Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(6), 1444-1452
Published Online November 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.146135

Hans汉斯

基于自表示学习的化工行业多源异构碳数据聚类分析

刘小楠[1*]，周　强[1]，黄　勇[1]，冯晶晶[2,3]，何陆灏[2]，张　娜[4#]

[1]四川轻化工大学化学工程学院，四川　自贡
[2]中山大学碳中和与绿色发展研究院，广东　广州
[3]广东埃文低碳科技股份有限公司，广东　广州
[4]广州南方学院，广东　广州

收稿日期：2025年8月12日；录用日期：2025年11月7日；发布日期：2025年11月17日

摘　要

在碳达峰碳中和战略驱动下，化工行业急需通过精准的碳排放数据聚类分析完成数据归类，帮助碳减排路径科学规划。现有研究虽在多模态数据融合方面取得进展，但针对化工行业的多源异构碳数据聚类分析能力不足。本文创造性地提出一种基于自表示学习，适配化工行业的多源异构碳数据集的聚类分析系统。该系统能够挖掘多源异构数据的互补性以及高阶流形数据结构，实现精准的聚类分析，为碳减排路径科学规划提供技术支撑。

关键词

双碳战略，化工行业，碳排放，多源异构数据，聚类分析

Clustering Analysis of Multi-Source Heterogeneous Carbon Data in the Chemical Industry Based on Self-Representation Learning

Xiaonan Liu[1*], Qiang Zhou[1], Yong Huang[1], Jingjing Feng[2,3], Luhao He[2], Na Zhang[4#]

[1]College of Chemical Engineering, Sichuan University of Science & Engineering, Zigong Sichuan
[2]Institute of Carbon Neutrality and Green Development, Sun Yat-sen University, Guangzhou Guangdong
[3]Guangdong Avi Low Carbon Technology Co., Ltd., Guangzhou Guangdong
[4]Nanfang College Guangzhou, Guangzhou Guangdong

Received: August 12, 2025; accepted: November 7, 2025; published: November 17, 2025

[*]第一作者。
[#]通讯作者。

文章引用: 刘小楠, 周强, 黄勇, 冯晶晶, 何陆灏, 张娜. 基于自表示学习的化工行业多源异构碳数据聚类分析[J]. 人工智能与机器人研究, 2025, 14(6): 1444-1452. DOI: 10.12677/airr.2025.146135

Abstract

Driven by the dual carbon strategy, chemical enterprises need accurate clustering system of carbon-emission data for better data analysis, assisting in the scientific planning of carbon emission reduction roadmaps. Although encourage studies on multi-source data fusion, methods for multi-source heterogeneous carbon datasets clustering are limited. This paper creatively proposes a self-representation clustering system for multi-source heterogeneous carbon data sets. The system is capable of mining the complementarity of multi-source heterogeneous data and leveraging higher-order manifold data structures to achieve precise clustering analysis, thereby providing technical support for the scientific planning of carbon emission reduction pathways.

Keywords

Dual Carbon Strategy, Chemical Industry, Carbon Emission, Multi-Source Heterogeneous Data, Clustering Analysis

1. 引言

为应对全球气候变暖问题，中国提出并践行碳中和碳达峰国家发展战略。中国计划在 2060 年前实现碳排放达峰并逐步实现碳中和。化工行业是国民经济的重要经济支柱，为各行业提供基础原材料和产品，同时也是温室气体排放的重点行业之一。化工行业碳排放量高，约占全国总排放量的 5.3%。因为它生产过程中的化学反应、能源消耗和原材料使用均产生大量二氧化碳等其他温室气体。因此，化工行业迫切需要对多源异构碳数据进行采集、管理和分析。基于数据分析结果，帮助制定科学有效的减排路径，助力国家实现碳中和碳达峰发展目标。

目前，针对碳排放数据多源异构融合分析方面，学界取得了相关成果。在碳数据融合算法方面，Dong 等从多维属性分析的角度，提出了多源异构数据融合方法，并将其应用于分析电力企业能源消耗和温室气体排放[1]。然而，但该方法并不适用于化工行业。因为化工行业包含化学反应与物料转化过程。Bruckner 等基于蚁群算法，并借助模糊理论，提出企业碳排放信息集成管理模型。该模型只能完成评估减排绩效，但并没将供应链过程中的排放数据纳入分析[2]。在工业数据平台设计方面，Xiong 等人通过借助 Redis 提高了实时碳数据缓存速度。但是，它只聚焦在设备监控，碳核算的标准化处理[3]。Zhao 等的工业大数据平台实现了多源多模态数据的接收与存储，但无法完成碳数据的专业化清洗与聚类分析[4]。在数据分析方面，Wang 等提出的工业大数据分析框架专注于优化通用生产参数，未融合化工行业特定标准(如《中国石油化工企业温室气体排放核算方法与报告指南》[5]。

