Statistics and Application 统计学与应用, 2025, 14(4), 201-215
Published Online April 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.144102

Hans汉斯

基于差异化创造性搜索算法的数据属性噪声修复方法及其应用

张冬青，韩冉冉，陈继强*

河北工程大学数理科学与工程学院，河北 邯郸

收稿日期：2025年3月18日；录用日期：2025年4月9日；发布日期：2025年4月22日

摘　要

水质数据在采集和传输过程中常受到属性噪声干扰，严重影响数据的准确性和可靠性，以致降低水质评价与分类的效果。为此，提出了一种基于差异化创造性搜索算法的数据属性噪声修复方法并将其应用于水质评价。该方法通过差异化创造性搜索理论，设计了一种高效的属性噪声修复框架，通过直觉模糊集识别属性噪声样本，利用熵权法计算每个属性的权重，得到了属性值的噪声分数，有效识别了数据中的属性噪声。为验证方法的有效性，在8个公共数据集上与其他6种属性噪声预处理方法进行了对比实验，并应用于黄河流域七里铺断面地表水水质评价问题。

关键词

分类，属性噪声，差异化创造性搜索，直觉模糊集，水质数据

Data Attribute Noise Repair Method Based on Differential Creative Search Algorithm and Its Applications

Dongqing Zhang, Ranran Han, Jiqiang Chen*

School of Mathematics and Physics, Hebei University of Engineering, Handan Hebei

Received: Mar. 18th, 2025; accepted: Apr. 9th, 2025; published: Apr. 22nd, 2025

Abstract

Water quality data are often interfered with attribute noise in the process of collection and

*通讯作者。

文章引用: 张冬青, 韩冉冉, 陈继强. 基于差异化创造性搜索算法的数据属性噪声修复方法及其应用[J]. 统计学与应用, 2025, 14(4): 201-215. DOI: 10.12677/sa.2025.144102

transmission, which seriously affects the accuracy and reliability of the data, and thus reduces the effectiveness of water quality evaluation and classification. Therefore, a data attribute noise repair method based on the Differential Creative Search (DCS) algorithm is proposed and applied to water quality evaluation. The method designs an efficient attribute noise repair framework through the theory of differentiated creative search, identifies attribute noise samples through intuitionistic fuzzy sets, calculates the weight of each attribute by using entropy weighting method, and obtains the noise scores of the attribute values, which efficiently identifies the attribute noise in the data. In order to verify the effectiveness of the method, comparative experiments with six other attribute noise preprocessing methods were conducted on eight public datasets and applied to the problem of surface water quality assessment in the Qilipu section of the Yellow River Basin.

Keywords

Classification, Attribute Noise, Differential Creative Search, Intuitionistic Fuzzy Sets, Water Quality Data

1. 引言

地表水作为城乡居民生活用水、工业和农业生产用水的主要来源，发挥着至关重要的作用。然而，随着人口增长和工业化加速推进，水污染问题日益严峻，成为全球面临的重大挑战之一[1]。

水质监测作为评估和管理水质的重要手段，水质数据的质量和可靠性对环境保护和污染治理至关重要[2]-[5]。然而，在数据采集、整理、分析过程中常常会引入属性噪声，导致水质监测数据质量参差不齐，从而导致对水质状况的误判或漏判，影响环境保护和污染治理效果。

当前水环境质量评价(即水质评价)方法主要有两种，一种是以水质物理化学指标实测值为依据的评价方式；另一种是以水生物种群与水质的关系为依据的生物学评价方式。第一种水环境质量评价方法运用较为普遍，例如单因子评价法[6]、主成分分析法[7]、灰色关联法[8]、模糊评价法[9]、神经网络[10]与支持向量机[11]等。然而，这些方法大都是基于水质数据直接评价，对水质数据进行噪声处理后再评价的研究相对较少。

为降低属性噪声的影响，很多学者提出了对数据属性噪声的处理方法。常见的处理方法是使用预处理技术，其中一类方法是去除噪声样本。编辑最近邻(Edited Nearest Neighbors, ENN) [12]是基于相似性滤波器的主要代表，它删除了那些类别与其最近邻居中的大多数不一致的样本。高一致性随机森林(High Agreement Random Forest, HARF) [13]会删除随机森林分类器实际类别标签置信度较低的噪声样本。然而，使用噪声滤波器在提高数据质量的同时，会使数据集中的有效信息流失。因此，一些学者提出了属性噪声修复的方法以提高数据集的质量。Teng [14]根据属性之间以及属性和类别(标签)之间的相互依赖性，利用剩余属性和类别的值来预测一个属性的值，从而替换可能的噪声值，这样基于修正后的数据构建分类器，以提高算法的预测能力。王石等[15]提出了一个基于聚类分析的数据属性噪声修复算法。该算法通过聚类分析确认数据产生噪声的具体属性，然后根据其余的干净数据来修复噪声，同时统计噪声在属性上的分布规律。Zhai 和 Zhang [16]提出了融合稀疏和低秩先验的鲁棒主成分分析方法，可更准确地识别异常值和噪声。Sáez 和 Corchado [17]提出了一种新的属性噪声修复方法。该方法计算数据集中每个属性值

的错误分数，然后通过优化过程来纠正其潜在的噪声。然而，该方法对每个属性值错误分数的计算中并未考虑到属性对分类的不同贡献。并且，采用的模拟退火算法是一种元启发式算法，在全局搜索方面表现出色，但其收敛速度相对较慢，尤其是在接近最优解时。

差异化创造性搜索(Differentiated Creative Search, DCS)算法[18]作为一种结合差异化知识获取、收敛思维与发散思维的智能优化算法，具有较快的收敛速度，能高效处理高维非线性数据、鲁棒的噪声识别与修复能力、可扩展性强、易于与其他方法结合等优势。

考虑到水质数据通常具有高维性和非线性特征，且常常包含由传感器故障、传输误差或环境干扰引起的属性噪声，严重影响了水质评价的准确性。为此，为进一步提高水质评价的准确性和可靠性，构建了一种基于差异化创造性搜索算法的数据属性噪声修复方法(Attribute Noise Repair Method Based on Differential Creative Search Algorithm, DCSANR)并将其应用于水质评价。该方法通过计算评价因子权重，并基于此权重得到数据每个属性值的含噪声分数。然后，对含噪声的属性值通过优化过程传递，利用差异化创造性搜索算法对数据属性噪声逐步修复。最后，探索了该方法在黄河流域七里铺断面地表水水质评价问题中的应用。

