ICS 35.100.05
CCS L 79

中华人民共和国国家标准

GB/T 45225—2025

人工智能　深度学习算法评估

Artificial intelligence—Deep learning algorithms evaluation

2025-01-24 发布　　2025-01-24 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前　言

本文件按照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。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件的起草单位：中国电子技术标准化研究院、中国科学院软件研究所、中科南京软件技术研究院、北京航空航天大学、北京软件产品质量检测检验中心有限公司、北京航天自动控制研究所、中国南方电网有限责任公司超高压输电公司、上海计算机软件技术开发中心、中国科学技术大学、北京眼神科技有限公司、上海商汤智能科技有限公司、电装智能科技(上海)有限公司、中电科大数据研究院有限公司、浪潮电子信息产业股份有限公司、中国移动通信集团有限公司、北京声智科技有限公司、广电运通集团股份有限公司、上海文鳐信息科技有限公司、杭州海康威视数字技术股份有限公司、卡斯柯信号有限公司、阿里云计算有限公司、天津(滨海)人工智能创新中心、中国兵器工业信息中心、上海燧原科技股份有限公司、上海市人工智能行业协会、深圳云天励飞技术股份有限公司、四川长虹电子控股集团有限公司、中国船舶集团有限公司综合技术经济研究院、北京计算机技术及应用研究所、香港科技大学、中国科学院空间应用工程与技术中心、浙江大学、中国航空工业集团公司沈阳飞机设计研究所、北京邮电大学、南瑞集团有限公司、重庆国科础智信息技术有限公司、国科础石(重庆)软件有限公司、重庆建设工业(集团)有限责任公司。

本文件主要起草人：鲍薇、叶珩、孟令中、薛云志、马骋昊、高卉、刘祥龙、孔昊、王洋、王宁、陈文捷、张兰、杨春林、吴庚、朱健、董乾、杨光、蔡惠民、杜国光、王珂琛、聂锦燃、陈孝良、徐天适、芮子文、任文奇、周庭梁、吴涛、史殿习、谢晚冬、梅敬青、陈曦、饶雪、曹钰、吴立金、徐哲炜、宋金珂、刘艾杉、郭晋阳、王金波、纪守领、温晓玲、程祥、陈徯、胡艳玲、罗勇军、张洋。

人工智能　深度学习算法评估

1　范围

本文件确立了人工智能深度学习算法的评估指标体系，描述了评估方法等内容。

本文件适用于指导深度学习算法开发方、用户方以及第三方等相关组织对深度学习算法及其训练得到的深度学习模型开展评估工作。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 35273—2020　信息安全技术　个人信息安全规范

GB/T 40660—2021　信息安全技术　生物特征识别信息保护基本要求

GB/T 41867—2022　信息技术　人工智能　术语

3　术语和定义

GB/T 41867—2022 界定的以及下列术语和定义适用于本文件。

3.1

深度学习　deep learning

通过训练具有许多隐藏层的神经网络来创建丰富层次表示的方法。

注：深度学习是机器学习的一个子集。

［来源：GB/T 41867—2022，3.2.27］

3.2

深度学习算法　deep learning algorithm

使用深度神经网络结构进行学习和推理、以完成特定功能的代码片段。

3.3

深度学习模型　deep learning model

基于输入数据或信息产生推理或预测结果的数学架构。

3.4

测试数据　test data

用于评估最终机器学习模型性能的数据。

［来源：GB/T 41867—2022，3.2.3］

3.5

对抗样本　adversarial examples

在数据集中添加细微干扰形成的输入样本，能以较高概率诱导深度学习算法给出错误的输出，甚至是给出特定结果。

4 评估指标体系

4.1 评估指标构成

深度学习算法的评估指标体系包括基础性能、效率、正确性、兼容性、可解释性、鲁棒性、安全性、公平性8个质量特性，见图1。在实施评估过程中，应根据不同类型的深度学习算法，在不同质量特性下设置具体评估指标。

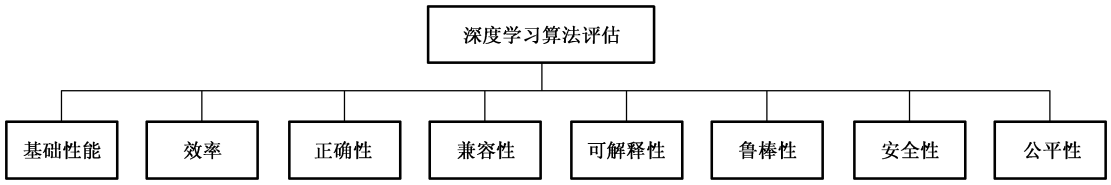

图1 深度学习算法评估指标体系

4.2 基础性能

基础性能指深度学习算法执行过程中的性能特征，不同算法可能涉及不同性能指标。基础性能的评估指标应包括但不限于以下。

a) 准确率：预测正确的样本数占总样本数的比率。
b) 精度：预测类别为正样本的集合中真实类别为正样本的比率。
c) 召回率：被正确预测的正样本占全部正样本的比率。
d) 错误率：对于给定的数据集，预测错误的样本占总样本的比率。
e) 精度和召回率的调和平均数(F1值)：衡量二分类模型精度的一种指标，兼顾了分类模型的精度和召回率。
f) 两个概率分布间的差异的非对称性度量(KL散度)：它比较了真实分布和理论(拟合)分布之间的差异。
g) 受试者工作特性曲线(ROC曲线)：由不同设定条件下的真正率和假正率值画出的响应曲线，是反映敏感性和特异性连续变量的综合指标。
h) 精度召回率曲线(PRC曲线)：一种同时显示不同阈值下深度学习算法精度和召回率的图形化方法。一般 x 轴表示召回率，y 轴表示精度。
i) 累积响应曲线(CRC曲线)：也称为增益曲线或增益图，是显示跨多个阈值的总数据中真阳性率和阳性预测百分比的图形方法。

附录A给出了深度学习算法针对不同任务选取的基础性能指标示例。

4.3 效率

效率指深度学习算法在达到给定性能目标时所消耗的资源与时间的多少。效率的评估指标应包括但不限于以下。

a) 时间特性：深度学习算法执行其功能时，响应时间、处理时间及吞吐率满足需求的程度。可使用平均响应时间、平均周转时间、平均吞吐量等指标来表示。
   1) 平均响应时间：响应一个用户任务的平均时间，计算方法见公式(1)。

$$\overline{T}=\sum_{i=1}^{n}(T_i)/n \qquad \cdots\cdots(1)$$

式中：

$\overline{T}$ ——平均响应时间；

T_i——第 i 次测量时算法的响应时间；

n ——测得的响应次数。

2） 平均周转时间：完成一个作业的平均时间，计算方法见公式(2)。

$$\widetilde{T}=\sum_{i=1}^{n}(\mathrm{TE}_i-\mathrm{TS}_i)/n \quad \cdots\cdots(2)$$

式中：

$\widetilde{T}$ ——平均周转时间；

TS_i ——作业进程 i 的开始时刻；

TE_i ——作业进程 i 的结束时刻；

n ——完整测得作业进程时间的次数。

3） 平均吞吐量：单位时间内完成作业的平均数量，计算方法见公式(3)。

$$\mathrm{tp}=\sum_{i=1}^{n}(W_i/\mathrm{Tt}_i)/n \quad \cdots\cdots(3)$$

式中：

tp ——平均吞吐量；

W_i ——第 i 次观察时间内完成的作业数量；

Tt_i ——第 i 次观察的时间周期；

n ——观察的次数。

b） 资源利用性：深度学习算法执行其功能时，使用资源数量和类型满足需求的程度。可使用处理器平均占用率、内存平均占用率等指标来表示。

1） 处理器平均占用率：执行一组给定任务，处理器所需要的时间与运行时间的平均比率，计算方法见公式(4)。

$$\mathrm{OC}=\sum_{i=1}^{n}(\mathrm{Tw}_i/\mathrm{TW}_i)/n \quad \cdots\cdots(4)$$

