# 面向无人智能装备的全链路质量特性评估指标体系与测评方法研究

罗 晗[1,2] 丰 瑞[1,2] 冯璐峰[1,2]

1.贵州航天计量测试技术研究所 贵州 贵阳 550009

2.贵州省基础零部件及装备质量检测与评价全省重点实验室 贵州 贵阳 550009

**【摘 要】**：针对无人智能装备向自动化、智能化、自主化过程中，算法质量评测标准不统一、评价维度不全、黑箱算法可验证性不足等突出问题，围绕算法、数据质量两个核心维度构建通用质量评估指标体系与配套测评方法。分层搭建多级量化评价指标，明确各指标数学测算公式，标准化全流程测评步骤。研究成果可支撑无人智能装备算法迭代优化与工程落地测试。

**【关键词】**：无人智能装备；全链路评测；质量特性；多级指标体系；量化测评

**DOI:10.12417/3041-0630.26.09.020**

## 引言

在当今科技迅猛发展的背景下，依托深度学习技术，无人系统具备环境感知、自主决策等能力。然而，智能产品的功能复杂度与日俱增，但其质量评估体系却存在显著滞后性[1]，

针对智能产品通用质量特征评估的研究，国内外学者已取得显著进展。然而仍面临三重挑战：各行业评测标准不统一，评估体系碎片化，产品横向对比困难；测试多局限于实验室环境，决策透明性相关评测；AI 黑箱特性导致传统评测手段适用性较差。美国国防部发布的《人工智能赋能系统研制测试与评估》指南，强调从算法透明性、数据可信度、系统鲁棒性三个维度构建评估框架[2]，为全球无人智能产品质量评估体系提供了重要参考。

综上，构建通用化无人智能产品质量评估指标体系十分必要。量化指标能够确立研发性能基准、定位技术缺陷，标准化评测方案可挖掘复杂工况下的系统隐患，为无人智能产品总体设计与技术优化提供依据。

## 1 评估指标体系构建

### 1.1 数据质量评估

#### 1.1.1 准确性

数据准确性包含数据内容准确性、标注一致性、标注准确性以及数据集构造合理性。

（1）数据内容准确性(Data Accuracy，DA)：数据集中所包含的图像信息是否能够准确、真实地反映其所描述的实际场景或对象，计算公式如下：

$$DA = \frac{N_{correct}}{N_{total}}$$

其中，DA 表示数据内容准确性，$N_{correct}$表示满足正确性要求的数据数量，$N_{total}$表示被评价的数据总量。

（2）标注一致性(Annotation consistency，AC)：数据集中相同实例的标注是否一致，计算公式如下：

$$AC = \frac{N_{consistensy}}{N_{total}}$$

其中，AC 表示标注一致性，$N_{consistensy}$表示标注一致的数据数量，$N_{total}$表示数据总量。

（3）标注准确性(Annotation Accuracy，AA)：标注信息是否能够正确反映真实情况，计算公式如下：

$$AA = \frac{N_{accurate}}{N_{total}}$$

其中，AA 表示标注准确性，$N_{consistensy}$表示标注一致的数据数量，$N_{total}$表示数据总量

（4）数据集构造合理性(Rationality of dataset composition，DCR)：数据集中关于测试集、验证集、训练集的分配是否合理，计算公式如下：