本文结构如下。第 2 节简要介绍了差异化创造性搜索算法(DCS)。第 3 节构建了基于 DCS 算法的数据属性噪声修复方法。第 4 节分析了基于 DCS 的数据属性噪声修复方法在水质数据属性噪声修复中的应用。第 5 节为结论与展望。

2. 差异化创造性搜索算法概述

2.1. 基本原理

差异化创造性搜索(DCS)是一种突破性的优化算法，它将独特的知识获取过程与创造性的现实主义范式相结合，从而改变优化策略。DCS 的主要目标是通过采用新提出的双重战略方法来提高决策效率，该方法在以团队为基础的框架内平衡发散和收敛思维，主要步骤包括团队初始化、差异化知识获取、收敛思维、发散思维、团队多元化、回顾性评估[18]。

2.2. 核心步骤

(1) 初始化

DCS 算法的初始步骤是采用随机初始化方式生成包含 NP 个成员的团队 X，每个成员对应问题的一个候选解，由 D 维向量 $x_i=\left(x_{i1},x_{i2},\cdots x_{iD}\right)$ 表示，第 i 个个体的第 j 维具体值通过式(1)确定：

$$x_{i,d}=lb_d+r_{i,d}\times\left(ub_d-lb_d\right),r_{i,d}\sim U\left(0,1\right) \tag{1}$$

其中 lb_d 和 ub_d 代表第 d 维的下界和上界，随机变量 $r_{i,d}$ 服从在 0 和 1 之间均匀分布确保每个成员的初始位置在可行空间内随机分布。

在初始化之后，评估每个 x_i 以产生其适应度，随后对 X 中的所有成员按适应度升序排序，较低的值表示有较好表现。

(2) 差异化知识获取

$$\begin{aligned}\eta_{i,t}&=\frac{1}{2}\left(\left[U\left(0,1\right)\times\varphi_{i,t}\right]+\begin{cases}1, & \text{若}U(0,1)\le\varphi_{i,t}\\ 0, & \text{其他}\end{cases}\right)\\ \varphi_{i,t}&=0.25+0.55\times\sqrt{R_{i,t}/N}\end{aligned} \tag{2}$$

其中 $R_{i,t}$ 为第 t 次迭代开始时 $\boldsymbol{x}_i$ 的排名，$\varphi_{i,t}$ 为第 t 次迭代时 x_i 的 φ 系数值。φ 系数是衡量某成员的知识

不完善程度的定量决定因素，φ 系数随知识差距的大小而变化；φ 值越高，意味着知识差距越大，表明成员更需要学习、吸收和整合新的知识或经验；反之，φ 值越小，知识不完善程度越小，说明成员的知识基础更全面、更扎实。

对一些成员而言，新知识可能会在某些方面带来比其他方面更大的变化。对其他成员而言，变化可能均匀地分布在所有维度上。即使知识的整体维度保持不变，这种可变性也使学习成为每个成员的独特体验。因此，差异化知识获取更类似于自然过程。第 t 次迭代时个体 x_i 的量化知识获取率的计算公式见式(2)。

差异化知识获取过程对每个 x_i 的作用可以使用式(3)执行

$$\begin{aligned} &j_{rand} \sim U\left(\{1,2,\ldots,D\}\right) \\ &v_{i,d} = \begin{cases} v_{i,d}, & \text{若} U(0,1) \le \eta_{i,t} \text{或} d = j_{rand} \\ x_{i,d}, & \text{其他} \end{cases} \end{aligned} \tag{3}$$

其中 $V_{i,t}$ 为第 t 次迭代的试验成员，$v_{i,d}$ 为试验成员 $V_{i,t}$ 的第 d 维度元素。j_{rand} 是从 1 到 D 之间随机选择的整数。

(3) 收敛思维

该策略依赖最佳执行者的知识基础，并结合了两个团队成员的随机贡献。x_i 通过整合团队成员 x_{r1} 和 x_{r2} 提供的信息，提炼团队领导者的知识 x_{best}。更新公式为

$$v_{i,d} = w \times x_{best,d} + \lambda_t \times \left(x_{r2,d} - x_{i,d}\right) + \omega_{i,t} \times \left(x_{r1,d} - x_{i,d}\right) \tag{4}$$

其中加权因子 w 调整最佳向量 x_{best} 的影响，默认值为 1。λ_t 是第 t 次迭代时成员 x_i 的 λ 系数值。λ_t 的计算公式为 $\lambda_t = 0.1 + 0.518 \times \left(1 - \sqrt{NFE_t / NFE_{\max}}\right)$，$NFE_t$ 表示当前函数评估的次数，$NFE_{\max}$ 表示函数评估的最大次数。$\omega_{i,t}$ 服从(0，1)上的均匀分布，表示成员的学习强度状态。x_{r1} 是从 $\{x_1, x_2, \cdots, x_{NP}\}$ 中随机选择的成员，且 $x_{r1} \ne x_i \ne x_{best}$。$x_{r2}$ 是从 $\{x_{ngs+1}, \cdots, x_{NP}\}$ 中随机选择的成员，并且满足 $x_{r2} \ne x_{r1} \ne x_i \ne x_{best}$。参数 ngs 表示高性能成员的数量。对于 30 人以下的群体，设置 ngs 的最小值为 6，对于较大的群体，$ngs = 100\% * p * NP$， p 设置为 0.2。

(4) 发散思维

在更新成员位置时，除了成员现有的知识外，还应考虑创造性的元素。因此成员的更新公式为

$$V_i = x_{r1} + Lk\left(\alpha, \sigma\right) \tag{5}$$

其中 $Lk(\alpha, \sigma)$ 是具有控制参数 α 和 σ 的 Linnik 分布随机数生成器。

(5) 边界处理

$$v_{i,d} = \begin{cases} \dfrac{x_{i,d} + lb_d}{2}, v_{id} < lb_d \\ \dfrac{x_{i,d} + ub_d}{2}, v_{id} > ub_d \end{cases} \tag{6}$$

其中 $v_{i,d}$ 表示第 t 次迭代的试验成员 $V_{i,t}$ 第 d 维度的元素。lb_d 和 ub_d 分别表示第 d 维的下限和上限。

(6) 团队多元化

对于低性能的团队成员使用新成员替换，公式如下：

$$V_{NP} = lb + U\left(0,1\right) \times \left(ub - lb\right) \tag{7}$$

其中 V_{NP} 表示第 NP 个试验成员。

(7) 回顾性评估

$$x_{i,t+1}=\begin{cases}V_{i,t}, & f\left(V_{i,t}\right)\le f\left(x_{i,t}\right)\\ x_{i,t}, & \text{其他}\end{cases} \tag{8}$$

