小样本数据下特种材料基因工程的数据扩充方法

杨涛[1]，张兆波[2]，郑添屹[3]，彭保[3,4]
1. 深圳市科荣软件股份有限公司，广东　深圳　518063；
2. 广东粤海珠三角供水有限公司，广东　广州　511455；
3. 华南师范大学华南先进光电子研究院，广东　广州　510006；
4. 深圳信息职业技术学院信息与通信学院，广东　深圳　518172

摘要
随着地下水利、水务管网对材料需求的多样性和复杂性日益加剧，通过机器学习高效便捷地设计满足个性化需求的特种材料成为人们关注的热点。传统监督学习方法均以大量数据训练建模为基础，但从深埋地下水务管网、高端军工设备等领域所需的特种材料，如稀贵高熵合金等获取大数据集，需要的成本极高且周期较长。为了解决该问题，提出了一种小样本扩充模型——RX-SMOGN，使用极致梯度提升模型和使用交叉验证的递归特征消除算法进行特征筛选，使用SMOGN算法扩充数据集。提出以高熵合金相结构为研究对象，训练传统机器学习模型对其进行预测以验证RX-SMOGN模型的有效性。由五折交叉验证及4个评价指标结果可知，RX-SMOGN模型充分提高了机器学习模型的性能，为合金材料设计提供了一种更便捷的方法，充分提高了合金材料设计的效率。

关键词
小样本扩充；特征工程；机器学习；高熵合金；稀贵金属

中图分类号：TP181；TG139　　文献标志码：A　　doi: 10.11959/j.issn.2096-0271.2024019

Data expansion method for genetic engineering of special materials with small sample data

YANG Tao[1], ZHANG Zhaobo[2], ZHENG Tianyi[3], PENG Bao[3,4]
1. Shenzhen Koron Soft Co., Ltd., Shenzhen 518063, China
2. GD Holdings Pearl River Delta Water Supply Co., Ltd., Guangzhou 511455, China
3. South China Academy of Advanced Optoelectronics, South China Normal University, Guangzhou 510006, China
4. School of Information and Communication, Shenzhen Institute of Information Technology, Shenzhen 518172, China

Abstract
With the increasing diversity and complexity of material requirements for underground water conservancy and water

pipeline networks, the efficient and convenient design of special materials to meet individual needs through machine learning has become a hot topic of concern. Traditional supervised learning methods are all based on a large dataset to train models, but obtaining large datasets for special materials required in deeply buried underground water pipeline networks and high-end military equipment, such as rare and high-entropy alloys, etc. requires extremely high cost and a long period. To solve this problem, we propose a small sample expansion model-RX-SMOGN, using XGBoost and RFECV algorithms for feature screening. We enrich the dataset with the SMOGN algorithm. In this paper, the phase structure of high-entropy alloys is used as the research object, and traditional machine learning models are trained to predict them to verify the effectiveness of the RX-SMOGN model. From the results of 5-fold cross-verification and 4 evaluation indicators, it can be seen that the RX-SMOGN model fully improves the performance of the machine learning model, provides a more convenient method for alloy material design, and fully improves the efficiency of alloy material design.

Key words
small sample expansion, feature engineering, machine learning, high-entropy alloy, rare precious metal

0 引言

珠江三角洲水资源配置工程设计使用寿命为100年，整个工程具有线路长、流速高、压力大、渗透性和腐蚀性强、维修成本高等特征显著，对地下管网性能的个性化要求很强，迫切需要满足工程需求的特种材料。同时，航空发动机[1]、导弹等军工领域也面临着类似的材料性能个性化定制需求[2]。材料基因成为突破这一共性难题最受关注的技术手段。但传统基于机器学习和深度学习的材料基因模型大多需要大量标注数据以进行模型训练，而实际中有效数据量较少，难以支撑模型的训练工作，导致机器学习在该领域中不能发挥其应有的作用。为此，如何利用小样本数据进行有效的模型训练便成为当前亟待解决的问题。

