\{B,C,D,...\}$均与某一工艺参数A强线性相关，则只保留工艺参数A作为关键工艺参数，并从数据集中删除其余工艺参数所对应的数据列。工艺参数A与$\{B,C,D,...\}$之间的线性依赖关系通过表格的形式记录，以便于后续工艺参数优化时进行查询。最终获得23个线性弱相关或不相关的工艺参数作为模型最终输入，如表2所示。

表2 降维后的压铸工艺参数列表

序号	参数名称	序号	参数名称
1	合型偏移时间	13	充填时间
2	压射补偿时间	14	增压时间
3	开模补偿时间	15	增压延时时间
4	取件机补偿时间	16	压射持续时间
5	合型时间	17	熔化定量持续时间
6	开模持续时间	18	闭环行程2
7	动模温度	19	动模压力
8	模温机入口管道1温度	20	最大压力
9	液管1温度	21	填充压力
10	压铸机油温	22	铸造压力
11	铝液液位	23	1号大杠锁模力中值
12	料饼厚度		

2.3 铸件质量参数（标签值）的前处理

一般来说，一个压铸件可能会同时存在多种影响品质的缺陷，缺陷种类越多，造成产品报废的风险越大，其品质越差。因此，本文将一个压铸件产生的缺陷种类数作为判定铸件品质优劣、评价等级的标准，对铸件进行分级。若质量等级为0级，代表该铸件产生的缺陷种类数为0级，即没有出现上述任何一种缺陷，质量最优；若质量等级为5级，代表该铸件在该工艺参数下同时产生的缺陷种类数为5种，即产生了欠铸、冷隔、裂纹、孔洞、拉模/粘模5种铸造缺陷，铸件质量最差。在本数据集中，没有同时产生4种或5种缺陷的压铸件样本，因此，质量等级范围为0~3级，共4类，在后文中将以第0、1、2、3类进行描述。将降维后的23种压铸工艺参数与对应的质量等级结合，形成用于机器学习的数据集，从而建立压铸件质量预测模型建立。

3 镁合金压铸件质量预测模型建立

3.1 随机森林模型建立

随机森林（Random Forest，RF）模型是一种常用的集成学习模型。一个随机森林模型由N个决策树模型组成，模型的输出为N个决策树模型输出的多数决定投票结果（分类问题）或平均值（回归问题）。随机森林模型的预测准确性与决策树数量、划分方法、最大深度、最大特征数、最小叶

节点样本数等超参数有关。其中,决策树数量N是对随机森林模型预测准确性影响最大的超参数,N过小会引起模型欠拟合,准确性不足,N过大则会增加模型复杂程度,造成过拟合并增加运算成本,因此,在模型建立时需对其进行优化。本文采用Python第三方库Scikit-Learn中的Random Forest Classifier函数建立随机森林模型,划分标准(Criterion)选择基尼系数(Gini),最大深度(Max_Depth)、最小叶节点样本数(Min_Samples_Split)、最大特征数(Max_Feature)等超参数设定为默认值,考察决策树数量N与模型训练损失之间的关系,如图2所示。可以发现,当决策树数量为1,即随机森林模型退化为普通决策树模型时,模型在数据集上的训练损失较大。随着决策树数量的增加,模型训练损失迅速下降。当N>150个时,训练损失趋于收敛,不再随着决策树数量增加而减小。因此,本文选定N=500个为最终决策树数量。

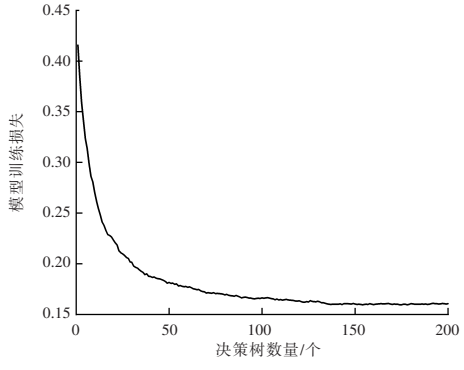

图2　决策树数量与RF模型预测误差间的关系

本文采用准确率(Accuracy)、平均真正率(Average True Positive Rate,ATPR)以及受试者工作特征曲线(Receiver Operating Characteristic Curve,ROC)下面积(Area Under ROC Curve,AUC)3个指标来表征模型预测的准确性。其中,准确率P_{Accuracy}代表模型总体预测能力,计算方法如下:

$$P_{\text{Accuracy}} = \frac{N_{\text{true}}}{N_{\text{total}}} \tag{2}$$

式中:N_{true}为预测正确的样本数,N_{total}为样本总数。

平均真正率P_{ATPR}代表每一类的检出能力的平均值:

$$P_{\text{ATPR}} = \frac{1}{n}\sum_{i=0}^{n}\frac{TP_i}{TP_i + FN_i} \tag{3}$$

式中:n为分类类别数,TP_i为类别i的预测真正数,FN_i为类别i的预测假负数。

AUC反映了模型总体预测性能的稳健性,通Scikit-Learn的Metrics子库中的roc_auc_score函数计算获得。结合3个不同的指标可对模型预测能力进行综合评估。