其中 $x_{i,t+1}$ 是第 t 次迭代的成员， $f(\cdot)$ 是对应的目标值。

回顾性评估是团队开发的关键工具，通过建立评估标准和分析历史绩效数据来识别趋势与改进方向，其选择机制与最优性能追踪可由式(8)表示。

最佳执行者的确定公式如下：

$$x_{best,t}=\begin{cases}x_{i,t+1}, & f\left(x_{i,t+1}\right)>f\left(x_{best,t}\right)\\ x_{best,t}, & \text{其他}\end{cases} \tag{9}$$

3. 基于 DCS 的数据属性噪声修复方法

为识别数据的属性噪声并对属性值进行修复，构建了如下基于 DCS 的数据属性噪声修复方法。

3.1. 数据集归一化

为确保所有属性在计算样本之间距离时具有相同的相关性，对数据集进行归一化处理。采用最大最小归一化方法，公式如下：

$$x_{ij}=\frac{x_{ij}-\min\left(x_{\cdot j}\right)}{\max\left(x_{\cdot j}\right)-\min\left(x_{\cdot j}\right)} \tag{10}$$

3.2. 基于直觉模糊集的噪声样本识别

假设给定输入空间上的训练数据集为 $D=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_n,y_n\right)\right\}$， $y_i\in\{+1,-1\}$ 是样本 $x_i\left(i=1,2,\cdots,n\right)$ 对应的类标签。任意样本 x_i 对正负类的隶属函数和非隶属函数定义如下。

(1) 对所有正负类样本，分别计算初始类中心

$$C_0^{\pm}=\frac{1}{l_0^{\pm}}\sum_{y_i=\pm1}x_i \tag{11}$$

其中 $C_0^{\pm}$ 分别表示正、负类的类中心， $l_0^{\pm}$ 分别表示正负类所包含样本点的数量。

(2) 计算初始距离

通过正负类样本点的初始类中心，分别计算正负类所有样本点到各自类中心的初始距离

$$d_{0,i}^{\pm}=\left\|\boldsymbol{x}_i-C_0^{\pm}\right\| \tag{12}$$

(3) 初步噪声样本识别

在噪声样本识别方法的基础上，选择四分位数法检测潜在噪声样本。箱线图中，在 $Q_3+1.5Q$ 和 $Q_1-1.5Q$ 处有两条线段，给出了判断是否为离群点的方式。由于样本与类中心的距离越近，属于该类的程度越大，因此距离大于 $Q_3+1.5Q$ 的样本被认为是噪声样本。如果样本与其类中心的初始距离满足式(13)，即认为该样本为初步噪声样本。

$$d_{0,i}>Q_3+1.5Q \tag{13}$$

其中 $Q=Q_3-Q_1$， Q_1 为 d_0 的下四分位数， Q_3 为 d_0 的上四分位数。

(4) 更新类中心和类半径

$$\begin{aligned} C^{\pm} &= \frac{1}{l^{\pm}} \sum_{y_i=\pm 1} x_i, \\ d^{\pm} &= \left\| x_i - C^{\pm} \right\|, \\ r^{+} &= \max_{y_i=+1} \left\| x_i - C^{+} \right\|, \\ r^{-} &= \max_{y_i=-1} \left\| x_i - C^{-} \right\|. \end{aligned} \tag{14}$$

将初步识别后满足条件式(13)的可能噪声样本进行删除，分别按式(14)重新计算剩余样本的类中心、样本点到其类中心的距离和类半径。

(5) 计算隶属函数

基于式(14)更新后的类中心和类半径，按式(15)计算样本 x_i 对正负类的隶属函数。

$$\mu(x_i) = \begin{cases} 1 - \dfrac{\left\| x_i - C^{+} \right\|}{r^{+} + \varepsilon}, & y_i = +1 \\ 1 - \dfrac{\left\| x_i - C^{-} \right\|}{r^{-} + \varepsilon}, & y_i = -1 \\ 0, & \text{初步噪声样本} \end{cases} \tag{15}$$

(6) 计算非隶属函数

正负类的非隶属函数可由基于样本点某个邻域内异类样本占所有样本的比例计算[19]，取值很大程度上依赖样本点邻域半径的选择。如果样本分布稀疏，并且选择了很小的邻域半径，会导致样本邻域内包含的样本过少，从而影响非隶属函数的准确性。考虑到核方法通过将原始特征空间映射到更高维特征空间，使得数据在新特征空间中更加密集，从而降低样本分布稀疏的影响。因此，采用核 K 近邻方法来计算正负类的非隶属函数。

对样本 x_i，通过核函数将其映射到高维特征空间，计算核空间中样本间的距离，找到其在核空间中的 k 近邻 $\{x_{i1}, x_{i2}, \cdots, x_{ik}\}$。将 $\{x_{i1}, x_{i2}, \cdots, x_{ik}\}$ 中与 x_i 异类的样本的个数记为 n_i。核空间中样本间的距离公式为：

$$d_{ij}^2 = K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j) \tag{16}$$

$$\rho(x_i) = \frac{n_i}{k} \tag{17}$$

正负类的非隶属函数为：

$$\upsilon(x_i) = (1 - \mu(x_i))\rho(x_i) \tag{18}$$

满足 $0 \le \upsilon(x_i) + \mu(x_i) \le 1$。

因此，通过改变隶属函数和非隶属函数的计算方法来实现对直觉模糊集的改进。若样本 x_i 对正负类的隶属度小于非隶属度($\mu(x_i) < \upsilon(x_i)$)则判定为含噪声的样本。这样，就建立了改进的基于直觉模糊集的含噪声数据识别方法。

根据初步识别的含噪声样本，从非噪声样本中随机选取总体的 20%作为测试集，其余样本作为训练集进行优化属性值。

在属性噪声优化中涉及 3 种数据集，分别为训练集 D_t、测试集 V_t、输出数据集 D_{out}，测试集保持不变，数据集 D_t， D_{out} 在开始迭代过程($t = 0$)之前按初始化为 $D_0 = D_{in}, D_{out} = D_{in}$。

3.3. 属性噪声检测与定位

为有效识别含有潜在噪声的属性值 x_{ij}，给每个属性值分配噪声分数。鉴于不同属性对分类任务的贡献度不同，且噪声水平与属性值紧密相关，计算属性值的噪声分数需综合考虑各属性的贡献度及其对应属性值的错误程度。

本文使用熵权法确定属性的贡献度。熵权法确定权重的步骤[20]如下：

(1) 计算第 j 项指标下第 i 个样本占该指标的比重：

$$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}},i=1,2,\cdots,n;j=1,2,\cdots,m \tag{19}$$

