Computer Science and Application 计算机科学与应用, 2025, 15(8), 138-150
Published Online August 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.158204

Hans汉斯

压敏胶剥离强度预测中的数据增强技术

郭　威，胡文军

湖州师范学院信息工程学院，浙江　湖州

收稿日期：2025年7月7日；录用日期：2025年8月8日；发布日期：2025年8月15日

摘　要

配方是决定紫外光固化压敏胶(UV-PSA)性能的关键所在，因研究配方的传统方法难以获得丰富数据，限制了计算机技术在这方面的应用。为此，提出自适应合成过采样算法解决该场景下的数据稀缺问题。首先，通过距离度量策略预处理原始数据，使其适用于回归任务；其次，结合近邻与远邻策略以及非线性插值技术，生成具有多样化和代表性的合成样本；最后，利用扩展后的样本建立泛化能力强的支持向量回归预测模型。实验结果表明，增强后的UV-PSA的数据集提升了包括支持向量回归在内的所有模型性能，验证了提出的数据增强技术在UV-PSA配方研究中的有效性。

关键词

数据增强，紫外光固化压敏胶，剥离强度，自适应合成过采样

Data Augmentation Techniques in the Prediction of Peel Strength of Pressure Sensitive Adhesives

Wei Guo, Wenjun Hu

School of Information Engineering, Huzhou University, Huzhou Zhejiang

Received: Jul. 7th, 2025; accepted: Aug. 8th, 2025; published: Aug. 15th, 2025

Abstract

Formulation is key to determining the performance of UV-PSA, and the use of computer technology in this area is limited by the lack of rich data available for traditional methods of studying formulations. Therefore, an adaptive synthesis oversampling algorithm was proposed to solve the problem of data scarcity in this scenario. Firstly, the distance measurement strategy was used to preprocess the raw data to make it suitable for regression tasks. Secondly, the nearest and far neighbor strategies

文章引用: 郭威, 胡文军. 压敏胶剥离强度预测中的数据增强技术[J]. 计算机科学与应用, 2025, 15(8): 138-150.
DOI: 10.12677/csa.2025.158204

and nonlinear interpolation techniques were combined to generate diverse and representative synthetic samples. Finally, the extended sample was used to establish a support vector regression prediction model with strong generalization ability. Experimental results show that the enhanced UV-PSA dataset improves the performance of all models, including support vector regression, and verifies the effectiveness of the proposed data augmentation technique in the study of UV-PSA formulations.

Keywords

Data Augmentation, UV-Curable Pressure Sensitive Adhesives, Peel Strength, Adaptive Synthetic

1. 引言

紫外光固化压敏胶(UV Cured Pressure Sensitive Adhesives, UV-PSA)因其具备快速固化、环保无溶剂和粘附性强等特点，已成为电子、医疗、汽车等多个工业领域中不可或缺的材料[1]-[3]。剥离强度作为其关键性能指标之一，直接影响产品的质量和可靠性[4]，因此，准确预测 UV-PSA 的剥离强度对于提高研发效率至关重要。然而，传统的实验方法通常面临高成本、操作复杂和测试周期长等挑战[5] [6]。例如，Pang 等[7]在研究萘基固化剂对紫外光固化压敏胶性能的影响时，仅一组实验的光照时间就需要 90 分钟以上。Liu 等[8]在离子液体表观摩尔体积的研究中，通过 X 射线衍射法进行测试，每组实验平均时间长达 20 小时。Li 等[9]指出，传统的实验室测试往往需要数天甚至数周的时间来完成样品准备、测试以及数据分析等步骤，并且在实验过程中需要大量的人员和设备投入。因此，亟需一种简便高效的方法，以补充或替代传统实验[10] [11]。

当前计算机技术的快速发展使得机器学习技术在金融、医学、气象等多个领域得到了广泛应用，并取得了显著成果[12]。计算机技术也同时推动了材料研发模式的转变：从传统的“经验 + 试错”方法，逐步向计算驱动的创新模式进行转变[13]。例如，Hart 等[14]系统性地总结了机器学习在合金研发中的应用，包括非晶合金、高熵合金、形状记忆合金、磁性材料以及超合金的性能优化。Lookman 等[15]基于自适应实验采样和贝叶斯优化，阐述了主动学习在加速新材料探索与设计中的潜力。Schmidt 等[16]则围绕材料计算与机器学习的结合，从基础算法、性能预测、新材料发现以及模型可解释性等方面进行了详尽的综述。Liu 等[17]基于材料基因工程的理论框架，提出了贯穿材料数据生命周期的研究方法，其中涵盖材料数据库构建、结构，性能关系预测以及新材料开发的实际应用。利用计算机技术不仅能够对材料成分、结构和性能定量预测，深入探究材料的机理特征，还可以为材料研究者们提供多尺度、多维度的研究视角。

然而，上述的研究大多依赖于大规模的样本数据，在像 UV-PSA 等特定领域，数据难以大规模获取，普遍存在数据稀缺的问题，限制了模型的性能和泛化能力[18]。为此，诸多研究者针对数据稀缺引发的小样本问题开展研究，提出了多种解决方案。例如，Li 等[19]将 GAN 的数据增强技术应用于数据的重构任务，该方法在提高数据质量和特征多样性方面获得成效。Wu 等[20]提出了通过构建目标金字塔，生成多尺度正样本的方法，缓解了目标尺度稀疏分布问题，但此类方法在低维且样本量极为有限。为了解决这一问题，Chao 等[21]提出了 h-SMOTE 方法，通过对少数类样本进行数据增强，有效平衡了数据集并提升