3.2　数据分布对模型预测能力的影响

对数据集进行分割,随机取其中20%的样本作为测试集,不参与模型训练,剩余80%的样本作为训练集与验证集参与模型训练。采用随机森林模型(N=500个)对训练集进行训练,验证方法为五折验证法,并在测试集上进行预测准确性测试,测试集混淆矩阵如表3所示。模型预测总准确率为93.67%,AUC为0.970 8(1.0为最优),预测准确率极高。然而由表3所示的混淆矩阵可知,在原始数据分布情况下,模型对第0类样本具有极高的预测准确性(真正率TPR高达99.42%,即9 788个第0类样本中有9 731个预测正确)。而对于其他3个类别的样本的预测真正率分别仅为48.35%、31.94%、18.18%,导致平均真正率A_{TPR}仅为49.47%。模型有极大的概率将此3类样本分为无缺陷的良品,存在严重的应用安全隐患。

表3　原始数据分布下的测试集混淆矩阵

	预测0/个	预测1/个	预测2/个	预测3/个	TPR/%
真实0	9 731	52	5	0	99.42
真实1	577	543	3	0	48.35
真实2	39	10	23	0	31.94
真实3	9	0	0	2	18.18

产生上述现象的原因为原始数据标签值分布不均匀。原始训练集的标签值分布如表4所示,第0类(即完全没有出现任何致命缺陷)的良品铸件有39 132个,远远大于第1类、第2类、第3类铸件的数量,第3类(即同时出现3种缺陷)铸件数量仅为34个。该种非均匀的标签分布称为“长尾分布”,在长尾分布数据集训练过程中,多数类会被过度学习,而少数类则得不到学习,从而导致模型趋向将所有输入均预测为多数类,少数类便无法被有效检出,对机器学习

模型的可靠性造成严重的负面影响。在压铸件质量预测问题中，只有将包含缺陷的产品（少数类）精确检出才能有效地对工艺参数进行智能评估与优化，并将报废件精确淘汰。因此，需调整原始数据分布，解决数据长尾分布带来的负面影响。

表4 原始训练集标签值分布情况

质量等级（缺陷种类数）	样本数/个
0	39 132
1	4 487
2	320
3	34

解决数据长尾分布的主要方法为降采样与过采样。降采样算法从多数类中筛选出合适数量的样本，减少多数类的数量，与少数类形成新的数据集。过采样算法通过一定的算法增加少数类的样本数，使其达到与多数类接近的数量。由于本文中原始数据集多数类与少数类样本数量差距较大，仅采用降采样算法会造成模型对数据集欠拟合，仅采用过采样算法会造成对数据集的过拟合。因此，本文采用降采样与过采样相结合的方法对数据分布进行均衡化处理。其中，采用随机降采样法（Random Down Sampling）对第0类的数据进行降采样，使其样本数降为10 000个。同时，采用合成少数过采样技术（Synthetic Minority Oversampling Technique，SMOTE）算法将第1类、第2类、第3类样本数分别提升至10 000个。SMOTE算法会根据原始数据之间的关系进行新样本的生成，所生成的新样本与原始数据不同，与大量产生重复数据的随机过采样法（Random Oversampling）相比，不易产生过拟合问题。

采用相同的随机森林模型对分布均衡化后的新训练集进行训练，并在相同的测试集上测试模型预测结果，混淆矩阵如表5所示。与原始数据分布情况下的模型相比，采用均衡数据进行训练的模型准确率下降为89.54%，但其AUC反而上升至了0.983 8，ATPR提升至87.65%，模型以第0类预测准确率的略微下降为代价，获得第1类、第2类、第3类预测准确率的大幅提升。对于第3类，尽管其测试集仅有11个样本，占总测试集样本数的0.1%，但其中10个样本被成功检出，可见数据分布均衡化后的模型具有对少数类极强的预测性能，可精确检出次品铸件、报废铸件。

表5 数据分布均衡后的测试集混淆矩阵

真实	预测0个	预测1个	预测2个	预测3个	TPR/%
真实0	8 737	995	55	1	89.26
真实1	72	1 041	10	0	92.70
真实2	4	12	56	0	77.77
真实3	0	0	1	10	90.90

