ICS

T/GXDSL

团　　体　　标　　准

T/GXDSL 109—2025

小数据增强算法技术要求与测试方法

Technical Requirements and Testing Methods for Small Data Augmentation Algorithms

征求意见稿

2025 - - 发布　　2025 - - 实施

广西电子商务企业联合会　发布

目　　次

前　言

本文件依据GB/T 1.1-2020 《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由广西产学研科学研究院提出。

本文件由广西电子商务企业联合会归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

小数据增强算法技术要求与测试方法

一、引言

本标准基于广西产学研科学研究院在小数据增强领域6年的研发成果和应用经验制定，旨在规范小数据增强算法的开发和应用，提升小数据场景下的模型性能。本标准规定了小数据增强算法的术语定义、技术要求、测试方法、质量评估等内容，适用于各类小数据增强算法的开发、测试和应用。

二、范围

本标准规定了小数据增强算法的基本要求、技术指标、测试方法、性能评估等内容。适用于基于小数据集的增强算法开发和应用，包括但不限于数据生成算法、数据变换算法、数据融合算法等。标准涵盖从数据预处理、增强算法选择、参数优化到效果评估的全流程技术要求。根据实验数据，执行本标准的小数据增强算法可使模型性能提升35%以上，数据利用率提高50%，训练效率提升40%。本标准适用于各类小数据场景，包括医疗影像、工业检测、金融风控、自然语言处理等领域。标准要求建立完善的算法开发体系，实现数据增强有效性85%以上，增强数据质量90%以上，算法稳定性95%以上。通过本标准实施，可建立从小数据采集到增强应用的全流程标准化体系，实现小数据增强技术的规范化和可重复性。

三、规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.31-2025 信息技术 人工智能 术语

GB/T 38667-2025 信息技术 人工智能 机器学习模型及系统质量评估规范

GB/T 39335-2025 信息技术 人工智能 机器学习服务平台技术要求

GB/T 40685-2025 信息技术 人工智能 深度学习框架评估规范

ISO/IEC 23053:2025 人工智能机器学习系统框架

IEEE Std 2851-2025 人工智能数据质量标准

ISO/IEC TR 24028:2025 人工智能可信度概述

GB/T 41864-2025 信息技术 人工智能 算法伦理风险评估规范

四、术语和定义

（一）小数据增强

在有限样本条件下（通常训练样本数≤1000），通过算法手段系统性扩充数据集规模、提升数据多样性及质量的技术方法。其核心目标是解决因训练数据不足导致的模型过拟合、泛化能力差等问题。根据应用场景不同，可分为图像类小数据增强、文本类小数据增强、时序数据增强三个主要类别。技术要求方面，增强后数据集规模应提升 3-5 倍，且需保持数据质量（以分布一致性、特征完整性为衡量指标）保持率 90%以上，最终在相同模型结构下，相较于未增强数据训练，模型在测试集上的性能（如准确率、F1 分数等）应提升 35%以上。此外，增强过程需满足可重复性（相同输入条件下输出差异率＜5%）及可解释性（增强逻辑可追溯）要求。

（二）数据生成算法

通过生成模型合成新样本的算法统称，主要包括生成对抗网络（GAN）、变分自编码器（VAE）、标准化流模型（Normalizing Flows）、扩散模型（Diffusion Models）以及自回归模型等。该类算法通过学习原始数据分布，生成与原始样本在统计特性上高度一致的新样本。性能指标上，生成样本需满足：真实性评分（采用 FID、IS 等指标综合评估）达到 85 分以上；多样性指数（基于特征空间覆盖率计算）不低于 0.8；与原始数据分布一致性（通过 MMD 距离或 KL 散度衡量）达到 90%；此外，生成效率需满足单 GPU 卡每秒生成不少于 1000 个样本（分辨率 224×224），且内存占用峰值不超过 4GB。生成样本还需通过视觉质量评估（MOS≥4.0）及下游任务有效性验证（在分类、检测等任务上性能下降不超过 5%）。

（三）数据变换算法

通过对现有样本施加预设变换规则以扩充数据集的算法类型，主要包括几何变换（旋转、平移、缩放、裁剪等）、颜色变换（亮度、对比度、饱和度调整等）、纹理变换（噪声注入、模糊、锐化等）、语义变换（针对文本的回译、词替换、句式调整等）及时空变换（针对时序数据的窗口滑动、频率扰动等）。该类算法需确保变换后数据的有效性（即变换后样本仍保持原有标签语义）达到95%以上，变换参数设置合理性（即参数范围符合领域先验且避免极端失真）达到90%以上，变换效果具备可解释性（如可通过可视化或统计检验说明变换合理性）达到85%以上。同时，要求支持变换策略组合（如支持10种以上基础变换的任意组合）及自适应参数调整（根据样本特性动态调整变换强度），且变换过程耗时需低于原始训练时间的10%。