了模型的学习能力。Jia 等[22]结合随机欠采样、SMOTE 技术与卷积神经网络，提出了一种用于小样本数据预处理的特征优化方法，显著提高了数据特征的提取与利用效率，但其性能高度依赖于参数选择，如 K 近邻数和合成样本生成比例。针对这一问题，Liu 等[23]通过自适应合成过采样(Adaptive Synthetic Sampling Approach for Imbalanced Learning, ADASYN)对不平衡数据进行过采样，并结合传统分类方法提升分类性能，有效改善了数据稀缺问题。尽管现有方法在缓解数据稀缺问题和提升模型性能方面已取得一定进展，但它在以下几个方面中仍存在不足：

1) 现有的 ADASYN 方法仅适用于分类任务，通过生成少数类样本来平衡类别分布，因此无法提供精确的回归预测值，限制了模型在回归任务中的预测能力。

2) 生成的新样本可能未准确反映原始数据的分布特征，导致模型在训练过程中过拟合于局部样本，忽视全局特性。

3) 当数据特征之间存在复杂交互或非线性关系时，算法常表现出较弱的适应性，往往导致模型性能的不稳定或下降。

为解决上述问题，本研究对现有的 ADASYN 算法进行优化。首先，利用度量学习计算每个样本与数据集中所有其他样本之间的距离。其次，分别计算每个样本与其最近邻和最远邻的距离总和，并将其与总体距离进行比值归一化，为每个样本分配一个合成数量的权重。接着，采用样条插值技术生成新样本的特征。最后，使用原始数据训练得到的预测模型为新样本生成标签。为验证改进方法的有效性，在实验中利用了均方误差等多个评价指标，对比了六种常见预测模型的性能。

2. 数据增强技术应用于 UV-PSA

实验的整体流程如图 1 所示。首先进行数据收集与预处理的过程，并对数据分布及物化关系进行先验分析，其次基于 ADASYN 改进数据增强方法，最后利用预测模型对增强后的数据进行训练，从而实现剥离强度的预测。

Figure 1. Overall model framework
图 1. 整体模型框架

2.1. 数据收集与预处理

通过 Google Scholar、Web of Science、Scopus 等数据库，使用“UV-curable PSA”“photoinitiator”“peel strength”等关键词检索相关文献，保留提供完整实验数据的文献，剔除低质量、不相关或不符合实验要求的数据。按 UV-PSA 配方的不同组分数据进行处理：预聚物的质量数据作为特征 1，如丙烯酸酯、环氧树脂等，用于形成 UV-PSA 的主体结构；光引发剂的质量数据作为特征 2，如自由基型、阳离子型光引发剂，在 UV 照射下触发聚合反应；添加剂的质量数据作为特征 3，如增粘剂、抗氧化剂、交联剂等，用于改善 PSA 性能；剥离强度作为标签，衡量 UV-PSA 的粘附性能。数据集样例如表 1 所示。

Table 1. Date example
表 1. 数据集样例

序号	预聚物 (wt%)	光引发剂 (wt%)	添加剂 (wt%)	剥离强度 (N/25mm)
1	0.660	0.329	0.009	0.430

续表

2	0.495	0.495	0.009	1.400
3	0.329	0.660	0.009	2.730
4	0.247	0.742	0.009	3.180
5	0.198	0.792	0.009	3.460
6	0.980	0.020	0.000	31.000

由于不同研究采用的实验方法和测量标准可能存在差异，为确保数据的可比性，进行了系统性的标准化处理，主要包括以下两步：

1) 换算单位，使所有剥离强度数据的单位一致，如有些研究的单位为 N/100mm，而有些为 N/25mm，为保证数据的完整性，试验对剥离强度数据进行统一换算。

2) 进行归一化处理。对各个配方变量进行 Min-Max 归一化，以消除不同变量的尺度差异，使数据适用于机器学习建模。

2.2. 数据先验分析

在模型训练前，通过描述统计量和相关系数对 UV-PSA 原始数据集进行分析，同时利用化学领域的物理化学知识探讨 PSA 配方组成成分的比例与剥离强度间的联系，旨在从化学与数据分析两方面进行全面考察数据中各变量间的关系，为后面的建模及数据预处理提供依据。

如表 2 所示，特征 1 与特征 2 呈极负相关，由此说明这两个特征的数据中可能存在一个负相关的物理化学关系。特征 3 的分布相对于数据来看比较集中(分布在 0.000~0.065 之间，幅度为 0.065)，变化比较小，可能不会对模型预测结果有太大的影响。

Table 2. Prior data analysis
表 2. 数据先验分析

	特征 1	特征 2	特征 3	标签
均值	0.929	0.057	0.014	7.600
标准差	0.149	0.151	0.016	8.210
值范围	0.198~0.998	0.000~0.792	0.000~0.065	0.080~35.000
集中区	高值区	低值区	低值区	/
与特征 2 相关系数	−0.994	/	/	/
与特征 3 相关系数	0.030	−0.128	/	/
与标签相关系数	0.077	−0.055	−0.215	/

统计分析可见，各特征值与标签的相关程度较低。虽然特征 1 与标签值的相关程度最大，但仍然偏低，说明特征 1 对剥离强度所起到的解释性作用也较低。当前特征对目标值标签的线性解释程度较低，可能是数据之间存在非线性关系或还需要进一步对特征进行处理来使其解释效果增强。

从物化原理的角度来看，剥离强度受到各种因素的综合影响。交联密度、高分子链结构和粘结剂与基材的润湿界面特性是影响剥离强度的主要因素[24]。交联强度和弹性主要取决于占胶粘剂主体成分的预聚物(特征 1)。光引发剂促使胶粘剂聚合反应进行，通过聚合反应的速率控制来调节固化时间和深度交

