

文章编号:1671-5896(2026)04-1008-07

融合 HHO 算法和混合聚类算法的多源异构数据自动分类

宁丽萍

(新疆维吾尔自治区人民医院 财务部, 乌鲁木齐 830001)

摘要: 针对多源异构数据的柔性重构需求需要处理数据细节和异常情况较多, 使其在对多源异构数据进行分类时, 难以通过捕获其隐藏特征, 整合多源异构数据的相似数据, 导致分类的准确性较低的问题, 提出融合 HHO (Harris Hawks Optimization) 算法和混合聚类算法的多源异构数据自动分类方法。利用低秩表示策略, 捕捉多源异构数据的隐藏特征, 获取其全局结构信息, 提高数据质量和一致性。组合蚁群聚类算法与 K -means 算法, 设计出混合聚类算法, 通过利用 K -means 算法修正组合蚁群聚类算法中的蚁巢消减误差, 实现多源异构数据的相似数据分组。将混合聚类算法与 HHO 算法结合, 通过更新个体适应度值, 确定 HHO 算法的最佳类中心, 实现相似分组后多源异构数据的自动分类。实验结果表明, 经分类结果的混淆矩阵和内聚度的评估结果共同验证, 所提方法不仅能自动分类多源异构数据, 且类别边界清晰, 数据点分布紧凑, 使混淆矩阵对角线上的真实值更高, 内聚度均值高达 98.86%, 具有较高的分类准确性。

关键词: 多源异构数据; 低秩表示; HHO 算法; 混合聚类算法; 自动分类

中图分类号: TP399

文献标志码: A

Automatic Classification of Multi-Source Heterogeneous Data by Integrating HHO Algorithm and Hybrid Clustering Algorithm

NING Liping

(Finance Department, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi 830001, China)

Abstract: Due to the flexible reconstruction requirements of multi-source heterogeneous data, which require handling of data details and abnormal situations, it is difficult to classify multi-source heterogeneous data by capturing its hidden features and integrating similar data, resulting in low classification accuracy. Therefore, a multi-source heterogeneous data automatic classification method that integrates HHO (Harris Hawks Optimization) algorithm and hybrid clustering algorithm is proposed. Using low rank representation strategy to capture hidden features of multi-source heterogeneous data, its global structural information is obtained, and data quality and consistency are improved. A hybrid clustering algorithm is designed by combining ant colony clustering algorithm and K -means algorithm. By using K -means algorithm to correct the ant nest reduction error in the combined ant colony clustering algorithm, similar data grouping of multi-source heterogeneous data is achieved. Combining hybrid clustering algorithm with HHO algorithm, by updating individual fitness values, the optimal class center of HHO algorithm is determined to achieve automatic classification of multi-source heterogeneous data after similar grouping. The experimental results show that the proposed method can automatically classify heterogeneous data from multiple sources and has clear category boundaries and compact data point distribution, resulting in higher

收稿日期: 2025-02-11

基金项目: 克拉玛依市创新环境建设计划(软科学)基金资助项目(2024hjrxx0028)

作者简介: 宁丽萍(1973—), 女, 湖南邵东人, 新疆维吾尔自治区人民医院讲师, 主要从事财务管理, 预算管理, 内部控制及成本管理、数据分析研究, (Tel)86-13899919282 (E-mail) ningliping9282@163.com。

true case values on the diagonal of the confusion matrix. The average cohesion is as high as 98.86%, demonstrating high classification accuracy.

Key words: multi-source heterogeneous data; low rank representation; Harris Hawks optimization (HHO) algorithm; hybrid clustering algorithm; automatic classification

0 引言

在当今数据爆炸时代，多源异构数据已成为信息社会的基石^[1]。随着云计算等技术的飞速发展，其产生速度、种类和规模都达到了前所未有的高度^[2]。理解和有效处理多源异构数据，对提升数据价值、优化资源配置具有重要意义。其中以数据分类最为基础且关键，其不仅可以提高信息的检索效率，而且还能挖掘出数据间的关联性和规律，为决策提供全面的数据支持，帮助决策者做出更加科学、准确的决策^[3]。然而，由于多源异构数据的来源多样化、数据结构复杂和数据量巨大等挑战性特点，使数据分类变得尤为复杂。因此，需要探索出能有效应对这些挑战的分类方法，为信息时代的数据处理提供新思路。

