面向计算机应用领域的高质量数据集构建方法研究

袁 敏
镇江市公共资源交易中心 江苏 镇江 212000

摘 要：本文面向计算机应用领域，针对现有数据集普遍存在的噪声过高、类别失衡、标注误差、适配性差四类共性质量问题，提出标准化数据集构建方法。首先，围绕具体落地场景开展场景化原始数据采集，按设备参数与环境条件进行同源采集，控制数据量级并完成初步归类。其次，通过数据清洗、归一化、均衡化三步预处理降低数据噪声，消除格式差异带来的计算误差。再次，采用人工初标与算法复检的双层标注模式，配合统一标注规范减少主观误差。同时，构建涵盖完整性、一致性、有效性与均衡性的多维质量评价体系，量化评估数据集质量等级。最后，针对高噪声、样本失衡、标注误差三类缺陷，分别提出滤波降噪、混合均衡优化、闭环修正等方法，从全链路提升数据集质量。

关键词：计算机应用；数据集构建；数据预处理；数据标注；数据质量优化

引言：计算机应用领域数据集质量直接决定模型训练效果与下游任务性能。当前开源数据集及项目自建数据集普遍存在原始数据噪声占比过高、类别分布严重失衡、人工标注误差突出、场景适配性较差四类共性缺陷，严重制约模型泛化能力与预测精度。上述问题根源于非标准化的数据构建流程，单一环节的局部优化难以从根本上解决。亟需建立覆盖采集、清洗、标注、验证全链路的规范化构建方法。本文围绕计算机视觉与自然语言处理两大应用场景，系统提出高质量数据集标准化构建流程与多维质量评价体系，为数据集质量提升提供完整解决方案。

1 计算机应用领域现有数据集存在的共性质量问题

面向计算机视觉图像识别与自然语言文本处理两大主流应用场景，当前开源数据集及项目自建数据集普遍存在四类固有质量缺陷，严重制约模型训练效果与下游任务性能。（1）原始数据噪声占比过高。图像数据中大量样本存在模糊失焦、强光反射、目标遮挡、分辨率不足等采集缺陷，部分样本因拍摄角度偏差导致目标形态失真；文本数据中乱码字符、无意义短句、重复冗余语句占比显著，无效数据大量占用存储与计算资源的同时，干扰模型对有效特征的学习与提取。（2）数据类别分布严重失衡。多数数据集呈现正样本过量、负样本稀缺的长尾分布特征，部分类别样本数量相差达数十倍甚至上百倍。模型训练过程中损失函数被高频类别主导，低频类别特征学习不充分，导致模型预测偏向性明显，泛化能力大幅下降。（3）人工标注误差突出。标注过程中标准不统一、漏标、错标、标注边界偏移等问题频发，尤其在目标边缘模糊、语义歧义场景下，不同标注人员的判断差异显著。标注结果与原始数据之间存在匹配偏差，直接降低了监督学习的标签可信度。（4）场景适配性较差。通用数据集数据来源单一、采集环境固定，与项目实际部署环境存在明显域偏差。模型在训练集上表现良好，但迁移至真实应用场景后，因数据分布不一致导致准确率快速下滑。上述问题均源于非标准化的数据构建流程，亟需建立覆盖采集、清洗、标注、验证全链路的规范化构建方法，从根本上提升数据集质量[1]。

2 面向计算机应用领域高质量数据集标准化构建流程

2.1 场景化原始数据采集

高质量数据集构建须摒弃通用数据集宽泛采集的模式，围绕计算机应用具体落地场景划定数据采集范围，确保原始数据与真实业务环境高度一致。（1）针对计算机视觉类应用，应严格按照实际部署的拍摄设备参数进行同源数据采集，包括设备分辨率、拍摄角度、光照条件及背景干扰元素等，最大限度缩小训练数据与实测数据之间的设备参数与环境参数差异。针对自然语言处理类应用，应定向采集对应领域的专业文本、对话语句及指令文本，严格剔除跨领域无关文本数据，确保语料与目标任务的语义一致性。（2）数据采集量级须遵循足量且不冗余原则，避免海量重复同源数据增加后续预处理的计算与存储压力。采集完成后须完成原始数据初步归类，按照数据类别与采集场景建立基础数据目录，形成结构化的原始数据档案，为后续清洗、标注、验证等环节提供规整的数据基础[2]。

2.2 多维度数据预处理

数据预处理是剔除无效数据、降低数据噪声的核心环节，分为数据清洗、数据归一化、数据均衡化三个细分步骤。（1）数据清洗主要剔除完全失效数据，对图像数据删除模糊失焦、过度曝光、目标完全遮挡等不可修复样本，对文本数据删除乱码、空文本、无意义字符及

