人工智能模型训练数据偏差检测方法研究

张静雯[1] 唐乐红[1*] 黄长泉[2] 阮煜鑫[1] 潘自坚[1]
1 阳光学院，福建 福州 350015
2 中科数创（厦门）智能科技研究院，福建 厦门 361015

[摘 要] 训练数据的样本失衡、历史偏见、标签噪声和代理变量关联会被人工智能模型学习并放大，仅依赖准确率或单一公平指标难以及时发现偏差。针对该问题，提出一种面向模型训练前审计的多维训练数据偏差检测方法。该方法从群体代表性、特征分布、标签一致性、群体公平和反事实敏感性五个维度构建指标，经过方向统一、标准化与权重学习形成综合偏差得分，同时输出偏差群体、关联特征和疑似异常标签。采用包含 1200 个数据集窗口的可控仿真进行验证，每个窗口含 600 个样本，并注入代表性偏差、特征偏移、标签偏差、代理偏差和混合偏差。结果表明，所提方法的精确率、召回率、F1 值和 AUC 分别达到 0.953、0.971、0.962 和 0.993，对多类型偏差的识别能力优于样本比例、统计均等和敏感系数等单指标方法。

[关键词] 人工智能；训练数据；数据偏差；算法公平；偏差检测；可信机器学习
DOI：10.64635/ja.2026.1447 中图分类号：TP181 文献标识码：A

Research on Bias Detection Methods for Artificial Intelligence Model Training Data

Zhang Jingwen[1], Tang Lehong[1] *, Huang Changquan[2], Ruan Yuxin[1], Pan Zijian[1]
1 Yango University, Fuzhou, Fujian 350015, China
2 Zhongke Shuchuang (Xiamen) Intelligent Technology Research Institute, Xiamen, Fujian 361015, China

Abstract: Sample imbalance, historical bias, label noise, and proxy-variable associations in training data can be learned and amplified by artificial intelligence models, and reliance on accuracy or a single fairness metric alone makes it difficult to detect bias in a timely manner. To address this problem, this paper proposes a multidimensional training data bias detection method for pre-training model audits. The method constructs indicators from five dimensions: group representativeness, feature distribution, label consistency, group fairness, and counterfactual sensitivity. Through direction unification, standardization, and weight learning, a comprehensive bias score is formed, while biased groups, associated features, and suspected abnormal labels are also output. Validation is conducted using a controllable simulation containing 1,200 dataset windows, each with 600 samples, with representativeness bias, feature shift, label bias, proxy bias, and mixed bias injected. The results show that the proposed method achieves a precision, recall, F1 score, and AUC of 0.953, 0.971, 0.962, and 0.993, respectively, and outperforms single-indicator methods such as sample proportion, statistical parity, and sensitivity coefficient in identifying multiple types of bias.

Keywords: artificial intelligence; training data; data bias; algorithmic fairness; bias detection; trustworthy machine learning

引言

人工智能系统的性能不仅取决于模型结构，也受到训练数据生成过程的直接约束。数据采集范围不足、特定群体样本缺失、历史决策结果被直接作为标签，以及人工标注标准不一致，都可能使模型形成系统性偏差。深度学习公平性研究表明，数据标注和模型设计中的偏见可能被模型强化，并在招聘、信贷、医疗和内容推荐等场景形成差异化影响[1]。因此，偏差治理不应只在模型上线后通过投诉或结果复核被动开展，而应前移到数据入库、标注验收和训练集版本发布阶段。

现有训练数据审计通常集中于缺失值、重复值、类别不平衡和异常值等一般质量问题，能够回答数据是否完整，却难以回答不同群体是否被同等代表、标签错误是否集中于某类人群、非敏感特征是否成为敏感属性的代理变量等问题[2]。

1 训练数据偏差类型与检测要求

训练数据偏差具有来源多样、相互耦合和场景依赖等特征。第一，代表性偏差表现为某些群体样本比例明显低于业务总体或目标人群基准，导致模型对少数群体学习不足。第二，特征分布偏差表现为同一业务含义的变量在群体间存在异常分布差异，可能来源

于采集设备、地区环境或流程规则不同。第三，标签偏差包括群体间不同的误标率、漏标率和标注尺度，尤其容易将历史制度偏差固化为监督信号。第四，代理偏差是指模型虽然未直接使用性别、年龄等敏感属性，但职业、地区、设备类型等变量与敏感属性高度关联，仍可能产生差异化结果。第五，交叉群体偏差由多个属性组合形成，单独考察某一属性时并不明显。

算法公平指标之间并非完全等价，不同指标可能给出不同甚至冲突的判断，单一指标难以覆盖全部偏差形态[3]。因此，检测方法应满足四项要求：一是同时覆盖样本、特征、标签和预测关联；二是支持多敏感属性及其组合；三是能够输出可定位的异常对象，而非只给出总体分数；四是允许根据业务风险调整阈值和指标权重。