$$DCR = \frac{1}{n}\sum W \times \left(1 - \frac{(E - A)^2}{E}\right)$$

其中，DCR 表示数据集构造合理性，n 表示数据集的数量，W 表示权重，E 表示期望比例，A 表示实际比例。

（项目基金项目支持：贵州省科技计划支撑。合同编号：ZSYS〔2025〕003）

#### 1.1.2 完整性

数据完整性包含元素完整性、标注完整性、类别完整性。

（1）元素完整性(Element completeness，EC)：数据集中所有必要字段或属性是否完整无缺，是否存在缺失值或空值的情况，计算公式如下：

$$EC = \frac{N_{complete}}{N_{total}}$$

其中，EC 为元素完整性，$N_{complete}$表示具备完整元素的数据数量，$N_{total}$表示数据总量。

（2）标注完整性(Annotation Completeness，AC)：数据集中每个样本是否都有对应的标注信息，是否存在未标注的样本，计算公式如下：

$$AC = \frac{N_{complete\_annot}}{N_{total}}$$

其中，AC 表示标注完整性，$N_{complete}$标注完整元素的数据数量，$N_{total}$表示数据总量。

（3）类别完整性(Category Completeness，CC)：数据集中所有类别标签是否全面覆盖了预期的数据类型，确保每个类别的样本都得到充分体现且无遗漏，计算公式如下：

$$CC = \frac{N_D}{N_{total}}$$

其中，CC 表示类别完整性，$N_D$表示数据集的类别数量，$N_{total}$表示总的类别数量。

#### 1.1.3 多样性

数据多样性用于量化数据集中样本在特征属性和类别分布上的差异程度，计算公式如下：

$$D = \frac{-\sum_{i=1}^{M} p_i \ln p_i}{\ln M}$$

其中，D 为多样性，M 表示元素的可能属性数量，$p_i$表示第 i 个属性的出现频率。

#### 1.1.4 唯一性

唯一性是指每个数据实体在数据集中不存在重复记录，用数据重复占(Duplicate Ratio，DR)来量化，计算公式如下：

$$DR = \frac{N_{duplicate}}{N_{total}}$$

其中，DR 为数据重复占比，$N_{duplicate}$表示重复数据数量，$N_{total}$表示数据总量。

#### 1.1.5 合规性

数据合规性是指数据集中的数据是否符合既定的格式和规范要求，用合规数据占比(Compliance Rate，CR)来量化，计算公式如下：

$$CR = \frac{N_{compliant}}{N_{total}}$$

其中，CR 表示合规数据占比，$N_{compliant}$表示通过合规性检查的数据数量，$N_{total}$表示总数据数量。

#### 1.1.6 安全性

安全性指数据集中包含中毒样本的比例及其潜在危害程度，反映了数据在面对恶意攻击或污染时的脆弱性和风险水平。安全性的量化指标为中毒数据占比，计算公式如下：

$$P_{poision} = \frac{N_{poisoned}}{N_{total}}$$

其中，$P_{poision}$表示中毒数据占比，$N_{poisoned}$表示被识别为中毒/污染的数据样本数量，$N_{total}$表示数据集中总样本数。

### 1.2 算法性能评估

#### 1.2.1 基础性

算法的基础指标[3]包括准确性、精度、召回率、错误率、F1 值、KL 散度、ROC 曲线，PRC 曲线、CRC 曲线等。

#### 1.2.2 收敛性

指在特定条件下，算法能够逐步接近问题的解，即随着迭代次数的增加，算法的输出逐渐趋于稳定。

#### 1.2.3 鲁棒性

鲁棒性的量化指标包含性能波动率、扰动稳定性。

（1）性能波动率(Performance Fluctuation Degree，PFD)：模型在原始测试集和经过非对抗扰动处理后的新测试集之间的性能差异。计算公式如下：

$$PFD = \frac{\left|P_{original} - P_{perturbed}\right|}{\left|p_{original}\right|}$$

其中，PFD 表示模型性能波动率，$P_{original}$表示在原始测试集上的性能指标，$P_{perturbed}$表示在添加扰动后的测试集上的性能指标。

（2）扰动稳定性(Perturbation Stability Distance，PSD)：模型在原始样本与添加非对抗扰动后的样本上，使得性能指标值发生变化的样本与其未受扰动的原始样本之间的最小距离。计算公式如下：

$$PSD_{\phi} = \min_{x \in X} [dist_{\phi}(x)]$$

其中，$PSD_{\phi}$为扰动稳定性，X 表示数据集，x 表示样本实例，$dist_{\phi}$表示在 ϕ 类型的扰动下样本与扰动样本的距离函数。具体计算方法如下：

$$dist_{\phi}(x) = \begin{cases} ||x - \bar{x}||_p, f(\bar{x}) \neq y \\ \infty, 其他 \end{cases}$$