式中：

OC ——处理器平均占用率；

Tw_i ——第 i 次观察中处理器执行一组给定任务所用的时间；

TW_i ——第 i 次观察中执行整体任务的运行时间；

n ——观察的次数。

2） 内存平均占用率：执行一组给定的任务所需要的内存与可用内存的平均比率，计算方法见公式(5)。

$$\mathrm{OD}=\sum_{i=1}^{n}(R_i/\mathrm{RW}_i)/n \quad \cdots\cdots(5)$$

式中：

OD ——内存平均占用率；

R_i ——第 i 次样本处理中执行一组给定任务所占用的实际内存大小；

RW_i ——第 i 次样本处理期间可用于执行任务的内存大小；

n ——处理的样本数。

4.4 正确性

正确性指深度学习算法代码、功能等方面开发设计的正确性。正确性的评估指标应包括但不限于：

a） 功能完备性：深度学习算法实现的功能达到所有指定任务和用户目标的程度。功能完备性可使用功能覆盖率来表示，计算方法见公式(6)。

$$F=1-\mathrm{Fm}/\mathrm{FW} \quad \cdots\cdots(6)$$

式中：

F ——功能覆盖率；

Fm——缺少的功能数量；

FW——指定的功能数量。

b) 功能正确性：深度学习算法提供具有所需精度的正确结果的程度，计算方法见公式(7)。

$$\mathrm{FR}=1-\mathrm{FE}/\mathrm{FW} \quad \cdots\cdots(7)$$

式中：

FR ——功能正确性；

FE ——功能不正确的数量；

FW——考虑的功能数量。

4.5 兼容性

兼容性指在相同软硬件环境下，深度学习算法能够与其他产品、系统或组件交换信息和/或执行其所需的功能的程度。兼容性的评估指标应包括但不限于：

a) 共存性：与其他产品共享通用环境和资源的条件下，深度学习算法能够有效执行其所需的功能并且不会对其他产品造成负面影响的程度，计算方法见公式(8)。

$$\mathrm{CO}=\mathrm{Cm}/\mathrm{CN} \quad \cdots\cdots(8)$$

式中：

CO ——共存性；

Cm ——与该深度学习算法可共存的其他规定的产品数量；

CN ——在运行环境中，该深度学习算法需要与其他产品共存的数量。

b) 硬件兼容性：深度学习算法在特定硬件上运行时，深度学习算法能够有效执行其所需的功能并且不会对其他产品造成负面影响的程度，计算方法见公式(9)。

$$C=\mathrm{Cdevice}/\mathrm{CCN} \quad \cdots\cdots(9)$$

式中：

C ——硬件兼容性；

Cdevice ——与该深度学习算法可兼容的计算处理器的数量；

CCN ——在运行环境中该深度学习算法需要兼容的计算处理器的数量。

4.6 可解释性

可解释性指深度学习算法对于结果的解释和理解能力。可解释性的评估指标应包括但不限于：

a) 解释一致性：深度学习算法决策结果与通过可解释性方法的输出结果具有一致性，可使用输出结果一致性进行评估。输出结果一致性是指通过计算输出结果的异众比率来表明数据的一致性，计算方法见公式(10)。

$$v_{\mathrm{r}}=\frac{\sum f_i-f_n}{\sum f_i} \quad \cdots\cdots(10)$$

式中：

v_{r} ——异众比率；

$\sum f_i$ ——变量值的总频数；

f_n ——众数组的频数；

n ——数组的数量。

b) 解释有效性：深度学习算法提供的解释能准确反映其决策逻辑，可使用判定系数进行评估。

判定系数又称 R^2 系数，是指反映因变量的全部扰动能通过回归关系被自变量解释的比例，计算方法见公式(11)。

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \quad \cdots\cdots(11)$$

式中：

R^2——判定系数；

注：R^2 值越接近于 1，回归拟合效果越好，一般认为超过 80%的模型拟合度比较高。

y_i——真实的观测值；

$\overline{y}$——真实观测值的平均值；

$\hat{y}_i$——预测值。

c) 解释因果性：生成的解释与待解释深度学习算法预测之间具有因果关系。解释因果性可使用特征贡献分数进行评估。特征贡献分数是指用来解释的重要性靠前的 k 个样本特征分数和与全部特征分数和的比值，计算方法见公式(12)。

$$f_{\text{score}}=\frac{\sum f_{\text{topk}}}{\sum f_i} \quad \cdots\cdots(12)$$

式中：

f_{score}——特征贡献分数；

$\sum f_{\text{topk}}$——用来解释的重要性靠前的前 k 个特征分数之和；

$\sum f_i$——全部特征分数和。

d) 解释充分性：生成的解释能够覆盖深度学习算法的整体功能，可使用离散系数进行评估。离散系数是指数据的标准差与平均数的比值，用来比较不同类别数据的离散程度，计算方法见公式(13)。

$$v_s=\frac{\sigma}{\overline{X}} \quad \cdots\cdots(13)$$

式中：

v_s——表示离散系数；

σ——数据的标准差；

$\overline{X}$——数据的平均数。

4.7 鲁棒性

鲁棒性指面对非对抗增广的样本时，深度学习算法保持与实验环境中测试性能相当的能力。鲁棒性的评估指标应包括但不限于以下。

a) 性能波动率：模型在原始测试数据集和经过非对抗扰动处理后的新测试数据集之间的性能差异，计算方法见公式(14)。性能波动率越小，模型面对扰动时稳定性越高。

$$\text{PFD}=\frac{|P_{\text{original}}-P_{\text{perturbed}}|}{|P_{\text{original}}|} \quad \cdots\cdots(14)$$

式中：

PFD——模型的性能波动率；

P_{original}——模型在原始测试数据集上的性能指标；

$P_{\text{perturbed}}$——模型在经过非对抗扰动后的新测试数据集上的性能指标。

对于多种扰动，模型的鲁棒性可通过公式(15)量化。

$$R=\sum_{i=1}^{N}w_i\times\text{PFD}_i \quad \cdots\cdots(15)$$

式中：

R ——模型鲁棒性；

w_i ——第 i 种扰动的权重；

N ——扰动方法总数；

PFD_i ——模型在第 i 种扰动下的性能波动率。

b) 扰动稳定性：模型在经历非对抗扰动后出现性能退化的样本与其对应的原始样本之间的最小距离，计算方法见公式(16)。扰动稳定性越大，模型在面对扰动时抵御能力越强。

$$\mathrm{PSD}_{\Phi}=\min_{x\in X}[\mathrm{dist}_{\phi}(x)] \quad \cdots\cdots(16)$$

式中：

PSD_{Φ} ——模型的扰动稳定性；

X ——数据集；

x ——样本实例；

dist_{ϕ} ——在 ϕ 类型的扰动下样本与扰动样本的距离函数，计算方式见公式(17)。

$$\mathrm{dist}_{\phi}(x)=\begin{cases}\|x-x'\|_p, if\ f(x')\neq y\\ \infty\end{cases} \quad \cdots\cdots(17)$$

式中：

$f(x')$——通过 ϕ 类型扰动生成的样本 x' 的判定结果；

y ——真实标签。

对于多种扰动，模型的鲁棒性可通过公式(18)量化。

$$R=\min_{x\in X}[\min_{\phi\in\Phi}\mathrm{dist}_{\phi}(x)] \quad \cdots\cdots(18)$$

式中：

R ——模型鲁棒性；

Φ ——扰动集合。

4.8 安全性

安全性指深度学习算法对对抗样本的防范能力。安全性的评估指标应包括但不限于：

a) 攻击成功率：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，计算方法见公式(19)。攻击成功率越小，模型在对攻击的抵抗能力越强。

$$\mathrm{ASR}=\frac{N_{\mathrm{adv}}}{N_{\mathrm{all}}} \quad \cdots\cdots(19)$$

式中：

ASR ——攻击成功率；

N_{all} ——样本总数；

N_{adv} ——预测失败的样本数。

b) 模型窃取程度：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，计算方法见公式(20)。模型窃取程度越大，代理模型与原始模型越相似。

$$\mathrm{MSD}=\frac{\sum_{x\in D}\delta(x)}{|D|} \quad \cdots\cdots(20)$$

式中：

MSD ——模型窃取程度；

$|D|$ ——数据集的样本总数；

$\delta(x)$ ——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。

注：当代理模型的预测结果与原始模型的预测结果的差值在设定区间内时，均为预测相同，赋值为 1。

c) 平均攻击查询次数：生成对抗样本所需的平均模型查询次数。平均攻击查询次数越少，模型越容易受到攻击。

d) 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，指标包括但不限于：均方误差、余弦相似度、$\mathcal{L}$-P 范数等。攻击隐蔽性越高，对抗攻击成功率越高更有效地欺骗模型。

