研究与开发

基于改进张量链分解的多聚类算法

张宏俊[1,2]，张泽宇[3]，张颖娇[4]，叶昊[5,6]，潘高军[7]
（1. 中国通信服务股份有限公司，北京 100073；
2. 南京邮电大学物联网学院，江苏 南京 210003；
3. 曼彻斯特大学人工智能学院，英国 曼彻斯特 M13 9PL；
4. 南京邮电大学现代邮政学院，江苏 南京 210003；
5. 华为技术有限公司南京研究所，江苏 南京 210012；
6. 北京外企德科人力资源服务上海有限公司，上海 200011；
7. 浙江省通信产业服务有限公司，浙江 杭州 310052）

摘　要：随着大数据时代的到来，高阶数据的有效表示和分析成为一项重大挑战。基于此，聚焦于张量分解技术在多聚类算法中的应用，特别是针对大型多源异构数据集的处理，深入研究并改进了张量链（tensor train，TT）分解方法，通过引入新的优化策略，显著提高了其在多聚类任务中的性能。创新主要体现在两个方面：一是提出了一种新的张量分解框架，该框架通过优化目标函数，有效降低了存储成本并提高了计算效率；二是将改进的张量分解技术应用于3种主要的多聚类算法中，包括自加权多视图聚类（self-weighted multi-view clustering，SwMC）、潜在多视图子空间聚类（latent multi-view subspace clustering，LMSC）和具有完整性感知相似性的多视图子空间聚类（multi-view subspace clustering with intactness-aware similarity，MSC IAS），显著提升了聚类的准确性和效率。为了验证方法的有效性，在7个真实的数据集上进行了全面的实验评估，包括准确性（accuracy，ACC）、归一化互信息（normalized mutual information，NMI）和纯度等3个指标。实验结果表明，所提出的方法在提取有意义的模式和提高聚类性能方面具有显著优势。

关键词：张量；多聚类算法；张量分解；多源异构数据；主成分分析
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025043

Multi-clustering algorithm based on improved tensor chain decomposition

ZHANG Hongjun[1,2], ZHANG Zeyu[3], ZHANG Yingjiao[4], YE Hao[5,6], PAN Gaojun[7]
1. China Communications Services Co., Ltd., Beijing 100073, China

收稿日期：2024−11−23；**修回日期：**2025−02−25
通信作者：潘高军，2021070702@njupt.edu.cn
基金项目：国家自然科学基金资助项目（No.61972208，No.62102194，No.62272239）；江苏省研究生科研与实践创新计划项目（No.KYCX22_1027，No.SJCX24_0339，No.SJCX24_0346）；南京邮电大学大学生创新训练计划项目（No.XZD2019116，No.XYB2019331）
Foundation Items：The National Natural Science Foundation of China (No.61972208, No.62102194, No.62272239), Jiangsu Graduate Research and Practice Innovation Program Project (No.KYCX22_1027, No.SJCX24_0339, SJCX24_0346), Nanjing University of Posts and Telecommunications (No.XZD2019116, No.XYB2019331)

2. School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
3. School of Artificial Intelligence, The University of Manchester, Manchester M13 9PL, UK
4. School of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
5. Nanjing Research Institute of Huawei Technologies Co., Ltd., Nanjing 210012, China
6. Beijing Adecco Human Resources Service Shanghai Co., Ltd., Shanghai 200011, China
7. Zhejiang Communication Industry Services Co., Ltd., Hangzhou 310052, China

Abstract: With the advent of the era of big data, the effective representation and analysis of high-level data has become a major challenge. Based on this, the application of tensor decomposition technology in multi-clustering algorithms was focused on especially for the processing of large multi-source heterogeneous datasets. The tensor train (TT) method was studied and improved in depth, which had significantly improved its performance in multi-clustering tasks by introducing a new optimization strategy. The innovations were mainly reflected in two aspects: firstly, a new tensor decomposition framework was proposed, which effectively reduced the storage cost and improved the computational efficiency by optimizing the objective function; secondly, the improved tensor decomposition technique was applied to three main multi-clustering algorithms, including self-weighted multi-view clustering (SwMC), latent multi-view subspace clustering (LMSC), and multi-view subspace clustering with intactness-aware similarity (MSC IAS), which significantly improved the accuracy and efficiency of clustering. To validate the effectiveness of the proposed methodology, comprehensive experiments were conducted on seven real-world datasets, including assessments of key metrics such as accuracy (ACC), normalized mutual information (NMI), and purity. Experimental results show that the proposed method has significant advantages in extracting meaningful patterns and improving clustering performance.

Key words: tensor, multi-clustering algorithm, tensor decomposition, multi-source heterogeneous data, principal component analysis

0 引言

多重聚类分析能够从多个角度发现大数据中潜在的数据模式，这使得它在自动化行业具有极高的应用价值[1]。然而，现有的多数方法在将异构数据根据应用需求分组到少数几个聚类中时面临挑战。张量作为矩阵的推广，能够表示现实世界数据集的复杂性，这使得多维数据聚类成为一个重要的研究课题[2]。尽管二维数据聚类技术已经取得了显著进展，但针对多维数据的子空间聚类方法的研究相对较少。

探索性数据分析（exploratory data analysis，EDA）是聚类技术应用的一个重要领域，它通过分析数据属性间的相互关系，将数据分解成易于处理的组块[3]。在张量分解中，选择合适的秩非常关键，尤其是在张量环（tensor ring，TR）分解时。一种新的秩选择方法可以自动确定接近最优的TR秩，这有助于降低存储成本，特别是对于那些具有复杂张量链（tensor train，TT）或TR结构的张量[4]。TR秩的选择通常在截断奇异值的分解（truncated singular value decomposition，t-SVD）之前进行，或通过t-SVD来确定。此外，还可以通过其他方式实现自适应的TR秩选择。

张量数据集可以通过Tucker张量分解来组织，Tucker张量能够描述完整或不完整的多维数据集[5]。根据Tucker分解，每个张量可以被分解为其核心张量和因子矩阵的线性组合。在处理密集张量时，其目标是设计高效的分布式实现方法。基于高阶正交迭代器（higher-order orthogonal iteration，HOOI）的方法使用张量与矩阵的乘积作为其基本操作[6]。张量分解技术，如标准格式和张量列格式，允许以较低的成本和复杂性，存储和处理

更高维度的信息，避免了信息存储和处理的指数级增长[7]。

张量分析在科学和工程的许多领域都有应用，包括医学中的脑电图信号处理、电磁学中的电磁传感器分析、黎曼几何、力学、弹性学以及相对论研究。近期研究表明，利用路径积分的张量网络分解被证明有助于开放量子系统的建模[8]。然而，这些方法的计算量随系统规模的扩大而急剧增加，使得在局部耗散环境中模拟大型量子系统的非平衡行为面临严峻挑战。

准确的多模式预测能够提升人们的判断力，帮助做出更好的决策。近年来，基于特征张量或Z特征向量的多元马尔可夫模型，被越来越多地应用于预测未来事件。然而，当许多传统的马尔可夫模型与基于张量的方法结合时，并不总能得出一致的结果。基于张量的估计算法在计算效率和响应时间上往往受到高阶张量带来的“维数诅咒”的制约[9]。当数据张量中存在有组织的缺失组件，如行、列、块或补丁的缺失，且这些缺失组件并非随机分布时，数据张量的补全工作变得更加复杂[10]。现有的张量补全技术尚无法有效处理这种有组织缺失数据的情况。

本文的创新点体现在以下3个方面：（1）全面分析和验证了与TT分解相关的亚线性预计算时间，确保了其作为所提出框架的基础的可行性；（2）深入研究了非结构化矩阵表示和量化张量链（quantics tensor train，QTT）分解的复杂性，为它们的压缩和反演技术提供了关键见解；（3）首次探讨了这些技术在线性问题中的适用性，揭示了它们在加速二阶优化方法中的潜力。

就解决方案而言，本文为更有效地处理线性系统提供了一种全面的方法。利用TT分解，可以设计出速度更快且更准确可靠的迭代求解器。将分层压缩技术集成到二阶优化方法中代表着该领域向前迈出了重要一步，为探索和创新开辟了新的途径。

1　相关工作

1.1　张量链分解

在数据科学和机器学习领域，高维数据的处理和分析是一个重要的研究课题。随着大数据技术的发展，数据的维度和复杂性不断增加，传统的数据处理方法，如主成分分析（principal component analysis，PCA）在处理高维数据时面临挑战。张量作为多维数据结构，能够更有效地表示和处理复杂的数据关系。TT分解是一种将高维张量分解为低维张量序列的方法，它在数据压缩、降维和模式识别等方面显示出显著的优势。

张量是一个多维数组，其中每个元素可以被索引为$(i_1,i_2,\cdots,i_d)$，其中d为张量的阶数（即维度的数量）。一个阶数为d的张量$\boldsymbol{\mathcal{A}}$可以表示为：$\boldsymbol{\mathcal{A}}\in\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$。

TT分解是一种特殊的张量分解方法，它将一个高阶张量分解为一系列低阶张量的乘积，这些低阶张量沿着链式结构排列。具体来说，一个阶数为d的张量$\boldsymbol{\mathcal{A}}$可以被分解为：

$$\boldsymbol{A}\approx\boldsymbol{\mathcal{G}}_1\times_1\boldsymbol{\mathcal{G}}_2\times_1\cdots\times_1\boldsymbol{\mathcal{G}}_d \tag{1}$$