数据的扩充可以充分提高模型的性能，Cong等人[3]提出使用改进的GAN对数据集添加编码模块，扩充人脸图像数据集，随后使用新数据集进行人脸检测，最终几种经典的人脸检测模型的检测精度均提高了3%以上。Zhang等人[4]提出训练WGAN生成水冷壁的缺陷数据扩充数据集，然后将数据集输入CNN中进行缺陷检测，结果表明精度得到了显著提高。Wan等人[5]使用InfoGAN生成了桥梁检测因子的新数据，极大地提高了模型的预测性能。此外，数据扩充的方法在语音识别、情感识别等领域应用也十分有效。Lee等人[6]提出了一种数据增强算法，该算法使用对声学频率加权的方法来增加数据集，充分提高了模型性能。Prayitno等人[7]提出了一种数据扩充的方法SRHA，通过重复语音数据中振幅最高的段落来扩充数据，最终该方法将语音情感识别的准确率从95.88%提高到98.16%。然而，以上这些方法多应用于图片、音频等数据的扩充，均不适用于材料领域中的数据。

综上所述，本文提出一种可用于合金材料领域的小样本扩充模型——RX-SMOGN，该方法首先使用RFECV算法配合XGBoost（extreme gradient boosting）模型确定扩充算法需要扩充的特征，随后使用SMOGN算法扩充数据，训练机器学习模型对高熵合金相结构进行预测以验证扩充算法结果。结果表明，RX-SMOGN模型能够提升传统机器学习模型的精度和性能，解决了机器学习在材料领域不能充分

发挥作用的问题，加速了新型高熵合金的设计工作，RX-SMOGN模型流程如图1所示。

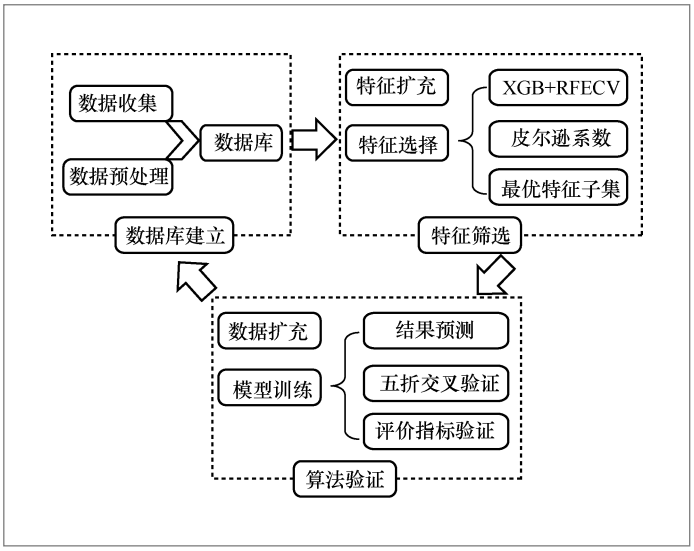

图1　RX-SMOGN 模型流程

1 方法介绍

1.1 特征筛选

本文使用XGBoost模型和RFECV算法组合的方式进行特征筛选。其中，XGBoost模型是一种基于梯度提升决策树的改进模型，常应用于监督学习。在XGBoost模型中，定义目标函数为训练损失和正则化之和，表达式如式(1)所示。

$$\mathrm{obj}(\theta)=L(\theta)+\Omega(\theta) \qquad (1)$$

其中，$L(\theta)$ 为训练损失项，衡量模型在训练数据上的预测准确度；$\Omega(\theta)$为正则化项，用来衡量模型的复杂度。根据基础目标函数，可以定义用于学习的目标函数的提升方式，具体的表达式如式(2)所示。

$$\mathrm{obj}=\sum_{i=1}^{n} I\left(y_i,\hat{y}_i^{(t)}\right)+\sum_{i=1}^{t}\Omega(f_i) \qquad (2)$$

其中，I 表示预测结果的损失函数，损失函数的确定是该算法比较重要的一步；t 表示提升的次数，$\hat{y}^{(t)}$ 表示第 t 次提升的预测结果，f_i 表示第 i 次提升涉及的训练函数。

RFECV算法是特征选择算法中的包裹式选择算法[8]。该算法通过求得最佳的交叉验证的得分来找到最优特征集合，以此来获得最佳的特征子集，该算法包含递归特征消除（recursive feature elimination，RFE）和交叉验证（cross-validation，CV）两个阶段。其中RFE阶段通过反复构建模型，逐次删除特征空间中模型评估重要度最低的特征以更新特征空间，直至得到所需数量的特征。CV阶段用于评估训练好的模型在新数据上的表现，可以在一定程度上降低过拟合的可能[9]。