(2) 计算第 j 项指标的熵值

$$e_j=-k\sum_{i=1}^{n}p_{ij}\ln\left(p_{ij}\right) \tag{20}$$

其中 $k=1/\ln(n)>0$，满足 $e_j\geq 0$。

(3) 计算信息熵冗余度

$$d_j=1-e_j \tag{21}$$

(4) 计算各评价因子的权值

$$w_j=\frac{d_j}{\sum_{j=1}^{m}d_j} \tag{21}$$

属性值 x_{ij} 的错误程度利用样本的 k 近邻 $x_{\eta_q}\left(q=1,2,\cdots,k\right)$ (实验中设置 $k=10$)计算，计算公式为

$$S_{error}\left(x_{ij}\right)=\left|x_{ij}-\frac{\sum_{q=1}^{k}x_{\eta_q j}}{k}\right| \tag{22}$$

因此，属性值 x_{ij} 的噪声分数为

$$S_{noise}\left(x_{ij}\right)=w_j\cdot S_{error}\left(x_{ij}\right) \tag{23}$$

根据属性值的噪声分数，按降序排序，选择前 $p\%$ (在本文的实验中 $p=5$)的属性值进行修复，从而允许从具有较大噪声分数的属性值到具有较小噪声分数的属性值逐步修复。

3.4. 属性值优化

利用 DCS 算法对 3.3 节中识别出的属性噪声进行优化，以提升数据集的质量。具体而言，将每个所需修复的属性值看作搜索空间中的一个独立维度，每个维度上的候选解即为可能的属性值。通过 DCS 算法的不断迭代和调整，可以在搜索空间中找到最优的属性值组合，从而实现对属性噪声的有效修复。

DCS 算法优化属性值的步骤如下：

(1) 属性值初始化

利用 DCS 算法的种群初始化步骤，生成一组初始属性值集合。搜索空间的上限和下限分别设为某样

本所在类的属性值最大值和最小值，以确保搜索范围合理且包含所有可能的解。

(2) 适应度评估

对每个属性值集合进行适应度评估，以衡量其表现优劣。适应度函数根据分类准确率来设计。通过计算的适应度值对属性值集合进行排序，以便后续步骤中的选择操作。

(3) 差异化知识获取

根据适应度排名，计算每个属性值集合的知识获取率。知识获取率反映了从当前状态向更优状态转变的潜力或速度。通过差异化知识获取策略，我们可以对不同表现的属性值集合采取不同的操作，以平衡探索和利用的关系。

(4) 属性值集合更新

利用收敛思维和发散思维等策略对属性值集合进行更新。发散思维则强调探索搜索空间中的不同区域，以增加找到全局最优解的可能性；收敛思维侧重于在当前最优解附近进行细致搜索，以寻找更优的解。结合交叉、变异等遗传操作，生成新的属性值集合，并替换全部旧集合，以形成新的种群。

(5) 评估与更新数据集。

将原数据集中含噪声的属性值替换为更新后属性值集合中的属性值。通过目标函数评估当前解的质量，并保留高质量的解。如果满足收敛条件或达到最大迭代次数，则停止迭代；否则，返回步骤(2)继续迭代优化。

通过上述步骤，DCS 算法能够逐步优化属性值，减少属性噪声对数据集质量的影响，从而提升分类等任务的性能。

3.5. 迭代检查

在更新迭代过程中，需要定义适应度函数来衡量数据集的质量，将适应度函数设置为

$$F\left(D_t,V_t\right)=ACC\left(k_2-NN,D_t,V_t\right) \tag{24}$$

其中 ACC 是分类准确率，k_2-NN 为分类器，D_t 为训练集，V_t 为测试集。$F(D_t,V_t)$ 越高，数据集 D_t 的质量越高。因此，初始数据集的质量为 $f_0=F(D_0,V_0)$。

对新数据集 D_t'，计算其适应度 f_t。如果适应度 f_t 大于最优解的适应度，则 $D_{t+1}=D_t',D_{out}=D_t'$。否则，$D_{out}$ 不更新，$D_{t+1}=D_t'$。

3.6. 停止准则

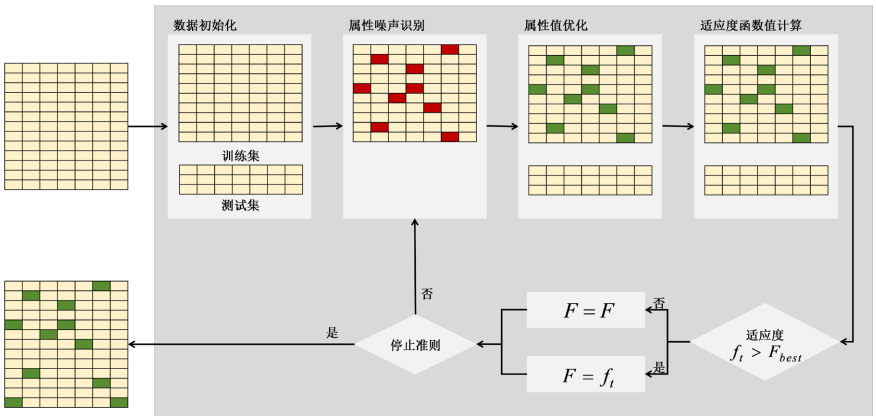

Figure 1. Main process of attribute noise repair method based on DCS
图 1. 基于 DCS 的属性噪声修复方法的主要流程

迭代过程的最后一步是检查停止准则。基于两个主要标准，如果满足其中任何一个，迭代过程就会停止，并返回最终数据集 D_{out}。停止准则的两个条件是：

(1) 最大迭代次数。这在执行的迭代次数 $t > t_{\max}$ 时发生。在实验中，设置 $t_{\max} = 20$，因为它提供了系统的良好性能和速度之间的平衡。

(2) 没有改进的迭代。若连续 t_{imp} 次数据集的质量无改进，即停止迭代。在实验中设置 $t_{imp} = 3$。

基于 DCS 的属性噪声修复方法的主要流程如图 1 所示。

4. 基于 DCS 的属性噪声修复方法在水质数据属性噪声修复中的应用

4.1. 数据来源

4.1.1. 公共数据集

为验证实验效果，选用来自 UCI (https://archive.ics.uci.edu/)中具有不同属性、不同样本数量和不同类别数量的 8 个公开数据集进行实验，数据集具体信息见表 1。在 8 个数据集上添加了不同比例的属性噪声，即从每个数据集中随机选择 5%、10%、15%、20%、25%、30%的属性值添加高斯噪声。