其中，$\boldsymbol{\mathcal{G}}_k$为一个$(r_{k-1}\times n_k\times r_k)$维的张量，$r_0=r_d=1$表示链的第$k$个节点的秩，$\times_k$表示第$k$个模式的张量乘积。

TT分解的优化通常涉及最小化重构误差，即最小化$\left|\left|\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{G}}_1\times_1\boldsymbol{\mathcal{G}}_2\times_1\cdots\times_1\boldsymbol{\mathcal{G}}_d\right|\right|_F^2$，其中$||\cdot||_F^2$为Frobenius范数，该范数定义为：

$$\left|\left|\mathcal{H}\right|\right|_F^2=\sqrt{<\mathcal{H},\mathcal{H}>}=\left(\sum_{i_1,i_2,\cdots,i_d}\mathcal{H}^2 i_1,i_2,\cdots,i_d\right)^{\frac{1}{2}} \tag{2}$$

针对TT分解，常用的优化算法包括交替最小二乘法（alternating least square，ALS），该算法每次固定其他张量，只优化一个$\boldsymbol{\mathcal{G}}_k$。此外，随

机梯度下降（stochastic gradient descent，SGD）通过随机选择$\mathcal{G}_k$进行优化，以提高收敛速度。

1.2 多视图聚类方法

本节深入探讨了多视图聚类的复杂性，并全面概述了问题描述、模型表示、数学公式、实验方法、理论推导以及所采用的数值方法等各个方面。首先，本节介绍了多视图聚类的基本概念和符号，并定义了一个k视图数据集，其中每个视图都使用自己的特征集来编码数据的独特视角。这些视图的维度可能会有所不同，这取决于所采用的特定算法。在分析中，本节特别指出了这一差异，并强调了在处理多视角数据时输入维度的重要性。

在模型表示部分，本节通过数学公式形式化了多视图聚类问题，使用矩阵和向量来表示数据集，并明确了不同视图和样本之间的关系。这一符号框架能够精确定义分析中所涉及的概念和操作。

在实验方法部分，本节概述了评估多视图聚类算法性能的策略。通过在真实数据集上进行严格的实验，本节比较了不同算法的表现，并深入分析了它们的优势和局限性。这一实证分析为所提出方法的实际应用和有效性提供了宝贵的见解。

理论推导在理解多视图聚类的基本原理方面发挥着至关重要的作用。本节进行了详细的数学推导，为所使用的算法奠定了理论基础。这些推导阐明了不同视图之间的关系、相似性矩阵的构建以及算法所追求的优化目标。

最后，本节讨论了用于解决由理论公式引起的优化问题的数值方法，利用先进的数值技术，如矩阵运算、迭代算法和优化求解器，高效地计算聚类结果。这些数值方法是为应对多视角数据带来的特定挑战而量身定制的，确保了解决方案的精确和可扩展。

1.2.1 基于图的模型

目前，张量分解常用的方法之一是基于图的聚类，其目的是生成数据相似性矩阵，再使用标准谱聚类技术或其他方法进行最终的标签分布。基于图的模型的创建也是多视图聚类的一部分。基于图的多视图聚类核心是为每个视图分配适当的权重，这是至关重要的一步。即使超参数的选择至关重要，其他研究者也会通过将不同的超参数组合纳入考量，评估每个参数设置的潜在价值。自动加权多图学习（auto-weighted multiple graph learning，AMGL）不需要任何额外的输入，采用传统的谱聚类方法来自动分配权重。其中，谱聚类表明其最终目标函数为：

$$\min_{E^{\mathrm{L}}E=\mathrm{c}} \lg(E^{\mathrm{L}}T_{\boldsymbol{a}}E) \tag{3}$$

其中，E表示损失函数，它衡量了模型预测与实际观测之间的差异。E^{L}为损失函数的变换形式，可能涉及某种特定的操作或变换，如拉普拉斯算子。$T_{\boldsymbol{a}}$是一个与矩阵或张量$\boldsymbol{A}$相关的变换操作，它定义了E与E^{L}之间如何相互作用。常数c用于调整损失函数的尺度或权重，以确保不同项在优化过程中的重要性。lg表示对数函数。整个表达式的目标是最小化经过变换后的损失函数的对数，以寻找最优的E值，从而使得模型的预测尽可能地接近真实数据。

根据式（3）提出了AMGL，其数学表达式为：

$$\min_{E^{\mathrm{L}}E=\mathrm{c}} \sum_{n=1}^{k}\sqrt{\lg\left(E^{\mathrm{L}}\boldsymbol{T}_a^nE\right)} \tag{4}$$

其中，n为索引变量，用于表示多个视图或多个图中的第n个元素，$\boldsymbol{T}_a^n$表示第n个视图的邻接矩阵或相似性矩阵。

为了构建式（4）的拉格朗日函数，需要求解E的额外偏导数，并将这些导数设为零。在这个过程中，权重因子将被整合进拉格朗日函数中。下文详细描述了整个求解过程中的两个关键步骤。

$$\sum_{n=1}^{k}\sqrt{\lg(E^{\mathrm{L}}E)+\partial(\Lambda,E)} \tag{5}$$

其中，Λ表示拉格朗日乘子。

$$\sum_{n=1}^{k} z^n \left(\frac{\partial L_o(E^{\mathrm{L}} \boldsymbol{T}_a^n E)}{\partial E} \right) + 2 \frac{\partial G(\Lambda, E)}{\partial E} = 0 \quad (6)$$

其中，$L_o(\cdot)$为原始目标函数。

式（4）的拉格朗日形式如式（5）所示，从式（6）中导出的形式化项由E表示。经推导，可得Z^n的数学表达式为：

$$Z^n = 1/2 \sqrt{\lg\left(E^{\mathrm{L}} \boldsymbol{T}_a^n E \right)} \quad (7)$$

z^n不是恒定值，它会随着E的变化而变化。然而，当将常数项纳入考虑时，式（7）转化为以下形式：

$$\min_{E^{\mathrm{L}} E = \boldsymbol{I}} \sum_{n=1}^{k} z^n \lg\left(E^{\mathrm{L}} \boldsymbol{T}_a^n E \right) \quad (8)$$

其中，$\boldsymbol{I}$表示单位矩阵，可以通过式（8）来计算E。根据式（7），z^n值会随着条件的变化而变化，因此可以通过迭代过程来寻找z^n和E的理想值。

综上可得，如果某个观点具有重大的重要性，它可能会对结果产生显著的影响。具体来说，当z^n乘以$z^n \lg(E^{\mathrm{L}} \boldsymbol{T}_a^n E)$时，如果该值显著增大，表明$z^n$相对较小，这与当前研究的情况相符合。

目标函数比较可用于计算AMGL和需要更多超参数的模型之间的误差，以计算AMGL和需要额外超参数的模型之间的差异。

$$\begin{aligned} &\min_{E^{\mathrm{L}} E = \boldsymbol{I}, z} \sum_{n=1}^{k} z^{n\,r} \lg\left(E^{\mathrm{L}} \boldsymbol{T}_a^n E \right) \\ &\text{s.t.} \sum_{n=\boldsymbol{I}}^{k} z^n = 1, k^n \geqslant 0 \end{aligned} \quad (9)$$

为了保持权重分布平滑，使用所谓的hyper参数，其值通常设置为“非负”，即使对算法参数的微小调整也可能对其性能产生重大影响。AMGL模型没有更多的参数，并且可以学习最佳的z^n和E值。虽然z^n不是完全独立的，但它的计算方法表明它与E值密切相关。无参数AMGL的基本相位如下所示。

输入： $\boldsymbol{Y} = \left\{ \boldsymbol{Y}^1, \boldsymbol{Y}^2, \cdots, \boldsymbol{Y}^k \right\}$，其中$\boldsymbol{Y}^n \in O^{v \times p^n}$为$v$行$p^n$列的正交矩阵），簇数$m$

输出： 指标矩阵$\boldsymbol{M}$

步骤1　初始化每个视图的权重$z^n = \frac{1}{k}$；计算每个视图对应的拉普拉斯矩阵$\boldsymbol{T}_a^n$；计算$\boldsymbol{T}_A = \sum_{n=1}^{z} z^n \boldsymbol{T}_a^n$；

步骤2　不收敛则继续执行；

步骤3　通过式（8）计算E；T_A的2～$(m+1)$个最小特征值；

步骤4　通过式（9）更新z^n；

步骤5　结束循环

1.2.2　自加权多视图聚类（self-weighted multiview clustering，SwMC）

在基于图的多视图聚类方法中，合理地为不同视图分配权重一直是一个挑战。尽管已有文献提出了多种解决方案，但这些方法要么需要人工干预，要么依赖于先验信息，都不能完全保证所发现的分布能真实反映每个视图对数据的贡献。约束拉普拉斯秩（constrained Laplacian rank，CLR）多视图聚类方法的优势在于，它能够避免后处理步骤。CLR通过在矩阵中加入秩的限制，生成了新的相似性矩阵$\boldsymbol{S}$，这个矩阵更加可靠，并且可以直接用于聚类分析。这个过程可以表示为：

$$\begin{aligned} &\boldsymbol{A}_{\mathrm{c}} = \boldsymbol{I}, \quad a_{\mathrm{ch},j} \geqslant \min(0, \mathrm{rank}(L_s)) = \\ &\qquad v - i \| \boldsymbol{A} - \boldsymbol{S} \|_E^2 \end{aligned} \quad (10)$$