联结构的形成，也控制胶粘剂的固化深度[25]。如果光引发剂加入过多会使得胶粘剂固化较深表面层固化过快，而过浅则使整体剥离强度受到影响。预聚物和光引发剂(特征 2)的相互作用也比较复杂，需要考虑反应的效率和粘度及固化深度的均衡[26]。添加剂(特征 3)通过对体系流变性、润湿性和韧性的影响进行调节，间接改善了剥离强度[28] [29]。

综上所述，剥离强度受到多种因素的联合作用，即某种单一成分含量的变化不能影响剥离强度的大小，交联度、高聚物结构及界面润湿性等因素共同作用影响到紫外光固化 PSA 的最终表现[27]。因此在配方设计中要考虑各因素间的相互作用，而非针对某单一成分含量的高低进行配方调整。

3. 数据增强

3.1. ADASYN 基础算法

现有的 ADASYN 算法[30]最初是针对分类任务设计的，流程如算法 3.1 所示。其核心机制依赖于离散的类别信息来指导合成样本的生成。然而，在回归任务中，目标变量为连续型数值，需精确预测而非简单分类，这一特性导致传统 ADASYN 算法在 UV-PSA 性能预测任务中面临显著局限性。

算法 3.1. 基础 ADASYN

输入：

训练样本集 $\mathbf{X}_{\text{tr}}$；

少数类样本数量为 m_s；

多数类样本数目为 m_l；

邻居数 K；

平衡参数 β。

步骤：

计算类别不平衡度 $d=\dfrac{m_s}{m_l}$；

计算合成样本数 $G=(m_l-m_s)\times\beta$；

每个少数类样本的最近邻样本中，计算属于多数类样本的数量 D_i；

计算比例 $r_i=\dfrac{D_i}{K}$；

对每个少数类样本的比值进行正则化计算 $r_i'=\dfrac{r_i}{\sum_{i=1}^{m_s}r_i}$；

对每个少数类样本计算需要合成的样本数量 $g_i=r_i'\times G$；

生成新样本：

$\mathbf{x}_{i,new}=\mathbf{x}_i+(\mathbf{x}_{zi}-\mathbf{x}_i)\times\lambda,\ \lambda\in[0,1]$。

输出：

增强后的样本集。

为解决上述问题，本研究提出一种改进的 ADASYN 算法，专门适配 UV-PSA 配方预测的回归任务需求。改进策略主要包含以下三个方面：

1) 以距离度量替代样本记数，根据数据分布调整样本生成密度，从而更准确反映样本的分布特点。

2) 引入混合近邻与远邻合成机制，结合 K 近邻和反向 K 近邻(Reverse KNN)策略，在局部生成多样性样本的同时，利用远邻关系捕捉数据的全局特性，避免过度依赖局部特征。

3) 采用非线性插值技术，根据样本密度动态调整权重，确保合成样本在反映局部趋势的同时，符合全局分布的连续性约束。

3.2. ADASYN 优化算法

传统 ADASYN 通过 KNN 来统计少数类中多数类样本数，进而计算合成样本数比例，其局限在于仅基于 KNN 采样生成合成样本，未充分考虑样本的全局分布，因此本实验进一步引入远邻策略。设训练样本集为 $\mathbf{X}_{tr}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_m\}$， m 为样本个数，对于每个样本 $\mathbf{x}_i$，计算与其他样本 $\mathbf{x}_j$ 的欧氏距离 $\mathbf{D}_{(i,j)}=\|\mathbf{x}_i-\mathbf{x}_j\|^2$，其中 $i,j\in[1,m]$ 且 $i\neq j$，得到距离矩阵 $\mathbf{D}$。

对于每个样本 $\mathbf{x}_i$ 通过对距离 $\mathbf{D}_{(i,:)}$ 进行升序排序，选择前 K 个近邻 $\mathbf{x}_{ner}=\{\mathbf{x}_{i,2},\mathbf{x}_{i,3},\cdots,\mathbf{x}_{i,K+1}\}$，计算其依据近邻的原理，需要生成样本占新样本总数的比例：

$$w_{ner,i}=\frac{\sum_{j=2}^{K+1}D_{i,j}}{\sum_{i=1}^{N}\sum_{j=1}^{N}D_{i,j}} \tag{1}$$

从距离矩阵 $\mathbf{D}$ 中选择最远的 K' 个邻居 $\mathbf{x}_{far}=\{\mathbf{x}_{i,m-K'},\mathbf{x}_{i,m-K'+1},\cdots,\mathbf{x}_{i,m-1}\}$，计算其依据远邻的原理，需要生成样本占新样本总数的比例：

$$w_{far,i}=\frac{\sum_{j=m-K'}^{m-1}D_{i,j}}{\sum_{i=1}^{N}\sum_{j=1}^{N}D_{i,j}} \tag{2}$$

结合了近邻与远邻样本，可以生成的合成样本不仅能够细化特征空间的局部细节，也有效扩展了全局覆盖范围。

自定义依据近邻的合成样本数量 G_{KNN} 和依据远邻的合成样本数量 G_{KFN}。接着对每个的生成样本的权重，进行动态归一化与自适应分配。

当 $G_{KNN}<m$ 时，为了增强稀疏区域的密度和平滑性，对近邻原理的权重 $w_{ner,i}$ 进行降序排序，选取前 $\frac{G_{KNN}}{2}$ 个具有最大权重的样本生成新样本，并对这些权重进行归一化处理：

$$\tilde{w}_{ner,i}=\frac{w_{ner,i}}{\sum_{j\in S_{ner}}w_{ner,j}},\quad i\in S_{ner} \tag{3}$$