Table 1. Dataset description
表 1. 数据集描述

数据集	样本数量	特征数量	类别数量	各类别样本数量
Haberman	306	3	2	225, 81
Diabetes	768	8	2	268, 500
Heart[1]	1190	11	2	561, 629
Wdbc	569	30	2	357, 212
Wine	178	13	3	59, 71, 48
Vowel	871	3	6	72, 89, 172, 151, 207, 180
Aggregation	788	2	7	170, 34, 273, 102, 130, 45, 34
Ecoli	336	8	8	143, 77, 2, 2, 259, 20, 5, 52

注：Heart 是由 Statlog、Cleveland 和 Hungarian 3 个心脏病数据集合并而成的。

4.1.2. 地表水水质数据

2023 年 1 月 1 日~2024 年 5 月 21 日中国环境监测总站发布的黄河流域七里铺断面地表水水质数据，剔除站点维护数据，共 2489 条数据，筛选后该站点水质类别共 5 类，分别为 II 类(764 条)、III 类(1466 条)、IV 类(223 条)、V 类(21 条)、劣 V 类(15 条)。根据污染源调查和监测数据，结合水质评价参数选择原则，选择水温、pH 值、溶解氧、电导率、浊度、高锰酸盐指数、氨氮、总磷、总氮 9 个水质指标作为评价因子，建立水质评价体系。根据《地表水环境质量标准》(GB3838-2002)，各项因子可分为 5 个等级，具体分级标准见表 2。

Table 2. Standard limits of some basic items of surface water environmental quality standards
表 2. 部分地表水环境质量标准基本项目标准限值

项目	I 类	II 类	III 类	IV 类	V 类
水温		人为造成的环境水温变化应限制在： 周平均最大温升 ≤1；周平均最大温降 ≤2			

续表

pH 值			6~9		
溶解氧≥	7.5	6	5	3	2
高锰酸盐指数≤	2	4	6	10	15
氨氮≤	0.15	0.5	1.0	1.5	2.0
总磷≤	0.02 (湖库 0.01)	0.1 (湖库 0.025)	0.2 (湖库 0.05)	0.3 (湖库 0.1)	0.4 (湖库 0.2)
总氮≤	0.2	0.5	1.0	1.5	2.0

4.2. 评价指标

混淆矩阵[21]常用于统计不同类别样本的分类情况，二分类混淆矩阵见表 3。TP 和 TN 分别表示正类(少数类)或负类(多数类)预测正确的样本数，FN 和 FP 分别表示正类或负类预测错误的样本数。

Table 3. Confusion matrix
表 3. 混淆矩阵

数据集	预测正类	预测负类
真实正类	TP (True Positive)	FN (False Negative)
真实负类	FP (False Positive)	TN (True Negative)

通过混淆矩阵，可计算评价分类算法性能的指标，如召回率(Recall)、准确率(Accuracy)、精确率(Precision)、F1-Score 等，计算公式分别如下：

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - Score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{25}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{26}$$

$$AUC = \frac{1}{2}\left(1 + \frac{TP}{TP + FN} - \frac{FP}{FP + TN}\right) \tag{27}$$

本文采用 F1-Score、受试者工作特征曲线下的面积 AUC (Area under curve, AUC)以及准确率 3 种常用的分类效果评价指标进行评价。F1-Score 同时考虑了精度和召回率，是二者的调和平均，可以全面评估分类器对少数类的识别情况。AUC 是受试者工作特征曲线下的面积，它与真正率和假正率指标有关，能够同时评估分类器对两类的分类能力。准确率表示算法正确分类的样本数占总样本数的比例，可直接地评价算法性能。

4.3. 结果分析

为验证本文构建的 DCSANR 方法的有效性，将 DCSANR 与 ANCES [17]、EF [22]、AENN [23]、BBNR [24]、ENN [12]、RPCA-SL [16]等 6 种噪声处理方法对数据进行处理后，采用 KNN 分类器进行分类。实验采用五折交叉验证，取实验结果的平均值作为最终结果。

4.3.1. 公共数据集结果

表 4 展示了各种方法在公开数据集上添加不同属性噪声水平下的分类准确率。从表中可以看出，随着噪声水平的增加，各方法的分类准确率整体呈现下降趋势。在 5%的噪声水平下，经过 DCSANR 方法对 8 个数据集的属性噪声进行修复后，有 6 个数据集的 KNN 分类准确率达到最高；在 10%、15%、20%、25%和 30%的噪声水平下，分别有 2、4、6、5 和 5 个数据集的 KNN 分类准确率取得最优值。在所有数据集上，DCSANR 对属性噪声修复后的数据 KNN 分类准确率最高，达到 0.8314。这表明 DCSANR 对含属性噪声的公共数据集有较好的修复效果。

Table 4. Classification accuracy of 7 methods under different noise levels
表 4. 不同噪声水平下 7 种方法的分类准确率

数据集	噪声水平	DCSANR	ANCES	EF	AENN	BBNR	ENN	RPCA-SL
Haberman	5%	0.8361	0.8033	0.6994	0.7287	0.7450	0.7516	0.7614
	10%	0.7049	0.6721	0.6994	0.7483	0.7581	0.7582	0.6602
	15%	0.7393	0.7230	0.7125	0.7190	0.7256	0.7354	0.6831
	20%	0.7213	0.7869	0.7092	0.7322	0.7354	0.7452	0.7284
	25%	0.8525	0.6721	0.6961	0.7386	0.7386	0.7386	0.6984
	30%	0.8033	0.6885	0.6964	0.7322	0.7418	0.7452	0.6768
Diabetes	5%	0.7078	0.7044	0.7019	0.7044	0.7044	0.7122	0.7435
	10%	0.7208	0.7013	0.6967	0.6992	0.7083	0.7070	0.6823
	15%	0.8117	0.7258	0.7058	0.7006	0.7032	0.7032	0.8138
	20%	0.7792	0.6897	0.6522	0.6641	0.6823	0.6862	0.7757
	25%	0.7468	0.6875	0.6810	0.6888	0.6875	0.6784	0.7281
	30%	0.7338	0.6766	0.6797	0.6706	0.6563	0.6316	0.6786
Heart	5%	0.8319	0.7649	0.6740	0.7009	0.7126	0.7059	0.8298
	10%	0.8025	0.7402	0.6773	0.7093	0.6992	0.6916	0.8135
	15%	0.8445	0.7257	0.6774	0.6824	0.6748	0.6782	0.7812
	20%	0.7899	0.7257	0.6261	0.6320	0.6404	0.6379	0.7221
	25%	0.8109	0.7165	0.6396	0.6169	0.6127	0.6043	0.7139
	30%	0.7815	0.6311	0.6412	0.6009	0.5892	0.5892	0.6979
Wdbc	5%	0.9561	0.9314	0.9296	0.9314	0.9314	0.9296	0.9404
	10%	0.9323	0.9198	0.9086	0.9139	0.9104	0.9086	0.9369
	15%	0.9316	0.916	0.9104	0.9069	0.9052	0.9261	0.9274
	20%	0.9018	0.8968	0.8981	0.9016	0.9016	0.9015	0.9012
	25%	0.8526	0.8874	0.8873	0.8856	0.8838	0.8541	0.8921
	30%	0.8404	0.8492	0.8908	0.8591	0.8591	0.8592	0.8824
Wine	5%	0.8722	0.8722	0.7144	0.6756	0.6865	0.6803	0.8683
	10%	0.8444	0.8689	0.7032	0.6975	0.6751	0.6743	0.8571
	15%	0.8167	0.7567	0.7483	0.6746	0.6635	0.6690	0.8079