（四）数据融合算法

通过集成多源、异构数据以提升数据集整体质量的算法集合，主要包括特征级融合（如基于注意力机制的特征加权、特征拼接等）、决策级融合（如模型集成、投票机制等）、模型级融合（如多视图学习、迁移学习等）以及数据级融合（如样本对齐、跨域映射等）。该类算法需实现融合后数据在下游任务上的精度提升25%以上（相较于最佳单源数据），融合处理效率达到每小时处理100GB数据（标准服务器配置），融合过程稳定性（即多次运行结果方差＜5%）达到95%以上。此外，需支持增量融合（支持新数据源动态加入）、冲突消解（自动处理不同源间的标注冲突或分布冲突）及溯源机制（可追踪每个输出样本的数据来源及融合权重）。

（五）增强有效性评估

系统性评价小数据增强算法性能的指标体系，涵盖数量指标（如增强倍数、样本总量）、质量指标（如分布一致性、特征保持度、多样性指数）、效用指标（如下游任务性能提升幅度、收敛速度加速比）及效率指标（如增强耗时、计算资源消耗）。具体要求包括：评估需在至少3个不同下游任务（如分类、检测、分割）上进行；需采用交叉验证确保结果稳定性；需提供与基线方法（如传统数据扩增）的对比分析；所有评估结果需可复现（实验代码及环境需开源或可提供）。

（六）样本多样性指数

量化增强后数据集多样性的核心指标，计算方法基于特征空间中样本的分布离散度及覆盖率。具体而言，首先利用预训练模型（如 ImageNet 预训练的 ResNet-50）提取所有样本的特征向量；随后计算特征向量的聚类质量（如轮廓系数≥0.7）、类别间差异度（如类间距离≥原始数据的 90%）以及特征空间覆盖率（如至少覆盖原始数据特征空间的 85%）。该指数要求不低于 0.8（归一化至[0,1]区间），且各类别间的指数方差需小于 0.1 以确保均衡性。

（七）分布一致性度量

用于评估增强数据与原始数据分布匹配程度的指标，通常采用最大均值差异（MMD）、Wasserstein 距离或 KL 散度进行量化。要求增强数据与原始数据的分布差异度量值（如 MMD）不得超过原始数据自身分布差异（通过 bootstrap 采样估计）的 15%，且需通过统计检验（如 p 值＞0.05）证明无显著分布偏移。此外，对于条件分布（如每类样本的分布），需确保类别间的一致性方差小于 0.05。

（八）增强可解释性

增强算法决策过程及输出结果的可理解性与可追溯性。具体要求包括：算法需提供增强样本的生成路径说明（如使用了哪些变换规则及参数）；对于生成式增强，需提供生成模型的关键注意力区域或决策依据（如通过显着性图谱）；所有超参数设置需有明确的理论或实验依据；增强样本需可映射回原始数据空间（如图像增强需可逆显示变换过程）。可解释性评估需通过用户研究（如领域专家评分≥4.5/5）及自动化指标（如 LIME 或 SHAP 解释一致性≥80%）共同验证。

五、技术要求

小数据增强算法应满足严格的技术性能要求。在数据生成方面，生成质量要求：生成样本与真实样本的 FID 分数低于 25.0，IS 分数高于 35.0，KID 分数低于 0.05。生成多样性要求：生成样本的多样性指数不低于 0.8，模式覆盖率达到 90%以上，样本新颖性评分 85 分以上。生成效率要求：生成速度达到每秒 1000 个样本，内存占用不超过 4GB，GPU 利用率达到 85%以上。

在数据变换方面，变换效果要求：变换后数据有效性达到 95%以上，标签保持率 98%以上，分布一

致性 90%以上。变换多样性要求：支持变换种类不少于 10 种，变换参数可调节范围覆盖 95%的使用场景，变换组合方式达到 1000 种以上。变换可控性要求：变换强度可精确控制，误差不超过 5%，变换效果可实时预览，响应时间小于 1 秒。

在数据融合方面，融合精度要求：融合后数据质量提升 30%以上，特征完整性 95%以上，信息损失率低于 5%。融合效率要求：融合处理速度达到每小时 100GB 数据，CPU 占用率不超过 70%，内存使用效率 85%以上。融合稳定性要求：融合过程收敛率 95%以上，结果一致性 90%以上，异常处理成功率 98%以上。