4.9 公平性

公平性指深度学习算法面向不同群体，保持相同输出质量的能力。公平性的评估指标应包括但不限于：

a) 敏感属性独立程度：不同敏感属性群体进行特定预测的比例之间的最大差异，计算方法见公式(21)。敏感属性独立程度越低，模型对不同群体的预测更加一致，公平性越高。

$$\mathrm{SAID}=\max_{a,b\in A,l\in L}\left|\frac{\mathrm{count}(\hat{Y}=l\mid A=a)}{\mathrm{count}(A=a)}-\frac{\mathrm{count}(\hat{Y}=l\mid A=b)}{\mathrm{count}(A=b)}\right| \quad \cdots\cdots(21)$$

式中：

SAID ——敏感属性独立程度；

A ——敏感属性集合；

L ——标签集合；

$\hat{Y}$ ——模型的预测结果；

$\mathrm{count}(x)$ ——计数函数。

b) 模型决策分离程度：真实类别为特定值时，模型在不同敏感属性群体之间做出错误预测的概率的差异，计算方法见公式(22)。

$$\mathrm{MDS}=\max_{a,b\in A,l\in L}\left|\frac{\sum_{l'\neq l}\mathrm{count}(\hat{Y}=l'\mid A=a,Y=l)}{\mathrm{count}(A=a)}-\frac{\sum_{l'\neq l}\mathrm{count}(\hat{Y}=l'\mid A=b,Y=l)}{\mathrm{count}(A=b)}\right| \quad \cdots\cdots(22)$$

式中：

MDS ——模型决策分离程度；

Y ——真实值，$l'\neq l$。

c) 模型决策充分程度：模型预测标签为特定值时，模型在不同敏感属性群体之间正确预测该标签的概率的差异，计算方法见公式(23)。

$$\mathrm{MDSF}=\max_{a,b\in A,l\in L}\left|\frac{\mathrm{count}(Y=l\mid A=a,\hat{Y}=l)}{\mathrm{count}(A=a)}-\frac{\mathrm{count}(Y=l\mid A=b,\hat{Y}=l)}{\mathrm{count}(A=b)}\right| \quad \cdots\cdots(23)$$

式中：

MDSF——模型决策充分程度。

5 评估等级

深度学习算法的评估结果分为优越级、进阶级、条件级、受限级四个等级。针对每一个算法失效，应基于确定的理由来预估潜在危险的严重性等级。深度学习算法失效的危险严重性等级如下。

——优越级：在该等级下深度学习算法的失效通常是一些小规模的问题，不会对整个系统或应用的性能造成严重威胁。例如，深度学习算法在某些特定情况下的性能略微下降，但不会导致显著问题，整体性能仍在可接受范围内。这类问题通常可以通过微小的调整、超参数优化或数据清洗来解决。

——进阶级：在该等级下深度学习算法的失效会对系统或应用的性能造成一定程度的负面影响，但

不至于导致严重问题。例如，深度学习算法的性能在某些关键任务中低于期望，但在其他任务上表现不错。解决这类问题可能需要更深入的研究、数据增强、迁移学习、模型选择等方法。

——条件级：在该等级下深度学习算法的失效会对整个系统或应用的性能产生重大影响，可能导致项目失败或严重损害用户体验。例如，深度学习算法的性能不稳定，导致无法在实际应用中可靠地使用。解决这类问题可能需要全面的重新设计、数据收集、模型选择等措施。

——受限级：在该等级下深度学习算法的失效可能对人们的生命、财产或安全构成直接威胁，可能导致法律问题或损害声誉。例如，自动驾驶汽车系统的算法失效，导致事故发生。解决这类问题可能需要紧急行动、彻底审查、法律干预等措施，需要综合考虑伦理、法规和道德问题。

根据算法失效的危险严重性等级，建立深度学习算法的等级目标，见表1。其中，等级目标从高到低依次分为优越级、进阶级、条件级、受限级四个级别。

表1 深度学习算法的等级目标

等级目标	等级目标说明
优越级	外部环境发生扰动或面对不友好的输入，不依赖利益相关方的管理和配置，能采取有效措施，按照预期完成工作，不影响算法结果
进阶级	外部环境发生扰动或面对不友好的输入，通过利益相关方的配置及管理，待评估算法能按照预期完成工作，不影响算法结果
条件级	在友好的外部环境及输入下，待评估算法可以按照预期完成工作； 外部环境发生扰动或面对不友好的输入，通过利益相关方的配置与管理，待评估算法能按预期完成工作，不对算法结果造成重大影响
受限级	在友好的外部环境及输入下，待评估算法能按照预期完成工作； 当外部环境发生扰动或面对不友好的输入，待评估算法不能按照预期完成工作，可能对算法结果造成重大影响

深度学习算法评估应面向不同等级目标，基于用户需求或过往经验，设定不同指标要求。深度学习算法评估时，可基于评估指标项得分所在区间，判定该指标项所处等级，详见附录A。