其中，相似性矩阵$\boldsymbol{S}$由原始数据导出，$a_{\mathrm{ch},j}$表示第j个视图的权重系数，L_s是基于相似性矩阵$\boldsymbol{S}$构造的拉普拉斯矩阵，v表示样本的索引。它提供了一个hyper参数来加以限制，同时使用CLR进行多视图集群。目标函数如下所示：

$$\begin{aligned} &\min_{\omega^v, \boldsymbol{S}} \sum_{n=1}^{k} \omega^v \| \boldsymbol{A} - \boldsymbol{S}^n \|_{E+r}^2 \| z_2^2 \| \\ &Lz^n \geqslant 0, w^L 1_m = \boldsymbol{I}, a_{\mathrm{ch}} \geqslant 0, \\ &A_i \boldsymbol{I}_v = 1, \mathrm{rank}(L_{\mathrm{s}}) = v - i \end{aligned} \quad (11)$$

其中，ω^v表示第v个样本的权重，r表示用于控

制权重分布均匀性的参数，a_{ch}表示所有视图的权重系数，$\boldsymbol{S}^n$表示第n个视图的相关相似性矩阵，Z为列向量$z^1,z^2,\cdots,z^k$中的整数数组，$z^k>0$。

此外，最终的约束条件确保了权重分布的均匀性。聚类精度受到某个值的显著影响，这个值可大可小，但它不会改变权重的分配方式。新的目标函数如下所示。

$$A_{c_h}\geqslant 0, a_c I_v=\min_{\mathrm{rank}(T_a)=v-c}\sum_{n=1}^{m}\omega^{v\|\boldsymbol{S}-\boldsymbol{A}^n\|_E} \quad (12)$$

式（12）是一个简洁而高效的表达式，但尚未定义权重项。应用拉格朗日乘数法进行推导和微调后，式（12）可转化为以下形式：

$$A_{c_h\geqslant R, I_n}=\min_{\mathrm{rank}(L_S)=v-i}\sum_{n=1}^{k}z^{n\boldsymbol{A}-\boldsymbol{S}^n}E \quad (13)$$

其中，$z^n=1/\left(2\|A-S^n\|_E\right)$，在给定矩阵$\boldsymbol{A}$的情况下，它被视为一个常数值。当矩阵$\boldsymbol{A}$被计算或更新时，$z^n$的值也会相应地自动更新。在迭代过程结束后，可以利用SwMC算法找到最优的$\boldsymbol{A}$和z^n值。SwMC的一般步骤如下所示。

输入：$S=\left\{S^1,S^2,\cdots,S^k\right\},S^n\in O^{v\times v}$，簇数$m$

输出：相似矩阵$\boldsymbol{A}\in O^{v\times v}$

步骤1　初始化每个视图$z^n=\dfrac{1}{k}$；

步骤2　不收敛则继续执行；

步骤3　通过式（11）计算$\boldsymbol{A}$；

步骤4　利用$z^n=1/\left(2\left\|\boldsymbol{A}-\boldsymbol{S}^n\right\|_E\right)$更新$z^n$；

步骤5　结束循环

1.2.3　潜在多视图子空间聚类（latent multi-view subspace clustering，LMSC）

LMSC是一种创新的多视图子空间聚类方法，该方法的提出是基于当前子空间聚类领域自表示技术的重要进展。该方法能够通过恢复数据的潜在形态来挖掘数据的子空间表示。当数据恢复和子空间挖掘这两个过程被整合到一个统一的算法框架中时，便形成了一种结合增广拉格朗日乘子（augmented Lagrangian multiplier）法和交替方向乘子法（alternating direction method of multipliers，ADMM）的算法[12]。本文还特别考虑了算法对天气变化的适应性，通过对算法进行噪声数据分析，提供了一个针对性的解决方案，以应对算法在处理噪声数据时可能遇到的问题。利用多视图数据集，研究者可以探索同一数据在不同潜在表示之间的多种映射关系。尽管存在许多不同的多视图子空间聚类算法，但LMSC与它们的主要区别在于方法论。LMSC不是依赖于单一视图的子空间重构，而是在综合了所有视图的信息之后进行子空间的重构。这种方法的核心目标是从尽可能多的数据源中收集信息，以期构建一个更全面、更精确的数据表示。多视图聚类方法的一个直观演示如图1所示，图1展示了如何将来自不同视图的信息融合，以实现更准确的聚类效果。这种方法不仅提高了聚类的质量，而且为多视图数据分析提供了一种新的视角。图1中，H^n表示第n个图的潜在表征，P^n表示第n个视图的分布模型，X^n表示第n个视图的原始输入数据。

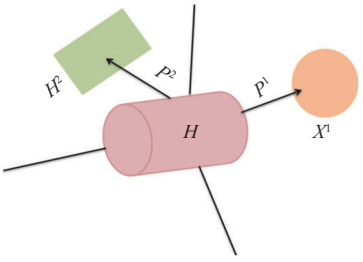

图1　多视图聚类方法的直观演示[13]

原始数据与其期望的潜在表示之间的联系，需要通过引入若干附加变量来明确。这些变量包括集合$\left\{D^1,\cdots,D^k\right\}$，其中$D^n\in O^{p^n\times o}$（假设此处的“$O$”实际上是指实数集“$\mathbf{R}$”）。每个视图都对应一个映射矩阵。$D^n$与潜在表示的乘积（乘以$J\in O^{o\times v}$）构成了对应视图的数据矩阵。需要预先确定的是，这些矩阵及其相互之间的关系需要被建立。J、D^n、Y^n之间的关系可以用式（14）来表示：

$$\min_{D,J}\ Y_j(Y,DJ) \tag{14}$$

Y和D通过垂直拼接的方式将两个矩阵集合$\left\{Y^1,\cdots,Y^k\right\}$和$\left\{D^1,\cdots,D^k\right\}$连接起来。$Y_j(\cdot,\cdot)$为潜在表示的损失函数。与其他多视图融合方法相比，这种方法使用权重系数来组合所有视图。

文献[13]使用了式（15）中的J作为数据特征的有效表示，并采用这种子空间聚类方法来研究理想的子空间表示。

$$\min_{W}\ T_o(J,JW)+\alpha\Omega(w) \tag{15}$$

其中，$T_o(\cdot,\cdot)$表示W的目标函数的解，$\Omega(\cdot)$是一个标量，用于正则化W，标量$\alpha>0$，用于平衡正则化项。值得注意的是，式（16）的形式受到了文献[14-16]内容的启发。如前所述，在考虑将子空间聚类与潜在表示学习相结合的3个因素之后，本文引入了额外的参数λ_1和λ_2来平衡式（15）。使用$l_{2,1}$范数来考查噪声对数据的影响，目标函数为：

$$\begin{aligned}&\min_{D,J,W,F_j,F_o}\ \|F_j\|_{2,1}+\lambda_1\|F_o\|_{2,1}+\lambda_2\|W\|_*\\&\text{s.t.}\,Y=DJ+F_j,J=JW+F_o\end{aligned} \tag{16}$$

其中，为了提高抗噪声性和列稀疏性，使用$\|\cdot\|_{2,1}$表示核矩阵$\|\cdot\|_*$。为了避免过拟合，将矩阵$\boldsymbol{W}$设为低秩，且存在D限制是为了防止J在整个计算过程中降至零，最后用约束等式进行检验。以下算法是基于J的同一学习过程$\boldsymbol{W}$中的几个视角，可以帮助理解如何获取潜在表示J和子空间表示，第3个视角确保$\boldsymbol{W}$的解更正常。SwMC的一般步骤如下所示。

输入：$y=\left\{y^1y^2,y^k\right\},y^n\in y^{p^n}\times v$，集群数量$k$，参数$\lambda$，潜在表示$H$的维数$r$

输出：E,J,D,F

步骤1　初始化$D=0,F=0,H=0,\boldsymbol{W}=0,X_1=0,X_2=0,X_3=0,\boldsymbol{W}=10^{-6},D=1.1,F=10^{-4},\max_w=10^6$，用随机值初始化$J$；

步骤2　若不收敛则继续执行；

步骤3　通过$D=\arg\min\frac{w}{2}\left\|\left(X+\frac{1}{w}x_1-F_J\right)-J^{-\mathrm{T}}D^{\mathrm{T}}\right\|_F^2$更新$D$；

步骤4　通过$SJ+JA=I$更新J

有$A=\boldsymbol{W}D^t,D,A=\boldsymbol{W}\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}-\boldsymbol{W}^{\mathrm{T}}+I\right);C=D^{\mathrm{T}}\boldsymbol{X}_1+\boldsymbol{X}_2\left(\boldsymbol{W}^{\mathrm{T}}-I\right)$

步骤5　通过$H=\left(J^{\mathrm{T}}H+I\right)^{-1}\left[\left(H+J^{\mathrm{T}}J-J^{\mathrm{T}}E_R+\left(X_3+J^{\mathrm{T}}X_2\right)/\boldsymbol{W}\right)\right]$更新$W$；

步骤6　通过$F=\arg\min_f\ \frac{1}{w}\|F\|_{2,1}+\left\|\frac{1}{2}\right\|\|F-J\|_F^2$更新$F$；