化工行业多源异构碳数据包含对观测对象的不同数据特征[6]。例如，碳数据文档则可以使用英语和中文等多种语言编写；碳数据包含结构化与非结构数据；碳数据的类型包含文档与图片。其次，由于化工产业工序的复杂性，工业软件的建设呈现多样性特征，例如实现技术多样化、存储设备不同、数据存储方式多样性等，导致出现大量多源异构数据，具有以下特点：数据异构性(包括结构化和非结构化)；数据多态性(静态数据和动态数据)；数据离散型(数据分布在不同的系统中)；以及数据量大。多源异构数据包含监测对

象的互补知识，互补性能够提升数据分析算法的学习性能，实现精准数据理解，从而帮助决策者制定科学减排路径。多源异构数据聚类分析能够将具有不同视角的样本准确归类为集群，通过将相似样本组织在一起并将差异较大的样本分配到不同组别。通常来说，多模态聚类分析方法比单模态聚类更具优势[7]。

为了实现精确多源异构数据聚类分析，近年来，学者多考虑引入数据的先验结构，指导聚类分析模型，提高模型的判别性能。Brbic [8]等人基于自表示模型，提出两个多源异构数据判别模型，分别使用一致性与核技术提高模型的性能。Abavisani [9]等人通过引入稀疏性与低秩性提高了判别模型的鲁棒性。Zhu [10]等人基于深度学习，提出稀疏与低秩深度模型。尽管这些模型都有不同程度的优越性，但是都存在一个共同的缺陷，即都是基于二阶矩阵方法。该类方法无法精准地挖掘多源异构数据的互补性。

另一方面，多源异构数据存在类别信息。通过引入这类信息，融入判别模型，能够提高判别模型的聚类性能。Xiao [11]等人基于张量自表示模型，将标签先验引入嵌入学习过程，实现了高精度聚类表现。Zhang [12]等人将标签先验构建为二阶矩阵并将其嵌入自表示模型，实现了无参数化学习。Tang [13]等人建模成对约束先验，实现了泛标签学习。这些研究成果都获得可观的聚类分析效果，但是都极度依赖先验知识，并且忽视了高阶流形数据结构。

综上分析，本文面向多源异构数据，拟解决多源异构数据的融合聚类分析。具体地，通过构建自表示三阶张量学习模型，挖掘多源异构数据间的互补性；同时，为了挖掘高阶流形数据结构，引入 Hessian 算子。本文算法基于张量学习框架，融合多源异构数据的互补性与高阶流形结构，实现高精度多源异构化工数据的聚类分析，助力碳中和发展精准决策指定。

2. 自表示三阶张量学习模型

2.1. 自表示学习

在机器学习和数据挖掘领域，自表示学习由于其能建模数据的相关性而被广泛应用于数据分析，主要用于处理高维数据的降维，特征提取和聚类问题。其核心思想是：将原始高维数据投影到一个低维的子空间中，低维表示旨在保留数据中的关键信息，同时去除噪声或冗余信息。低维表示具有更简化的数据结构、低计算复杂度，能够提高数据分析四任务(如分类、聚类、回归等)的性能。

2.2. 自表示学习模型

给定采集数据 $V \in R_{m\times n}^{+}$，自表示学习的目的是找到一个表示矩阵 $Q \in R_{n\times n}^{+}$，使得两者的乘积 VQ 能够很好的近似原始的采集数据 V。其数学形式如下：

$$V \approx VQ \tag{2.1}$$

通常使用 Frobenius 范数作为式(2.1)近似函数的代价函数。因此，自表示可以通过最小化以下代价函数来实现：

$$\min_{Q} \frac{1}{2}\left\|V - VQ\right\|_F^2 \tag{2.2}$$

其中 $\left\|\cdot\right\|_F$ 表示 Frobenius 范数。式(2.2)将输入碳数据矩阵 V 分解成了基矩阵 V 和系数矩阵 Q 的乘积。具体地，V 的每一列表示一个数据样本，Q 的每一行表示对应数据样本的权重系数。因此自表示学习的本质就是将原始的数据矩阵表示成原始输入矩阵的每一列的线性组合，而系数矩阵的每一行则对应线性组合中权重系数。