林少娃等^[4]应用电力领域的诉求文本、用户相关等多源异构数据，以卷积神经网络为基本算法，构建分类模型。但该方法并未考虑多源异构数据的结构、存储格式、访问方式等的差异，导致数据的质量下降，分类准确度降低。韩璐等^[5]就计算机领域中多源异构的分布式报送信息，利用经过多吸引子与遗传算法优化的元胞自动机，完成数据分类。但该方法并未考虑到多源异构数据在不同数据源中不同的表示方式，导致分类的准确性不高。李莉等^[6]基于电力领域的电力气象类多源异构数据，通过改进支持向量机算法，得到最小二乘法支持向量机算法，用于分类多源异构的相关电力气象数据。但该方法所用的鲸鱼优化算法分类的多元异构数据间的相似性并不明显，类别划分模糊，分类的准确性较低。梁越等^[7]基于法律科技领域的控申业务类多源异构数据，利用 Text Rank 算法、变换器和虚拟对抗训练、Focal 损失函数等多种策略，实现控申数据分类。但 Text Rank 算法受分词、文本清洗、高频词、文本长度等影响较大，导致数据的分类结果准确性降低。

根据分类多源异构数据的结构差异大、量大且冗余等特性，哈里斯鹰优化算法 (HHO: Harris Hawks Optimization)^[8] 凭借其较强的全局搜索能力，能在大规模、复杂的数据集中找到全局最优解。因此，结合设计的混合聚类算法，笔者提出一种用于多源异构数据的自动分类方法。通过全局搜索和局部精细调整相结合，能快速找到最优分类方案，进而提高自动分类的准确性。

1 多源异构数据自动分类方法

1.1 低秩表示下多源异构数据全局结构信息捕获

由于多源异构数据的结构、存储格式、访问方式等各不相同，使数据集成会因模式匹配、数据转换和数据冲突等操作产生数据质量下降问题，导致数据分类的质量下降。为此，笔者利用低秩表示策略，通过在多子空间内的近似采样^[9-11]，捕捉多源异构数据的隐藏特征，进而获取其全局结构信息，提高数据质量和一致性。

已知多源异构数据 x ，低秩矩阵 Z 和稀疏矩阵 E ，则三者间的关系式为

$$x = \|Z\| + \|E\|。 \tag{1}$$

根据 x 解值具有低秩性的约束条件，将式(1)作为其凸优化目标函数的约束。并在函数中添加核范数和迹范数^[12-13]，以应对联合子空间内的特征。若两范数分别是 $\|E\|_h$ 、 $\|E\|_g$ ，平衡 Z 与 E 的因子是 α ，则 x 的凸优化目标函数 $f(x)$ 及其约束方程组为

$$f(x) = \min \|Z\| + \alpha \|E\|_h \|E\|_g。$$

(2)

通过优化目标函数，获得由各模态纯净数据 $x'_j(j \in n)$ 构成的数据集 x' :

$$x' = f(x) \{x'_1, x_2, \cdots, x'_n\}。$$

(3)

利用

$$\mathbf{X} = \omega_1 x'_1 + \omega_2 x'_2 + \cdots + \omega_n x'_n \quad (4)$$

加权求和 x' , 建立低秩表示融合矩阵 $\mathbf{X}$, 获得多源异构数据的全局结构信息, 其中 ω_j 为各模态纯净数据的对应权值。

多源异构数据的全局结构信息有助于快速理解数据集的整体结构和特征, 从而更有效地进行模式匹配和数据转换, 减少数据集成过程中的复杂性和错误率。

1.2 在多源异构数据全局结构信息下的相似数据分组

考虑在多源异构数据中, 相同的概念在不同数据源中有不同的表示方式, 导致全局结构信息无法完全准确地表达数据的语义, 影响数据的准确整合和分类。为此, 笔者采用组合蚁群聚类算法^[14]与 K -means 算法^[15], 设计混合聚类算法, 实现多源异构数据的相似数据分组。并通过 K -means 算法修正组合蚁群聚类算法中的蚁巢消减误差, 将多源异构数据中相似的数据条目精准整合到同一个集合中, 进而实现数据的分组。