续表

	20%	0.7333	0.7161	0.6749	0.6746	0.6356	0.6579	0.7499
	25%	0.6889	0.6389	0.6913	0.6802	0.7081	0.7024	0.7091
	30%	0.6980	0.6247	0.6971	0.6859	0.6748	0.7081	0.6887
Vowel	5%	0.8908	0.8803	0.8601	0.8508	0.8508	0.8462	0.8542
	10%	0.8563	0.8656	0.8566	0.8497	0.8462	0.8496	0.853
	15%	0.8506	0.8618	0.8589	0.8520	0.8543	0.8485	0.8482
	20%	0.8621	0.8360	0.8463	0.8360	0.8359	0.8302	0.8299
	25%	0.8391	0.8242	0.8221	0.8175	0.8222	0.8245	0.8404
	30%	0.8218	0.7981	0.8142	0.8108	0.8085	0.7981	0.8094
Aggregation	5%	0.9962	0.9975	0.9962	0.9962	0.9962	0.9975	0.9924
	10%	0.9921	0.9892	0.9822	0.9873	0.9898	0.9906	0.9898
	15%	0.9886	0.9814	0.9822	0.9885	0.9885	0.9911	0.9837
	20%	0.9753	0.9627	0.9530	0.9657	0.9695	0.9720	0.9624
	25%	0.9573	0.9414	0.9462	0.9562	0.9549	0.9549	0.9401
	30%	0.9386	0.9305	0.9361	0.9373	0.9373	0.9386	0.9357
Ecoli	5%	0.9104	0.8735	0.8483	0.8513	0.8601	0.8424	0.9020
	10%	0.8209	0.8533	0.8362	0.8362	0.8302	0.8097	0.8282
	15%	0.7910	0.8059	0.8244	0.7887	0.7798	0.7799	0.8036
	20%	0.8060	0.7680	0.8040	0.7680	0.7560	0.7589	0.7802
	25%	0.8060	0.7648	0.7654	0.7801	0.7591	0.7382	0.7579
	30%	0.7164	0.7649	0.7411	0.7322	0.7145	0.6880	0.7145
Mean		0.8314	0.8003	0.7832	0.7825	0.7813	0.7798	0.8120

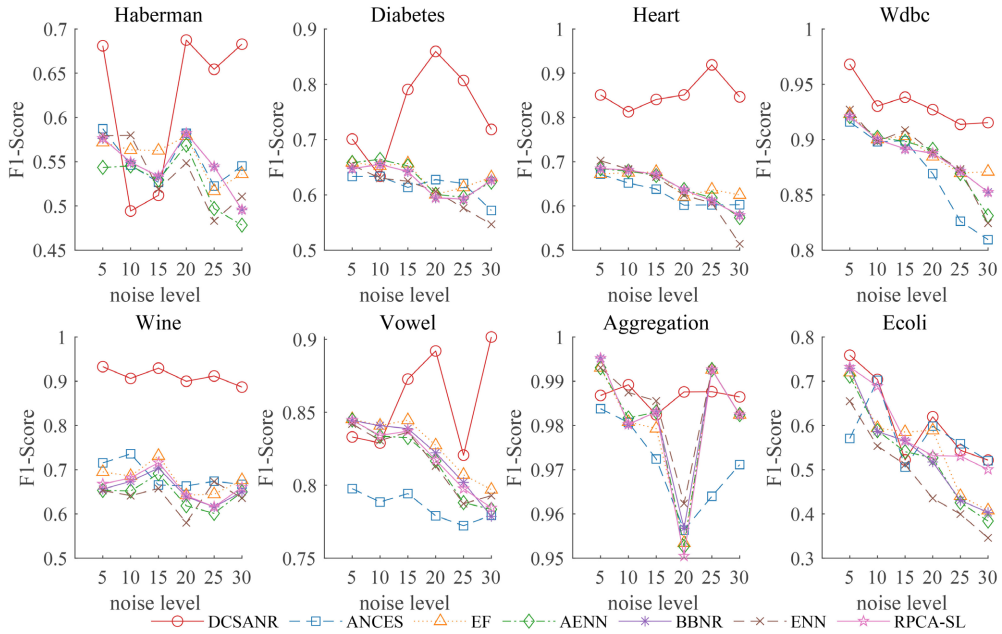

Figure 2. F1-Score comparison of 7 methods under different noise levels
图 2. 不同噪声水平下 7 种方法的 F1-Score 对比图

图 2 和图 3 展示了 7 种方法在不同噪声水平下公共数据集上的 F1-Score 和 AUC 值。由图 2 可以看出，DCSANR 在 F1-Score 上表现在多个数据集上明显优于其他过处理方法，表明 DCSANR 对属性噪声有较好的修复效果，并且具有最优的分类效果。由图 3 可以看出，相较于其他方法，DCSANR 在 Heart、Wdbc、Wine 数据集都获得最高的 AUC 值，更大程度提升了分类准确率，且在不同噪声水平的数据集上都能有效提高分类效果。

4.3.2. 地表水水质评价结果

表 5 展示了基于 KNN 分类器上各种噪声处理方法对黄河流域七里铺断面地表水水质数据(https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/Main.html)进行处理的分类准确率、F1-Score 和 AUC 三个指标的结果。DCSANR 的表现最优，分类准确率比 ANCES 提高了 9.3278%，F1-Score 和 AUC 也均为最高。这表明 DCSANR 对含属性噪声的地表水水质数据有较好的修复效果。