极端异常离群点数据，从源头消除无效数据对模型训练的干扰。（2）数据归一化统一所有数据的格式与参数标准。图像数据统一调整至相同分辨率、色彩空间与尺寸规格，消除因采集设备差异导致的像素值分布不一致；文本数据统一编码格式，去除特殊符号、多余空格及非标准字符，确保输入特征维度一致，消除格式差异带来的模型计算误差。（3）数据均衡化针对类别分布失衡问题，采用随机欠采样与小样本过采样两种策略，平衡各分类样本数量，缓解数据集类别倾斜问题。处理过程中须严格控制采样比例，确保原始数据特征信息不丢失，提升数据集整体分布的合理性与模型训练的稳定性。

2.3 双层校验精细化数据标注

标注质量直接决定数据集可用度，本文采用人工初标与算法复检相结合的双层标注模式，有效规避单一人工标注的主观误差。（1）标注前须制定统一细化的标注规范，明确标注边界判定标准、标注类别定义、标注输出格式及忽略区域范围，确保所有标注人员执行一致的操作标准。第一层开展人工标注，完成图像目标框选标注、文本实体识别标注、语义分类标注等基础工作，标注人员严格依据规范对每条数据进行逐条标注。第二层依托轻量化校验算法进行自动复检，批量识别漏标、错标、标注边界偏移等异常样本，将标记异常的数据返还人工进行二次修正，直至标注结果通过算法校验。（2）该双层标注模式保留了人工标注对复杂语义场景、模糊边界样本的判断能力，同时借助算法批量筛查降低人工复检工作量，从标注流程层面全方位减少标注误差，提升标注数据的准确性与一致性。

3 高质量数据集多维质量评价体系构建

3.1 数据完整性评价指标

数据完整性是衡量数据集质量的基础维度，主要评价样本缺失与特征缺失情况，包含单样本完整性与整体数据集完整性两个层面。（1）单样本完整性用于检测单条数据是否存在特征缺失或内容残缺问题。图像数据中表现为局部像素丢失、关键区域被遮挡或图像边缘信息不完整；文本数据中表现为语句残缺不全、字段缺失或编码异常导致的内容截断。该指标通过逐条扫描数据样本，统计存在上述缺陷的样本数量及占比。（2）整体完整性用于检测数据集各类别样本的覆盖程度，重点判断是否遗漏场景细分样本与边缘工况样本。若某一类别下的特殊场景、极端工况数据缺失，模型在该类场景下的预测能力将显著下降。评价时需统计各细分场景的样本覆盖率，确保数据集在类别维度与场景维度上均具备充分代表性。（3）量化指标方面，一般要求高质量数据集残缺数据占比低于1%，即每一百条数据中缺陷样本不超过一条，保证每一条有效数据均可完整提供模型训练所需的全部特征信息[3]。

3.2 数据一致性评价指标

数据一致性是衡量数据集内部标准统一程度的核心指标，分为格式一致性与标注一致性两个评价维度。（1）格式一致性用于检验全部数据的尺寸规格、编码方式、存储格式是否统一。图像数据需确保分辨率、色彩空间、通道数一致；文本数据需确保字符编码、字段分隔符、文件格式统一。若数据集中存在格式不一致问题，模型读取数据时将出现格式报错、特征解析失败等异常，直接中断训练流程。（2）标注一致性用于检验同类目标的标注标准是否统一。同一特征在不同样本中须采用相同的标注类别、标注边界与标注格式，防止同一目标出现多种不同标注结果，导致模型对特征产生混淆，训练收敛方向偏移。（3）评价实施层面，通过批量比对工具对全数据集进行一致性检测，自动识别格式异常样本与标注冲突样本，统一数据格式与标注口径，从根本上消除数据集内部标准混乱问题，确保模型输入数据的规范性与可解析性。

3.3 数据有效性与均衡性评价指标

数据有效性与均衡性是评价数据集能否支撑模型高效训练的两项关键量化指标，二者结合可客观判断数据集整体质量等级。（1）有效性用于衡量有效特征数据在整体数据中的占比。通过统计噪声数据、无效冗余数据的数量及占比，计算有效数据比率。图像数据中的模糊、遮挡、过曝样本，文本数据中的乱码、空文本、重复语句均属于无效数据。有效数据占比越高，模型训练过程中可提取的有效特征越充分，训练效率与收敛速度越优。（2）均衡性用于衡量不同类别样本数量的分布合理性。通过计算各类别样本数量差值，得出类别均衡系数。该系数取值范围为0至1，数值越趋近于1，说明各类别样本数量越接近，数据集类别分布越合理。若均衡系数偏低，表明数据集存在明显的长尾分布问题，模型训练将偏向高频类别，低频类别识别能力显著不足。两项指标从数据纯度与分布结构两个维度对数据集进行量化评价，为数据集质量分级提供客观依据。