由蚁群聚类算法得到多源异构数据的初始分组结果。具体流程如下。

1) 设置迭代与碰面数。若蚂蚁数量为 N_{an} , 迭代总数为 N_{zd} , 则初始迭代次数 N_0 由

$$N_0 = \frac{N_{zd} N_{an}}{2X} \quad (5)$$

解得。由此推导出蚂蚁的随机碰面次数 N_{pm} 如下:

$$N_{pm} = \frac{N_{zd} N_0}{2} \quad (6)$$

2) 初始化接受阈值。针对蚂蚁 $i(i \in N_{an})$ 的 N_{pm} 次碰面, 基于相似度^[16]得到最大值 $\max\{\text{sim}(i, *)\}$ 与均值 $\overline{\text{sim}(i, *)}$, 则通过

$$\frac{\overline{\text{sim}(i, *)} + N_{pm} \max\{\text{sim}(i, *)\}}{2} \rightarrow t_{\text{thr}_i} \quad (7)$$

初始化蚂蚁 i 的接受阈值 t_{thr_i} 。

3) 参数置零。令蚂蚁的标签、年龄等参数为 0。

4) 参数修整。根据接受阈值下蚂蚁的新巢创建、缺省规则^[17-18]等多条行为规则, 调整 t_{thr_i} 参数, 得到初始相似数据分组结果。

利用 K -means 算法, 修正因随机性产生的蚁巢消减误差, 得到多源异构数据最终的相似数据分组结果, 如下:

$$\mu = J \times \max\{\text{sim}(i, *)\} t_{\text{thr}_i}$$

其中 J 为蚁巢消减误差。

在对多元异构数据进行相似数据分组后, 组内数据的相似性有助于更准确地了解数据中的规律, 从而提高后续分类结果的准确性。

1.3 基于相似多源异构数据组的最佳类中心自动分类

考虑到在某些情况下, 数据间的相似性可能并不明显, 导致类别划分模糊, 难以进行准确的自动分类。为此, 笔者在混合聚类算法中融入 HHO 算法, 通过更新个体适应度值, 确定算法的最佳类中心, 实现相似分组后多源异构数据的自动分类。

1) 根据混合聚类算法得到的蚁巢量 K , 结合数据维度 W , 设定 HHO 算法的哈里斯鹰个体维数为 $K \times W$ 。则个体位置(即类中心) c_{**} 可表示为 $K \times W$ 的行向量 $(c_{11}, \cdots, c_{1W}, \cdots, c_{K1}, \cdots, c_{KW})$ 。

2) 已知数据 x'_j 属于第 $k(k \in K)$ 类 C_k , 其中心是 c_k , 则个体的适应度 f_{fit} 由

$$f_{\text{fit}} = \mu \sum_{k=1}^K \sum_{\forall x'_j \in C_k} \text{Dis}(x'_j, c_k) \quad (8)$$

解得。其中 $\text{Dis}(x'_j, c_k)$ 为 x'_j 与 c_k 的距离, 基于两者的余弦相似性^[19]解得到:

$$\text{Dis}(x'_j, c_k) = 1 - \cos(x'_j, c_k), \quad (9)$$

由此推导出猎物的当前位置即是适应度最大值对应的个体的位置。

- 3) 降序排列个体适应度值，保留前几个较优个体。
- 4) 重复混合聚类算法操作后，基于余弦函数^[20-21]，更新猎物的逃逸能量，进而获取个体的新方位。若猎物初始能量为 E_0 ，能量调节系数为 β ，则猎物逃逸能量 E 的更新如下：

$$E = \beta E_0 \cos\left(2\pi + \frac{t}{T}\right) f_{\text{fit}},$$

(10)

其中 t 、 T 为当前迭代序数与最大数。

- 5) 横向交叉处理哈里斯鹰群中位置更新后同维度的个体后，纵向交叉处理不同维度的个体，生成子代个体。据此求解各个体新的适应度值，筛选出种群中的较优个体。
- 6) 若 HHO 算法未终止，则不断更新个体适应度值；反之，则算法结束时的猎物当前位置即是混合聚类算法的最佳类中心。
- 7) 基于所得最佳类中心，重复混合聚类算法对多源异构数据的操作，完成自动分类。

2 实 验

以某医院的财务多源异构数据为研究对象，进行数据预处理。采用 Python 编程语言、Hive 数据分析工具编写、运行 HHO 算法与混合聚类算法的相应代码，调整 HHO 算法和混合聚类算法的参数，所得最佳参数值如表 1 所示。