1.2 数据扩充

本文使用SMOGN[10]算法进行数据扩充，该算法结合了随机欠采样和两种过采样技术，通过引入高斯噪声限制风险的产生，同时通过随机噪声尽可能使结果具有多样性。

SMOGN算法首先构建包含考虑目标变量值的数据分区，分区包括稀有分区、重要分区、正常分区和不太重要的分区。稀有分区和重要分区包含较高相关性的数据，即相关性高于预定义阈值的示例，而正常分区和不太重要的分区包含相关性较低的数据，即相关性示例目标变量值的分数低于设置的阈值。SMOGN算法对不同分区采用不同的采样方法，正常分区和不太重要的分区采用随机欠采样方法进行数据扩充，稀有分区和重要分区采用过采样的方法。SMOGN的算法流程如下。

算法SMOGN算法

输入：D-具有目标连续变量的数据集Y

t_R-数据集Y的相关性阈值

$\%u$-欠采样百分比

$\%o$-过采样百分比

K-最近邻的数量

dist-距离度量

输出：new$D$-新修改的数据集

ord$D$←通过Y的升序对D进行排序

$\phi()$←从Y分布得到的相关性函数

Bins_N←连续样本$<X_i,y_i>$的分区，$<X_i,y_i>\in$Ord$D$，使得$\phi(y_i)<t_R$

Bins_R←连续样本$<X_i,y_i>$的分区，$<X_i,y_i>\in$Ord$D$，使得$\phi(y_i)\geqslant t_R$

new$D$←Bins_R

foreach：$B\in$Bins_N执行

selNormCases←$\%u\times B$中的数据$|B|$

new$D$←new$D\cup$selNorm Cases

end

foreach：$B\in$Bins_R执行

ng←$\%o\times B$中的数据$|B|$

foreach：case$\in B$执行

nns←kNN(k,case,B,dist)

Dist$M$←示例case和B中样本的距离

max$D$←(Dist$M$)/2

for i←1 to ng执行

x←从nns中随机选取一个数据

if Dist$M(x)$<max$D$执行

new←使用Smote$R$来插入x和示例case

else

pert←min(max$D$,0.02)

new←在示例case中引入高斯噪声

end

new$D$←new$D\cup${new}

end

end

end

return new$D$

本实验中数据集的相关性阈值为0.9，欠采样百分比为1，过采样百分比为0.1，最近邻数量为10，距离度量采用明斯基距离。

2 实验和结果

2.1 数据收集

本文数据集由异系多主元高熵合金组成，其中包含已被证实报道的典型多主元高熵合金及已发表论文中的样本数据[11]，共150条数据：56个五元合金样本、48个六元合金样本、20个七元合金样本、19个四元合金样本、4个九元合金样本、3个三元合金样本。数据样本包含了16种元素（Nb、Mo、Ti、V、Ta、Al、Zr、Hf、W、Cr、Ni、Fe、Si、Co、Cu、Re），各元素的数量分布如**图2**所示。

2.2 实验过程及结果

2.2.1 特征扩充及数据预处理

本文首先对样本数据进行特征扩充，使用HEAPS软件计算了每一条样本数据的31条特征属性，充分扩大了数据特征维度，为后续特征筛选创造更大的搜索空间。此外，因本文数据集采集自多份文献，所以需要对其数据可用性进行筛查。本文对扩充特征后的每一条样本数据进行了筛查，检查其是否符合多主元高熵合金的标准，是否存在空值等异常值，经查本文数据均无异常值。

2.2.2 特征筛选

本文使用XGBoost模型和RFECV算法组合的方法对数据集中31条特征进行筛选，得到T_m、C_p、VEC、Δx^P等8个特征，

之后绘制了皮尔逊系数矩阵图，结果如**图3(a)**所示，由**图3(a)**可知，任意两个特征之间没有强相关性（P>0.95），但C_p、Δk二者与其他特征相关性极小，并且二者对高熵合金相结构的影响很小，故本文剔除了这两个特征，剔除后特征之间的相关性如**图3(b)**所示。

2.2.3 数据扩充及模型验证

本文使用SMONG算法将原有的150条数据集扩充至300条，扩充前后数据分布对比如**图4**所示。

由**图4**可知，本文扩充算法可以实现数据扩充且对数据具有一定的平衡作用，如原始数据集中面心立方晶格（FCC）数据最少（仅占总样本的21.2%），体心立方晶格（BCC）数据最多（占总样本的43.7%），扩充后的数据集中FCC相数据相对增加（占总样本的28.2%），BCC相数据有一定程度的减少（占总样本的22.9%），由此可知本文算法能够捕捉到原始数据中的样本不平衡现象，并进行一定程度的修正。