$\left\|\frac{1}{2}\right\|H-\left(W7:\text{通过}H=\frac{\lambda}{\boldsymbol{W}}\|H\|_*+X_3/\boldsymbol{W}^2{}_F\right)$更新$H$；

步骤7　通过$\begin{cases}\boldsymbol{X}_1=\boldsymbol{X}_1+\boldsymbol{W}\left(Y-DJ-F_J\right)\\\boldsymbol{X}_2=\boldsymbol{X}_2+\boldsymbol{W}\left(J-JW-F_O\right)\\\boldsymbol{X}_3=\boldsymbol{X}_3+\boldsymbol{W}\left(J-Z\right)\end{cases}$更新$\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$；

步骤8　通过$W=\min\left(dw;\min_w\right)$更新$\boldsymbol{W}$;

步骤9　验证是否满足循环结束的条件：

$$\|Y-DJ-F_J\|_\infty<\epsilon,\|J-JW-F_O)\|_\infty<\epsilon\text{and}\|H-W\|_\infty<\epsilon;$$

1.2.4　具有完整性感知相似性的多视图子空间聚类

本文所采用的多聚类算法，包括SwMC、LMSC和具有完整性感知相似性的多视图子空间聚类（multi-view subspace clustering with intactness-aware similarity，MSC IAS），均构建于优化理论的基础之上。这些算法通过迭代优化策略逐步逼近全局或局部最优解。具体来说，每个算法都设定了明确的停止准则，包括但不限于达到预设的迭代次数、目标函数的梯度下降到一个足够小的阈值，或目标函数值的改善低于特定的精度界限。

此外，实验跟踪了算法性能随迭代次数的变化，结果表明，随着迭代的进行，目标函数值逐渐稳定，从而验证了算法的收敛行为。这种对收敛性的监控和保证，确保了聚类结果既可靠又有效。

在基于图的聚类技术中，由于数据维度巨大且存在大量冗余和无用特征，相似性矩阵的构建常常是不准确的。如果从多个角度查看数据，会进一步加剧混乱。MSC IAS是一种针对多视图数据的新型子空间聚类方法。由于IAS使用了完整性空间学习，因此它可以提供一个更可靠的相似性矩阵，从而有助于聚类。

在具有完整性意识的情况下获得相似矩阵后，将归一化割（normalized cut，Ncut）方法用于相似矩阵。具体而言，作者的“完整空间”概念是指数据表示保留其所有信息，同时以协调的方式减少数据量的空间。因此，它可以包含形成相似矩阵所需的属性。MSC IAS的基本结构如图2所示。其中，X^n表示第n个视图的原始数据输入，L^n表示第n个视图的低维嵌入。

2 常用张量链方法

本文使用类似于广义奇异值分解的TT分解来压缩张量。TT分解在张量化向量和矩阵近似中的应用提供了它们的指数的系统细分方法，这种方法被定义为QTT，这将是本文关注的主题。基于此，矩阵可以被认为是对某种张量化向量进行运算的张量化算子。利用这一理解，本文展示了如何有效地将TT分解用作结构化矩阵的分层压缩和逆技术。采用传统方法对来自庞大数据集的异构数据进行分组是困难的。张量分解的特点在于，它可以成功地从大数据集中提取结构信息，从而有助于对数据进行分组和压缩。

2.1 CP分解法

PCA的高阶阵列被扩展为包括作为d模式张量的CP分解（CANDECOMP/PARAFAC decomposition，CPD）法。

$$\boldsymbol{Y}=\sum_{o=1}^{O}\lambda_o s_o{}^{(1)}\otimes s_o{}^{(2)}\otimes\cdots\otimes s_o{}^{(p)} \tag{17}$$

其中，O通常代表一个正整数，λ_o表示第r个秩因子中的权重，$s_o{}^{(c)}\in O^{v_c}$表示第c个模式的一个因子，其范数为1，$c\in[p]$且$o\in[O]$。

$$\boldsymbol{Y}=A\times_1 S^{(1)}\times_2 S^{(2)}\cdots\times_p S^{(p)} \tag{18}$$

其中，S表示对角核心张量与$\boldsymbol{X}$的模式积。因子矩阵$\boldsymbol{s}^{(c)}=\left[s_1{}^{(c)}s_2{}^{(c)}\cdots s_O{}^{(c)}\right]$，$c\in[p]$。

PARAFAC模型的一个基本限制是，不同模式中的成分仅与因子相互作用。例如，在3模式

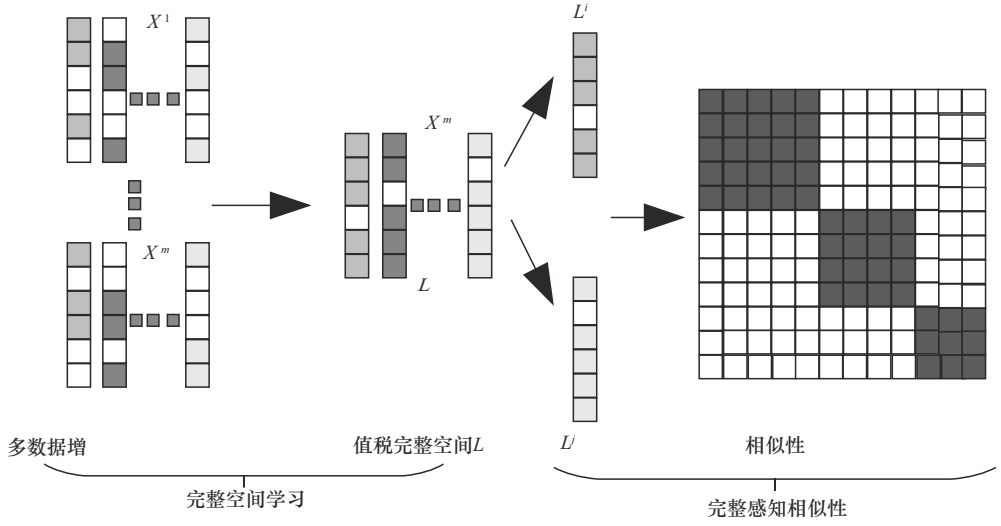

图2 MSC IAS的基本结构[17]

张量中，第一模式的第i个因子不会与其他模式中的第i个因子发生交互作用。对于d阶张量的Rank-R近似$O^{v_1\times v_2\times\cdots\times v_p}$，当对展开矩阵应用PCA时，所需的参数数量通常少于对展开矩阵应用CPD时的参数数量。文献[16]提供了两种替代证明，阐述PARAFAC模型在文献[17]这篇综述研究中的不相似性。在共同置换下，张量$\boldsymbol{Y}$的PARAFAC分解中的因子矩阵在秩为O的情况下基本上是唯一的，而列缩放对于所陈述的词汇量是唯一的。Kruskal使用矩阵$k-$秩得出了关于3模式CPD唯一性的结果。

$$m_{S^{(1)}}+m_{S^{(2)}}+m_{S^{(3)}}\geqslant 2O+2 \tag{19}$$

其中，$m_{S^{(c)}}$是$S(c)$的任意m列线性独立的最高k值[19]。文献[20]将这一结论推广到$p-$模张量为：

$$\sum_{c=1}^{p} m_{S^{(c)}}\geqslant 2O+p-1 \tag{20}$$

假设在评估第二模式中未知的需求集合之前，已经识别了第一的组件，对于每种模式和每次迭代，输入张量与CPD近似之间的差值的Frobenius范数会减小。ALS的优点在于，它保证了解决方案会随着迭代次数的增加而改进。然而，在实际应用中，较大的噪声或高阶模型可能会阻止ALS达到全局最小值，或迫使ALS进行数百次重复迭代[21-23]。

为了提高CPD算法的性能并加速收敛速度，已经有学者设计了多种解决方案[24-25]。线性搜索外推方法[26-27]和压缩方法[28]是两种具体的策略示例。OPT算法[29]、非负CP的梯度下降算法[30]、PMF3、阻尼高斯-牛顿（dGN）算法[31]以及快速dGN算法[32]都是为了解决在某些情况下ALS收敛缓慢的问题而进行的研究。此外，还需要考虑CP分解的联合对角化问题[33-34]。

2.2　Tucker分解与高阶奇异值分解

通过Tucker分解对d模式张量进行分解，该方法是通过将每个模式与一个核心张量相乘，然后再乘以一个矩阵来实现的。Tucker分解将d模式张量$\boldsymbol{X}$分解为：

$$\begin{aligned}\boldsymbol{Y}&=\sum_{c_1=1}^{v_1}\cdots\sum_{c_p=1}^{v_p}A_{c_1,c_2,\cdots,c_p}\\&\left(w^{(1)}{}_{c_1}\otimes w^{(2)}{}_{c_2}\otimes\cdots\otimes w^{(p)}{}_{c_p}\right)\\Y&=A\times{}_1\boldsymbol{W}^{(1)}\times 2\boldsymbol{W}^{(2)}\cdots\times{}_p\boldsymbol{W}^{(p)}\end{aligned} \tag{21}$$

其中，矩阵$\boldsymbol{W}^{(c)}=\left[w_1{}^{(c)}w_2{}^{(c)}\cdots w_{v_p}{}^{(c)}\right]$是平方因子矩阵。$A=Y\times{}_1\boldsymbol{W}^{(1),L}\times{}_2\boldsymbol{W}^{(2),L}\cdots\times{}_p\boldsymbol{W}^{(p),L}$，$\boldsymbol{W}^{(c),L}$表示沿每个模式的因子矩阵的转置。Tucker分解通常假设$U(i)_s$的秩小于n_i。