Table 5. Evaluation results of surface water quality of Qilipu section in the Yellow River Basin
表 5. 黄河流域七里铺断面地表水水质评价结果

方法	DCSANR	ANCES	EF	AENN	BBNR	ENN	RPCA-SL
准确率	0.8896	0.8137	0.7425	0.7449	0.7453	0.7493	0.8507
F1-Score	0.7916	0.7269	0.6270	0.6133	0.6141	0.6125	0.7531
AUC	0.8631	0.7452	0.6573	0.6535	0.6546	0.6514	0.7840

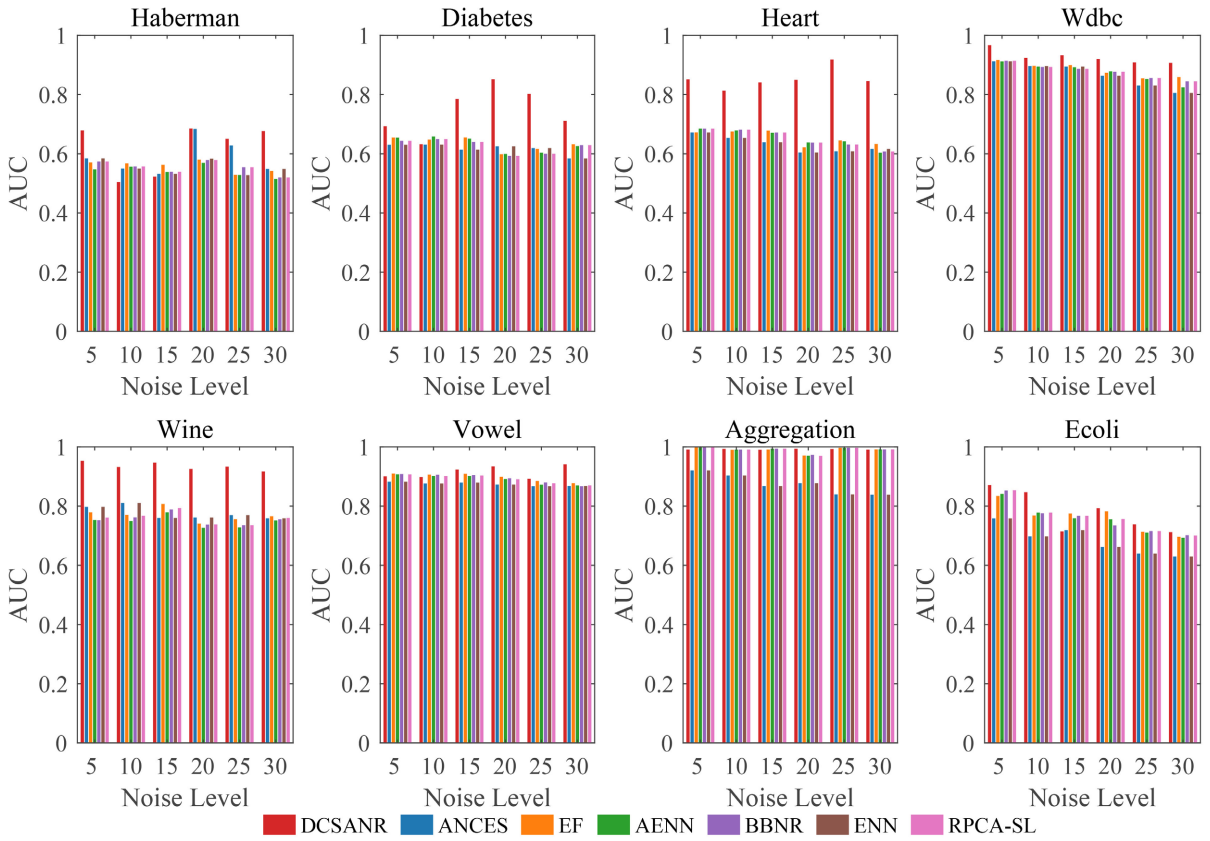

Figure 3. AUC comparison of 7 methods under different noise levels
图 3. 不同噪声水平下 7 种方法的 AUC 对比图

4.3.3. 统计检验

为进一步说明不同分类方法实验结果的统计特性，采用 Friedman 检验[25]和 Nemenyi 事后检验[25]来验证所构建的 DCSANR 方法与其他方法的性能差异。

Table 6. Friedman test and Nemenyi post hoc test results
表 6. Friedman 检验和 Nemenyi 事后检验结果

算法	平均序值	平均序值差	决策
DCSANR	2.1771		
ANCES	3.7500	1.5729	拒绝
EF	4.604	2.4269	拒绝
AENN	4.5313	2.3542	拒绝
BBNR	4.6875	2.5104	拒绝
ENN	4.6042	2.4271	拒绝
GE	3.6458	1.4687	拒绝

首先，根据平均分类准确率值对方法进行排序，确定其在处理数据集中的相对性能。数据集上表现最差的方法被分配一个更高的排名，而表现最好的方法被分配一个更低的排名，结果见表 6，若两个方法的排名相同，则取平均值。DCSANR 的平均序值为 2.1771，排名最低，因此具有最好的性能。

其次，使用 Friedman 检验来确定方法之间是否存在显著差异。原假设为所有方法在噪声处理方面的性能相同。实验部分，在 8 个公开数据集上分别引入了 5%、10%、15%、20%、25%和 30%共 6 种不同噪声水平的属性噪声，从而构建了 48 个独立的噪声数据集。基于这些数据集，对 7 种方法的分类性能进行了对比分析，并记录了各方法在不同噪声水平下的平均分类准确率。$F\left(M-1,\left(M-1\right)*\left(N-1\right)\right)$ 中，M 为模型数量，N 为数据集数量。因此，在 5%显著性水平下，$F\left(M-1,\left(M-1\right)*\left(N-1\right)\right)=F\left(6,282\right)=2.1308$。通过 Friedman 检验得到检验统计量 $T_F=10.2071>2.1308$，因此不能接受原假设，方法之间存在显著差异。

最后，采用 Nemenyi 事后检验来评估模型之间的两两差异。临界差 $C.D.=q_\alpha\sqrt{M(M+1)/6N}=1.3004$。以 DCSANR 为主要控制方法，与对比算法的平均序值差均大于临界差。根据 Nemenyi 事后检验，本文提出的 DCSANR 与其他方法差异显著，性能优于现有方法。具体结果见表 6，其中“接受”表示在 5%显著性水平下两种算法在数据集上的性能没有显著性差异，“拒绝”表示在 5%显著性水平下两种算法在数据集上的性能有显著性差异。