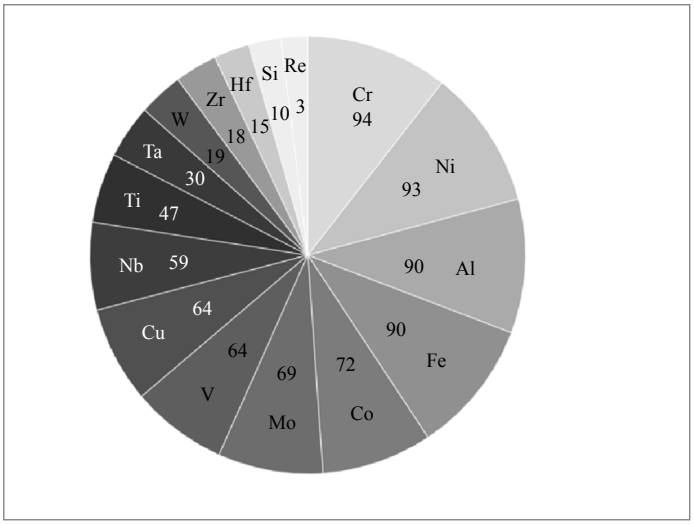

图 2　数据集中各元素的数量分布

为了证明RX-SMOGN模型的有效性，本文使用扩充前后两组数据集训练了随机森林等常见机器学习模型，五折交叉验证均值结果如**图5**所示，由**图5**可知，使用扩充后数据集训练的模型结果均优于原始数据集训练的结果。