4 数据集构建过程中常见缺陷及针对性优化方法

4.1 高噪声数据缺陷及降噪优化方法

数据集噪声是影响模型特征提取精度的关键因素，主要分为图像噪声与文本噪声两类。图像噪声包括高斯噪声、椒盐噪声及环境光照干扰噪声，表现为像素值异常波动、局部灰度突变或亮度不均匀；文本噪声包括乱

码字符、口语冗余助词及无关干扰词汇，表现为语义无效信息占比过高、文本结构混乱。针对图像噪声，采用中值滤波与高斯滤波算法完成图像平滑处理。中值滤波对椒盐噪声具有较强抑制能力，高斯滤波则有效消除高斯噪声与光照不均带来的像素波动，二者在保留图像目标边缘特征与纹理信息的前提下完成降噪。针对文本噪声，首先构建领域停用词词库，自动过滤无效助词、特殊符号与无关语句；其次结合上下文语义分析模型，对轻度残缺文本进行语义补全与结构修复。该优化方式仅清除无效干扰信息，不改动有效数据的特征与语义结构，在不损失数据可用性的基础上提升数据纯净度[4]。

4.2 类别样本失衡缺陷及均衡优化方法

类别样本失衡是数据集构建中的常见缺陷，小样本类别数据不足直接导致模型对该类别的识别精度偏低。单纯过采样易引发模型过拟合，单纯欠采样则会丢失有效特征信息，两种单一策略均存在明显局限。采用混合均衡优化方案，针对不同类别分别施策。针对大样本类别，采用随机欠采样策略剔除高度相似的重复样本，降低冗余数据对模型训练的主导影响，同时保留类别核心特征。针对小样本类别，采用无损数据增强方式扩充样本规模。图像数据通过旋转、裁剪、亮度微调等几何变换与像素变换实现样本扩充；文本数据通过同义语句改写、句式重组等方式实现样本扩容。该方案全程不生成虚假无效数据，所有扩充样本均基于原始数据的合理变换产生，确保扩充后数据贴合真实应用场景，在平衡各类别样本数量的同时，有效避免模型过拟合问题，提升模型对小样本类别的泛化识别能力。

4.3 标注误差缺陷及闭环修正优化方法

人工标注不可避免存在主观误差，单一复检机制无法彻底消除标注问题，因此需构建标注-复检-测试-修正的闭环优化机制。在双层标注完成后，将数据集投入简易基础模型进行预训练，通过模型训练损失值反向判断标注缺陷。若局部样本损失值异常偏高，说明该区域存在集中标注错误，据此定位异常样本并批量修正。该方法利用模型对标注噪声的敏感性，精准识别人工标注难以发现的系统性误差。同时留存所有标注修改记录，分析错误类型与分布规律，迭代优化标注规范，逐步降低后续批次数据的标注错误率。该闭环机制将标注质量控制从单次检验升级为持续改进流程，有效提升标注精准度与数据集整体可信度[5]。

结束语

高质量数据集是计算机应用领域模型训练的核心基础。本文系统构建了覆盖场景化采集、多维度预处理、双层校验标注、多维质量评价及缺陷优化的全链路标准化体系，针对噪声过高、类别失衡、标注误差、适配性差四类共性缺陷，提出了降噪滤波、混合均衡、闭环修正等针对性方法。各环节相互衔接、协同约束，从根本上提升了数据集的完整性、一致性、有效性与均衡性。后续可在此框架基础上，结合自动化采集工具与在线质量监控平台，进一步提升构建效率与质量可控性，为模型训练提供高可信度数据支撑。

参考文献

[1]梁琛,马天鸣.计算机图像处理技术在建筑工程中的应用[J].工程抗震与加固改造,2024,46(01):189.

[2]田文涛.计算机图像处理技术在网页设计中的应用[J].集成电路应用,2024,41(01):422-424.

[3]吴学栋.计算机图像处理技术在纺织工业的应用研究[J].化纤与纺织技术,2023,52(12):50-52.

[4]彭楠.新形势下计算机应用技术创新实践探究[J].信息记录材料,2021,22(08):68-70.

[5]白平.计算机应用的发展现状和发展趋势创新研究[J].数字通信世界,2021,(04):130-131.