6 评估流程

6.1 概述

深度学习算法的评估流程分为黑盒评估和白盒评估，包括评估准备、评估执行、分析评估、评估结论等四大步骤。黑盒评估流程如图2所示，白盒评估流程如图3所示。

其中：

——评估准备，包括输入测试数据集及算法参数、测试数据集质量审查、深度学习框架/模型的漏洞检查、选择质量特性、选择评估指标、构建评估模型等子步骤；

——评估执行，包括推理阶段的执行、获取训练/推理阶段的数据、计算测试指标等子步骤；

——分析评估，包括算法质量评估（多次评估）、分值计算、等级评估等子步骤；

——评估结论，包括编制评估报告等。

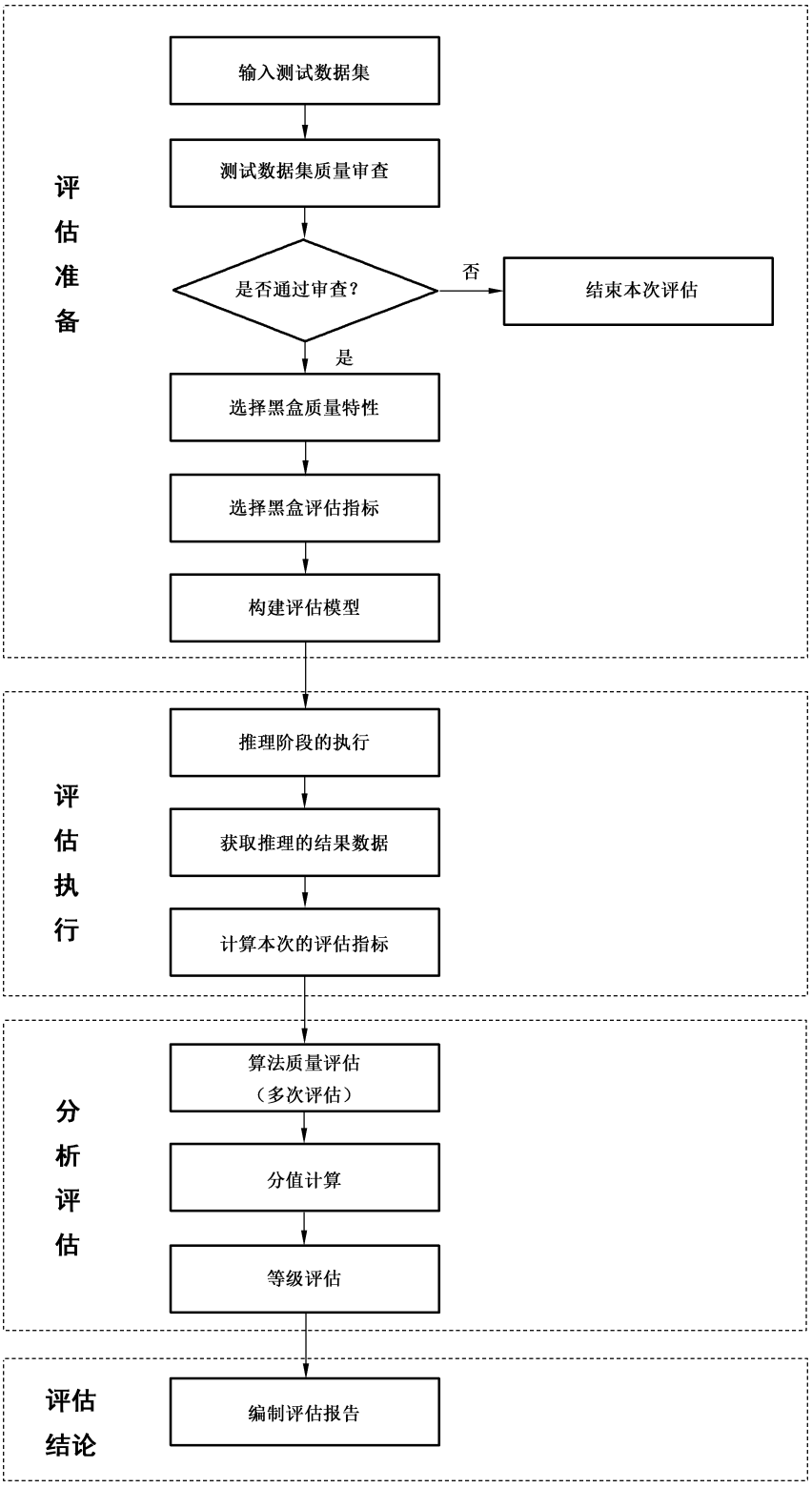

图 2　深度学习算法的黑盒评估流程

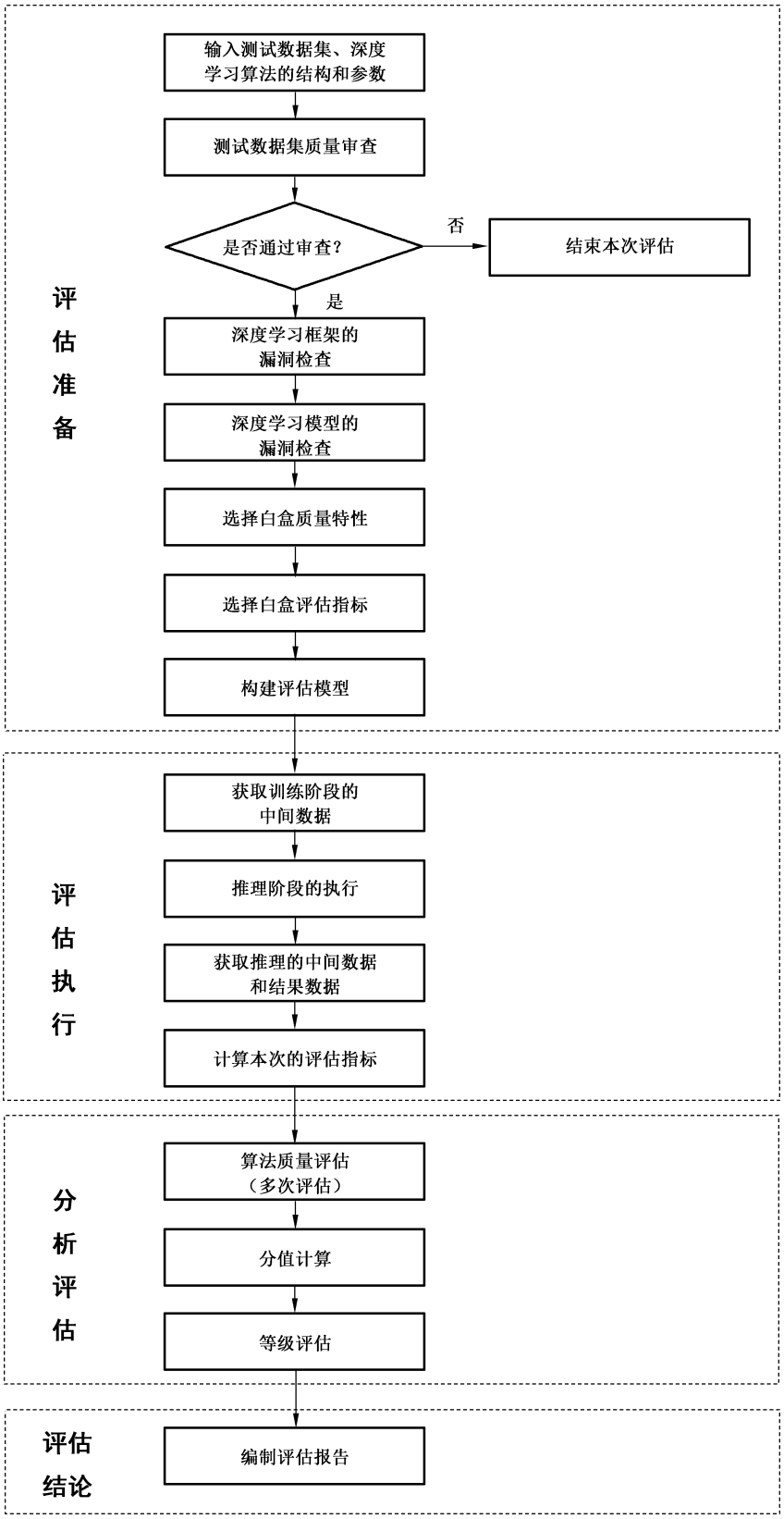

图 3　深度学习算法的白盒评估流程

6.2 评估准备

6.2.1 输入测试数据集及算法参数

对于深度学习算法的黑盒评估，其输入包括一个或多个测试数据集。

对于深度学习算法的白盒评估，其输入包括但不限于一个或多个测试数据集、深度学习算法的测试参数或配置、测试的约束条件等。

6.2.2 测试数据集质量审查

被测方应按质量要求提供测试数据集，质量审查应包括：

a) 对数据的完整性进行审查，评估数据是否存在缺失值、异常值或未标记的数据点；

b) 对数据的准确性进行审查，与数据采集、标注等环节的实际情况进行比对验证，或通过领域专家的评估进行验证，评估数据的时间戳、标签或其他标识是否有误；

c) 对数据的一致性进行审查，评估数据是否具备相同的格式；

d) 对数据的重复性进行审查，评估数据是否存在重复记录或者冗余特征；

e) 对数据的偏差进行审查，评估数据是否存在数据分布和标签分布不均的情况；

f) 对数据的可用性进行审查，评估数据是否具备参考文档以及数据是否受法律或者隐私限制。

当测试数据集质量通过审查，则进入“选择质量特性”步骤；否则，结束本次评估。

6.2.3 深度学习框架的漏洞检查

深度学习框架的漏洞检查仅针对深度学习算法的白盒评估，评估应包括如下。

a) 机器学习框架的模型构建及训练组件提供抑制训练流程模板，通过去除模型中不重要的参数，降低通过模型逆向攻击获取原始样本的风险，包括：

   1) 抑制隐私基类，用于执行模型训练过程；

   2) 抑制隐私监视器基类，用于执行模型参数置零操作。

b) 机器学习框架的模型推理及部署组件具备面向模型攻击模拟的模型安全测试组件，支持：

   1) 攻击模拟工具，如：成员推理攻击等；

   2) 提供模型信息安全评测指标，如：隐私泄露程度等。

c) 机器学习框架的模型推理及部署组件提供模型加密解密组件，支持：

   1) 部署前加密，利用加密算法对参数文件或推理模型加密；

   2) 训练时加密，部署运行时自动解密。

6.2.4 深度学习模型的漏洞检查

深度学习模型的漏洞检查仅针对深度学习算法的白盒评估，评估应包括：

a) 深度学习模型相关训练数据集和测试数据集满足 GB/T 35273—2020 中对个人信息控制者和 GB/T 40660—2021 中对生物特征识别信息控制者的所有要求；

b) 在深度学习模型的全生命周期中，确保模型和相关数据的完整性，如采用身份验证或权限控制等方式防御模型开发阶段的非授权访问和窃取；

c) 在深度学习模型的全生命周期中，确保模型和相关数据的可用性，如采用差分隐私、联邦学习等隐私计算技术处理后的模型和数据可被正常使用，加密信息不应被删除或损坏；

d) 在深度学习模型的全生命周期中，确保模型和相关数据的保密性，如在模型训练或推理过程中通过安全加密、可信执行环境等技术避免后门攻击、模型反演等恶意攻击得到模型参数和数据；

e) 根据深度学习模型对用户的干预程度进行分类分级管理；

f) 保护用户的操作记录等数据，防止数据泄密或作他用。

6.2.5 选择质量特性和评估指标

选择评估指标包括质量特性选择、评估指标选择两个部分。

——质量特性选择包括基础性能、效率、正确性、兼容性、可解释性、鲁棒性、安全性、公平性 8 个选项。例如，对于黑盒评估流程，可选择基础性能、效率、正确性、兼容性等；对于白盒评估流程，8 个质量特性均适用，可根据评估需求进行选择。