其中，S_{ner} 表示近邻权重最高的样本索引集合。由此计算 $\mathbf{x}_i$ 依据近邻原理，所需要生成的新样本数量 $g_{ner,i}$：

$$g_{ner,i}=\mathrm{round}\left(\tilde{w}_{ner,i}\times G_{KNN}\right),\quad i\in S_{ner} \tag{4}$$

当 $G_{KNN}\geq m$ 时，对所有样本的近邻权重进行归一化处理，并计算合成样本数：

$$\tilde{w}_{ner,i}=\frac{w_{ner,i}}{\sum_{j=1}^{m}w_{ner,j}},\quad i\in\{1,2,\cdots,m\} \tag{5}$$

$$g_{ner,i}=\mathrm{round}\left(\tilde{w}_{ner,i}\times G_{KNN}\right),\quad i\in\{1,2,\cdots,m\} \tag{6}$$

当 $G_{KFN}<m$ 时，为增强特征空间的全局覆盖和样本多样性，对依据远邻原理的权重 $w_{far,i}$ 进行降序排序，选取前 $\frac{G_{KFN}}{2}$ 个具有最大权重的样本进行生成，并对这些权重进行归一化处理，得到：

$$\tilde{w}_{far,i}=\frac{w_{far,i}}{\sum_{j\in S_{far}}w_{far,j}},\quad i\in S_{far} \tag{7}$$

其中，S_{far} 表示远邻权重最高的样本索引集合。由此计算依据远邻原理的每个样本生成数量：

$$g_{far,i}=\text{round}\left(\tilde{w}_{far,i}\times G_{KFN}\right),\quad i\in S_{far} \tag{8}$$

当 $G_{KFN}\geq m$ 时，对所有样本的远邻权重进行归一化处理，并计算合成样本数：

$$\tilde{w}_{far,i}=\frac{w_{far,i}}{\sum_{j=1}^{m}w_{far,j}},\quad i\in\{1,2,\cdots,m\} \tag{9}$$

$$g_{far,i}=\text{round}\left(\tilde{w}_{far,i}\times G_{KFN}\right),\quad i\in\{1,2,\cdots,m\} \tag{10}$$

传统的 ADASYN 使用线性插值生成合成样本，存在生成样本分布单一的问题。为使生成样本能更好地捕捉特征之间的非线性关系，对于样本 $\mathbf{x}_i$ 和选取的邻居之间使用样条插值，用于生成特征的非线性组合。这种插值策略使得生成样本更符合原始数据的复杂特征分布。

对于每个 $\mathbf{x}_i$ 样本，生成 $g_{ner,i}$ 个基于近邻样本 $\mathbf{x}_{ner}$ 的合成样本 $\mathbf{x}_{ner,syn}$：

$$\mathbf{x}_{ner,syn}=\mathbf{x}_i+\lambda\left(\mathbf{x}_{ner}-\mathbf{x}_i\right) \tag{11}$$

对于每个 $\mathbf{x}_i$ 样本，生成 $g_{far,i}$ 个基于远邻样本 $\mathbf{x}_{far}$ 的合成样本 $\mathbf{x}_{far,syn}$：

$$\mathbf{x}_{far,syn}=\mathbf{x}_i+\lambda\left(\mathbf{x}_{far}-\mathbf{x}_i\right) \tag{12}$$

其中 λ 为样条条插值权重，将生成的样本与原始样本 $\mathbf{x}_O$ 进行合并，最终得到增强样本集 $\mathbf{x}_E$：

$$\mathbf{x}_E=\mathbf{x}_O\bigcup\mathbf{x}_{ner,syn}\bigcup\mathbf{x}_{far,syn} \tag{13}$$

改进算法的步骤概述如算法 3.2 所示：

算法 3.2. 改进 ADASYN 算法流程

输入：

训练样本集 $\mathbf{x}_{tr}$；

邻居数 K；

近邻生成样本数 $g_{nea,i}=\text{round}\left(\tilde{w}_{ner,i}\cdot G_{KNN}\right)$；

远邻生成样本数 G_{KFN}。

步骤：

计算与所有样本的欧氏距离：$\mathbf{D}_{(i,j)}=\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2$；

