

基于信息熵 -D-S 证据理论的用户画像多源异构数据融合方法

汪 钊¹
WANG Zhao

摘 要

在融合互联网平台、物联网的异构网络环境下, 用户行为数据源自高维稀疏的射频指纹、非结构化的多模态时序信号, 呈现出典型的语义鸿沟与特征碎片化现象, 在数据融合过程中易因分布漂移与量纲异构引发实体混淆, 进而导致属性失配等问题, 严重影响用户画像多源异构数据的融合质量。为了缓解这一问题, 文章提出了基于信息熵 -D-S 证据理论的用户画像多源异构数据融合方法。经过归一化处理后, 通过联合信息熵的计算, 量化用户画像数据源的不确定性。结合 D-S 证据理论, 为量化后不同信息熵的数据源分配权重参数, 利用 Dempster-Shafer 组合规则, 建立用户画像多源异构数据的融合规则, 将多源异构数据的语义不确定性问题转化为可度量的信任决策过程, 降低用户行为数据在证据理论框架下的属性失配冲突。在马氏距离的评估下, 采用 GK 模糊聚类算法实现用户画像多源异构数据的融合处理, 实验结果表明, 该方法所得出的数据融合结果, 表现出的 MAE 值平均仅为 1.2%, 覆盖率达到 95% 以上, 数据本体对齐率达到 98%, 有着较高的数据融合质量。

关键词

数据融合; 多源异构数据; 用户画像; 信息熵; D-S 证据理论; 用户画像数据

doi: 10.3969/j.issn.1672-9528.2026.05.012

0 引言

用户画像数据分散在互联网平台、物联网设备等多种数据来源, 呈现出典型的多源异构性。这一特性使用户画像的构建较为困难, 容易出现数据孤岛的问题, 影响了用户画像构建的效果。在这样的情况下, 采用更加有效的手段对用户画像的多源异构数据进行融合分析。对此, 领域内很多研究人员提出了多源异构数据融合方法。

当前多数学者已对网络多源异构数据的融合问题进行了深入分析。胡靖阳等^[1]构建深度学习捕捉多源异构数据的多尺度特征, 并在特征级上, 完成多源异构数据融合处理。但该方法未对不确定性数据进行量化处理, 在面临多源异构数据的特征空间差异较大情况时, 会影响其实际的数据融合质量。杨童博等^[2]采用最小二乘法估计近似失效点, 建立真实的失效分布函数, 通过分布拟合, 将多源异构数据融合。该方法泛化能力较强, 能够应对概念漂移问题。但该方法在实际应用过程中存在冷启动问题, 如果缺乏归一化后的数据作为基础数据, 实际的数据融合质量可能下降。唐伟云^[3]在数据空间将多源异构数据动态排序, 并采用深度学习模型提取

数据特征。在改进聚类中心的基础上, 利用改进的模糊聚类算法实现多源异构数据融合。该方法能够在数据加密条件下完成融合, 数据安全性较高。但该方法的线性距离度量无法准确衡量多源异构数据的语义相似性, 聚类性能不理想, 导致数据融合质量较低。周璇^[4]提取多源异构数据的实体与关系, 建立相应的知识图谱, 实现多源异构数据的全方位融合。但该方法若没有归一化后的数据作为数据基础, 可能会影响知识图谱的实体对齐性能, 导致实际的多源异构数据融合质量较低。

因此, 本研究提出了一种基于信息熵 -D-S 证据理论的用户画像多源异构数据融合方法。

1 用户画像多源异构数据融合方法设计

1.1 基于信息熵的用户画像数据源不确定数据量化方法

在构建用户画像时, 由于用户的行为具有一定的偶然性, 使得不同来源的用户数据存在不确定性^[5]。为了避免这一问题对多源异构数据融合质量的影响, 采用信息熵对用户画像数据源不确定性进行了量化分析。考虑到多源异构的用户画像数据存在量纲不一致的问题, 采用极差法进行归一化^[6]处理, 用公式表示为:

$$x = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}} \quad (1)$$

1. 陕西能源职业技术学院 陕西咸阳 712000

[基金项目] 陕西能源职业技术学院一般专项自然科学研究计划项目 (2024KYZRP03QN)