与PARAFAC不同，Tucker模型能够实现跨模式收集的因子之间的相互作用，这些相互作用的强度被包含在核心张量中。图嵌入方法通常用于降低数据的维度，同时保持图的数据结构，以便更好地对数据进行分类并识别目标数据。最后但同样重要的是，CPD和Tucker分解都是基于它们外部乘积之和的模型表示，其中一个模型的最一般版本包含了另一个模型。然而，正是它们的独特性使它们在实际应用中有所不同。Tucker分解与高阶奇异值分解（higher-order singular value decomposition，HoSVD）是Tucker分解的一种形式，它通过限制成分矩阵来实现正交性。左系统是一种约束条件，要求每个降低的张量$\boldsymbol{X}$都是HoSVD中的因子矩阵$\boldsymbol{U}(i)$。通过对HoSVD的正交因子矩阵进行子集划分，可以产生截断的HoSVD，它具有较低的n秩并近似于张量$\boldsymbol{X}$。由于核心张量的正交性，对于给定的多线性秩，HoSVD是唯一的。与矩阵的SVD不同，HoSVD的$(R1,R2,\cdots,Rd)$截断并不是$\boldsymbol{X}$的最佳$(R_1,R_2,\cdots,R_d)$近似。解决以下优化问题可以得到最佳的$(R_1,R_2,\cdots,R_d)$秩近似。

使用非负分解方法时，Tucker模型被用来寻找张量中潜在的非负实数模式[36-39]。张量的非负张量分解（non-negative tensor decomposition，NTD）可以通过解决以下问题来计算：

$$\min_{S, W^{(1)}, W^{(2)}, \cdots, W^{(p)}} Y - A \times_1 W^{(1)} \times_2 W^{(2)} \cdots \times_p W^{(p)} \quad (22)$$

属于 $A \in O_+^{O_1 \times O_2 \times \cdots \times O_p}, W^{(c)} \in O_+^{v_1 \times O_2}; c \in [p]$

为了解决这一优化问题，可以使用非负ALS以及在每次迭代中利用多种更新方法[39-40]或低秩NMF[41-42]来修改核心张量$\boldsymbol{S}$和因子矩阵$\boldsymbol{U}(i)$。

3 张量网

像PARAFAC和Tucker这样的张量分解方法用于将复杂的有效数据张量分解为基本张量和矩阵。一方面，张量网络（tensor network，TN）具有一个更高级别的张量作为核心，这在计算和存储方面提供了便利[43-45]。当网络中的一个或多个张量被压缩时，所形成的结构就被称为TN。当TN与指定的开放索引收缩时，会创建一个新的张量。对于给定的张量，有许多不同的TN表示，确定收缩指数的最佳顺序对TN分解效率至关重要。由于优化了拓扑结构，高级张量数据的图形表示变得简单而直观[45-46]。常见的TN拓扑如树张量网络状态（tree tensor network state，TTNS）、TT、投影纠缠对态（projected entangled pair states，PEPS）和投影纠缠对算子（projected entangled pair operators，PEPO）等。

3.1 层次张量的分解

Tucker分解已被扩展为使用层次Tucker（hierarchical tucker，HT）分解（也称为层次张量表示）来减少内存需求[48-50]。HT分解通过基于层次结构迭代的划分模式，为每个节点[51]创建具有一组特定维度模式$T[d]$的树状Tucker分解。

3.2 TT的分解

TT分解可以被视为HT的一个特定实例，其中所有节点都是相连的。底层的张量网络以列车或级联的方式链接。在分解高阶张量时，模型参数的数量不会随着张量维度的增加而呈指数增长，因此可以将巨大的张量数据压缩为更小的核心张量[52]。这种方法避免了Tucker模型的指数增长，并提供了更有效的存储复杂性。一个张量的TT分解$Y \in O^{v_1 \times v_2 \times \cdots \times v_p}$可以表示为：

$$Y_{c_1, \ldots, c_p} = J_1(c_1) \cdot J_2(c_2) \cdot \cdots J_p(c_p) \quad (23)$$

一系列SVD用于推导$\boldsymbol{X}$的TT分解。首先，对$\boldsymbol{X}$的第一个维度矩阵化进行SVD得到第一个维度矩阵化后的结果G_1。

在量子物理学领域，TT形式被定义为具有开放边界条件（open boundary condition，OBC）的矩阵乘积态（matrix product state，MPS）表示[53]。相较于HT模型，TT/MPS模型展现出几个优势，包括更简单的实际实现、更高的计算效率（线性与张量阶相关）以及计算的简便性。尽管TT形式在信号分析和机器学习中被广泛应用，但它仍存在不足。为了初始化TT模型，需要对边界因子进行秩–1约束，这意味着它们需要以矩阵的形式存在。TT核心乘法不具备置换不变性，因此需要采用如互信息估计[54-55]等优化技术。为了克服这些挑战，TR分解应运而生[56-57]，通过去除边界核心的单位秩限制，并用跟踪操作替换它们，从而去除了核心顺序的依赖性。

3.3 截断奇异值的分解

t乘积[58]被用于定义三阶张量的截断奇异值的分解（truncated singular value decomposition，t-SVD）。与标准的多线性代数相反，实现t-SVD的代数是建立在三阶张量上定义的线性运算上的。三阶张量可分解为：

$$\boldsymbol{Y} = W \times A \times N^{\mathrm{T}} \quad (24)$$

其中，t乘积表示将相同大小的模式3纤维按循环方式排列。这种分解可以使用傅里叶级数矩阵SVD来完成。$\boldsymbol{X}$的管秩是通过高阶奇异值分解确定的，它等同于$\boldsymbol{S}$中具有显著非零奇异值的独立管子的数量。此外，类似于CPD和Tucker模型，具有一定秩的t-SVD可以被证明是在Frobenius范数意义下减少误差的最佳近似。

4　实验结果分析

本次实验的实验环境设置见表 1。

本节使用 7 个公开可用的数据集对上述方法进行实证评估。实验在受控环境中进行，并配备了高性能计算资源，以确保获得准确可靠的结果。实验设置包括使用最先进的硬件服务器，这些服务器配备了强大的处理器、充足的内存和专用的图形处理单元（graphics processing unit，GPU），以处理与多视图聚类相关的计算密集型任务。这种强大的基础设施能够助力高效地处理大型数据集并及时获得结果。为了基准测试提出方法的性能，本文将其与广泛使用的 *k*-means 聚类方法进行了比较。两种方法都在流行的编程语言中实现，并利用优化的库和算法来最大化计算效率。实验环境由先进的硬件和软件资源组成，经过精心配置，以支持对本研究中描述的多视图聚类方法进行严格测试和准确评估。然而，直接应用多视图聚类方法并不适用多数情况。因此，本文将多个视图的元素组合成一个单独的视图进行处理。为了比较上述两种方法之间的性能差异，本文选择了精确度（accuracy，ACC）、归一化互信息（normalized mutual information，NMI）和纯度 3 个指标。ACC 是衡量有效算法评估多维数据能力的一个指标，是通过计算被正确放置在同一簇中的配对所占的百分比得出的。NMI 是一种评估群组发现方法执行网络划分效果的指标。纯度是衡量一个簇中包含多少单一类别数据点的指标，其计算可以概念化为：对每个簇，计算构成该簇多数的类别的数据点数量所占的比例。

多维数据库为高效分析数据和生成解决方案提供了强大能力。与关系型数据库相比，多维数据库能够更快速地压缩数据，并支持对多个产品维度的数据进行模拟和查看，这对许多行业尤其有利。然而，由于其复杂性，只有具备相关知识的专家才能全面理解和分析这些数据。

此外，本节还比较了不同分解方法的数据压缩率和归一化重构误差。各种编码方法在不同数据集上的输入参数值范围见表 2。

4.1　张量数据的高效编码

数据压缩，即通过减少表示数据所需的位数，可以有效降低网络带宽需求，加速文件传输，并节省存储空间。PIE 数据集包含了 138 张照片，这些照片分别是在 6 个不同的视角和 6 种

表 1　实验环境设置

实验环境		配置细节
硬件环境	服务器型号	Dell PowerEdge R740
	处理器	2x Intel Xeon Gold 6258R (24 核, 3.0 GHz)
	内存	512 GB DDR4 RAM
	存储	8 TB SSD（固态硬盘）
	GPU	NVIDIA Tesla V100 32 GB
	网络	10 Gbit/s 高速网络连接
软件环境	操作系统	Ubuntu 20.04 LTS
	编程语言	Python 3.8
	数值计算库	NumPy 1.21.0, SciPy 1.7.0, Scikit-learn 0.24.2, TensorFlow 2.5.0
	图形处理库	CUDA 11.2, cuDNN 8.1

表2　各种编码方法在不同数据集上的输入参数值范围

编码方法	PIE数据集	HIS (a&b)数据集		COIL-100数据集
图形编码	4	5	6	7
TT（误差）	0.001～0.500	0.001～0.500	0.001～0.555	0.210～0.500
HT（最大H秩）	2～300	2～960	100～960	1～80
HoSVD（阈值T）	0.20～0.99	0.1500～0.999 9	0.884 1～0.999 9	0.40～0.76
HOOI（阈值T）	0.20～0.99	0.1500～0.999 9	0.884 1～0.999 9	0.40～0.76
CPD（秩R）	1～300	3～767	—	4～200