对于(2.2)最优化问题，通常采用梯度下降法。但是传统地梯度下降方法其收敛速度慢，并且对步长很敏感。为了克服这些缺点，研究者提出投影梯度下降法，它克服乘法法则难收敛的问题，同时能够通过采用策略在每步的迭代中优化步长来逼近最优解。投影梯度方法用于边界约束的最优化问题具体表述如下。

我们考虑以下标准形式的边界约束优化问题：

$$\min_{o\in R^n} f(x),\quad \text{s.t. } p_i < o_i < q_i,\ i=1,\cdots,n \tag{2.3}$$

其中 $f(o)$ 是函数连续可微，p 和 q 分别是上下边界的临界值。投影梯度法通过如下的迭代规则从 o^k 计算 o^{k+1}：

$$o^{k+1} = L\left[o^k - \alpha^k \nabla f\left(o^k\right)\right] \tag{2.4}$$

其中

$$L[o_i] = \begin{cases} o_i, & p_i < o_i < q_i \\ q_i, & o_i \geq q_i \\ p_i, & o_i \leq p_i \end{cases} \tag{2.5}$$

$L[o_i]$ 算子将 o_i 映射到特定的区间内。不用策略下的投影梯度方法采取不同方式来计算步长 α^k。

我们采取最优梯度投影梯度方法用于对自表示最优化问题进行的求解，其具体描述如下：

(1) 给定任意的 $0<\alpha<1$，$0<\beta<1$，随机初始化 o^1。

(2) 对于 $k=1,2,\cdots$

$$o^{k+1} = L\left[o^k - \alpha_k \nabla f\left(o^k\right)\right] \tag{2.6}$$

$$f\left(o^{k+1}\right) - f\left(o^k\right) \leq \sigma f\left(o^k\right)^{\mathrm{T}}\left(o^{k+1} - o^k\right) \tag{2.7}$$

其中 $\alpha_k = \beta^{t_k}$，t_k 是非负整数，$\sigma = 0.01$。最优步长能够保证每次迭代能够逼近最优值。当步长太大，可能达不到最优解；当步长太小，收敛速度太慢。在迭代中，我们通过找到满足式(2.6)最大的 β^{t_k} 作为步长 α_k 来逼近最优解。

Q 的偏导数计算如下：

$$\nabla f(Q) = V^{\mathrm{T}}(VQ - V) \tag{2.8}$$

步长，我们需要选择合适的 α 来确保下式(2.9)满足式(2.6)。

$$\tilde{Q} \equiv L\left[Q - \alpha \nabla f(Q)\right] \tag{2.9}$$

其中 $\tilde{Q}$ 是优化的结果，最小化代价函数：

对于函数 $f(o)$ 和任意的向量 d：

$$f(o+d) = f(o) + \nabla f(o)^{\mathrm{T}} d + \frac{1}{2} d^{\mathrm{T}} \nabla^2 f(o) d \tag{2.10}$$

因此对于两个连续的迭代 o 和 $\tilde{o}$，式(2.7)可以写成：

$$(1-\sigma)\nabla f(o)^{\mathrm{T}}(\tilde{o} - o) + \frac{1}{2}(\tilde{o} - o)^{\mathrm{T}} \nabla^2 f(o)(\tilde{o} - o) \leq 0 \tag{2.11}$$

因此，根据上式我们可以得到：

$$(1-\sigma)\left\langle \nabla f(Q), \tilde{Q} - Q\right\rangle + \frac{1}{2}\left\langle \tilde{Q} - Q, \left(V^{\mathrm{T}}V\right)\left(\tilde{Q} - Q\right)\right\rangle \leq 0 \tag{2.12}$$

其中 $\sigma = 0.01$，$\langle\cdot,\cdot\rangle$ 表示两个矩阵对应元素的乘积和。经过 k 轮迭代后满足的 $\tilde{Q}$ 就是最终的最优解。

2.3. 海森算子

海森算子建模了数据流形结构中的平滑结构信息[14]。通过海森算子对自表示学习进行约束能够保

证学习结果保存数据的平滑流形结构。海森算子的理论基础是 Eell 能量，计算如下：

$$S_{Eells}(f)=\int_M \left\|\nabla_a\nabla_b f\right\|^2_{T_x^*M\otimes T_x^*M}\,\mathrm{d}V(x), \tag{2.13}$$