2 多维训练数据偏差检测方法

2.1 总体流程

方法总体流程如图 1 所示。首先接入训练样本、标签、敏感属性、数据字典和业务基准，完成缺失、重复、类型及取值范围审计；随后并行计算分布偏差、标签偏差、群体公平和反事实敏感性指标；再对指标进行同向化与标准化，通过验证集学习融合权重并形成综合偏差得分；最后依据阈值输出风险等级，同时定位异常群体、关键特征和疑似错误标签。检测结果反馈到采样、标注与数据版本治理环节，形成“检测－修正－复检”的闭环[4]。

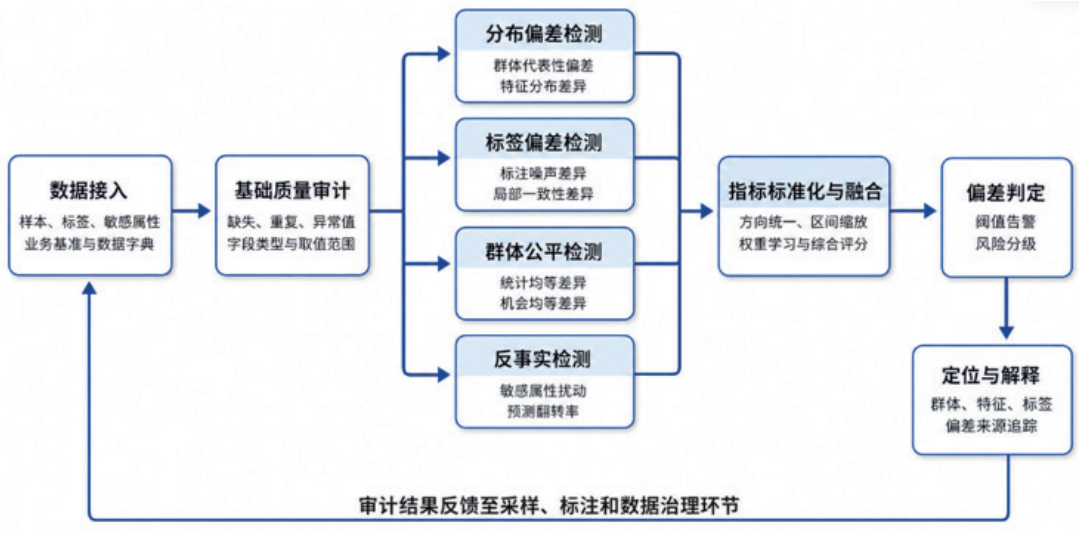

图 1 人工智能模型训练数据偏差检测总体流程

2.2 指标构建

（1）群体代表性指标。设敏感群体g在训练集中的比例为p_g，业务基准比例为q_g，群体集合为G，且$\sum_{g\in G} p_g = 1$、$\sum_{g\in G} q_g = 1$，则代表性偏差定义为$B_r = \frac{1}{2}\sum_{g\in G}|p_g - q_g|$。其中$B_r \in [0,1]$，该指标越大，说明训练集结构与目标总体偏离越明显。

（2）特征分布指标。对连续特征计算群体间标准化均值差或分布距离，对类别特征计算条件频率差。将特征 j 在群体 g 与参考群体之间的归一化分布差异记为$d_{jg} \in [0,1]$。为减少量纲影响并避免少量无关特征稀释关键偏移，取全部差异值的 95% 分位数作为 $B_f = Q_{0.95}(\{d_{jg}\})$。

（3）标签一致性指标。先利用非敏感特征训练轻量基准模型，并采用交叉验证获得样本i的预测概率$\hat{p}_i$。样本级不一致度定义为$r_i = \frac{1}{2}\left(|y_i - \hat{p}_i| + \left|y_i - \frac{1}{K}\sum_{j\in N_K(i)} y_j\right|\right)$。群体 g 的平均不一致度记为$\bar{r}_g$，标签一致性指标定义为$B_l = \max_{g,h\in G}|\bar{r}_g - \bar{r}_h|$。该指标越大，说明疑似误标越可能集中于特定群体。

（4）群体公平指标。对分类任务计算统计均等差异和机会均等差异。统计均等差异记为$B_{SP} = \max_{g,h\in G}|P(\hat{Y}=1 \mid A=g) - P(\hat{Y}=1 \mid A=h)|$，机会均等差异记为$B_{EO} = \max_{g,h\in G}|P(\hat{Y}=1 \mid Y=1, A=g) - P(\hat{Y}=1 \mid Y=1, A=h)|$，并定义综合群体公平指标 $B_{fair} = \max(B_{SP}, B_{EO})$。该指标用于衡量不同群体的正类预测比例差和真实正类样本识别机会差。