选择近邻 K 个样本：S_{ner}；

选择远邻 K 个样本：S_{far}；

依据公式 1 计算近邻权重 $w_{ner,i}$；

依据公式 2 计算远邻权重 $w_{far,i}$；

If: $G_{KNN}<m$，$G_{KFN}<m$

对 $w_{ner,i}$ 按降序排序，选取前 $\frac{G_{KNN}}{2}$ 个样本 $\mathbf{x}_{ner}$；

对 $w_{far,i}$ 按降序排序，选取前 $\frac{G_{KFN}}{2}$ 个样本 $\mathbf{x}_{far}$；

依据公式 3 归一化依据近邻样本的合成数量权重：$\tilde{w}_{ner,i}$；

依据公式 4 计算依据远邻合成样本数量：$g_{ner,i}$；

依据公式 7 归一化依据远邻样本的合成数量权重：$\tilde{w}_{far,i}$；

依据公式 8 计算依据近邻合成样本数量：$g_{far,i}$；

Else: $G_{KNN}\geq m$，$G_{KFN}\geq m$

依据公式 5 归一化依据近邻样本的合成数量权重：$\tilde{w}_{ner,i}$

续表

依据公式 6 计算依据近邻合成样本数量： $g_{ner,i}$；

依据公式 9 归一化依据远邻样本的合成数量权重： $\tilde{w}_{far,i}$

依据公式 10 计算依据远邻合成样本数量： $g_{far,i}$；

基于近邻样本的 $g_{ner,i}$，依据公式 11 生成合成样本： $\mathbf{x}_{ner,syn}$；

基于远邻样本的 $g_{far,i}$，依据公式 12 生成合成样本： $\mathbf{x}_{far,syn}$；

依据公式 13 合并原始样本与生成样本 $\mathbf{x}_{new}$。

输出：

增强后的样本集 $\mathbf{x}_{new}$。

3.3. 复杂度分析

改进 ADASYN 算法在基础 ADASYN 的 K 近邻采样机制上，引入了全局远邻策略和非线性插值，增强了样本多样性和分布适应性，但计算复杂度有所增加。基础 ADASYN 的时间复杂度主要由 K 近邻搜索决定，为 $O(m \cdot N \log N)$，其中 m 为少数类样本数，N 为总样本数；而改进算法需计算全样本距离矩阵并进行排序，时间复杂度升至 $O(N^2 \log N)$。空间复杂度方面，基础算法仅需存储 K 近邻信息 $O(m \cdot K \cdot d)$，其中 d 为特征维度，而改进算法因存储全局距离矩阵，空间占用增至 $O(N^2)$。该改进以更高的计算代价换取了更优的数据增强效果，适用于中小规模数据集或可接受离线计算的场景。

4. 实验

4.1. 评价指标

为系统评估改良后的 ADASYN 在 PSA 剥离强度的预测任务中的有效性，选取均方误差(Mean Squared Error, *MSE*)、平均绝对误差(Mean Absolute error, *MAE*)、决定系数(Coefficient of Determination, R^2)、k 折交叉验证的 MSE 标准差 σ_{MSE}、k 折交叉验证的 *MSE* 变异系数 CV_{MSE} 和 k 折交叉验证的 *MSE* 波动范围 ΔMSE 这 6 种评价指标做评判，具体计算方法呈现在表 3 中。

Table 3. Evaluation metrics
表 3. 评价指标

评价指标	计算公式
MSE	$\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$
MAE	$\frac{1}{m}\sum_{i=1}^{m}\lvert y_i-\hat{y}_i\rvert$
R^2	$1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2}$
σ_{MSE}	$\sqrt{\frac{1}{k-1}\sum_{j=1}^{k}\left(MSE_j-\overline{MSE}\right)^2}$
CV_{MSE}	$\frac{\sigma_{MSE}}{\overline{MSE}}\times 100\%$
ΔMSE	$\max_{1\le j\le k}(MSE_j)-\min_{1\le j\le k}(MSE_j)$

m 为样本数量； y_i 为真实标签值； $\hat{y}_i$ 为模型预测标签值； k 为交叉验证折数。

4.2. 对比试验设计

如图 2 所示，实验设计分为两组，第一组实验对未经任何处理的原始样本数据进行预测，通过交叉验证和验证集评估模型的性能，记录模型在原始数据上的评价指标，以此作为基准性能。为提高结论的可靠性，SVR 作为核心基础模型的同时，选取了决策树(Decision Tree, DT)、随机森林(Random Forest, RF)、线性回归(Linear Regression, LR)、多项式回归(Polynomial Regression, PR)和岭回归(Ridge Regression, RR)做平行对比。

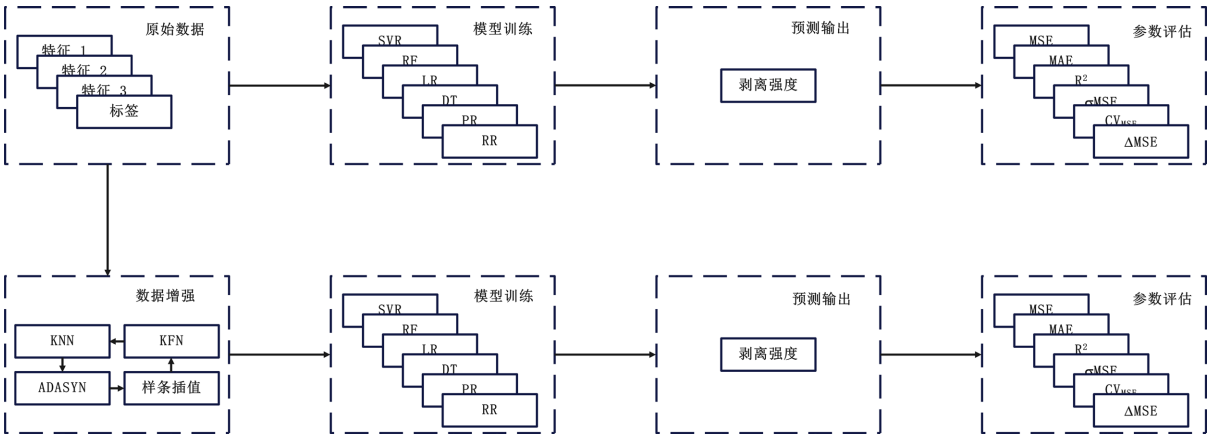

Figure 2. Performance comparison flowchart for data augmentation
图 2. 数据增强性能比较流程图

在第二组实验应用改进的 ADASYN 算法进行数据增强。将增强后的数据用于模型训练，随后同样采用交叉验证和验证集评估模型性能，记录各个模型在增强数据下的评价指标。