其中，$\nabla_a\nabla_b$ 是投影函数 f 的偏导。

当海森算子用来约束时，我们构建如下约束算子：

$$\begin{aligned}\hat{S}_{Hess}(f)&=\sum_{i=1}^{n}\sum_{r,s=1}^{d}\left\|\frac{\partial^2 f}{\partial x_r\partial x_s}x_i\right\|^2\\&=\sum_{i=1}^{n}\sum_{x_\alpha\in N_k(x_i)}\sum_{x_\beta\in N_k(x_i)}\mathbf{f}_\alpha\mathbf{f}_\beta\mathbf{H}^{(i)}_{\alpha\beta}\\&=\mathbf{f}\mathbf{H}\mathbf{f}^{\mathrm{T}}\end{aligned} \tag{2.14}$$

其中，$\mathbf{H}$ 是海森矩阵，$\boldsymbol{f}$ 是被约束因子。通过优化(2.14)能够保证 $\boldsymbol{f}$ 挖掘平滑流形结构信息。

2.4. 海森算子约束三阶张量自表示学习模型

基于自表示学习模型，我们提出海森算子约束三阶张量自表示学习来解决多源异构碳数据聚类分析的问题。具体的，针对多源异构数据的互补性，我们采用三阶张量表示学习来挖掘；针对多源异构数据的高阶流形结构信息，我们采用海森算子来建模。具体模型如下：

$$\begin{aligned}&\min_{Z^{(v)},E}\left\|\mathcal{Z}\right\|_*+\alpha\left\|E\right\|_{2,1}+\underbrace{\beta\sum_{v=1}^{V}tr\left(Q^{(v)}H^{(v)}Q^{(v)^{\mathrm{T}}}\right)}_{\text{Hessian regularization}}\\&\text{s.t. }\ X^{(v)}=X^{(v)}Q^{(v)}+E^{(v)},v=1,2,\cdots,V\\&\qquad\mathcal{Z}=\Phi\left(Q^{(1)},Q^{(2)},\cdots,Q^{(V)}\right)\\&\qquad E=\left[E^{(1)};E^{(2)};\cdots;E^{(V)}\right],\end{aligned} \tag{2.15}$$

其中，$X^{(v)}$ 代表监测 v 号信息源下的碳数据，$Q^{(v)}$ 代表 v 号信息源下的自表示结果，$E^{(v)}$ 是 v 号信息源下的自表示误差矩阵。$H^{(v)}$ 是 v 号信息源下的海森约束矩阵，$\mathcal{Z}$ 是多源异构数据的三阶张量表示。以上模型解决了多源异构碳数据的互补性与高阶流形结构数据挖掘问题。

2.5. 聚类实验

我们将对所提算法在实际场景中的应用展开实验和讨论。首先我们介绍对比算法，描述实验所采取的数据集和评价标准，聚类分析结果以及算法运行时间。

2.5.1. 对比算法

我们参与评估的算法如下：

(1) PMLRSSC [8]：成对相关性多源稀疏低秩自表示学习算法；

(2) CMLRSSC [8]：中心相关性多源稀疏低秩自表示学习算法；

(3) T-SVDMSC [15]：三阶张量多源异构数据低秩表示；

(4) JLMVC [16]：协同学习多源异构数据自表示学习；

(5) GLTA [17]：图驱动多源异构数据自表示学习；

(6) NLRTGC [18]：先验驱动多源异构数据自表示学习。

(7) TMSRL [13]：强连接驱动三阶张量多源异构自表示学习。

(8) CTLR [12]：成对约束驱动多源异构数据自表示学习。

(9) Proposed：本文所提方法。

2.5.2. 实验数据与评价指标

本文在七个主流的多源异构数据集上测试本文所提方法以及对比方法的聚类表现。实验重复 20 次，取平均值并记录。特别的，对比算法的参数都设置为原文参数。本实验采用五种标准评价指标来评估算法的性能，分别是：ACC，NMI，AR，F-score 以及 Precision [16]。具体实验数据的描述如下表 1：

Table 1. Specific parameters of experimental data
表 1. 实验数据具体参数

Datasets	Views	Classes	Size	Type
Politicsie	9	7	348	text
3Sources	3	6	169	text
Extented YaleB	3	38	2414	image
Prokaryotic	2	4	551	prokaryotic
Flowers	3	68	1360	image
Scene-15	3	15	4485	image
MITIndoor	4	67	5360	image