其中，$f(\bar{x})$表示通过 ϕ 类型的扰动生成样本$\bar{\phi}$的性能指标值，y 表示原始样本在模型上的性能指标值。

#### 1.2.4 安全性

算法安全性的量化指标为攻击成功率、抗攻击次数比例、抗攻击类型比例、攻击隐蔽性、模型窃取程度及平均查询次数。

（1）攻击成功率(Attack Success Rate，ASR)：描述在经过攻击方法构建的新测试数据集中，模型预测失败的样本数与总样本数之间的比率。计算公式如下：

$$ASR = \frac{N_{adv}}{N_{all}}$$

其中，ASR 为攻击成功率，$N_{adv}$为预测失败的样本数，$N_{all}$为样本总数。

（2）抗攻击次数比例（Anti-attack Num Ratio，ANR）：对于某一类攻击，模型在原始样本集与添加对抗扰动后的对抗样本集上未产生识别结果偏移的样本数量与对抗样本集数量的比例，计算公式如下：

$$ANR = 1 - ASR$$

其中：ASR 为攻击成功率。

（3）抗攻击类型比例（Anti-attack Type Ratio，ATR）：模型遭受对抗攻击时，能够成功将其攻破的攻击类型数量占所有攻击类型总数量的比 ATR 例。计算公式如下：

$$ATR = 1 - \frac{N_{breach}}{N_{test}}$$

其中，$N_{breach}$表示攻破系统的攻击类型数量，$N_{test}$表示攻击类型数量。

（4）攻击隐蔽性：对抗攻击生成的对抗样本与原始样本之间的平均相似程度。计算公式如下：

$$\cos(A,B)=(A \cdot B)/(\|A\| \times \|B\|)$$

其中：A，B 分别表示对抗样本与原始样本。

（5）模型窃取程度(Model Stealing Degree，MSD)：描述通过如模型蒸馏或其他方法构建的代理模型与原始模型之间的性能差异。计算公式如下：

$$MSD = \frac{\sum_{x \in D} \delta(x)}{|D|}$$

其中：MSD 表示模型窃取程度，|D|表示数据集样本总数，δ(·)为指示函数，当代理模型的预测与原始模型的预测相同时为 1，否则为 0。

（6）平均查询次数：用来衡量生成对抗样本所需的平均模型查询次数。

## 2 评估流程

### 2.1 数据质量评估流程

首先明确定义质量评估需求，确定关注的数据质量维度与目标；建立数据规范，制定数据采集、存储等环节的标准规则；确定评价指标，筛选或设计可量化的精准衡量指标；实施评价，运用评价指标的定义及评估方法获取质量结果。

### 2.2 算法性能评估流程

在开展算法评估时，需遵循“先评数据质量，再评算法”的逻辑。当通过评估准备阶段得到可靠的数据后，用其开展算法评估，依次经构建模型、选算法指标、运行任务、计算与分析指标，判断算法是否达标，决定交付或优化算法。

## 3 总结

本研究针对无人智能产品构建了一套通用质量评估指标体系。研究从算法、数据质量两个维度出发，建立了包含多层级指标的评价体系，规范了评估的基本流程，实现了定性与定量相结合的综合评价方法，该指标体系可以为解决当前无人智能产品质量评估中的标准不统一、维度不完善和黑箱模型验证困难等问题提供了解决方案。

## 参考文献：

[1] 权晓伟,张灏龙,龚茂华,等.美军军事智能试验鉴定技术发展态势研究[J].中国航天,2021,(02):62-64.

[2] Office of the Under Secretary of Defense for Research and Engineering,U.S.Department of Defense..Developmental Test and Evaluation of Artificial Intelligence-Enabled Systems Guidebook[R].Washington,DC:U.S.Department of Defense.2025.

[3] 张龙,王数,雷震,等.AIGC 军事大模型评估体系框架研究[J].战术导弹技术,2025(1):42-52.