式中: x 表示归一化的多源异构数据, x_{ij} 表示用户画像数据中第 i 个数据样本在第 j 个维度的数据值, $\max x_{ij}$ 和 $\min x_{ij}$ 分别表示用户画像数据的最大值和最小值。消除量纲差异的干扰后, 引入信息熵的概念, 对用户画像数据源的不确定性进行量化计算^[7]。用公式表示为:

$$H(x) = -\sum_{x \in C} P(x) \cdot \log_2(x) \quad (2)$$

式中: $H(x)$ 表示用户画像数据源的信息熵, C 表示用户画像数据的取值范围, $P(x)$ 表示用户画像数据在某一数值的取值概率。

考虑到部分用户画像数据为连续型的随机变量, 本研究对式 (2) 进行了优化, 采用联合分布概率密度函数计算用户画像数据的信息熵^[8], 用公式表示为:

$$H(x, y) = -\iint P(x, y) \cdot \log_2(x, y) dx dy \quad (3)$$

式中: x 、 y 表示两个连续型的用户画像数据随机变量, $H(x, y)$ 表示两者的联合信息熵, $P(x, y)$ 表示联合分布概率密度函数, dx 和 dy 表示两个随机变量的变化值。在信息熵的计算下, 量化用户画像数据源的不确定性, 分析多源异构数据的信息熵冗余度, 对用户画像多源异构数据的差异程度进行量化表达。用公式表示为:

$$d_j = 1 - H(x, y) \quad (4)$$

式中: d_j 表示多源异构数据在第 j 个维度的信息熵冗余度。

1.2 基于 D-S 证据理论的多源异构数据融合方法

异构网络环境中, 用户行为数据来源复杂, 涵盖射频指纹、多模态时序信号等高维稀疏特征, 易出现语义鸿沟与特征碎片化现象, 进而引发属性失配与语义歧义问题。为了能够将用户画像多源异构数据进行有效融合, 本研究基于引入 D-S 证据理论, 判断数据融合的支持度, 建立多源异构数据融合的基本规则^[9], 从而缓解因数据分布漂移或语义异构所导致的属性匹配偏差, 降低属性失配的可能性。定义多源异构数据融合规则的识别框架, 用公式表示为:

$$\Theta = \{\emptyset, \{A(x)\}\} \quad (5)$$

式中: Θ 表示数据融合可行规则的有限集合, $\emptyset$ 表示空集, $\{A(x)\}$ 表示识别框架的幂值。根据信息熵计算所量化的数据源不确定性, 为每个用户画像数据的数据源分配一个权重参数, 作为证据理论的信息效用值^[10]。权重参数的计算方法用公式表示为:

$$w_j = d_j / \sum_{j=1}^m d_j \quad (6)$$

式中: w_j 表示数据源不确定性器权重参数, m 表示用户画像多源异构数据的维度数。根据该权重参数定义的信息效用

值, 构造数据融合 D-S 证据理论的基本概率分配函数 (basic probability assignment, BPA)^[11]。则数据融合的信任值计算方法用公式表示为:

$$B(x, y) = \sum_{A \in \Theta} w_j \cdot 2^{\Theta} M(x_j) \quad (7)$$

式中: $B(x, y)$ 表示数据融合信任值, M 表示基本概率指派的 BPA 函数。在此基础上, 利用 D-S 证据理论中的 Dempster-Shafer¹ 组合规则, 建立用户画像多源异构数据的融合规则。用公式表示为:

$$U(x, y) = \frac{n}{1-K} \sum_{i \in \Theta} w_i B(x, y) \quad (8)$$

式中: $U(x, y)$ 表示多源异构数据融合规则, n 表示用户画像数据样本量, K 表示证据融合的冲突程度。

通过上述步骤, 在 D-S 证据理论的分析下, 建立得到用户画像多源异构数据的融合规则, 作为融合过程的基准, 降低因数据分布漂移或语义异构所导致的属性失配问题。基于所建立的数据融合规则, 通过度量用户画像多源异构数据与融合规则之间的相似性, 采用 GK 模糊聚类算法进行多源异构数据的融合处理。根据马氏距离^[12] 评估用户画像多源异构数据与融合规则之间的相似性, 用公式表示为:

$$S(x, U) = U(x, y) \sqrt{(x - U) E^{-1}(x, U)} \quad (9)$$

式中: $S(x, U)$ 表示融合规则相似性的量化计算值, E^{-1} 表示逆矩阵。将式 (9) 的计算结果作为多源异构数据聚类融合的隶属度值, 由此构建数据融合过程, 用公式表示为:

$$X = \alpha \sum_{i=1}^n S(V)^N \quad (10)$$

式中: X 表示用户画像多源异构数据的模糊聚类融合, V 表示聚类中心, N 表示聚类数目, α 表示模糊系数。为了确保这一融合过程的有效性, 为其定义一个目标函数^[13]。采用拉格朗日乘数法建立数据融合的目标函数, 用公式表示为:

$$J(X, V, U) = \sum_{i=1}^n \sum_{j=1}^m S(x, U)^{\sigma} \quad (11)$$

式中: J 表示目标函数, σ 表示拉格朗日乘子。

在目标函数的指导下, 逼近最优聚类状态, 实现用户画像多源异构数据的融合处理。

2 实验

2.1 实验准备

为了检验本文方法在实际应用中的数据融合性能, 本次实验面向多个领域, 包括电力领域、电商领域等, 采用 Python 工具在多个数据源中爬取实验数据。实验环境如图 1 所示。本次实验共采集了 100 位用户的多来源数据作为实验数据。对所获取的用户画像数据进行初步处理, 去除其中的无关数据与重复数据, 构建本次测试的实验数据集。

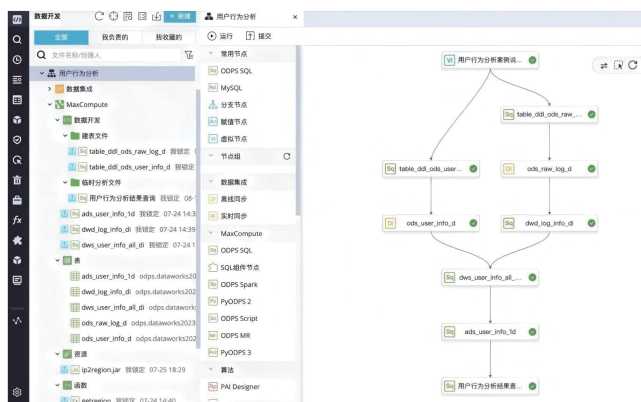


图1 用户画像异构数据融合实验环境

2.2 测试结果

2.2.1 MAE 指标评估结果分析

利用测试集中的数据,采用 MAE 指标测试并统计不同方法的归一化值,得到对比结果如图 2 所示。

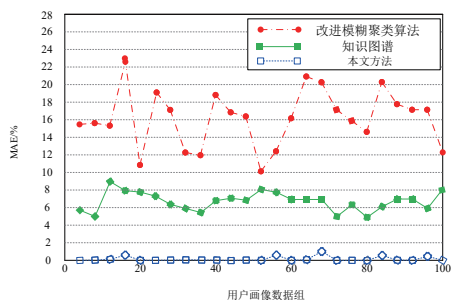


图2 数据归一化 MAE

从图2可以明显看出,本文方法表现出了更低的MAE值,平均仅为1.2%。因为本文方法采用信息熵对用户画像数据源不确定性进行量化分析,并通过极差法消除量纲差异。这种量化方式能够更准确地把握数据源的内在特征和不确定性程度,使本文方法在预测归一化值时能够更贴合真实值,从而降低了 MAE 值。