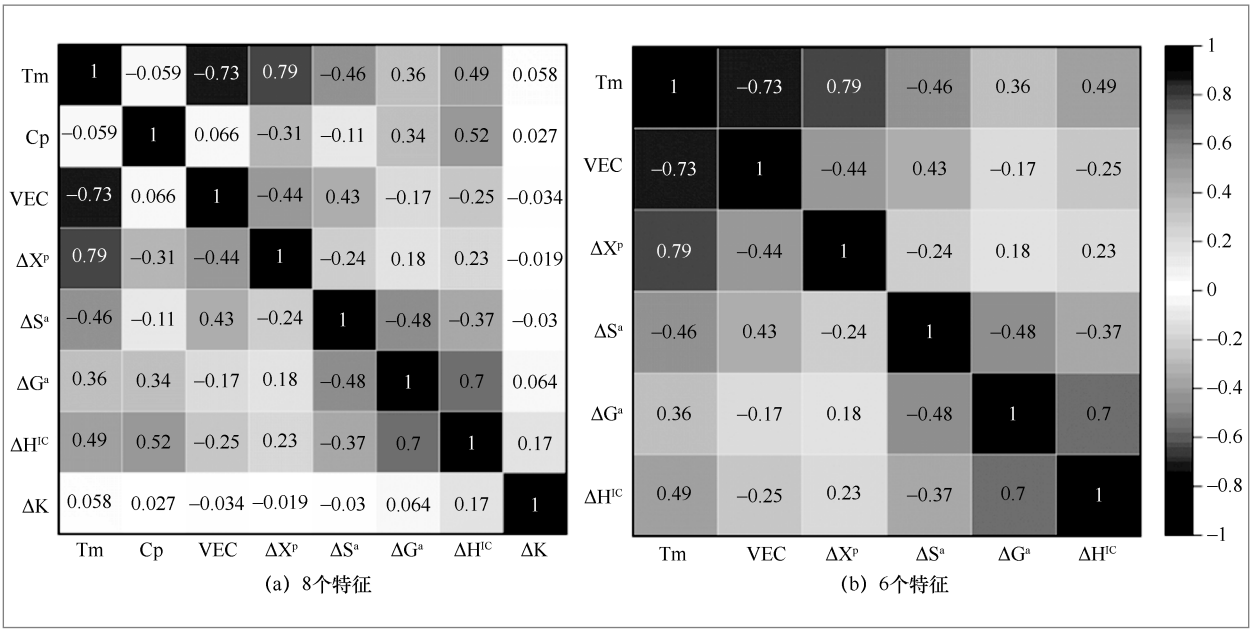

图 3　特征皮尔逊系数矩阵图

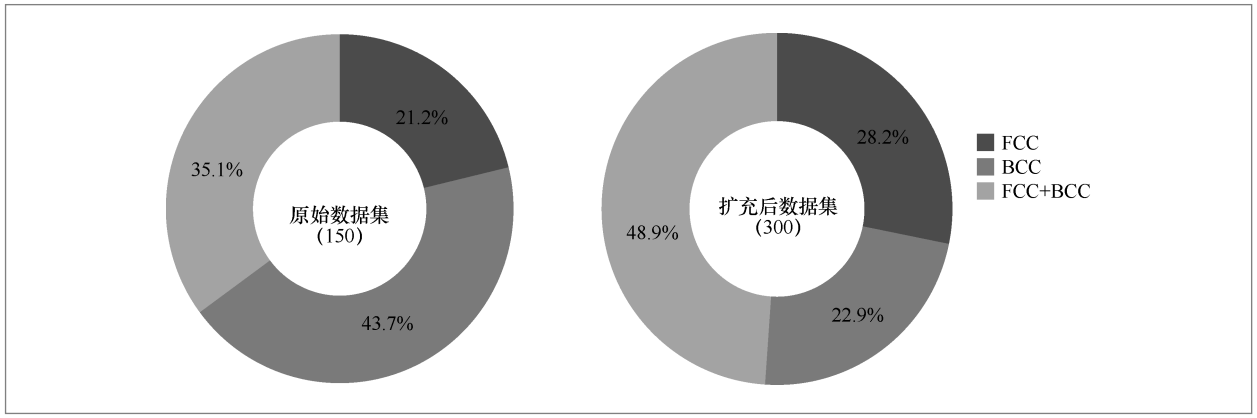

图 4 扩充前后数据分布对比

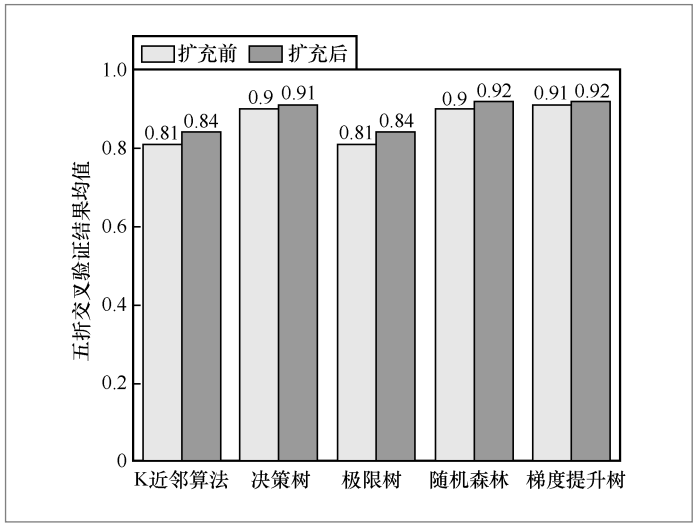

图 5 五折交叉验证结果对比

为了验证RX-SMOGN的性能，本文选用4个评价指标对样本扩充后数据的质量进行评价，4个评价指标为准确率（ACC）、ROC曲线下方面积大小（AUC）、召回率（Recall）、F1分数，5个模型各指标对比结果如**图6**所示。由**图6**可知，各模型在训练扩充后的数据集时，均表现出更好的预测效果。其原因在于，RX-SMOGN结合了过采样和欠采样技术，并通过生成合成样本和清除噪声样本的方式来平衡不同类别的数据量。此外，还可以减少类别不平衡带来的偏差，提高对少数类的识别能力，提升模型的泛化性能。具体来说，当训练数据集存在类别不平衡问题时，模型可能会更倾向于预测多数类别，而忽略少数类别，导致模型的性能下降。使用RX-SMOGN算法可以通过生成更多的合成样本和删除噪声样本的方式平衡不同类别的数据量，从而使模型在训练和测试阶段都能够得到更好的表现。因此得出结论，RX-SMOGN模型能够学习原数据集的内在规律并扩充出有效的新数据，对机器学习模型的性能有一定的提升。