——评估指标选择是指在每个质量特性下，选择若干个评估指标。

不同任务类型（图像分类、目标检测、语音识别、文本情感分析、文本命名实体识别等）的深度学习算法选取的评估指标要求不同，因此在面向算法的评估过程中应确定与之对应的评估指标要求。

注：不同应用场景的深度学习算法评估指标可能不同，同一指标的基准分值也可能不尽相同。例如，图像识别算法在公安办案场景多选择准确率作为评估指标，在门禁场景多选择召回率作为评估指标；图像识别算法在医疗场景的准确率等级阈值设置一般高于手写体识别场景。附录 A 给出了深度学习算法评估不同指标的推荐阈值。

6.2.6 构建评估模型

图 4 给出了由指标参数体系和指标权重体系构成的深度学习算法评估模型。其中，指标参数体系包含的各质量特性项（1 到 M）是所关联的评估指标项的父辈评估参数（第 i 个质量特性项的权重为 $X_i\%$，含 p_i 个评估指标项），评估指标项是所关联的一级评估指标子项的父辈评估参数（第 i 个质量特性项的第 j 个评估指标项的权重为 $Y_{ij}\%$，含 q_{ij} 个评估指标子项），如果有更多层级，对各层级的理解依次类推。图 4 中用虚线表示可能经多次分解而形成的评估指标子项和指标权重。$X_i\%$、$Y_{ij}\%$、$Z_{ijk}\%$ 表示指标参数体系中各指标参数在本级的指标权重值，$\sum X_i\%$、$\sum Y_{ij}\%$、$\sum Z_{ijk}\%$ 均为 1。