2.5.3. 实验结果

本文将实验结果记录在表 2~5。从实验结果中，我们得出了以下结论：

(1) 所提方法在所有数据集上都优于对比算法。具体的，在数据集 Politicsie 上，所提方法在五个评价指标下比第二优秀的方法分别获得了 2.1%，1.9%，2.0%，4.4%，和 3.3%的性能提升。

(2) 与先验驱动的方法相比，例如：NLRTGC，TMSRL 和 CTRL，所提方法表现优异。主要是因为所提方法采用了复权策略。该策略能够提高特征的挖掘性能，最终提高聚类模型的判别性，从而提高聚类精度。

(3) 与矩阵方法相比，例如 PMLRSSC 和 CMLRSSC，所提方法优越性显著。主要因为本文所提方法建立于三阶张量学习模型。张量学习能够高精度地挖掘多源异构数据的互补性。

表 6 记录了所提方法的运行时间，从结果可知：所提算法由于挖掘了高阶流形数据结构，需要较高的时间成本。

Table 2. Experimental results of the algorithm on the Poloticsie and 3sources datasets
表 2. 算法在 Poloticsie 和 3sources 数据集上的实验结果

Dataset	Politicsie					3sources				
Method	Acc	NMI	AR	F-score	Precision	Acc	NMI	AR	F-score	Precision
PMLRSSC	0.556	0.433	0.284	0.455	0.521	0.603	0.625	0.432	0.568	0.625
CMLRSSC	0.532	0.426	0.268	0.334	0.508	0.595	0.624	0.458	0.530	0.628
T-SVDMSC	0.872	0.819	0.852	0.898	0.905	0.765	0.667	0.646	0.728	0.665
JLMVC	0.886	0.836	0.865	0.902	0.913	0.836	0.738	0.675	0.751	0.809
GLTA	0.908	0.828	0.876	0.897	0.926	0.849	0.748	0.713	0.767	0.837
NLRTGC	0.925	0.918	0.905	0.924	0.918	0.841	0.813	0.803	0.768	0.795
TMSRL	0.896-	0.859	0.896	0.889	0.925	0.855	0.831	0.818	0.802	0.816
CTRL	0.931	0.926	0.917	0.908	0.932	0.878	0.865	0.825	0.815	0.829
Proposed	0.952	0.945	0.937	0.952	0.965	0.907	0.873	0.845	0.848	0.854

Table 3. Experimental results of the algorithm on Extended YaleB and Prokaryotic datasets
表 3. 算法在 Extended YaleB 和 Prokaryotic 数据集上的实验结果

Dataset	Extended YaleB					Prokaryotic				
Method	Acc	NMI	AR	F-score	Precision	Acc	NMI	AR	F-score	Precision
PMLRSSC	0.226	0.189	0.165	0.262	0.131	0.407	0.427	0.388	0.501	0.415

续表

CMLRSSC	0.208	0.216	0.259	0.182	0.158	0.414	0.434	0.374	0.512	0.402
T-SVDMSC	0.642	0.657	0.520	0.550	0.525	0.533	0.507	0.457	0.545	0.456
JLMVC	0.616	0.618	0.645	0.546	0.525	0.633	0.465	0.445	0.548	0.545
GLTA	0.624	0.621	0.446	0.482	0.463	0.541	0.538	0.435	0.525	0.537
NLRTGC	0.679	0.703	0.663	0.645	0.627	0.654	0.505	0.456	0.559	0.492
TMSRL	0.669	0.722	0.713	0.658	0.595	0.578	0.468	0.468	0.522	0.478
CTRL	0.703	0.715	0.708	0.636	0.548	0.588	0.524	0.478	0.559	0.502
Proposed	0.725	0.742	0.736	0.698	0.652	0.694	0.548	0.509	0.598	0.587

Table 4. Experimental results of the algorithm on the Flowers and Scene-15 datasets
表 4. 算法在 Flowers 和 Scene-15 数据集上的实验结果