不同的照明条件下拍摄的[59]。PIE数据集在不同编码方法下的编码效果对比如图3所示。

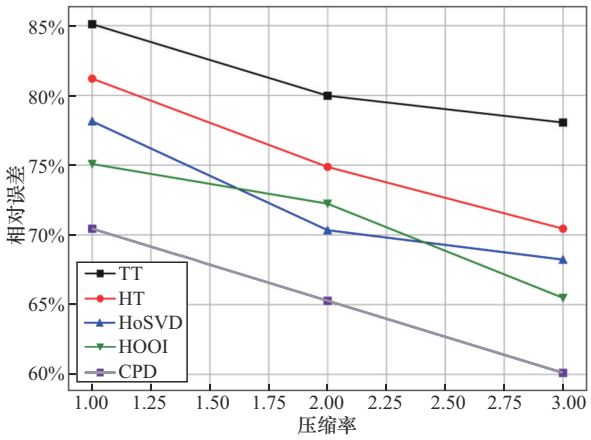

图3　PIE数据集在不同编码方法下的编码效果对比

图3对比了原始数据与经过各种编码方法处理后的数据存储需求，展示了每种编码技术在减少数据位数方面的效果。由图3可知，随着压缩率的逐渐增加，各方法的相对误差均有所下降。压缩率的对比揭示了各编码方法在压缩数据大小方面的表现。此外，图3还展示了压缩数据的重构质量，通过对比原始图像与重构图像来评估数据质量的保留情况。

4.2　HSI数据的张量压缩与聚类效果对比

压缩是指使用比原始信息更少的比特数对信息进行编码的过程。数据压缩可以节省磁盘空间、降低输入/输出（input/output，I/O）需求，或在传输数据时提高带宽利用率。HSI数据集包含了100张图像，每张图像均在148个不同波长下捕获。这种多光谱数据集提供了丰富的波长信息，使得每张图像不仅包含空间维度的信息，还包括光谱维度的数据。这种高维数据可以揭示目标物体的详细光谱特征，有助于进行更精确的图像分析和处理。对这些数据进行压缩和聚类分析，可以有效提取有用信息并减少存储和计算负担。现有方法和拟议方法在HIS（a）数据集和HIS（b）数据集上的对比分别如图4、图5所示。

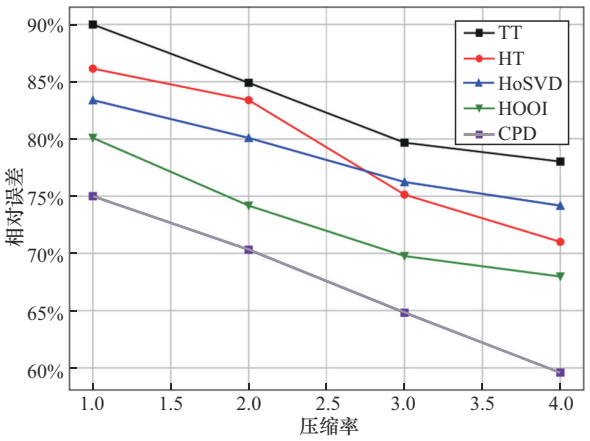

图4　现有方法和拟议方法在HIS(a)数据集上的对比

由图4和图5可知，随着压缩率的提升，算法的相对误差逐渐下降。这表明，压缩过程有效地减小了数据的冗余，同时保留了足够的关键信息，从而实现了较高的准确性。在HSI数据集的处理过程中，随着压缩率的增加，虽然压缩了数

据量，但由于压缩算法能够有效保留重要的空间和光谱信息，相对误差有所减少。这不仅能减少存储需求，还能提升数据传输效率和处理速度，同时不损失太多的分析精度。

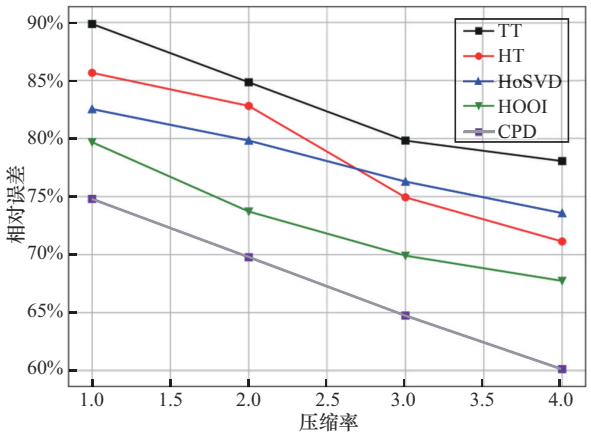

图5　现有方法和拟议方法在HIS（b）数据集上的对比

4.3　COIL数据的张量压缩与聚类分析

在聚类算法的性能评估中，MAME（maximum-mean）聚类有效性指标引入了一种新的视角。该指标旨在改善现有聚类有效性指标在类内紧致性和类间分离性度量上的不足。该指标综合考虑整个数据集的特征，计算了分为K类和1类情况下的比值，并提出了新的模糊紧致性度量表达式。此外，MAME指标还引入了最大聚类中心距离和平均聚类中心距离，为分离性的度量提供了新的方法。在5个UCI数据集和6个人工数据集上的对比实验显示，与9个其他聚类有效性指标相比，MAME指标展现了更高的准确性和稳定性，证明了其良好的鲁棒性。

本实验也采用了MAME指标来评估本文提出的聚类算法。结果显示，MAME指标相较于其他传统指标，提供了更加全面和准确的评估结果，这进一步验证了聚类方法的有效性和优越性。

测量的绝对误差与实际测量值的比值被称为相对误差。压缩率和相对误差作为参数。COIL-100数据集中包括来自100个对象的7 200张图像。每个对象的图像是从72个不同角度拍摄的，每个图像包含128个像素，且每个角度相隔5°。

COIL数据集上不同算法相对误差的对比如图6所示。实验结果表明，在处理大量数据集时，当输出表的输出与大小的比值（O/Size）较高时，压缩效果最好。当它收敛时，压缩性能最佳。对于COIL-100数据集，当压缩率超过2.0时，TT和HT虽然持续恶化，但依然优于其他算法。COIL数据集的结果表明，大多数情况下，TT不仅在较低的压缩率上优于其他算法，在更高的压缩率下依然优于其他算法。

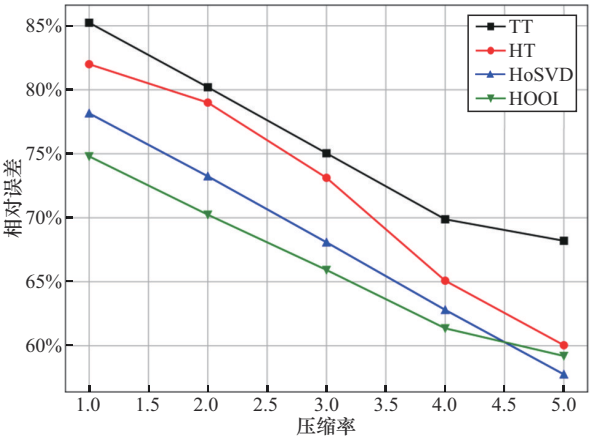

图6　COIL数据集上不同算法相对误差的对比

4.4　张量聚类性能评估

张量聚类性能评估旨在全面测量算法在实际应用中的表现。首先，计算ACC、NMI和纯度，评估算法在PIE、HSI和COIL数据集上的聚类准确性。其次，记录算法的运行时间和内存使用量，以测试其计算效率。再次，在数据中引入不同水平的噪声或缺失值，评估算法的鲁棒性。最后，测试算法在处理大规模数据集时的可扩展性，并通过可视化工具展示聚类结果的清晰度和准确性，从而综合评价算法的实际效果。

聚类算法散点图如图7所示。在聚类分析中，首先生成了3个不同簇的数据，这些数据点在

二维空间中形成了明显的群体。使用 Matplotlib 绘制散点图，直观地展示了这些数据点的聚类效果。散点图中通过不同颜色表示不同的簇标签，使得每个簇的分布和分界清晰可见。这种可视化呈现有助于理解数据在空间中的分布及聚类算法的效果。

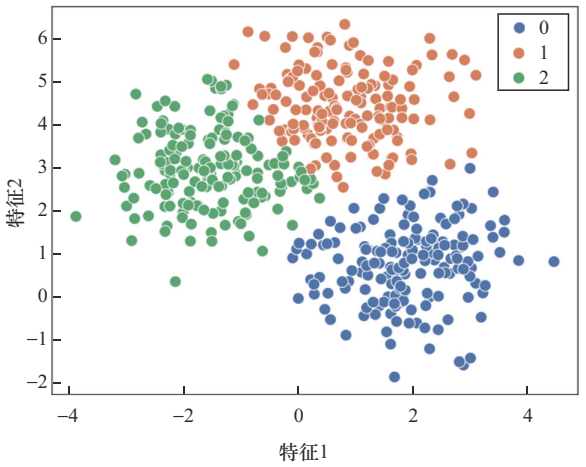

图7　聚类算法散点图

经过张量聚类分析，得到具体数据结果，见表3。

表3　具体数据结果

指标名称	结果
ACC	78%
NMI	0.59
纯度	78%

原始数据与相对应的噪声干扰情况如图9所示，图8的上半部分展示了原始数据的分布情况，每个簇的数据点用不同的颜色表示，下半部分则展示了带噪声的数据及其 *k*-means 聚类结果。不同噪声水平下的图形直观显示了噪声如何影响聚类效果，噪声水平越高，聚类结果的分散性和准确性可能越差。从图8中可以观察到噪声对聚类算法鲁棒性的具体影响。