表 1 算法最优参数设置

Tab. 1 Optimal parameter settings for algorithm

算法	名	值	算法	名	值
HHO 算法	哈里斯鹰/只	60		初始碰面蚂蚁/只	15
	最大迭代/次	500		聚类碰面蚂蚁/只	22
	猎物能量初始状态	0.76		最大迭代/次	300
	猎物能量最终状态	-0.14		初始中心点/个	25
	混合聚类算法	聚类/类			
		6			

选取基于元胞自动机多源异构数据分类方法(文献[5]方法)与基于 Text Rank 算法的多源异构数据分类方法(文献[7]方法)作为对比方法，用于与所提方法进行对比实验验证。

采用笔者方法对所选医院财务的多源异构数据集进行相似性分组，所得分组结果可视化图像如图 1 所示。

从图 1 可看出，所提方法精准地将相似性数据进行了分组，表现出了优异的性能。

针对 3 种方法得到的医院财务数据集进行分类，通过多分类混淆矩阵来评估各方法的分类性能，如图 2 所示。

从图 2 可看出，所提方法在各个类别上的分类性能更优，对角线上的真实值均显著高于对比方法，即使是在处理容易混淆的数据类别时，依然表现出色。表明所提方法凭借混合聚类算法中 K-means 算法对蚁群聚类算法的修正，及 HHO 算法对混合聚类算法的进一步优化，有效改善了中心敏感问题，不仅可以自动分类医院财务的多源异构数据，而且分类的准确性更高。

针对 3 种方法得到的目标数据集分类结果，采用数据的分类内聚度指标评估各方法的分类的准确性，对比结果如图 3 所示。

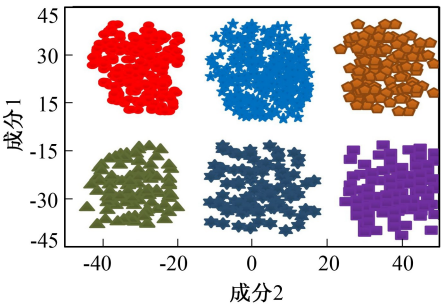


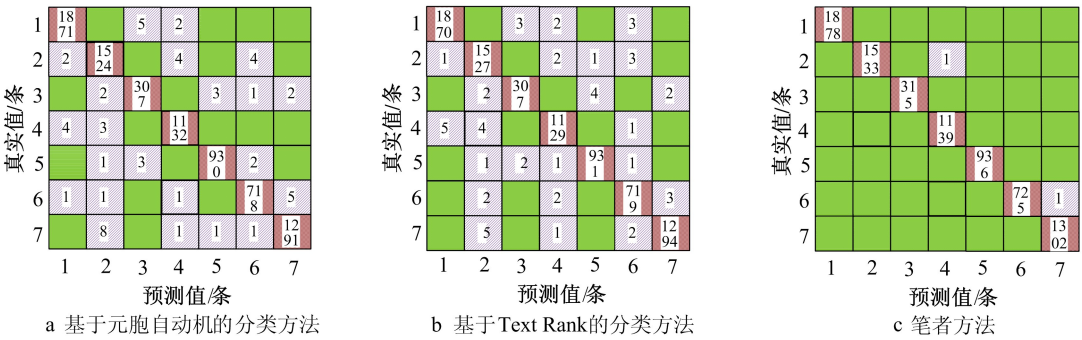
图 1 医院财务多源异构数相似性

分组结果可视图

Fig. 1 Visualization of the grouping results

of multi-source heterogeneous number

similarity in hospital finance



1. 现金管理, 2. 应收账款, 3. 负债, 4. 医疗收入, 5. 医疗成本, 6. 医疗行业, 7. 投资与融资

图2 3种方法的医院财务多源异构数据的分类混淆矩阵对比结果

Fig.2 Comparison results of classification confusion matrix for multi-source heterogeneous data in hospital finance using three methods

从图3可看出,所提方法各类别的数据内聚度均明显高于对比方法,特别是第6个类别内的数据点,内聚度值更接近100°,表明所提方法的分类准确性更高。这是因为所提方法通过混合聚类算法中K-means算法对蚁群聚类算法的修正,不仅能将相似的数据条目精准分组到同一个集合中,而且能更有效地将目标数据集内最相似的多源异构数据点聚集在一起,形成更加紧凑的类别簇,进而得到更好的分类效果。