最后，通过对比两组实验结果的评价指标，分析改进的 ADASYN 对模型性能的具体影响，考察其在提高模型预测精度、增强泛化能力以及改善数据分布覆盖等方面的优势。

4.3. 验证集结果分析

原始样本随机分为 80%的训练集和 20%的验证集，横向对比六类模型的表现，结果如下表 4 所示。从整体上看，不同回归模型之间，RF 表现出显著的优势，其 *MSE* 值和 *MAE* 值分别为 19.412 和 3.248，远低于其他模型，同时其 R^2 值达到 0.708，表明模型能够较好地拟合数据，并具有较强的预测能力。此外，RF 的 *MSE* 值波动范围较小，模型表现较为稳定。相比之下，SVR 和 DT 的表现相对一般。SVR 的 *MSE* 和 *MAE* 值分别为 24.380 和 3.861，R^2 值为 0.052，尽管误差相对较小，但其对数据整体趋势的捕捉能力有限。DT 的 *MSE* 值为 50.121，*MAE* 值为 5.225，R^2 值仅为 0.037，模型的预测精度和稳定性均不如 RF。LR 和 PR 的性能则较为欠缺，两者的 R^2 值分别为−0.294 和 0.104，均未能展现出对数据的有效拟合。LR 的 *MSE* 值高达 125.257，*MAE* 值为 8.285，表现出较大的预测误差。而 PR 尽管略好于 LR，但其 *MSE* 值为 59.759，仍表明模型的预测性能不佳，且不稳定。RR 的表现与 SVR 和 PR 相似，*MSE* 为 63.153，*MAE* 为 5.964，R^2 值仅为 0.053，表明模型的预测误差较大，整体表现不理想。

Table 4. Performance evaluation form of the original sample validation set
表 4. 原始样本验证集性能评价表

模型	*MSE*	*MAE*	R^2
SVR	24.380	3.861	0.052

续表

DT	50.121	5.225	0.037
RF	19.412	3.248	0.708
LR	125.257	8.285	−0.294
PR	59.759	5.737	0.104
RR	63.153	5.964	0.053

利用改进的合成过采样方法得到的增强样本集，对应六种回归模型横向比较模型效果如表 5 所示。整体来看，增强样本集在 DT 和 SVR 两个模型展现了良好的适应性。其中，DT 的 *MSE* 为 22.136，*MAE* 为 3.375，R^2 值为 0.442，在多个指标中表现最佳。SVR 紧随其后，*MSE* 和 *MAE* 分别为 12.089 和 2.961，R^2 值达到 0.339，虽然在误差控制上略优于 DT，但在整体拟合性能上稍显不足。RF 的表现也令人满意，*MSE* 为 25.084，*MAE* 为 3.770，R^2 值为 0.420，展现了稳定性方面的优势。相比之下，线性模型如 LR 和 PR 的性能则受到一定限制，其 *MSE* 值分别为 55.489 和 46.118，R^2 值为 0.096 和 0.176，表明传统方法在应对改进后的数据特性时效果有限，误差较大且稳定性较低。

Table 5. Enhanced sample validation set performance evaluation form
表 5. 增强样本验证集性能评价表

模型	*MSE*	*MAE*	R^2
SVR	12.089	2.961	0.339
DT	22.136	3.375	0.442
RF	25.084	3.770	0.420
LR	55.489	5.577	0.096
PR	46.118	5.022	0.176
RR	59.746	5.674	0.173

验证集在 SOMO [31]方法上的性能对比结果如表 6 所示，改进的 ADASYN 在 *MSE* 和 *MAE* 指标上表现更优，但在 R^2指标上略低，表明其在降低预测误差方面更有效，而 SOMO 在模型解释力上稍具优势。

Table 6. Comparative performance evaluation table
表 6. 对比性能评价表

	MSE	*MAE*	R^2
原始样本	24.380	3.861	0.052
SOMO	13.499	3.042	0.378
改进 ADASYN	12.089	2.961	0.339

4.4. 交叉验证结果分析

为验证改进算法的鲁棒性与泛化能力，实验采用五重交叉验证对原始数据集进行评估。结果如表 7 所示。表中 $\overline{MSE}$ 、 $\overline{MAE}$ 、 $\overline{R^2}$ 均为五重交叉验证的平均值，反映了不同模型在整体预测能力上的表现。

Table 7. Five-fold crossover performance evaluation form for the original sample
表 7. 原始样本五重交叉性能评价表

模型	$\overline{MSE}$	$\overline{MAE}$	$\overline{R^2}$	σ_{MSE}	CV_{MSE}	ΔMSE
SVR	59.086	4.944	0.136	27.286	0.461	86.372~31.800
DT	50.257	4.925	0.086	16.926	0.3367	67.183~33.330
RF	57.785	5.573	0.072	34.703	0.600	92.488~23.081
LR	78.070	6.437	−0.182	51.542	0.660	129.612~26.527
PR	76.467	6.504	−0.214	35.866	0.469	112.334~40.600
RR	68.156	6.267	−0.169	41.358	0.606	109.514~26.797

从结果来看，DT 模型在各项指标上表现相对较优，其 *MSE* 和 *MAE* 值分别为 50.257 和 4.925，均为最低，表明其预测误差较小。同时，DT 的 $\overline{R^2}$ 值为 0.086，尽管不高，但在所有模型中处于领先地位。此外，DT 的 ΔMSE (33.330~67.183)和 σ_{MSE} (16.926)均较小，表明其在交叉验证中的性能较为稳定。相比之下，LR 和 PR 模型表现较差。两者的 $\overline{R^2}$ 值分别为−0.182 和−0.214，均为负值，说明其预测性能甚至低于简单均值模型。同时，二者的 *MSE* 值分别为 78.070 和 76.467，*MAE* 值分别为 6.437 和 6.504，均表明误差较大，且不稳定。SVR 和 RF 的表现相对接近，其 *MSE* 值分别为 59.086 和 57.785，*MAE* 值分别为 4.944 和 5.573，尽管在误差控制上有所提升，但其 $\overline{R^2}$ 值分别为 0.136 和 0.072，表明模型的预测能力仍需进一步优化。RR 的 *MSE* 值为 68.156，$\overline{R^2}$ 值为−0.169，表现相对较弱。