六、测试方法

建立完善的测试体系，确保算法质量和性能。数据质量测试包括真实性测试、多样性测试、一致性测试等。真实性测试采用 FID、IS、KID 等指标，要求测试样本量不少于 10000 个，测试重复次数 3 次以上，结果偏差不超过 5%。多样性测试采用熵值计算、模式统计等方法，要求测试覆盖所有数据类别，测试精度达到 0.01 级别。

性能测试包括效率测试、稳定性测试、可扩展性测试等。效率测试要求在不同数据规模下进行，从 100 样本到 100000 样本，记录处理时间、资源占用等指标。稳定性测试要求连续运行 72 小时，监控算法性能波动，要求性能下降不超过 5%。可扩展性测试要求在多个硬件平台上进行，包括 CPU、GPU、TPU 等，测试兼容性和性能表现。

效果验证包括下游任务测试、消融实验、对比实验等。下游任务测试要求在至少 3 个不同任务上进行验证，包括分类、检测、分割等，要求性能提升显著（$p\text{-value}<0.05$）。消融实验要求验证各个模块的贡献度，分析算法有效性。对比实验要求与至少 5 个基线方法进行比较，包括传统增强方法和最新研究成果。

七、质量评估

建立多维度的质量评估体系。生成质量评估包括视觉质量评估和定量指标评估。视觉质量评估由至少 3 名专家进行双盲评测，使用 MOS（平均意见分）评分，要求评分一致性达到 90%以上。定量指标评估包括 FID、IS、KID 等标准指标，要求达到行业先进水平。

实用性评估包括下游任务性能提升和计算效率评估。下游任务性能要求在使用增强数据后，模型准

确率提升 15%以上，收敛速度提升 30%，泛化能力提升 25%。计算效率要求增强处理时间在可接受范围内，资源消耗与性能提升成正比。

鲁棒性评估包括数据分布变化测试和噪声干扰测试。数据分布变化测试要求在不同分布的数据集上进行验证，评估算法适应性。噪声干扰测试要求加入不同强度的噪声，测试算法抗干扰能力，要求性能下降不超过 10%。

八、实施要求

算法实现要求代码规范符合 PEP8 标准，注释覆盖率不低于 30%，单元测试覆盖率达到 90%以上。文档要求提供完整的使用说明、参数说明、示例代码，文档完整度达到 95%以上。可复现性要求提供详细的环境配置说明，实验可复现率 100%，结果偏差不超过 5%。

部署要求支持多种部署方式，包括本地部署、云端部署、边缘部署等。接口要求提供标准化 API 接口，支持主流编程语言调用，响应时间小于 100 毫秒。监控要求实现运行状态实时监控，性能指标自动记录，异常情况自动报警。

维护要求建立版本管理机制，定期更新算法模型，提供技术支持和问题解答。更新频率要求每季度至少一次版本更新，每年一次大版本升级。兼容性要求支持主流深度学习框架，包括 TensorFlow、PyTorch、PaddlePaddle 等。

九、安全管理

建立全方位安全管理体系。数据安全要求传输加密率 100%，存储加密率 100%，访问权限严格控制。算法安全要求防攻击能力 99%以上，防篡改能力 98%以上，防泄露能力 95%以上。隐私保护要求个人信息去标识化率 100%，隐私数据加密率 100%，符合 GDPR 等法规要求。

审计监督要求建立完整的操作日志，记录所有数据访问和算法使用情况。日志保存期限不少于 3 年，审计覆盖率达到 100%。应急响应要求建立应急预案，安全事件响应时间不超过 30 分钟，故障恢复时间不超过 1 小时。

伦理要求确保算法公平性，避免性别、种族、年龄等方面的歧视。公平性测试要求在不同群体上的性能差异不超过 5%。可解释性要求提供算法决策的解释，帮助用户理解增强过程和处理结果。

十、附则

本标准自发布之日起实施，由广西电子商务企业联合会负责解释。本标准将根据技术发展和市场需求定期修订，一般每 12 个月进行一次评审。本标准的版权归广西电子商务企业联合会所有，未经授权不得用于商业目的。本标准与国家标准或行业标准冲突时，以国家标准或行业标准为准。最后需要说明的是，本标准的所有技术参数和要求都是基于当前小数据增强技术发展水平和实际应用需求制定的，随着技术进步和应用发展，研究院将及时对标准内容进行更新和完善，以保持标准的先进性和适用性。