为了进一步验证RX-SMOGN模型在同类模型中的性能，本文选用SMOTE算法作为对照组。SMOTE算法是一种综合采样人工合成数据算法，可用于小样本扩充。本文选用4个评价指标来对比两种算法的性能表现，指标分别为ACC、AUC、Recall和F1分数。首先，分别使用两个算法将150条原始数据扩充至300条。其次，使用扩充后的数据集训练随机森林等常用算法的相分类能力。最后，各指标结果如**图7**所示。如**图7**可知，RX-SMOGN模型相比于SMOTE算法给模型带来了更好的指标表现。从稳定性的角度分析，SMOTE算法通过对少数类样本进行插值，生成新的合成样本来增加数据集的多样性。但是，由于SMOTE算法

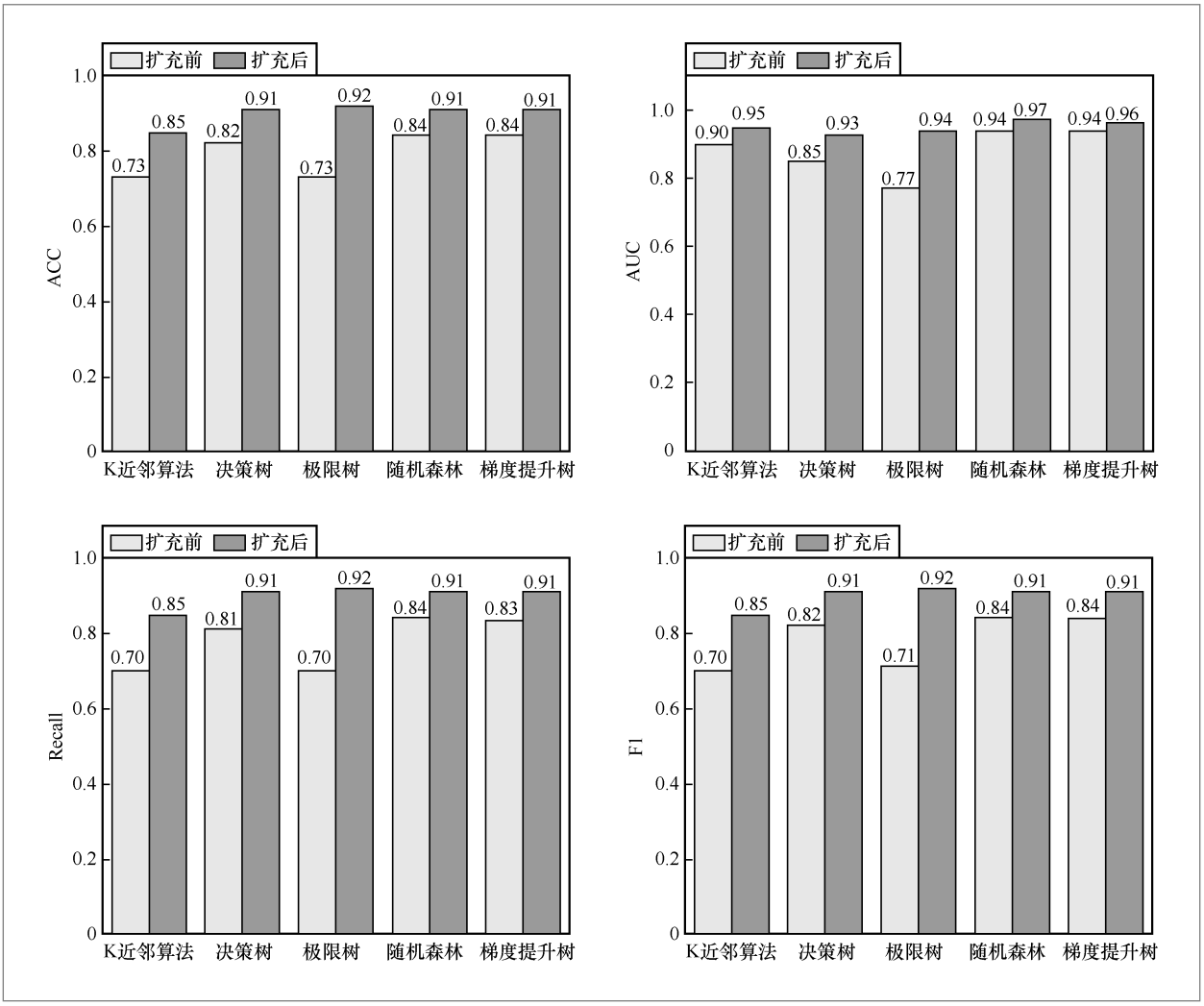

图 6　5 个模型的各指标对比

没有考虑到样本之间的密度分布，可能会导致生成的合成样本过于聚集在原始样本的区域，从而提升模型过拟合的风险，稳定性欠佳。而RX-SMOGN模型通过计算样本的密度将少数类样本分成密度分布不同的区域，然后根据区域的密度分别进行过采样和欠采样操作，从而增加数据的多样性，提高模型的泛化能力和稳定性。从模型表现能力方面分析，由于SMOGN算法考虑了样本的密度分布，可以生成更加具有多样性和代表性的合成样本，从而提高了模型的泛化能力。而SMOTE算法生成的合成样本过于聚集在原始样本的区域，不能很好地表示数据的多样性和复杂性，难以提升模型的性能表现。由此可知，RX-SMOGN模型的性能表现优于SMOTE算法。