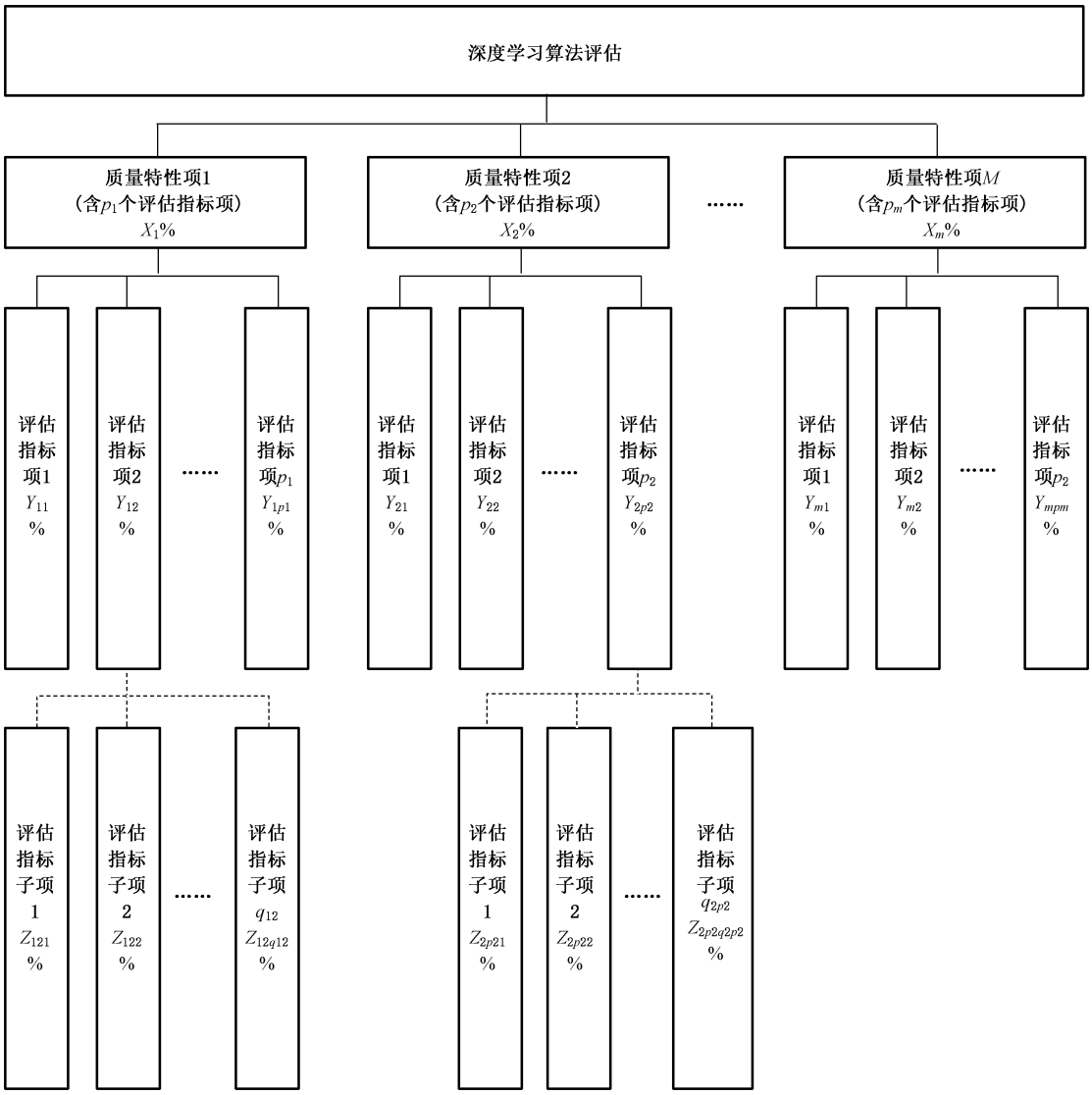

图 4　深度学习算法的评估模型

对于评估指标的权重，可使用权重计算的评估模型得到。附录 B 给出了权重计算方法的示例。

6.3　评估执行

6.3.1　推理阶段的执行

运行一次评估任务包括算法测试环境部署、被测算法加载、被测算法测试执行等三个部分。

——测试环境部署：包括硬件环境（如服务器）搭建、软件环境（如操作系统、数据库）搭建、兼容性测试、网络环境部署等。

——被测算法加载：读取或解析被测算法的文件（如程序包或数据包）。

——被测算法测试执行：使用选定的被测数据集运行算法文件。

6.3.2　评估指标的获取和计算

深度学习算法评估测试数据集与相关质量特性及评估指标的关联性如图 5 所示。

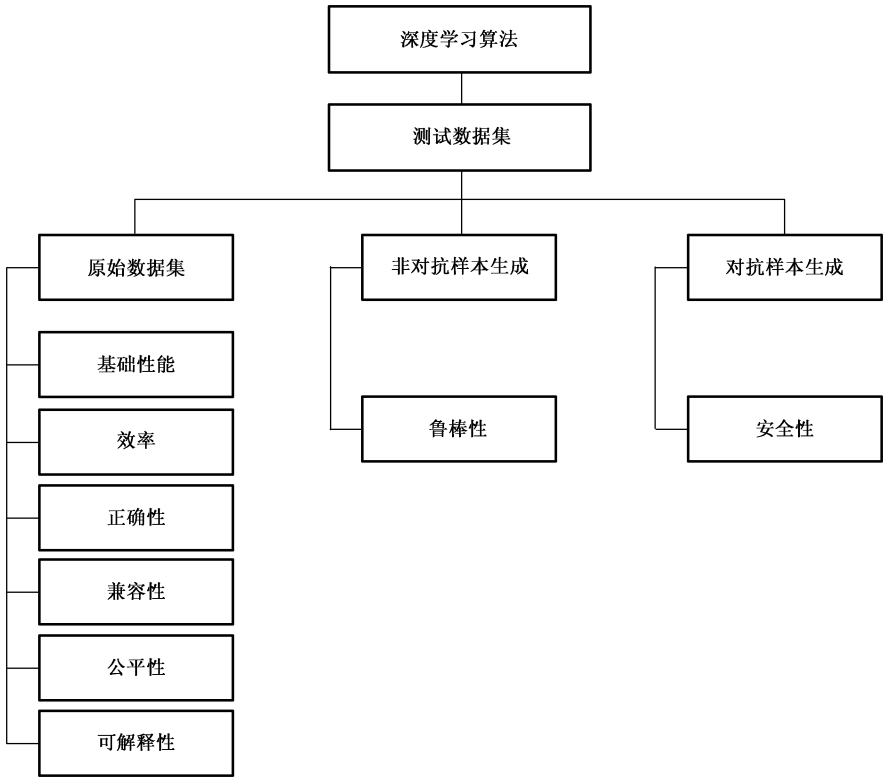

图 5　深度学习算法评估测试数据集与相关评估指标的关联性

其中：

——测试数据集的输出，包含样本及其标签，用于数据集质量评估、深度学习算法结果预测、非对抗样本生成及对抗样本生成；

——原始数据集的输出，包含样本及预测值，用于基础性能、效率、正确性、兼容性、公平性、可解释性等质量特性及其评估指标的计算；

——非对抗样本生成的输出，包含增广样本及其标签，用于鲁棒性等质量特性及其评估指标的计算；

——对抗样本生成的输出，包含对抗样本，用于安全性等质量特性及其评估指标的计算。

6.4　分析评估

6.4.1　算法质量评估

算法质量评估是指同一被测深度学习算法使用一个或多个测试数据集进行测试，每个测试数据集执行一次算法测试过程，得到相应的测试数据，基于测试数据进行算法质量综合评价。

6.4.2　分值计算

深度学习算法评估得分计算方式见公式(24)。

$$T_{\text{total}} = \sum_{i=1}^{M} [X_i \times \sum_{j=1}^{N} (Y_{ij} \times S_{ij})] \times 100\% \quad \cdots\cdots (24)$$

式中：

T_{total}——深度学习算法最终评估得分；

M　——指标参数体系包含的质量特性项的总项数；

N　——第 i 个质量特性项的评估指标项的总项数；

X_i ——第 i 个质量特性项的指标权重值；

Y_{ij} ——第 i 个质量特性项的第 j 个评估指标项的指标权重值；

S_{ij} ——第 i 个质量特性项的第 j 个评估指标项的得分。

第 i 个质量特性项的第 j 个评估指标项的得分的计算方法见公式(25)。

$$S_{ij}=\sum_{k=1}^{P}(Z_{ijk}\times S_{ijk})\times 100\% \quad \cdots\cdots(25)$$

式中：

S_{ij} ——第 i 个质量特性项的第 j 个评估指标项得分；

P ——第 i 个质量特性项的第 j 个评估指标项包含的一级评估指标子项的总项数；

Z_{ijk} ——第 i 个质量特性项的第 j 个评估指标项的第 k 个一级评估指标子项的指标权重值；

S_{ijk} ——第 i 个质量特性项的第 j 个评估指标项的第 k 个一级评估指标子项的得分。

对于有二级、三级、四级等评估指标子项体系的得分计算方式，依次往下类推。

6.4.3 等级评估

在对深度学习算法的等级判定中，可使用评分表等级判定和边界评估模型等级判定等方法。

评分表等级判定，即在评分表中同时考察算法评估的总基准分值和每个质量特性的基准分值。算法达到对应级别应同时大于或等于算法评估的总基准分值和各质量特性的基准分值的要求，如表 2 所示。

表 2 深度学习算法的等级判定

等级	算法评估总基准分值	质量特性的基准分值			
		评估指标项 1	评估指标项 2	……	评估指标项 n
优越级	T_1%	T_{1_1}%	T_{1_2}%	……	T_{1_n}%
进阶级	T_2%	T_{2_1}%	T_{2_2}%	……	T_{2_n}%
条件级	T_3%	T_{3_1}%	T_{3_2}%	……	T_{3_n}%
受限级	T_4%	T_{4_1}%	T_{4_2}%	……	T_{4_n}%
注：T_a%表示第 a 级算法评估的总基准分值；T_{a_b}%表示第 a 级对第 b 项评估指标的基准分值。					

边界评估模型等级判定，即根据深度学习算法在原始数据集和生成数据集下的算法完成情况是否达到四个等级阈值的分布情况自适应地绘制出算法的评估等级边界，生成边界评估判定图，判定算法的总分值在哪个等级区间内，如图 6 所示。

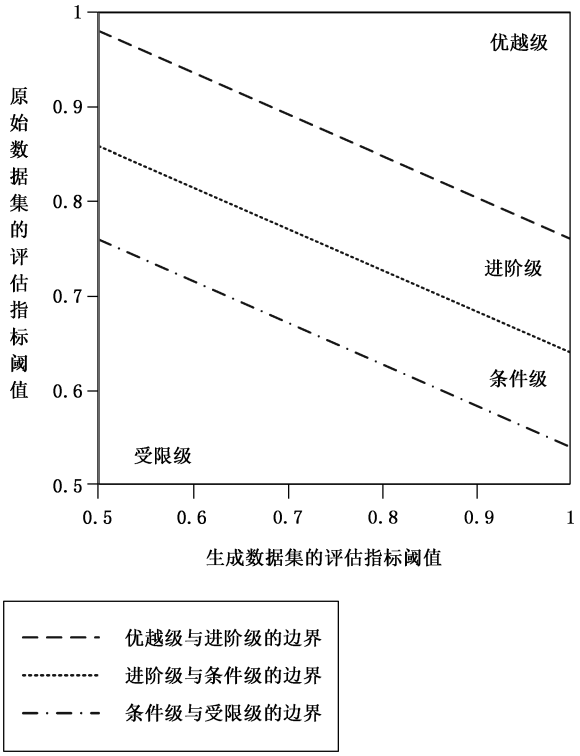

图 6　深度学习算法边界评估模型示例

其中，生成数据集的评估指标包括鲁棒性和安全性，原始数据集的评估指标包括基础性能、效率、正确性、兼容性、公平性、可解释性。生成数据集和原始数据集的评估指标阈值，是由选定的质量特性的四个阈值加权求和得到的，评估指标的权重可使用权重评估模型，见 6.2.6。

附录 C 给出了深度学习算法评估实施案例。

6.5　评估结论

根据评分表，最终判定被评估的深度学习算法的等级，并编制评估报告。报告应包括被测深度学习算法的说明、测试数据集的说明、评估总体结论、各质量特性详细评估结果等。评估报告应满足评估机构及评估管理机构的编制要求。

附 录 A
（资料性）
深度学习算法评估指标选取和阈值设定

A.1 基于历史数据的统计分析

通过历史数据的收集和分析，可以了解各个指标的分布情况和变化趋势，进而确定各个指标的阈值，与阈值偏离较大的情况将被视为异常或不符合预期。

例如，网站访问量的阈值可以基于过去一年访问量数据的计算平均值和标准差，根据正态分布的性质来确定。超过阈值的访问量将被视为异常情况。

A.2 基于专家意见的主观判断

某些情况下，存在数据缺乏或数据分布不规律等问题，难以使用基于历史的统计方法来确定阈值。此时，可依靠领域内的专家经验和知识，通过讨论和协商的方式来确定阈值。

例如，在医学领域中，确定某个指标的正常范围时，可根据医生的专业知识和经验，以及患者的年龄、性别、病史等因素，判断医学指标数值是否正常。

A.3 基于业务需求的目标设定

某些情况下，需要通过设定阈值以实现提高生产效率、降低成本或增加收益等特定目标。此时，可根据目标要求来设定阈值。

例如，在生产过程中，可根据自身业务需求设定合适的阈值来整体约束产品的质量。如果产品的某个指标与阈值有出入，可以及时采取措施，调整生产流程，以确保产品质量的稳定性和一致性。