Table 8. Enhanced sample five-fold crossover performance evaluation form
表 8. 增强样本五重交叉性能评价表

模型	$\overline{MSE}$	$\overline{MAE}$	$\overline{R^2}$	σ_{MSE}	CV_{MSE}	ΔMSE
SVR	34.183	3.765	0.314	13.181	0.209	46.032~29.669
DT	37.069	3.868	0.410	17.481	0.471	54.550~19.587
RF	36.151	4.791	0.348	18.384	0.508	54.536~17.766
LR	62.110	6.063	0.0421	26.298	0.423	135.811~88.408
PR	53.274	5.612	0.111	17.874	0.335	71.149~35.400
RR	73.422	5.925	0.127	32.901	0.448	106.323~40.520

增强样本集的五重交叉验证结果如表 8 所示。从表中可以看出，增强样本显著优化了模型的预测性能。其中，SVR 的提升尤为显著，其 *MSE* 值从 59.086 降低至 34.183，下降了 42.13%，*MAE* 值从 4.944 降低至 3.765，减少了 23.86%，$\overline{R^2}$ 值从 0.136 提升至 0.314，增幅达到 130.88%。这一结果表明，SVR 在增强数据集上展现了更强的预测能力和稳定性。DT 模型在改进后依然保持出色的性能，其 *MSE* 和 *MAE* 值分别为 37.069 和 3.868，较原始数据集结果分别下降 26.28%和 21.43%，$\overline{R^2}$ 值提升至 0.410，较原始值增加了 376.74%，表现最佳。此外，DT 的 ΔMSE (19.587~54.550)和 σ_{MSE} (17.481)相对较小，展现了良好的稳定性。RF 在增强数据集上的表现亦有所提升，*MSE* 值下降 37.46%，*MAE* 值下降 14.00%，$\overline{R^2}$ 值提升至 0.348，表明其能够有效捕捉增强数据的模式。但与 DT 相比，RF 的误差波动范围和标准差稍大。相比之下，LR 和 PR 模型的表现尽管有所提升($\overline{R^2}$ 值分别从−0.182 和−0.214 提升至 0.042 和 0.111)，但其 *MSE* 和 *MAE* 值仍较高，表明其在建模复杂非线性关系时的能力不足。RR 的表现也有一定改善，但提升幅度相对有限。

参考文献

[1] Bae, M., Ahn, S., You, S., Kim, J., Kim, D., Kim, H., *et al.* (2024) Expanded Illite Filler in UV-Curable Polymer Electrolytes for All-Solid-State Li-Ion Batteries. *Coatings*, **14**, Article 1158. https://doi.org/10.3390/coatings14091158

[2] Kowalczyk, A., Kowalczyk, K., Gruszecki, J., Idzik, T.J. and Sośnicki, J.G. (2024) Thermally Stable UV-Curable Pressure-Sensitive Adhesives Based on Silicon-Acrylate Telomers and Selected Adhesion Promoters. *Polymers*, **16**, Article 2178. https://doi.org/10.3390/polym16152178

[3] 李石开, 曾东鳌, 杜方舟, 等. 血管化类器官的构建方法及生物材料[J]. 合成生物学, 2024, 5(4): 851-866.

[4] Lv, J., Wang, Q. and Xie, H. (2024) Isosorbide-based Acrylic Pressure-Sensitive Adhesives through UV-Cured Crosslinking with a Balance between Adhesion and Cohesion. *Polymers*, **16**, Article 3178. https://doi.org/10.3390/polym16223178

[5] Liu, Z., Yuan, G., Yang, H., Wang, K., Tao, Y., Wu, K., *et al.* (2025) High-Performance PSA with Dual-Network Structure for Enhanced Cohesion. *Colloids and Surfaces A*: *Physicochemical and Engineering Aspects*, **711**, Article ID: 136271. https://doi.org/10.1016/j.colsurfa.2025.136271

[6] Zhang, K., Yang, X., Wang, Y., Yu, Y., Huang, N., Li, G., *et al.* (2025) Artificial Intelligence in Drug Development. *Nature Medicine*, **31**, 45-59. https://doi.org/10.1038/s41591-024-03434-4

[7] Pang, B., Ryu, C. and Kim, H. (2014) Effect of Naphthyl Curing Agent Having Thermally Stable Structure on Properties of UV-Cured Pressure Sensitive Adhesive. *Journal of Industrial and Engineering Chemistry*, **20**, 3195-3200. https://doi.org/10.1016/j.jiec.2013.11.065

[8] Liu, Q., Chu, J., Yang, X., Huang, Y., Zhao, M., Zheng, Q., *et al.* (2021) Study of Apparent Molar Volumes of Ether Functionalized Ionic Liquids with Three Ether Solvents. *Journal of Molecular Liquids*, **333**, Article ID: 115958. https://doi.org/10.1016/j.molliq.2021.115958

[9] Li, Z., Yoon, J., Zhang, R., Rajabipour, F., Srubar III, W.V., Dabo, I., *et al.* (2022) Machine Learning in Concrete Science: Applications, Challenges, and Best Practices. *npj Computational Materials*, **8**, Article No. 127. https://doi.org/10.1038/s41524-022-00810-x

[10] Papadimitriou, I., Gialampoukidis, I., Vrochidis, S. and Kompatsiaris, I. (2024) AI Methods in Materials Design, Discovery and Manufacturing: A Review. *Computational Materials Science*, **235**, Article ID: 112793. https://doi.org/10.1016/j.commatsci.2024.112793

[11] 李勇. UV 光固化压敏胶的制备方法及研究进展[J]. 中国胶粘剂, 2019, 28(9): 43-46.