3 结束语

综上所述，本文提出了一种可用于合金材料领域的小样本数据扩充模型——RX-SMOGN。该方法首先使用RFECV算法等进行特征筛选，之后使用SMOGN

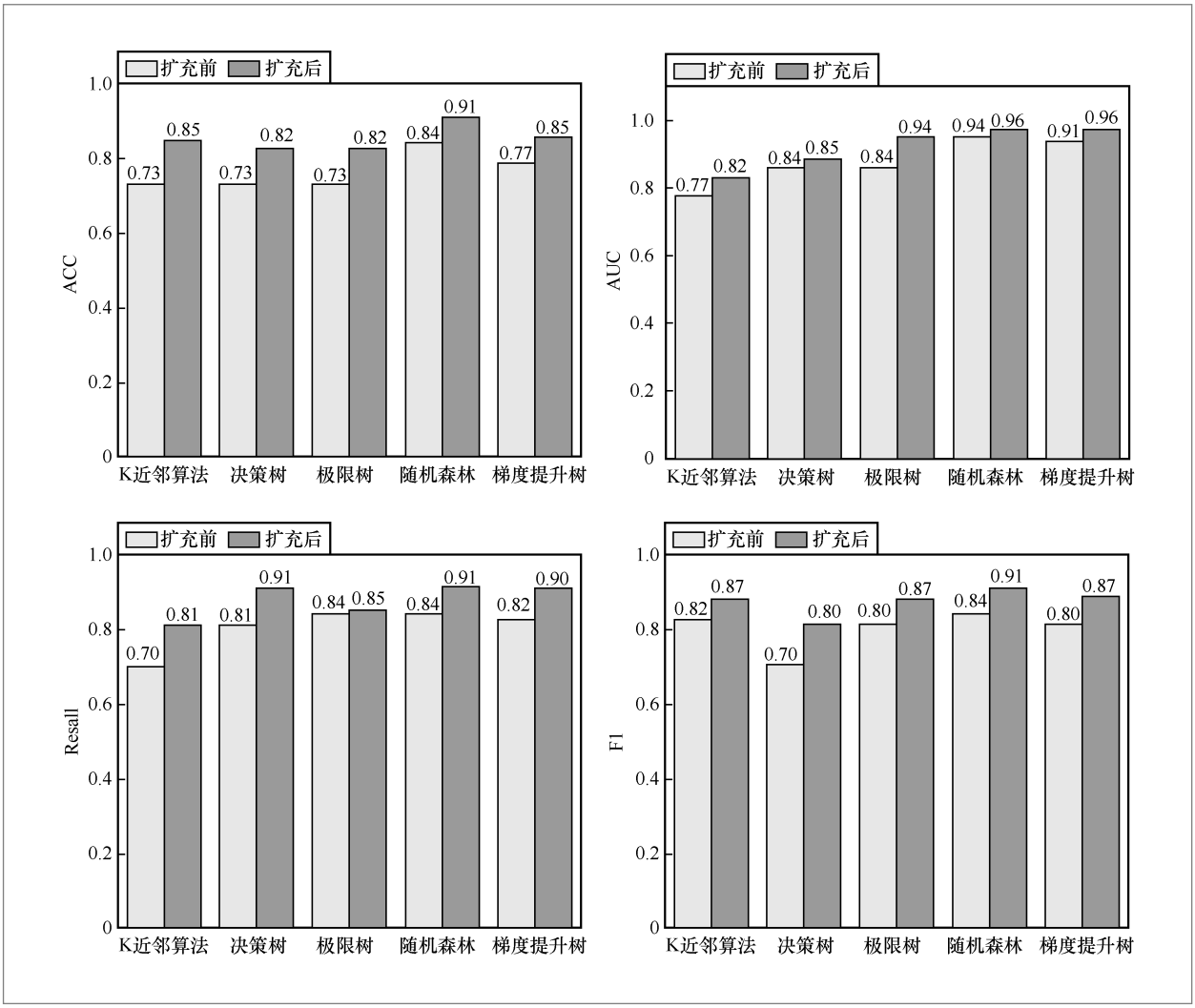

图 7 模型性能对比

算法扩充数据，最后以高熵合金相结构为研究对象对其进行预测。通过五折交叉验证结果及4个评价指标的对比可证明RX-SMOGN模型能够提升传统机器学习模型的精度，解决了机器学习在材料领域由可用数据集较少导致的预测准确率低的问题，加速了新型高熵合金的设计工作。但本文方法亦有局限，在扩充数据时，算法还不能完美地平衡不同类别样本的数量，且本文特征筛选方法效率仍有提升空间，在后续研究中可以针对以上两点进行挖掘。

参考文献:

[1] YAO H W, QIAO J W, HAWK J A, et al. Mechanical properties of refractory high-entropy alloys: experiments and modeling[J]. Journal of Alloys and Compounds, 2017, 696: 1139-1150.

[2] GUO S, NG C, LIU C T. Anomalous solidification microstructures in Co-free $Al_xCrCuFeNi_2$ high-entropy alloys[J]. Journal of Alloys and Compounds, 2013, 557: 77-81.

[3] CONG K L, ZHOU M. Face dataset augmentation with generative adversarial network[J]. Journal of Physics: Conference Series, 2022, 2218(1): 012035.
[4] ZHANG Y, LU L, WANG Y W, et al. An improved defect detection method of water walls using the WGAN[J]. Journal of Physics: Conference Series, 2020, 1626(1): 012152.
[5] WAN P, HE H L, GUO L, et al. InfoGAN-MSF: a data augmentation approach for correlative bridge monitoring factors[J]. Measurement Science and Technology, 2021, 32(11): 114008.
[6] LEE S G, LEE S. Data augmentation for DNN-based speech enhancement[J]. Journal of Korea Multimedia Society, 2019, 22(7): 749-758.