（5）反事实敏感性指标。在保持非敏感特征不变的条件下，将样本敏感属性a_i替换为另一可行取值a_i'，统计基准模型预测发生翻转的比例：$B_c = \frac{1}{n}\sum_{i=1}^{n} I\left[\hat{y}(x_i, a_i) \neq \hat{y}(x_i, a_i')\right]$，其中$I$为示性函数。该指标是对反事实敏感性的操作化近似，数值越大说

明数据中越可能存在直接或间接歧视关联。

2.3 综合评分与定位

将五类指标通过稳健标准化转换为$z_k = \frac{B_k - \text{median}(B_k)}{\text{IQR}(B_k)+\varepsilon}$，其中IQR为四分位距。以数据窗口是否存在系统性偏差作为二分类标签，使用逻辑回归在权重学习集上学习权重w_k，综合得分为$B = \sigma\left(\sum_{k=1}^{5} w_k z_k + b\right)$，其中$\sigma(t) = \frac{1}{1+e^{-t}}$。当$B$超过预警阈值时，根据加权贡献 $C_k = w_k z_k$确定主要偏差来源。样本级定位采用“高标签不一致度 + 高反事实翻转贡献”的联合规则，特征级定位则按分布差异和敏感属性关联强度排序。可信机器学习强调公平性应与可解释性、鲁棒性和隐私保护共同纳入模型全生命周期管理[5]。

3 仿真实验与结果分析

3.1 实验设置

为验证方法对不同偏差类型的覆盖能力，构建1200 个二分类数据集窗口，每个窗口包含 600 个样本、2 个业务特征、1 个敏感属性和 1 个二分类标签。其中 600 个窗口不注入系统性偏差，其余窗口随机注入代表性偏差、特征偏移、群体差异化标签噪声、敏感属性直接影响和混合偏差五种情形。数据按 65% 和35% 划分为权重学习集与测试集。比较方法包括样本比例检测、统计均等检测、敏感系数检测和本文多维方法。各单指标方法在学习集上选择使 F1 值最大的阈值，所有方法均在同一测试集上评价。

3.2 结果分析

表 1 显示，样本比例检测能够发现代表性偏差，但对标签偏差和代理偏差不敏感，召回率较高而精确率仅为 0.531。统计均等检测的 F1 值达到 0.779，但特征偏移未必立即导致正类比例变化，仍存在漏检。敏感系数检测容易把随机相关识别为偏差，精确率较低。本文方法同时利用分布、标签和反事实信息，精确率与召回率分别达到0.953 和0.971，F1 值为0.962，AUC 达到0.993。结果说明，多维指标融合可以降低单一统计差异造成的误判，并提高混合偏差的识别稳定性。

表 1 不同偏差检测方法的仿真结果

检测方法	精确率	召回率	F1 值	AUC
样本比例检测	0.531	0.948	0.680	0.729
统计均等检测	0.871	0.705	0.779	0.850
敏感系数检测	0.506	0.976	0.667	0.756
本文多维方法	0.953	0.971	0.962	0.993

进一步分析发现，代表性指标对样本结构异常贡献最大，特征分布和反事实指标对代理偏差更敏感，标签一致性指标能够提供异常样本定位线索。不过，综合得分不等同于伦理或法律结论，其作用是筛选需要复核的数据版本和群体。人工智能嵌入数据治理时可能同时带来效率提升与算法歧视风险，治理措施需要结合规则、问责和数据环境优化共同实施[6]。

4 结论

本文针对训练数据偏差难以通过一般数据质量检查发现的问题，提出由群体代表性、特征分布、标签一致性、群体公平和反事实敏感性组成的多维检测方法。可控仿真结果表明，该方法能够同时识别多类偏差，并在精确率、召回率、F1 值和 AUC 方面优于单指标检测。方法的主要价值在于将偏差审计前移到训练前阶段，并把总体风险判断与群体、特征、标签定位结合，为补充采样、重标注和公平训练提供可执行依据。实际应用中仍需结合业务目标、法律规范和专家判断，避免将统计差异机械等同于歧视结论。

[参考文献]

[1] 陈晋音，陈奕芃，陈一鸣，等．面向深度学习的公平性研究综述[J]．计算机研究与发展，2021,58(2):264-280.

[2] 古天龙，李龙，常亮，等．公平机器学习：概念、分析与设计[J]．计算机学报，2022,45(5):1018-1051.

[3] 范卓娅，孟小峰．算法公平与公平计算[J]．计算机研究与发展，2023,60(9):2048-2066.

[4] 王艳，侯哲，黄滟鸿，等．基于概率模型检查的树模型公平性验证方法[J]．软件学报，2022,33(7):2482-2498.

[5] 陈彩华，余程熙，王庆阳．可信机器学习综述[J]．工业工程，2024,27(2):14-26.

[6] 彭丽徽，张琼，李天一．人工智能嵌入政府数据治理的算法歧视风险及其防控策略研究[J]．农业图书情报学报，2024.