3 结 语

笔者组合蚁群聚类算法与K-means算法,设计出混合聚类算法,提出一种多源异构数据自动分类方法,实现多源异构数据的自动分类,提高分类结果的准确性。为进行有效的隐私保护,未来可以考虑引入深度学习等技术,进一步优化HHO算法和混合聚类算法在隐私保护方面的能力,确保数据的安全性和隐私性。

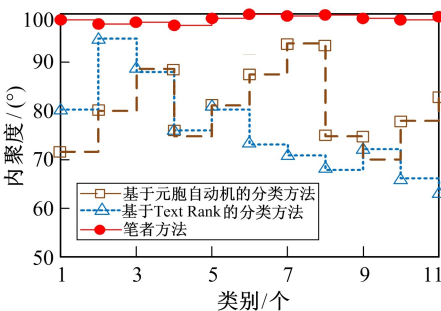


图3 3种方法的医院财务多源异构数据分类内聚度对比结果

Fig.3 Comparison results of cohesion of multi-source heterogeneous data classification in hospital finance using three methods

参考文献:

[1] KAMIL M Z, KHAN F, AHMED A S. Multi-Source Heterogeneous Data Integration for Incident Likelihood Analysis [J]. Computers & Chemical Engineering: An International Journal of Computer Applications in Chemical Engineering, 2024, 185 (6): 1047-1072.

[2] TAVAKOLI S S, MOZAFFARI A, DANAEI A, et al. Explaining the Effect of Artificial Intelligence on the Technology Acceptance Model in Media: A Cloud Computing Approach [J]. The Electronic Library, 2023, 41(1): 1-29.

[3] KENGER M N, OZCEYLAN E. A Hybrid Approach Based on Mathematical Modelling and Improved Online Learning Algorithm for Data Classification [J]. Expert Systems with Applications, 2023, 218: 1112-1127.

[4] 林少娃, 陈奕汝, 顾洁, 等. 基于隐含狄利克雷分布主题模型和特征级异构数据融合的电力故障主动性预警研究 [J]. 电子器件, 2022, 45(2): 432-438.

LIN S W, CHEN Y R, GU J, et al. Proactive Warning System Based on Electronic Power User Interactive Complaint Text and Multi-Source Heterogeneous Big Data Analysis [J]. Chinese Journal of Electron Devices, 2022, 45 (2): 432-438.

[5] 韩璐, 陈威宇, 张斐, 等. 差异化需求下的非关系型分布式报送信息大数据分类方法 [J]. 电信科学, 2023, 39(6): 114-121.

HAN L, CHEN W Y, ZHANG F, et al. Big Data Classification Method of Non Relational Distributed Submission Information under Differentiated Requirements [J]. Telecommunications Science, 2023, 39(6): 114-121.