[7] PRAYITNO B A, SUYANTO S. Segment repetition based on high amplitude to enhance a speech emotion recognition[J]. Procedia Computer Science, 2019, 157: 420-426.
[8] SAMANT R, RAO S. A study on Feature Selection Methods in Medical Decision Support Systems[J]. Esrsa Publications, 2013(11).
[9] WANG S H, CHEN S N. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling[J]. Journal of Petroleum Science and Engineering, 2019, 174: 682-695.
[10] BRANCO P, TORGO L, RIBEIRO R P. SMOGN: a pre-processing approach for imbalanced regression[J]. Proceedings of Machine Learning Research[J], 2017, 74: 36-50.
[11] GRUBER G C, LASSNIG A, ZAK S, et al. Synthesis and structure of refractory high entropy alloy thin films based on the MoNbTaW system[J]. Surface and Coatings Technology, 2022, 439: 128446.

作者简介

杨涛(1984-),男,深圳市科荣软件股份有限公司工程师、董事长,主要研究方向为水务信息化、人工智能技术。

张兆波(1984-),男,广东粤海珠三角供水有限公司高级工程师、经理,主要研究方向为智能水务、人工智能技术。

郑添屹(1998-),男,华南师范大学华南先进光电子研究院硕士生,主要研究方向为工业大数据、材料基因技术。

彭保（1979– ），男，博士，深圳信息职业技术学院教授、研究员，主要研究方向为工业大数据、材料基因技术。

收稿日期： 2023–07–17

基金项目： 深圳大学稳定保障计划项目（No. 20200829114939001）；深圳信息职业技术学院校级创新科研团队项目（No. TD2020E001）；珠三角水资源配置工程科研项目（No. CD88–QT01–2022–0068）

Foundation Items: Shenzhen University Stability Support Plan (No. 20200829114939001), Project of Shenzhen Institute of Information Technology School–level Innovative, Scientific Research Team (No. TD2020E001), The Pearl River Delta Water Resources Allocation Engineering Scientific Research Project (No. CD88–QT01–2022–0068)