2.2.2 覆盖率指标评估

为了进一步测试所提方法的数据融合质量,将覆盖率为实验指标,测试结果如图 3 所示。

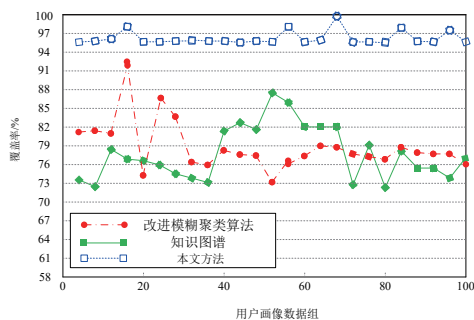


图3 数据融合质量的覆盖率

由图3可知,本文方法融合后与原始数据之间的重合度

更高,覆盖率达到 95% 以上,表明该方法的融合效果更好。因为所提方法在建立数据融合规则后,采用 GK 模糊聚类算法进行多源异构数据的融合处理,并根据马氏距离评估相似性,以计算结果作为聚类融合的隶属度值。这种融合方式能够更合理地整合数据,保证融合后数据的完整性和准确性,进而提高了覆盖率。

2.2.3 对齐率指标评估

将数据本体对齐率作为测试指标,分析数据融合效果,如图 4 所示。

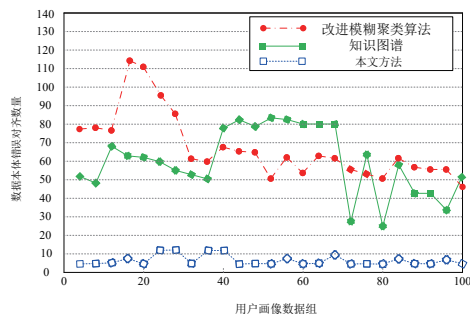


图4 数据融合质量的数据本体对齐率

从图4能够看到,本文方法的数据本体的错误对齐数量较少,平均不到10个,对齐率达到98%,具有较高的数据融合质量。因为所提方法利用 GK 模糊聚类算法在规则基础上进行数据融合处理,使数据本体对齐更加准确,错误对齐数量减少,对齐率提高。

3 结语

本研究设计的基于信息熵-D-S 证据理论的用户画像多源异构数据融合方法,通过信息熵的计算,能够处理用户画像数据源的不确定性问题,量化表达多源异构数据之间的差异程度,有效增强了多源异构数据融合过程中,区分数据样本差异性的能力。结合 D-S 证据理论所建立的融合规则,实现了质量较高的多源异构数据融合处理,能够为用户画像构建过程提供准确且全面的用户数据信息,应对数字化时代多渠道数据量爆发的问题,满足用户画像分析的企业运营需求。

参考文献:

- [1] 胡靖阳,陈靖辉.基于深度学习的多源数据融合在医疗数据分析中的应用[J].信息与电脑,2025,37(7):9-11.
- [2] 杨童博,魏政,周坤.基于近似失效点的多源异构数据融合可靠性评估模型[J].机械强度,2025,47(3):136-142.
- [3] 唐伟云.基于改进模糊聚类算法的多源异构数据安全聚合研究[J].国外电子测量技术,2025,44(2):148-154.
- [4] 周璇.知识图谱技术下智慧城市多源异构大数据融合方法研究[J].信息记录材料,2025,26(1):232-234.