原始数据与相对应的聚类情况如图9所示，原始数据的散点图清晰地展示了数据簇的真实分布，反映了张量聚类算法在理想条件下的性能。通过对图9的观察，可以看到每个簇的形状和分布情况，这对于理解张量数据中的簇结构至关重要。带噪声数据的散点图揭示了噪声对聚类结果的影响。噪声的引入模拟了实际应用中数据的复

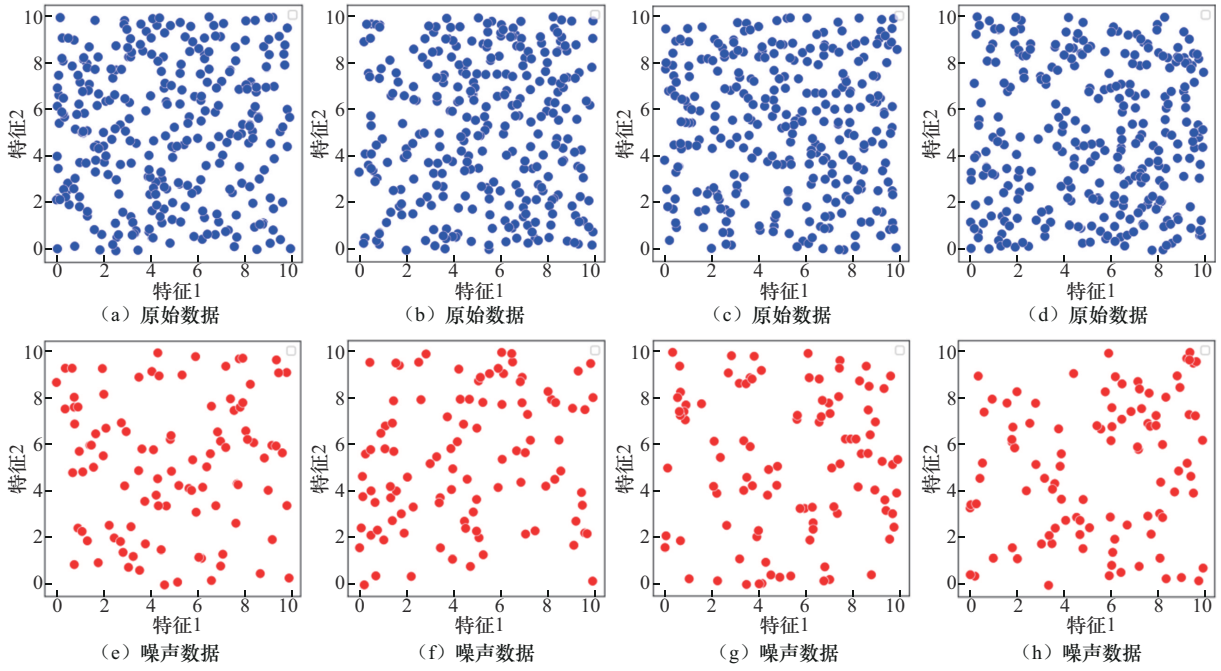

图8　原始数据与相对应的噪声干扰情况

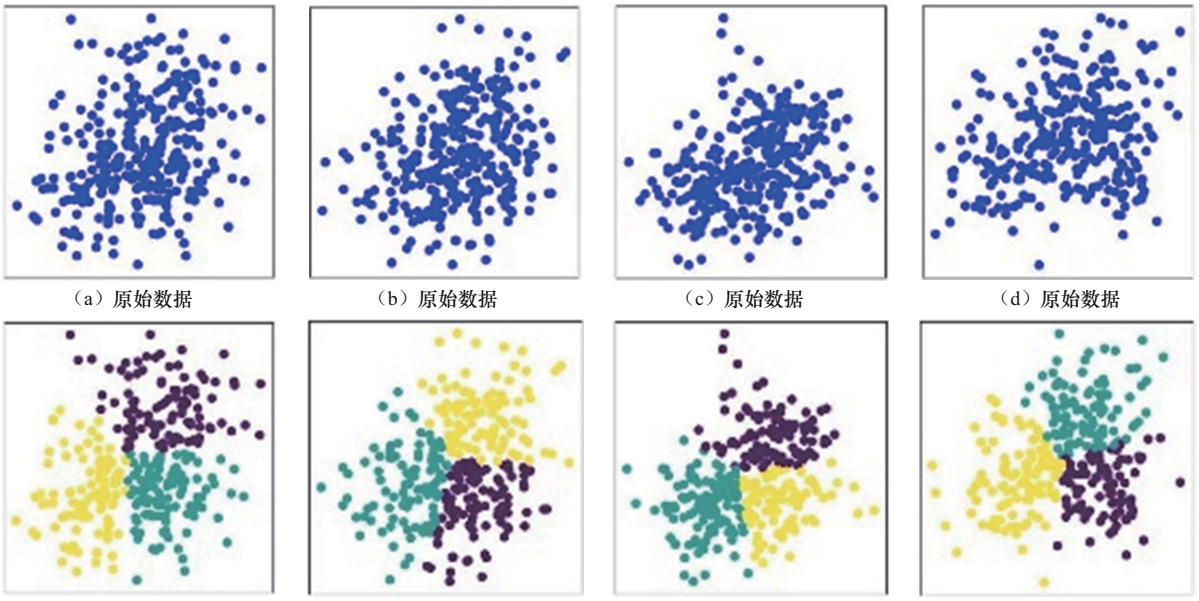

图9　原始数据与相对应的聚类情况

杂性，有助于评估张量聚类算法在面对数据噪声时的稳定性和准确性。

5　结束语

本文对8种多视图聚类方法进行了全面分析，并在7个不同的数据集上评估了它们的性能。通过采用ACC、NMI和纯度等指标进行量化，为这些技术的优势和局限性提供了有价值的见解。随着张量类型数据复杂性和维度的不断增加，层次张量分解方法正逐渐成为可视化和表示目的的关键工具。张量聚类的应用跨越了多个学科领域，包括深度学习、本体论、功能性磁共振成像分析、大数据管理、信息检索、非线性系统识别和知识发现等。传统聚类方法一直依赖于二元分类方法，在处理多个分类问题时可能会变得烦琐。与传统聚类方法相比，张量多聚类方法展示了更高的准确性，为处理复杂数据集提供了更高效的解决方案。张量分解在简化秩一分解和压缩大型数据集方面发挥着关键作用，能够用简化的表示形式替换复杂的系数。这不仅降低了计算复杂性，而且以一种直观的方式揭示了数据内的结构连接。展望未来，张量聚类方法的进一步发展将提高多维数据分析的准确性并提升整体性能。利用张量分解和多视图聚类技术的优势，可以从复杂数据集中解锁新的见解，推动各领域的创新和进步。

参考文献：

[1] ZHAO Y L, YANG L T, ZHANG R H. A tensor-based multiple clustering approach with its applications in automation systems[J]. IEEE Transactions on Industrial Informatics, 2018, 14(1): 283-291.

[2] MAI Q, ZHANG X, PAN Y Q, et al. A doubly enhanced EM algorithm for model-based tensor clustering[J]. Journal of the American Statistical Association, 2022, 117(540): 2120-2134.

[3] KOWALSKI P A, JECZMIONEK E. Parallel complete gradient clustering algorithm and its properties[J]. Information Sciences, 2022, 600: 155-169.

[4] SEDIGHIN F, CICHOCKI A, PHAN A H. Adaptive rank selection for tensor ring decomposition[J]. IEEE Journal of Selected Topics in Signal Processing, 2021, 15(3): 454-463.

[5] TICHAVSKÝ P, PHAN A H, CICHOCKI A. Krylov-levenberg-marquardt algorithm for structured tucker tensor decompositions[J]. IEEE Journal of Selected Topics in Signal Processing, 2021, 15(3): 550-559.

[6] ZHENG Y L, LIU Q S. A review of distributed optimization:

problems, models and algorithms[J]. Neurocomputing, 2022, 483: 446-459.

[7] MICKELIN O, KARAMAN S. On algorithms for and computing with the tensor ring decomposition[J]. Numerical Linear Algebra with Applications, 2020, 27(3): e2289.

[8] BOSE A, WALTERS P L. A multisite decomposition of the tensor network path integrals[J]. The Journal of Chemical Physics, 2022, 156(2): 024101.

[9] LIU H Z, YANG L T, YAO T, et al. Tensor-train-based higher order dominant Z-eigen decomposition for multi-modal prediction and its cloud/edge implementation[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(2): 1353-1366.

[10] ASANTE-MENSAH M G, AHMADI-ASL S, CICHOCKI A. Matrix and tensor completion using tensor ring decomposition with sparse representation[J]. Machine Learning: Science and Technology, 2021, 2(3): 035008.

[11] CHEN H Y, AHMAD F, VOROBYOV S, et al. Tensor decompositions in wireless communications and MIMO radar[J]. IEEE Journal of Selected Topics in Signal Processing, 2021, 15(3): 438-453.