Dataset	Flowers					Scene-15				
Method	Acc	NMI	AR	F-score	Precision	Acc	NMI	AR	F-score	Precision
PMLRSSC	0.507	0.617	0.358	0.431	0.337	0.411	0.415	0.488	0.368	0.352
CMLRSSC	0.515	0.625	0.365	0.455	0.328	0.422	0.436	0.465	0.385	0.324
T-SVDMSC	0.742	0.765	0.757	0.702	0.764	0.813	0.807	0.776	0.782	0.748
JLMVC	0.715	0.728	0.788	0.716	0.782	0.804	0.825	0.811	0.805	0.786
GLTA	0.758	0.782	0.805	0.732	0.805	0.823	0.806	0.795	0.822	0.815
NLRTGC	0.808	0.825	0.778	0.809	0.762	0.735	0.796	0.805	0.835	0.821
TMSRL	0.821	0.809	0.805	0.798	0.809	0.815	0.828	0.788	0.818	0.806
CTRL	0.837	0.796	0.816	0.769	0.772	0.828	0.816	0.815	0.806	0.838
Proposed	0.848	0.845	0.827	0.835	0.812	0.855	0.848	0.839	0.849	0.859

Table 5. Experimental results of the algorithm on the MITindoor dataset
表 5. 算法在 MITindoor 数据集上的实验结果

Method	ACC	NMI	AR	F-score	Precision
PMLRSSC	0.425	0.542	0.268	0.273	0.259
CMLRSSC	0.415	0.508	0.240	0.224	0.235
T-SVDMSC	0.725	0.783	0.604	0.598	0.585
JLMVC	0.734	0.714	0.611	0.627	0.638
GLTA	0.742	0.727	0.648	0.635	0.656
NLRTGC	0.769	0.761	0.665	0.648	0.728
TMSRL	0.755	0.814	0.677	0.665	0.674
CTRL	0.768	0.826	0.688	0.676	0.655
Proposed	0.788	0.845	0.705	0.716	0.748

Table 6. Running time results (in seconds) of algorithms NLRTGC, TMSRL, CTRL, and Proposed on various datasets
表 6. 算法 NLRTGC，TMSRL，CTRL 和 Proposed 在各数据集上的运行时间结果(单位：秒)

Method	Politicsie	3Sources	YaleB	Prokaryotic	Flowers	Scene-15	MITindoor
NLRTGC	49.90	57.04	97.54	88.45	103.04	2214.45	3537.68
TMSRL	43.89	52.71	83.75	53.08	112.45	1757.46	2638.26
CTRL	54.73	45.56	91.54	50.82	91.74	2048.32	3418.73
Proposed	55.62	71.54	220.68	125.89	136.78	2285.58	3932.47

3. 结论与展望

本研究针对多源异构碳数据集的聚类分析展开研究，提出了新型基于自表示子空间学习的聚类分析

方法，该方法能够有效地挖掘多源异构数据的互补性以及高阶流形结构信息。方法在七个标准多源异构数据集上进行了测试，相比于其它对比算法，本文所提算法优越性得到证实。并取得了重要实践成果。基于多源异构数据融合、子空间学习理论，张量学习框架，数据流形理论和梯度优化理论，构建了涵盖数据融合、清洗和聚类分析的架构。该架构能够融合化工生产、能源消耗、供应链等环节涉及的多源异构数据，运用先进技术挖掘多源异构数据的互补性，子空间结构，流形结构，提高数据聚类分析的精度。总体来说，本研究成果紧密贴合化工企业碳中和碳达峰实践需求，为化工企业践行精准碳排放管控、科学制定减排策略提供了坚实的技术支撑，助力推动该行业绿色低碳成功转型。

未来，化工行业多源异构碳数据聚类技术发展将呈现多维度发展趋势。一是针对化工行业多源异构数据的噪音处理，可通过降维分析理论，对数据进行降维去噪，提高数据质量，优化数据理解。二是针对大数据多源异构数据的冗余性，可通过哈希学习理论对数据进行哈希域投影学习，实现数据的快速处理。三是引入多源异构数据的标签属性，融入数据学习模型，提高聚类模型的判别性能，实现高精度聚类分析结果，助力碳中和政策决策。

基金项目

四川省自然科学基金青年基金项目(2025ZNSFSC1264)；四川轻化工大学“652”科研创新团队(SUSE652A003)；广西重点研发计划项目(桂科 AB24010156)。

参考文献

[1] Dong, F., Zhu, J., Li, Y., Chen, Y., Gao, Y., Hu, M., *et al.* (2022) How Green Technology Innovation Affects Carbon Emission Efficiency: Evidence from Developed Countries Proposing Carbon Neutrality Targets. *Environmental Science and Pollution Research*, **29**, 35780-35799. https://doi.org/10.1007/s11356-022-18581-9

[2] Bruckner, B., Hubacek, K., Shan, Y., Zhong, H. and Feng, K. (2022) Impacts of Poverty Alleviation on National and Global Carbon Emissions. *Nature Sustainability*, **5**, 311-320. https://doi.org/10.1038/s41893-021-00842-z

[3] 熊肖磊, 王春伟, 赵炯, 等. 基于 Redis 与 SSM 的大型设备数据运用系统设计[J]. 现代机械, 2018(6): 29-34.