5. 结论与展望

本文提出了一种基于差异化创造性搜索(DCS)的数据属性噪声修复方法并应用于水质评价，旨在解决水质数据中噪声干扰导致的准确性和可靠性问题。通过结合差异化创造性搜索优化算法，设计了一种高效的噪声修复框架。结合直觉模糊集理论，提升了噪声样本识别的准确性和鲁棒性。属性噪声识别过程中通过熵权法为评价因子赋予权重，并基于此权重和数据集中每个属性值的错误程度得到属性值含噪声分数。在属性优化过程中使用 DCS 算法提升数据处理效率。通过对比实验验证了所构建方法的优越性，在多个公共数据集和黄河流域七里铺断面地表水水质数据上均取得了显著效果。

致　谢

作者衷心感谢为本研究提供数据和资源的各组织机构。同时，特别感谢课题组全体成员的宝贵建议与支持，他们的贡献对本研究的顺利开展起到了重要作用。

参考文献

[1] 左其亭. 黄河流域生态保护和高质量发展研究框架[J]. 人民黄河, 2019, 41(11): 1-6, 16.

[2] 杨 程, 郭亚坤, 郑兰香, 等. T-S 模糊神经网络模型训练样本构建及其在鸣翠湖水质评价中的应用[J]. 水动力学研究与进展, 2020, 35(3): 356-366.

[3] 郑培超, 周椿棪, 王金梅, 等. 基于 KPCA-PSO-ELM 算法的地表水化学需氧量紫外-可见吸收光谱检测研究[J]. 光谱学与光谱分析, 2024, 44(3): 707-713.

[4] 朱勇杰, 席晓勇, 李欣甜, 等. 预处理 + 臭氧 + AO + MBR 组合工艺在船舶油污水处理的应用[J]. 水处理技术, 2024, 50(3): 142-147.

[5] 窦皓. 河清岸绿, 让城市生活更美好[N]. 人民日报, 2024-03-07(017).

[6] 孙悦, 李再兴, 张艺冉, 等. 雄安新区——白洋淀冰封期水体污染特征及水质评价[J]. 湖泊科学, 2020, 32(4): 952-963.

[7] 田福金, 马青山, 张明, 等. 基于主成分分析和熵权法的新安江流域水质评价[J]. 中国地质, 2023, 50(2): 495-505.

[8] 杜浩田, 栾建勤, 刘霞. 基于灰色关联法的潇河上游水质评价[J]. 人民黄河, 2022, 44(S2): 157-158.

[9] Hu, G., Mian, H.R., Abedin, Z., Li, J., Hewage, K. and Sadiq, R. (2022) Integrated Probabilistic-Fuzzy Synthetic Evaluation of Drinking Water Quality in Rural and Remote Communities. *Journal of Environmental Management*, **301**, Article ID: 113937. https://doi.org/10.1016/j.jenvman.2021.113937

[10] García-Alba, J., Bárcena, J.F., Ugarteburu, C. and García, A. (2019) Artificial Neural Networks as Emulators of Process-Based Models to Analyse Bathing Water Quality in Estuaries. *Water Research*, **150**, 283-295. https://doi.org/10.1016/j.watres.2018.11.063

[11] Xu, T., Coco, G. and Neale, M. (2020) A Predictive Model of Recreational Water Quality Based on Adaptive Synthetic Sampling Algorithms and Machine Learning. *Water Research*, **177**, Article ID: 115788. https://doi.org/10.1016/j.watres.2020.115788

[12] Devroye, L., Györfi, L. and Lugosi, G. (1996) A Probabilistic Theory of Pattern Recognition. Springer.

[13] Sluban, B., Gamberger, D. and Lavra, N. (2010) Advances in Class Noise Detection. In: Coelho, H., Studer, R. and Wooldridge, M., Eds., *Proceedings of the Nineteenth European Conference on Artificial Intelligence*, IOS Press, 1105-1106.

[14] Teng, C.M. (1999) Correcting Noisy Data. In: Bratko, I. and Dzeroski, S., Eds., *Proceedings of the Sixteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., 239-248.

[15] 王石, 李玉忱, 刘乃丽, 等. 在属性级别上处理噪声数据的数据清洗算法[J]. 计算机工程, 2005, 31(9): 86-87, 227.

[16] Zhai, W. and Zhang, F. (2024) Robust Principal Component Analysis Integrating Sparse and Low-Rank Priors. *Journal of Computer and Communications*, **12**, 1-13. https://doi.org/10.4236/jcc.2024.124001

[17] Sáez, J.A. and Corchado, E. (2022) ANCES: A Novel Method to Repair Attribute Noise in Classification Problems. *Pattern Recognition*, **121**, Article ID: 108198. https://doi.org/10.1016/j.patcog.2021.108198

[18] Duankhan, P., Sunat, K., Chiewchanwattana, S. and Nasa-Ngium, P. (2024) The Differentiated Creative Search (DCS): Leveraging Differentiated Knowledge-Acquisition and Creative Realism to Address Complex Optimization Problems. *Expert Systems with Applications*, **252**, Article ID: 123734. https://doi.org/10.1016/j.eswa.2024.123734

[19] Ha, M., Wang, C. and Chen, J. (2012) The Support Vector Machine Based on Intuitionistic Fuzzy Number and Kernel Function. *Soft Computing*, **17**, 635-641. https://doi.org/10.1007/s00500-012-0937-y

[20] 孟朝霞, 尹萍, 贾宏恩. 基于熵权——云模型的某水库水质评价研究[J]. 应用数学进展, 2022, 11(10): 7161-7172.

[21] Yang, X., Huang, P., An, L., Feng, P., Wei, B., He, P., *et al.* (2022) A Growing Model-Based OCSVM for Abnormal Student Activity Detection from Daily Campus Consumption. *New Generation Computing*, **40**, 915-933. https://doi.org/10.1007/s00354-022-00193-z

[22] Brodley, C.E. and Friedl, M.A. (1999) Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research*, **11**, 131-167. https://doi.org/10.1613/jair.606

[23] Tomek, I. (1976) An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems and Man and Cybernetics*, **6**, 448-452.

[24] Delany, S.J. and Cunningham, P. (2004) An Analysis of Case-Base Editing in a Spam Filtering System. In: Funk, P. and González Calero, P.A., Eds., *Advances in Case-Based Reasoning*, Springer, 128-141. https://doi.org/10.1007/978-3-540-28631-8_11

[25] Demiar, J. and Schuurmans, D. (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, **7**, 1-30.