[12] LIN Z C, LIU R S, SU Z X, et al. Linearized alternating direction method with adaptive penalty for low-rank representation[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: ACM Press, 2011: 612-620.

[13] ZHANG C Q, HU Q H, FU H Z, et al. Latent multi-view subspace clustering[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 4333-4341.

[14] CHENG Y, ZHAO R L. Multiview spectral clustering *via* ensemble[C]//Proceedings of the 2009 IEEE International Conference on Granular Computing. Piscataway: IEEE Press, 2009: 101-106.

[15] ELHAMIFAR E, VIDAL R. Sparse subspace clustering: algorithm, theory, and applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2765-2781.

[16] HU H, LIN Z C, FENG J J, et al. Smooth representation clustering[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 3834-3841.

[17] WANG X B, LEI Z, GUO X J, et al. Multi-view subspace clustering with intactness-aware similarity[J]. Pattern Recognition, 2019, 88: 50-63.

[18] WHITE M, YU Y L, ZHANG X H, et al. Convex multi-view subspace learning[C]//Proceedings of the 25th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA: NIPS, 2012, 1673−1681.

[19] KRUSKAL J B. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics[J]. Linear Algebra and Its Applications, 1977, 18(2): 95-138.

[20] SIDIROPOULOS N D, BRO R. On the uniqueness of multilinear decomposition of N-way arrays[J]. Journal of Chemometrics, 2000, 14(3): 229-239.

[21] CICHOCKI A, MANDIC D, DE LATHAUWER L, et al. Tensor decompositions for signal processing applications: from two-way to multiway component analysis[J]. IEEE Signal Processing Magazine, 2015, 32(2): 145-163.

[22] KOLDA T G, BADER B W. Tensor decompositions and applications[J]. SIAM Review, 2009, 51(3): 455-500.

[23] GRASEDYCK L, KRESSNER D, TOBLER C. A literature survey of low-rank tensor approximation techniques[J]. GAMM-Mitteilungen, 2013, 36(1): 53-78.

[24] PHAN A H, TICHAVSKÝ P, CICHOCKI A. Fast alternating LS algorithms for high order CANDECOMP/PARAFAC tensor factorizations[J]. IEEE Transactions on Signal Processing, 2013, 61(19): 4834-4846.

[25] CHEN D, HU Y Y, WANG L Z, et al. H-PARAFAC: hierarchical parallel factor analysis of multidimensional big data[J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 28(4): 1091-1104.

[26] ANDERSSON C A, BRO R. The N-way toolbox for MATLAB[J]. Chemometrics and Intelligent Laboratory Systems, 2000, 52(1): 1-4.

[27] CHEN Y N, HAN D R, QI L Q. New ALS methods with extrapolating search directions and optimal step size for complex-valued tensor decompositions[J]. IEEE Transactions on Signal Processing, 2011, 59(12): 5888-5898.

[28] KIERS H A L. A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity[J]. Journal of Chemometrics, 1998, 12(3): 155-171.

[29] ACAR E, DUNLAVY D M, KOLDA T G. A scalable optimization approach for fitting canonical tensor decompositions[J]. Journal of Chemometrics, 2011, 25(2): 67-86.

[30] COHEN J, FARIAS R C, COMON P. Fast decomposition of large nonnegative tensors[J]. IEEE Signal Processing Letters, 2015, 22(7): 862-866.

[31] PAATERO P. A weighted non-negative least squares algorithm for three-way ‘PARAFAC’ factor analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1997, 38(2): 223-242.

[32] PHAN A H, TICHAVSKÝ P, CICHOCKI A. Low complexity damped Gauss: Newton algorithms for CANDECOMP/PARAFAC[J]. SIAM Journal on Matrix Analysis and Applications, 2013, 34(1): 126-147.

[33] DE LATHAUWER L. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization[J]. SIAM Journal on Matrix Analysis and Applications, 2006, 28(3): 642-666.

[34] DE LATHAUWER L, CASTAING J. Blind identification of underdetermined mixtures by simultaneous matrix diagonalization[J]. IEEE Transactions on Signal Processing, 2008, 56(3): 1096-1105.

[35] DE LATHAUWER L, DE MOOR B, VANDEWALLE J. A multilinear singular value decomposition[J]. SIAM Journal on Matrix Analysis and Applications, 2000, 21(4): 1253-1278.

[36] CICHOCKI A, ZDUNEK R, PHAN A H, et al. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation[M]. New Jersey: Wiley, 2009.

[37] MØRUP M, HANSEN L K, ARNFRED S M. Algorithms for sparse nonnegative tucker decompositions[J]. Neural Computation, 2008, 20(8): 2112-2131.

[38] KIM Y D, CHOI S. Nonnegative tucker decomposition[C]//Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2007: 1-8.

[39] CICHOCKI A, ZDUNEK R, CHOI S, et al. Non-negative tensor factorization using alpha and beta divergences[C]//Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. Piscataway: IEEE Press, 2007: III-1393-III-1396.

[40] AHMADI-ASL S, ABUKHOVICH S, ASANTE-MENSAH M G, et al. Randomized algorithms for computation of tucker decomposition and higher order SVD (HoSVD)[J]. IEEE Access, 2021, 9: 28684-28706.

[41] LI H C, LIU S, FENG X R, et al. Sparsity-constrained coupled nonnegative matrix-tensor factorization for hyperspectral unmixing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 5061-5073.

[42] MARMORET A, COHEN J E, BERTIN N, et al. Uncovering audio patterns in music with nonnegative tucker decomposition for structural segmentation[EB]. 2021.

[43] HAGHSHENAS R, GRAY J, POTTER A C, et al. Variational power of quantum circuit tensor networks[J]. Physical Review X, 2022, 12: 011047.

[44] CICHOCKI A, LEE N, OSELEDETS I, et al. Tensor networks for dimensionality reduction and large-scale optimization: part 1 low-rank tensor decompositions[J]. Foundations and Trends® in Machine Learning, 2016, 9(4/5): 249-429.

[45] ORÚS R. A practical introduction to tensor networks: matrix product states and projected entangled pair states[J]. Annals of Physics, 2014, 349: 117-158.

[46] HANDSCHUH S. Changing the topology of tensor networks[J]. arXiv preprint, 2014: 1203.1503.

[47] HÜBENER R, NEBENDAHL V, DÜR W. Concatenated tensor network states[J]. New Journal of Physics, 2010, 12(2): 025004.

[48] GRASEDYCK L. Hierarchical singular value decomposition of tensors[J]. SIAM Journal on Matrix Analysis and Applications, 2010, 31(4): 2029-2054.

[49] ZHAO Y L, YANG L T, ZHANG R. Tensor-based multiple clustering approaches for cyber-physical-social applications[J]. IEEE Transactions on Emerging Topics in Computing, 2018, 8(1): 69-81.

[50] PAN F, ZHANG P. Simulation of quantum circuits using the big-batch tensor network method[J]. Physical Review Letters, 2022, 128(3): 030501.

[51] 张宏俊, 李鹏, 王汝传, 等. 一种基于张量的多维印章数据处理方法: CN117592951A[P]. 2024-02-23.
ZHANG H J, LI P, WANG R C, et al. A tensor-based multi-dimensional seal data processing method: CN117592951A[P]. 2024-02-23.

[52] OSELEDETS I V. Tensor-train decomposition[J]. SIAM Journal on Scientific Computing, 2011, 33(5): 2295-2317.

[53] QIN Y L, TANG Z J, WU H Z, et al. Flexible tensor learning for multi-view clustering with Markov chain[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 26(4): 1552-1565.

[54] BARCZA G, LEGEZA, MARTI K H, et al. Quantum-information analysis of electronic states of different molecular structures[J]. Physical Review A, 2011, 83: 012508.

[55] EHLERS G, SÓLYOM J, LEGEZA, et al. Entanglement structure of the Hubbard model in momentum space[J]. Physical Review B, 2015, 92(23): 235116.

[56] ZHAO Q, ZHOU G, XIE S, et al. Tensor ring decomposition[J]. arXiv preprint, 2016: 1606.05535.

[57] WANG W Q, AGGARWAL V, AERON S. Efficient low rank tensor ring completion[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5698-5706.

[58] STAHLG, STAHLG. Contributions to a theoretical framework for CSCL[C]//Proceedings of the Conference on Computer Sup-

port for Collaborative Learning: Foundations for a CSCL Community. New York: ACM Press, 2002: 62-71.

[59] SHAIBU M. Assessment of physicochemical parameters and organochlorine pesticide residues in selected vegetable farmlands soil in Zamfara state, Nigeria[J]. Science Progress and Research, 2022, 2(2): 559-566.

[作者简介]

张宏俊（1985– ），男，博士，中国通信服务股份有限公司正高级工程师，主要研究方向为高性能计算、大数据关键技术。

张泽宇（2000– ），男，曼彻斯特大学人工智能学院硕士生，主要研究方向为机器学习、数据挖掘以及图像识别。

张颖娇（2000– ），女，南京邮电大学现代邮政学院硕士生，主要研究方向为网络安全、物流隐私自动化。

叶昊（1999– ），男，华为技术有限公司南京研究所助理工程师，主要研究方向为智能制造、自动化容器编排平台、云服务安全测试以及数据挖掘。

潘高军（1977– ），男，浙江省通信产业服务有限公司高级工程师，主要研究方向为物联网、大数据关键技术、人工智能、区块链。