[4] 赵德基, 王力, 狄军峰. 基于 Dubbo + NoSQL 的工业领域大数据平台研究[J]. 数字技术与应用, 2017(7): 64-67.

[5] 王宏志, 梁志宇, 李建中, 等. 工业大数据分析综述: 模型与算法[J]. 大数据, 2018, 4(5): 62-79.

[6] Yin, M., Gao, J., Xie, S. and Guo, Y. (2019) Multiview Subspace Clustering via Tensorial T-Product Representation. *IEEE Transactions on Neural Networks and Learning Systems*, **30**, 851-864. https://doi.org/10.1109/tnnls.2018.2851444

[7] Yang, Z., Liang, N., Yan, W., *et al.* (2020) Uniform Distribution Non-Negative Matrix Factorization for Multiview Clustering. *IEEE Transactions on Cybernetics*, **99**, 1-14.

[8] Brbić, M. and Kopriva, I. (2018) Multi-View Low-Rank Sparse Subspace Clustering. *Pattern Recognition*, **73**, 247-258. https://doi.org/10.1016/j.patcog.2017.08.024

[9] Abavisani, M. and Patel, V.M. (2018) Multimodal Sparse and Low-Rank Subspace Clustering. *Information Fusion*, **39**, 168-177. https://doi.org/10.1016/j.inffus.2017.05.002

[10] Zhu, W. and Peng, B. (2020) Sparse and Low-Rank Regularized Deep Subspace Clustering. *Knowledge-Based Systems*, **204**, Article ID: 106199. https://doi.org/10.1016/j.knosys.2020.106199

[11] Xiao, X., Chen, Y., Gong, Y. and Zhou, Y. (2021) Prior Knowledge Regularized Multiview Self-Representation and Its Applications. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 1325-1338. https://doi.org/10.1109/tnnls.2020.2984625

[12] Zhang, C., Fu, H., Wang, J., Li, W., Cao, X. and Hu, Q. (2020) Tensorized Multi-View Subspace Representation Learning. *International Journal of Computer Vision*, **128**, 2344-2361. https://doi.org/10.1007/s11263-020-01307-0

[13] Tang, Y., Xie, Y., Zhang, C. and Zhang, W. (2022) Constrained Tensor Representation Learning for Multi-View Semi-Supervised Subspace Clustering. *IEEE Transactions on Multimedia*, **24**, 3920-3933. https://doi.org/10.1109/tmm.2021.3110098

[14] Lee, J.M. (2006) Riemannian Manifolds: An Introduction to Curvature. Springer Science & Business Media.

[15] Xie, Y., Tao, D., Zhang, W., Liu, Y., Zhang, L. and Qu, Y. (2018) On Unifying Multi-View Self-Representations for

Clustering by Tensor Multi-Rank Minimization. *International Journal of Computer Vision*, **126**, 1157-1179. https://doi.org/10.1007/s11263-018-1086-2

[16] Chen, Y., Xiao, X. and Zhou, Y. (2020) Jointly Learning Kernel Representation Tensor and Affinity Matrix for Multi-View Clustering. *IEEE Transactions on Multimedia*, **22**, 1985-1997. https://doi.org/10.1109/tmm.2019.2952984

[17] Chen, Y., Xiao, X. and Zhou, Y. (2020) Multi-View Subspace Clustering via Simultaneously Learning the Representation Tensor and Affinity Matrix. *Pattern Recognition*, **106**, Article ID: 107441. https://doi.org/10.1016/j.patcog.2020.107441

[18] Pan, B., Li, C. and Che, H. (2023) Nonconvex Low-Rank Tensor Approximation with Graph and Consistent Regularizations for Multi-View Subspace Learning. *Neural Networks*, **161**, 638-658. https://doi.org/10.1016/j.neunet.2023.02.016