基于改进 YOLOv11 的轻量化玉米病虫害智能识别系统

高 俊¹ 李博慧¹ 程湘瑜¹ 景思源¹ 吴翰如¹ 张芸博¹ 翟树杰¹ 刘钟泽¹

GAO Jun LI Bohui CHENG Xiangyu JING Siyuan WU Hanru ZHANG Yunbo ZHAI Shujie LIU Zhongze

摘 要

针对复杂农田场景中出现的由于玉米病虫害目标尺度小、背景干扰大所造成的检测精度低、模型部署困难等问题,文章通过在核心模块中加入 YOLOv11 检测模型,并在 YOLOv11 主干网络引入注意力机制模块来加强小目标病害的特征提取与表征,还使用了通道剪枝以及 INT8 量化方法来减小模型体积和计算量。从实验结果可知,改进后的模型不仅有效提高了复杂环境下小目标病虫害的检测精度和识别实时性,也为轻量化智能模型在智慧农业边缘端的应用提供了可靠的工程技术方案和实践依据。

关键词

YOLOv11; 病虫害检测; 模型压缩; 边缘计算; 智慧农业

doi: 10.3969/j.issn.1672-9528.2026.05.013

0 引言

早期智慧农业存在缺乏统一规划、数据标准混乱、技术适配性差、算法性能低等众多局限,随着政策和科技的共同推动,数据同产业相连接,为技术大规模应用打下了良好的基础。然而智慧农业在实践过程中还存在着数据精准度不高、算法反应慢、技术集成度低等问题。特别是把通用 YOLO 模型迁移到大田场景的时候,会遇到复杂的背景干扰、小目标

检测误差以及边缘设备算力不足三大难题^[1]。尽管已有研究对于轻量化注意力机制、农业背景下深度学习的应用进行了大量的研究,但是对于 YOLOv11 在农业特殊场景中的适配性和边缘部署的轻量化还存在不足。

针对玉米病虫害智能检测中遇到的问题,以改进 YOLOv11 为基础,自主研发了一套名为“智链农芯—多擎驱动的智慧农业系统”,为算法的改进与部署提供了高效、可拓展的平台^[2]。

为了解决农业中对病虫害识别目标小、背景复杂、需要实时处理的难题,本文提出了一种基于改进 YOLOv11 轻量级病虫害识别算法。采用高效的注意力机制提高细微病斑的特征提取,用剪枝和量化的方法压缩模型,在保证精度的同

1. 山西工学院信息工程学院 山西朔州 036000

[基金项目] 山西省教育厅山西省高等学校科技创新项目 (2023L439)

- [5] 李莉,孙世军,朱坤双,等.面向电力气象数据的多源异构数据融合方法研究[J].电子设计工程,2024,32(16): 178-182.
- [6] 李洵,纪元,张盛春.面向数字孪生的输变电设备多源异构数据自动采集方法[J].自动化技术与应用,2024,43(12): 139-142.
- [7] 杨秋菊.基于模糊理论的多源异构传感器数据融合模型[J].吉林大学学报(工学版),2024,54(10): 3058-3063.
- [8] 刘静,张耀辉.“5G+智慧教育”背景下教学过程多源异构数据融合研究[J].湖南邮电职业技术学院学报,2024,23(3): 61-64.
- [9] 陶迎松.云计算环境下基于 X-means 优化聚类的海量多源异构数据分类[J].北部湾大学学报,2024,39(2): 40-46.
- [10] 李高扬,王宁,高若田,等.面向新型电力系统智能化提效的多源异构数据融合技术研究[J].电测与仪表,2024,

61(11): 116-121.

- [11] 陈鼎,彭长根,谭伟杰.基于联邦学习的多源数据用户画像设计方案[J].南京邮电大学学报(自然科学版),2023,43(5): 83-91.
- [12] 唐圣学,刘成龙,王鸿凡.基于马氏距离与 WOA-LSTM RCP 的早期故障诊断[J].计算机仿真,2025,42(2): 546-552.
- [13] 苗光尧,安静,黄小花,等.基于多源数据融合的电力用户画像构建方法研究[J].自动化技术与应用,2022,41(8): 93-96.

【作者简介】

汪钊(1998—),男,陕西宝鸡人,硕士,助教,研究方向:数据挖掘。

(收稿日期: 2025-10-09 修回日期: 2026-05-12)