- [6] 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.
LI L, SUN S J, ZHU K S, et al. Research on Multi-Source Heterogeneous Data Fusion Method for Electric Meteorological Data [J]. Electronic Design Engineering, 2024, 32(16): 178-182.
- [7] 梁越, 刘晓峰, 李权树, 等. 面向司法文本的不均衡小样本数据分类方法 [J]. 计算机应用, 2022, 42(S2): 118-122.
LIANG Y, LIU X F, LI Q S, et al. Classification Method for Unbalanced and Small Sample Data in Judicial Documents [J]. Journal of Computer Applications, 2022, 42(S2): 118-122.
- [8] KARNAVAS Y L, NIVOLIANITI E. Harris Hawks Optimization Algorithm for Load Frequency Control of Isolated Multi-Source Power Generating Systems [J]. International Journal of Emerging Electric Power Systems, 2024, 25(4): 555-571.
- [9] 范莉莉, 卢桂馥, 唐肝翌, 等. 基于 Hessian 正则化和非负约束的低秩表示子空间聚类算法 [J]. 计算机应用, 2022, 42(1): 115-122.
FAN L L, LU G F, TANG G Y, et al. Low-Rank Representation Subspace Clustering Method Based on Hessian Regularization and Non-Negative Constraint [J]. Journal of Computer Applications, 2022, 42(1): 115-122.
- [10] 罗曦, 熊贤祝, 刘勇进. 求解低秩密度矩阵约束最小二乘问题的优函数罚方法 [J]. 福州大学学报(自然科学版), 2024, 52(2): 127-133.
LUO X, XIONG X Z, LIU Y J. Majorized Penalty Algorithm for the Least Squares Problem with the Low Rank Density Matrix Constraint [J]. Journal of Fuzhou University (Natural Science Edition), 2024, 52(2): 127-133.
- [11] 邵磊, 许昊天. 一种监测母线槽的无线传感器信号采样策略 [J]. 天津理工大学学报, 2024, 40(4): 83-89.
SHAO L, XU H T. A Wireless Sensor Signal Sampling Strategy to Monitor the Parent Line Slot [J]. Journal of Tianjin University of Technology, 2024, 40(4): 83-89.
- [12] 赵柏, 林敏, 肖圣杰, 等. 基于速率分割的可重构智能表面辅助星地融合网络鲁棒安全传输方案 [J]. 通信学报, 2023, 44(12): 50-60.
ZHAO B, LIN M, XIAO S J, et al. Rate Splitting Based Robust Secure Transmission Scheme in RIS-Assisted Satellite-Terrestrial Integrated Network [J]. Journal on Communications, 2023, 44(12): 50-60.
- [13] 张红梅, 武江南, 赵永梅, 等. 基于截断 p -Shrinkage 范数的航空发动机数据重构 [J]. 北京航空航天大学学报, 2024, 50(1): 39-47.
ZHANG H M, WU J N, ZHAO Y M, et al. Aero-Engine Data Reconstruction Based on Truncated p -Shrinkage Norm [J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(1): 39-47.
- [14] 朱承元, 孙辰欣, 赵立刚. 基于蚁群聚类算法的扇区复杂性分析 [J]. 计算机仿真, 2022, 39(7): 81-85.
ZHU C Y, SUN C X, ZHAO L G. Analysis and Verification of Sector Complexity Based on Ant Colony Clustering Algorithm [J]. Computer Simulation, 2022, 39(7): 81-85.
- [15] 程亚南, 王晓峰, 刘淞佐, 等. 求解多起点多旅行商问题的 K -Means 聚类信息传播算法 [J]. 科学技术与工程, 2022, 22(23): 10146-10154.
CHENG Y N, WANG X F, LIU S Z, et al. K -Means Clustering Information Propagation Algorithm for Multiple Depots Multiple Traveling Salesman Problem [J]. Science Technology and Engineering, 2022, 22(23): 10146-10154.
- [16] 孙辉, 钟诚. 融合过滤和相似度计算的高错误率基因组数据敏感序列识别 [J]. 小型微型计算机系统, 2023, 44(6): 1227-1235.
SUN H, ZHONG C. Recognizing Sensitive Sequences from Genomic Data with High Error Rate Integrating Filter and Similarity Calculation [J]. Journal of Chinese Computer Systems, 2023, 44(6): 1227-1235.
- [17] 李炜卓, 卢冰洁, 杨骏铭, 等. 基于网络表示学习的车险欺诈溯源分析研究 [J]. 计算机科学, 2023, 50(2): 300-309.
LI W Z, LU B J, YANG J M, et al. Study on Abductive Analysis of Auto Insurance Fraud Based on Network Representation Learning [J]. Computer Science, 2023, 50(2): 300-309.
- [18] 吕英杰. 基于蚁群算法和冲突消除的多机器人移动路径优化研究 [J]. 机械设计与制造工程, 2023, 52(12): 55-61.
LÜ Y J. Research on Multi-Robot Moving Path Optimization Based on Ant Colony Algorithm and Conflict Elimination [J]. Machine Design and Manufacturing Engineering, 2023, 52(12): 55-61.

- [19]PARK S, KANG D, PAIK J. Cosine Similarity-Guided Knowledge Distillation for Robust Object Detectors [J]. Scientific Reports, 2024, 14(1): 1432-1461.
- [20]SHWETA R, S. S, ARK K K. Design of an IoT Platform for Data Analytics Based Fault Detection and Classification in Solar PV Power Plants Using CFKC and ODENN [J]. International Journal of Modeling, Simulation, and Scientific Computing, 2023, 14(2): 429-441.
- [21]李颜平, 吴刚. 基于典型数据集的数据预处理方法对比分析 [J]. 沈阳工业大学学报, 2022, 44(2): 185-192.
LI Y P, WU G. Comparative Analysis of Data Preprocessing Methods Based on Typical Data Set [J]. Journal of Shenyang University of Technology, 2022, 44(2): 185-192.

(责任编辑: 刘东亮)