表 A.1 给出了不同类型深度学习算法的评估指标在四个评估等级的阈值设定示例。

表 A.1　评估指标的阈值设定示例

算法类型		图像分类				语音识别				文本识别			
评估等级		受限级	条件级	进阶级	优越级	受限级	条件级	进阶级	优越级	受限级	条件级	进阶级	优越级
基础性能	准确率	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	精度	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	召回率	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	F1 值	＜0.8	≥0.8	≥0.9	≥0.99	＜0.8	≥0.8	≥0.9	≥0.99	＜0.8	≥0.8	≥0.9	≥0.99
	AUC	＜0.8	≥0.8	≥0.9	≥0.99	＜0.8	≥0.8	≥0.9	≥0.99	＜0.8	≥0.8	≥0.9	≥0.99
	……												
效率	平均响应时间	＞7.5 s	≤7.5 s	≤3 s	≤0.75 s	＞7.5 s	≤7.5 s	≤3 s	≤0.75 s	＞7.5 s	≤7.5 s	≤3 s	≤0.75 s
	平均周转时间	＞50 ms	≤50 ms	≤20 ms	≤5 ms	＞50 ms	≤50 ms	≤20 ms	≤5 ms	＞50 ms	≤50 ms	≤20 ms	≤5 ms
	平均吞吐量	＜20	≥20	≥50	≥200	＜20	≥20	≥50	≥200	＜20	≥20	≥50	≥200
	处理器平均占用率	＞5％	≤5％	≤3％	≤1％	＞5％	≤5％	≤3％	≤1％	＞5％	≤5％	≤3％	≤1％
	内存平均占用率	＞5％	≤5％	≤3％	≤1％	＞5％	≤5％	≤3％	≤1％	＞5％	≤5％	≤3％	≤1％
正确性	功能完备性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	功能正确性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
兼容性	共存性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	硬件兼容性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
可解释性	解释一致性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	解释有效性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	解释因果性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
	解释充分性	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％	＜80％	≥80％	≥90％	≥99％
鲁棒性	性能波动率	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％
	扰动稳定性	＞0.95	≤0.95	≤0.85	≤0.75	＞0.95	≤0.95	≤0.85	≤0.75	＞0.95	≤0.95	≤0.85	≤0.75

表 A.1　评估指标的阈值设定示例（续）

算法类型		图像分类				语音识别				文本识别			
评估等级		受限级	条件级	进阶级	优越级	受限级	条件级	进阶级	优越级	受限级	条件级	进阶级	优越级
安全性	攻击成功率	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％
	模型窃取程度	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％
	平均攻击查询次数	＜2	≥2	≥5	≥10	＜2	≥2	≥5	≥10	＜2	≥2	≥5	≥10
	攻击隐蔽性	＞0.95	≤0.95	≤0.85	≤0.75	＞0.95	≤0.95	≤0.85	≤0.75	＞0.95	≤0.95	≤0.85	≤0.75
公平性	敏感属性独立程度	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％
	模型决策分离程度	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％
	模型决策充分程度	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％	＞20％	≤20％	≤10％	≤5％
不同算法类型的基础性能质量特性将由不同评估指标组成，具体评估时应面向待评估深度学习算法类型选取对应的基础性能指标进行评估。如，图像分类选取准确率、召回率、F1 分数等；语音识别选取字错误率、句错误率、字匹配率等；文本识别选取字符识别准确率、字符识别召回率等 **注 1**：平均响应时间的阈值设定是基于包含 100 个样本的用户任务。 **注 2**：效率的阈值设定的硬件配置。CPU：主频 3.00 GHz，RAM：DDR432.0 GB，GPU：显存 24 GB、82.6 TFLOPS（FP16）。												

表 A.2 给出了根据深度学习算法不同任务类型，选择不同的基础性能评估指标的示例。

表 A.2 基础性能评估指标的选取示例

测试数据集类型	深度学习算法的任务类型	基础性能的评估指标
图像	分类(二分类)	F1 分数、准确率、精确率、召回率、G-mean、特异度、误诊率、错误率等
	分类(多分类)	加权平均精确率、加权平均召回率、加权平均 F1 分数、宏观平均精确率、宏观平均召回率、宏观平均 F1 分数、微观平均精确率、微观平均召回率、微观平均 F1 分数、准确率、召回率、F1 分数等
	目标检测(单类/多类)	IoU、mAP、AP 明细、置信度等
	目标跟踪(单类/多类)	IoU、MOTA、MOTP、IDP1、IDP、IDR、主要跟踪目标数量、主要丢失目标数量、部分跟踪目标数量、MT、ML、PT、IDSW、碎片总数、mAP、AP 等
	语义分割	像素准确率、类别平均像素准确率、类别像素准确率、IoU、MIoU 等
	姿态估计	OKS、mAP、AP、置信度
视频	目标检测(单类/多类)	IoU、mAP、AP 明细、置信度
	目标跟踪(单类/多类)	IoU、MOTA、MOTP、IDP1、IDP、IDR、主要跟踪目标数量、主要丢失目标数量、部分跟踪目标数量、MT、ML、PT、IDSW、碎片总数、mAP、AP
文本	情感分析	F1 分数、准确率、精确率、召回率、G-mean、特异度、误诊率、错误率
	命名实体识别	加权平均精确率、加权平均召回率、加权平均 F1 分数、宏观平均精确率、宏观平均召回率、宏观平均 F1 分数、微观平均精确率、微观平均召回率、微观平均 F1 分数、准确率召回率、F1 分数
语音	语音识别	平均词错误率、平均词信息丢失率、平均匹配错误率、平均字符错误率、平均词信息保留
结构化数据	分类	F1 分数、准确率、精确率、召回率、G-mean、特异度、误诊率、错误率
	回归	平均绝对误差、平均均方误差、平均绝对百分比误差、决定系数
	聚类	轮廓系数、DBI、方差比准则

附 录 B
（资料性）
深度学习算法评估指标权重计算方法

B.1 熵权法-TOPSIS 模型

熵权法和 TOPSIS 模型是用于综合评价的模型，其中熵权法的主要目的是对指标体系进行赋权，借鉴了信息熵思想，通过计算指标的信息熵，根据指标的相对变化程度对系统整体的影响来决定指标的权重，即根据各个指标标志值的差异程度来进行赋权，从而得出各个指标相应的权重，相对变化程度大的指标具有较大的权重。TOPSIS 模型是通过逼近理想解的程度来评估各个样本的优劣等级，在归一化的原始数据矩阵中，找到有限方案中的最优方案和最劣方案，然后分别计算评价对象和最优方案和最劣方案之间的距离，并以此作为依据来评价样本的优劣等级。熵权法-TOPSIS 模型则是借鉴了两者特点的模型。

假设目前是对深度学习算法进行评估分析，已选择共 n 个评估指标，使用 3 组不同的测试数据集进行多轮测试：

$$X=\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ & \vdots & \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

其中，x_{ij} 表示深度学习算法评估的第 i 个评估指标使用第 j 个测试数据集得到的测试结果。

熵权法-TOPTSIS 模型的建立主要分为以下几步。

a） 求评估指标在质量特性中的比值。计算第 j 个评估指标中，第 i 个质量特性的比重。

$$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}}$$

所得结果构建得到数据的比重矩阵：

$$P=\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ & \vdots & \\ p_{n1} & p_{n2} & p_{n3} \end{bmatrix}$$

b） 计算各项指标的熵值。计算对于第 j 个评估指标的熵值大小 e_j。

$$\mathrm{e}_j=-\frac{1}{\ln n}\sum_{i=1}^{n}p_{ij}\ln p_{ij}$$

c） 计算信息冗余度。计算对于第 j 个评估指标的信息冗余度大小 d_j。

$$d_j=1-\mathrm{e}_j$$

d） 定权重。根据信息冗余度确定各个指标的权重 w_j。

$$w_j=\frac{d_j}{\sum_{j=1}^{3}d_j}$$

e） 归一化。对原始数据进行归一化处理得到 $\widetilde{x_{ij}}$。

$$\widetilde{x_{ij}}=\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}x_{ij}^{2}}}$$