[12] Eyring, V., Collins, W.D., Gentine, P., Barnes, E.A., Barreiro, M., Beucler, T., *et al.* (2024) Pushing the Frontiers in Climate Modelling and Analysis with Machine Learning. *Nature Climate Change*, **14**, 916-928. https://doi.org/10.1038/s41558-024-02095-y

[13] Agrawal, A. and Choudhary, A. (2016) Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Materials*, **4**, Article ID: 053208. https://doi.org/10.1063/1.4946894

[14] Hart, G.L.W., Mueller, T., Toher, C. and Curtarolo, S. (2021) Machine Learning for Alloys. *Nature Reviews Materials*, **6**, 730-755. https://doi.org/10.1038/s41578-021-00340-w

[15] Lookman, T., Balachandran, P.V., Xue, D. and Yuan, R. (2019) Active Learning in Materials Science with Emphasis on Adaptive Sampling Using Uncertainties for Targeted Design. *NPJ Computational Materials*, **5**, Article No. 21. https://doi.org/10.1038/s41524-019-0153-8

[16] Schmidt, J., Marques, M.R.G., Botti, S. and Marques, M.A.L. (2019) Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Computational Materials*, **5**, Article No. 83. https://doi.org/10.1038/s41524-019-0221-0

[17] Liu, Y., Niu, C., Wang, Z., Gan, Y., Zhu, Y., Sun, S., *et al.* (2020) Machine Learning in Materials Genome Initiative: A Review. *Journal of Materials Science & Technology*, **57**, 113-122. https://doi.org/10.1016/j.jmst.2020.01.067

[18] 赵凯琳, 靳小龙, 王元卓. 小样本学习研究综述[J]. 软件学报, 2021, 32(2): 349-369.

[19] Li, Y., Xiao, N. and Ouyang, W. (2019) Improved Generative Adversarial Networks with Reconstruction Loss. *Neurocomputing*, **323**, 363-372. https://doi.org/10.1016/j.neucom.2018.10.014

[20] Wu, J., Liu, S., Huang, D. and Wang, Y. (2020) Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020. *ECCV* 2020, Springer, 456-472. https://doi.org/10.1007/978-3-030-58517-4_27

[21] Chao, X. and Zhang, L. (2021) Few-Shot Imbalanced Classification Based on Data Augmentation. *Multimedia Systems*, **29**, 2843-2851. https://doi.org/10.1007/s00530-021-00827-0

[22] Jia, B., Tian, Y., Zhao, D., Wang, X., Li, C., Niu, W., *et al.* (2021) Bidirectional RNN-Based Few-Shot Training for Detecting Multi-Stage Attack. In: Wu, Y. and Yung, M., Eds., *Information Security and Cryptology*, Springer, 37-52.

https://doi.org/10.1007/978-3-030-71852-7_3

[23] Liu, J., Chen, Y., Lan, L., Lin, B., Chen, W., Wang, M., *et al.* (2018) Prediction of Rupture Risk in Anterior Communicating Artery Aneurysms with a Feed-Forward Artificial Neural Network. *European Radiology*, **28**, 3268-3275. https://doi.org/10.1007/s00330-017-5300-3

[24] Feng, X. and Li, G. (2022) UV Curable, Flame Retardant, and Pressure-Sensitive Adhesives with Two-Way Shape Memory Effect. *Polymer*, **249**, Article ID: 124835. https://doi.org/10.1016/j.polymer.2022.124835

[25] Paul, R., John, B. and Sahoo, S.K. (2022) UV-Curable Bio-Based Pressure-Sensitive Adhesives: Tuning the Properties by Incorporating Liquid-Phase Alkali Lignin-Acrylates. *Biomacromolecules*, **23**, 816-828. https://doi.org/10.1021/acs.biomac.1c01249

[26] Czech, Z., Bartkowiak, M. and Mozelewska, K. (2021) Influence of Unsaturated Photoinitiators on Acrylic Pressure-Sensitive Adhesive Properties. *International Journal of Adhesion and Adhesives*, **107**, Article ID: 102846. https://doi.org/10.1016/j.ijadhadh.2021.102846

[27] Dana, S.F., Nguyen, D., Kochhar, J.S., Liu, X. and Kang, L. (2013) UV-Curable Pressure Sensitive Adhesive Films: Effects of Biocompatible Plasticizers on Mechanical and Adhesion Properties. *Soft Matter*, **9**, 6270-6281. https://doi.org/10.1039/c3sm50879j

[28] Antosik, A.K., Mozelewska, K., Musik, M., Miądlicki, P. and Wilpiszewska, K. (2023) Influence of Illite and Its Amine Modifications on the Self-Adhesive Properties of Silicone Pressure-Sensitive Adhesives. *Materials*, **16**, Article 2879. https://doi.org/10.3390/ma16072879

[29] Huang, J., Fu, P., Li, W., Xiao, L., Chen, J. and Nie, X. (2022) Influence of Crosslinking Density on the Mechanical and Thermal Properties of Plant Oil-Based Epoxy Resin. *RSC Advances*, **12**, 23048-23056. https://doi.org/10.1039/d2ra04206a

[30] He, H.B., Bai, Y., Garcia, E.A. and Li, S.T. (2008) ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 2008 *IEEE International Joint Conference on Neural Networks* (*IEEE World Congress on Computational Intelligence*), Hong Kong, 1-8 June 2008, 1322-1328. https://doi.org/10.1109/ijcnn.2008.4633969

[31] Tonin, F., Patrinos, P. and Suykens, J.A.K. (2021) Unsupervised Learning of Disentangled Representations in Deep Restricted Kernel Machines with Orthogonality Constraints. *Neural Networks*, **142**, 661-679. https://doi.org/10.1016/j.neunet.2021.07.023