3.3 镁合金压铸工艺参数的重要性排序

随机森林分类模型可通过基尼重要性、平均降低准确性、排列重要性、夏普利边际贡献值（Shapley Additive Explanation，SHAP）重要性等算法实现输入参数的重要性计算与排序，从而获得对模型输出结果影响最大的关键输入参数。由于本文采用的随机森林模型采用了基尼系数作为划分方法，可直接通过模型训练时每个特征在决策树中的平均基尼不纯度减少值来获得23种压铸工艺参数的基尼重要性，按从大到小的顺序排列后形成相对重要程度（Relative Importance），如图3所示。比压、填充时间等重要工艺参数与金属液填充模具时的熔体流态紧密相关，会影响冷隔、气孔等缺陷的产生概率。模温机温度与模具温度对金属液凝固过程具有重要的影响，温度过高或过低会造成冷隔、缩孔、粘模。模温机部分区域的温度则代表了模具温度与热流的不均匀性，与模具总体平均温度相比，对压铸件局部区域凝固缺陷的产生概率具有更大的影响。由于本模型的输入工艺参数在前期降维时通过相关性计算删去了大量线性相关工艺参数，因此，该相对重要性排序可扩展为71种工艺参数的重要性排序，其中，线性相关工艺参数间重要性相同，在工艺参数优化时选择其中可调节的工艺参数进行微调。该研究结果为压铸缺陷形成机制研究以及压铸工艺参数智能优化提供了关键工艺参数选择。

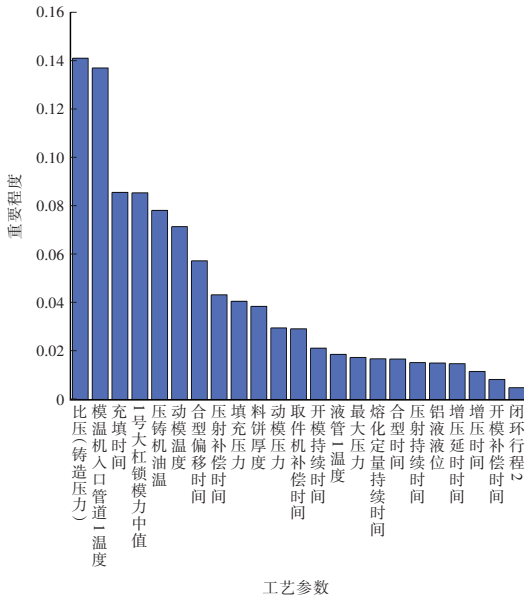

图3　23种镁合金压铸件关键工艺参数的相对重要性

4　结束语

本文建立了镁合金压铸“工艺参数-质量参数”工业大数据集,通过相关性分析有效实现了工艺参数降维,大幅降低了机器学习模型的规模。工业大数据的长尾分布现象会使模型预测结果严重偏离正确值,降低含缺陷样品的检出能力,采用随机降采样与SMOTE过采样相结合的方法进行数据分布均衡化,可有效提升含缺陷样本的检出能力。通过建立随机森林模型与超参数优化,最终获得准确率为89.54%、AUC为0.983 8、P_{ATPR}为87.65%的镁合金压铸件质量预测模型,实现次品、报废品的精确预测与识别。同时,利用基尼重要性算法对影响铸件质量的关键工艺参数进行了重要性排序,为压铸工艺参数智能反向优化控制提供了研究基础与理论依据。

参考文献:

[1] MA C S, YU W B, ZHANG T T, et al. The Effect of Slow Shot Speed and Casting Pressure on the 3D Microstructure of High Pressure Die Casting AE44 Magnesium Alloy[J]. Journal of Magnesium and Alloys, 2023, 11(2): 753-761.

[2] HOU Y Y, WU M W, HUANG F, et al. Defect Band Formation in High Pressure Die Casting AE44 Magnesium Alloy[J]. China Foundary, 2022, 19(3): 1-10.

[3] XIE Q, SUVARNA M, LI J L, et al. Online Prediction of Mechanical Properties of Hot Rolled Steel Plate Using Machine Learning[J]. Materials Design, 2021, 197.

[4] XU X N, WANG L Y, ZHU G M, et al. Predicting Tensile Properties of AZ31 Magnesium Alloys by Machine Learning[J]. JOM, 2020, 72(11): 1-8.

[5] CHAUDRY U M, HAMAD K, ABUHMED T. Machine Learning-Aided Design of Aluminum Alloys with High Performance[J]. Materrial Today Communications, 2021, 26.

[6] OH J M, NARAYANA P L, HONG J K, et al. Property Optimization of TRIP Ti Alloys Based on Artificial Neural Network[J]. Journal of Alloys and Compounds, 2021, 884.

[7] 刘彬, 汤爱涛, 潘复生, 等. 基于参数优化的人工神经网络的AZ31镁合金力学性能预测模型[J]. 重庆大学学报, 2011, 34(3): 44-49.

[8] LI N, ZHAO S Y, ZHANG Z G. Property Prediction of Medical Magnesium Alloy Based on Machine Learning[C]// 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), , 2021.

[9] CAO X Y, ZHANG Y B, CHEN H. Predicting Mechanical Properties and Corrosion Resistance of Heat-Treated 7N01 Aluminum Alloy by Machine Learning Methods[J]. IOP Conference Series: Materials Science and Engineering, 2020, 774.

[10] 郝永志, 赵海东, 林嘉华. 基于机器学习的挤压铸造铝合金力学性能预测[J]. 特种铸造及有色合金, 2019, 39(8): 859-862.