得到归一化矩阵：

$$X=\begin{bmatrix}\widetilde{x_{11}} & \widetilde{x_{12}} & \widetilde{x_{13}}\\ \widetilde{x_{21}} & \widetilde{x_{22}} & \widetilde{x_{23}}\\ & \vdots & \\ \widetilde{x_{n1}} & \widetilde{x_{n2}} & \widetilde{x_{n3}}\end{bmatrix}$$

f） 构造加权矩阵。根据归一化矩阵和指标权重计算得到加权矩阵 Z。

$$Z=\begin{bmatrix}\widetilde{x_{11}}w_1 & \widetilde{x_{12}}w_2 & \widetilde{x_{13}}w_3\\ \widetilde{x_{21}}w_1 & \widetilde{x_{22}}w_2 & \widetilde{x_{23}}w_3\\ & \vdots & \\ \widetilde{x_{n1}}w_1 & \widetilde{x_{n2}}w_2 & \widetilde{x_{n3}}w_3\end{bmatrix}$$

g） 寻找最优劣方案。根据加权矩阵计算结果，得到最优方案 z_j^+ 和最劣方案 z_j^-。

$$\begin{cases}z_j^+=\max(z_{1j},z_{2j},\ldots,z_{nj})\\ z_j^-=\min(z_{1j},z_{2j},\ldots,z_{nj})\end{cases}$$

h） 计算最优劣距离。计算各个样本与最优方案距离 D_i^+ 和最劣方案距离 D_i^-。

$$\begin{cases}D_i^+=\sqrt{\sum_j(z_{ij}-z_j^+)^2}\\ D_i^-=\sqrt{\sum_j(z_{ij}-z_j^-)^2}\end{cases}$$

i） 构造相对接近度。根据各个样本的最优方案距离和最劣方案距离计算相对接近度 C_i。

$$C_i=\frac{D_i^-}{D_i^++D_i^-}$$

j） 排序。根据各个样本的相对接近度进行排序，得到各个样本的优劣程度排名。

B.2 CRITIC 法

CRITIC 法是一种基于评价指标的对比强度和指标之间的冲突性来综合衡量指标的客观权重。考虑指标变异性大小的同时兼顾指标之间的相关性，并非数字越大就说明越重要，完全利用数据自身的客观属性进行科学评价。其中，对比强度是指同一个指标各个评价方案之间取值差距的大小，以标准差的形式来表现。标准差越大，说明波动越大，即各方案之间的取值差距越大，权重会越高。另外，指标之间的冲突性，用相关系数进行表示，若两个指标之间具有较强的正相关，说明冲突性越小，权重会越低。

假设目前是对深度学习算法进行评估分析，已选择共 n 个评估指标，使用 3 组不同的测试数据集进行多轮测试：

$$X=\begin{bmatrix}x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\\ & \vdots & \\ x_{n1} & x_{n2} & x_{n3}\end{bmatrix}$$

其中：

x_{ij}——深度学习算法评估的第 i 个评估指标使用第 j 个测试数据集得到的测试结果。

CRITIC 法的建立主要分为以下几步。

a） 无量纲化处理。可使用正向化处理，将非极大型指标进行极大型处理，消除不同量纲对评估结果的影响。

$$\widetilde{x_{ij}}=\frac{x_{ij\max}-x_{ij}}{x_{ij\max}-x_{ij\min}}$$

b） 指标变异性计算。可使用标准差方法界定指标的差异波动情况，指标权重正比于其标准差

大小。

$$\begin{cases} \overline{x_j} = \dfrac{1}{n}\sum_{i=1}^{n} \widetilde{x_{ij}} \\ S_j = \sqrt{\dfrac{\sum_{i=1}^{n} (\widetilde{x_{ij}} - \overline{x_j})^2}{n-1}} \end{cases}$$

其中：

S_j——经过极大值变换后的第 j 个指标的标准差。

c) 指标冲突性计算。可使用相关系数表示不同指标间的相关性，相关系数越高，该指标与其他指标的相关性越强，其重复程度越高。此时，减少对该指标分配的权重以降低相似指标对评估结果的重复影响。

$$R_j = \sum_{i=1}^{3} (1 - r_{ij})$$

其中：

R_j——第 j 个指标与其他指标之间的冲突性；

r_{ij}——是第 i 个指标和第 j 个指标之间的相关系数。

d) 信息量计算。可通过以下公式计算指标的信息量大小。

$$C_j = S_j R_j$$

其中：

C_j——第 j 个指标的信息量大小；

S_j——经过极大值变换后的第 j 个指标的标准差；

R_j——第 j 个指标与其他指标之间的冲突性。

某个指标的信息量，由该指标的标准差和指标冲突性的乘积来计算，信息量越大表示该指标在整个评估指标体系中的作用越大。

e) 确定客观权重。根据各指标信息量大小分配指标客观权重。

$$w_j = \frac{C_j}{\sum_{i=1}^{3} C_j}$$

其中：

w_j——第 j 个指标的权重大小。

附 录 C
（资料性）
深度学习算法评估实施案例

C.1 深度学习算法说明

深度学习图像分类算法是一种利用深度神经网络对图像进行自动分类的技术。这类算法通过训练模型来识别图像中的关键特征，并将图像归类到预定义的类别中。

C.2 评估准备

评估准备包括测试数据集质量审查、选择质量特性和评估指标、构建评估模型3个流程。

a） 测试数据集质量审查

测试数据集为图像分类数据集。

b） 选择质量特性和评估指标

质量特性选择基础性能和可解释性；其中，基础性能的评估指标选择F1分数、准确率、精确率、召回率、错误率等，见表C.1。可解释性的评估指标选择解释一致性、解释有效性、解释因果性、解释充分性等，见表C.2。

表C.1 基础性能的评估指标

评估指标	评估结果	评估得分	测试结果	等级说明
F1分数				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
准确率				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
精确率				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
召回率				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
错误率				优越级≤10%，进阶级≤20%，条件级≤30%，受限级>30%
总评				优越级[75，100]，进阶级[50，75)，条件级[25，50)，受限级[0，25)
注：基础性能计算的权重分配为，基础性能＝20% F1分数＋20% 准确率＋20% 精确率＋20% 召回率＋20% 错误率。				

表C.2 可解释性的评估指标

评估指标	评估结果	评估得分	测试结果	等级说明
解释一致性				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
解释有效性				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
解释因果性				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
解释充分性				优越级≥99%，进阶级≥90%，条件级≥80%，受限级<80%
总评				优越级[75，100]，进阶级[50，75)，条件级[25，50)，受限级[0，25)
注：可解释性指标计算的权重分配为，可解释性＝25% 解释一致性＋25% 解释有效性＋25% 解释因果性＋25% 解释充分性。				

c) 构建评估模型

设定本轮算法评估的评估分支计算公式为，评估分值＝75％ 基础性能＋25％ 可解释性。

C.3 评估执行

运行深度学习图像分类算法，获取推理结果，基于推理结果计算基础性能和可解释性评估指标的测试结果。

C.4 分析评估

将测试结果计算填入表 C.1 和表 C.2 中，得到该深度学习图像分类算法的评估等级，如表 C.3 和表 C.4 所示。

表 C.3 基础性能的评估指标（含评估结果）

评估指标	评估结果	评估得分	测试结果	等级说明
F1 分数	优越级	98	0.98	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
准确率	优越级	99.87	99.87％	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
精确率	进阶级	92	92％	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
召回率	优越级	98	98％	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
错误率	进阶级	87	13％	优越级≤10％，进阶级≤20％，条件级≤30％，受限级＞30％
总评	优越级	94.97	—	优越级[75，100]，进阶级[50，75)，条件级[25，50)，受限级[0，25)
注：评估得分表示每个指标的测试结果对应的评估得分，100 分制，0 为最低分，100 为最高分，精度为小数点后两位。对于正向指标，如 F1 分数、准确率、精确率、召回率，评估得分＝测试结果×100。对于反向指标，如错误率，评估得分＝(1－测试结果)×100。				

表 C.4 可解释性的评估指标（含评估结果）

评估指标	评估结果	评估得分	测试结果	等级说明
解释一致性	优越级	99	99％	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
解释有效性	优越级	89	89％	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
解释因果性	进阶级	81	81％	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
解释充分性	优越级	97	97％	优越级≥99％，进阶级≥90％，条件级≥80％，受限级＜80％
总评	优越级	91.5	—	优越级[75，100]，进阶级[50，75)，条件级[25，50)，受限级[0，25)
注：解释一致性、解释有效性、解释因果性、解释充分性均为正向指标，其评估得分＝测试结果×100。				

C.5 评估结论

综合基础性能和可解释性的评估结果，计算得出本次深度学习图像分类算法的评估等级为优越级。

参 考 文 献

[1] GB/T 25000.23—2019 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第23部分:系统与软件产品质量测量

[2] GB/T 43437—2023 信息技术 信息产品研发能力评估模型

[3] ISO/IEC TS 4213:2022 Information technology—Artificial intelligence—Assessment of machine learning classification performance

[4] Barocas S, Hardt M, Narayanan A. Fairness and machine learning: Limitations and opportunities[M]. MIT press, 2023