Statistics and Application 统计学与应用, 2026, 15(3), 12-25
Published Online March 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.153051

Hans汉斯

基于流形学习的全球创新指数降维与聚类分析

何飞雪

重庆工商大学数学与统计学院，重庆

收稿日期：2026年2月2日；录用日期：2026年2月24日；发布日期：2026年3月4日

摘　要

全球创新指数(Global Innovation Index, GII)作为衡量国家创新能力的高维多指标体系，其复杂性对深入分析与直观可视化构成了挑战。传统线性降维方法在处理其非线性数据结构时存在局限，而现有流形学习方法未充分考虑特征重要性差异。为此，本研究提出一种融合特征权重的改进UMAP方法，旨在更有效地揭示全球创新格局的内在结构与集群特征。以2013~2022年118个经济体的GII数据为基础，首先通过熵权法计算特征权重，并将其融入UMAP的距离度量中以构建加权降维模型；进而使用K-Means聚类，结合多种评估指标量化聚类效果，最终采用TOPSIS方法进行综合评价排序。实验结果显示，熵权UMAP在聚类数为5时取得最优综合性能，其TOPSIS排名第一，较采用的PCA降维方法具有更优的结构识别能力，为全球创新格局分析提供了更鲁棒的数据处理工具，也为类似多指标综合评价体系的降维与聚类问题提供了新的方法参考。

关键词

全球创新指数，流形学习，熵权法，UMAP，聚类分析，TOPSIS

Dimensionality Reduction and Clustering Analysis of the Global Innovation Index Based on Manifold Learning

Feixue He

School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing

Received: February 2, 2026; accepted: February 24, 2026; published: March 4, 2026

Abstract

As a high-dimensional, multi-indicator system for measuring national innovation capabilities, the Global Innovation Index (GII) poses challenges for in-depth analysis and intuitive visualization due to

文章引用: 何飞雪. 基于流形学习的全球创新指数降维与聚类分析[J]. 统计学与应用, 2026, 15(3): 12-25.
DOI: 10.12677/sa.2026.153051

its complexity. Traditional linear dimensionality reduction methods have limitations in handling its non-linear data structure, while existing manifold learning approaches have not fully considered differences in feature importance. To address this, this study proposes an improved UMAP method that integrates feature weights, aiming to more effectively reveal the intrinsic structure and cluster characteristics of the global innovation landscape. Based on GII data from 118 economies between 2013 and 2022, feature weights are first calculated using the entropy weight method and incorporated into UMAP's distance metric to construct a weighted dimensionality reduction model. Subsequently, K-Means clustering is applied, and multiple evaluation metrics are used to quantify clustering performance. Finally, the TOPSIS method is employed for comprehensive evaluation and ranking. Experimental results show that entropy-weighted UMAP achieves optimal comprehensive performance when the number of clusters is set to 5, ranking first in the TOPSIS evaluation. Compared to the PCA dimensionality reduction method used, it demonstrates superior structural recognition capabilities. This study provides a more robust data processing tool for analyzing the global innovation landscape and offers a new methodological reference for dimensionality reduction and clustering in similar multi-indicator comprehensive evaluation systems.

Keywords

Global Innovation Index, Manifold Learning, Entropy Weight Method, UMAP, Cluster Analysis, TOPSIS

1. 引言

在全球创新竞争日趋激烈的背景下，各国创新能力的科学评估与量化分析已成为制定科技政策、优化创新资源配置的重要依据。全球创新指数作为衡量国家创新综合表现的核心指标体系，不仅反映了各国在创新投入、产出与效率上的差异，也为研究创新驱动因素提供了多维度的数据基础。近年来，相关研究围绕 GII 展开了深入探讨：例如，《2025 年全球创新指数》[1]报告揭示了全球创新呈现“强技术、慢落地”的结构性分化趋势，中国首次进入全球创新力前十，并在创新集群数量上位列第一；孙玛媛等[2]从国家创新体系视角出发，系统评估了中国在知识产出与无形资产等方面的优势与短板，指出中国需在创新监管、高等教育与开放合作等方面持续提升；Nasir 等[3]则通过面板数据方法，验证了创新投入、产出与效率对 GII 的显著正向影响，尤其在高分位点更为明显；Ma 等[4]基于机器学习构建了全球智能创新指数，揭示了高收入并不必然伴随高智能创新水平的结论，为智能创新评估提供了新工具；Rodrigues [5]整理了 2011~2022 年 GII 七大支柱的纵向面板数据，为跨国创新比较提供了结构化支持；Huarng 等[6]通过结构定性关联分析，深入挖掘了 GII 多层变量间的整体关系，为理解国家创新排名的形成机制提供了新的系统分析视角；Yu 等[7]进一步运用模糊集定性比较分析，揭示了高制度水平、人力资本与研究、基础设施等多因素共同构成高创新得分的因果组合；Crespo 等[8]通过模糊集分析，区分了高收入与低收入国家实现高创新绩效的差异化路径；El B 等[9]则聚焦低收入国家，强调应优先支持技术引进与本土应用能力建设，而非盲目追求研发投入；Eufrazio 等[10]基于 GII 面板数据，采用主成分分析与 K-Means 聚类，构建了涵盖 118 个经济体的创新模式分类框架。这些研究在丰富 GII 理论体系的同时，也凸显了当前创新评估中面临的高维数据冗余、非线性结构复杂以及多指标协同分析困难等共性挑战，尤其在全球创新数据不断累积与多维化的趋势下，如何高效提取关键创新特征、降数据维度，成为提升评估效率与

解释力的关键问题。

而在应对高维数据降维与结构保持的挑战中，流形学习方法因其在非线性数据降维中的优越表现，近年来在多个领域得到广泛应用与拓展。其中，t-SNE 及其优化变体在保持高维数据局部结构方面表现突出：Allaoui 等[11]提出 t-SNE-PSO 方法，通过粒子群优化改进 t-SNE 的梯度下降过程，有效缓解局部最小值与拥挤问题，提升降维与可视化效果；谢斌等[12]则针对彩色图像灰度化任务，提出基于 t-SNE 最大化的自适应方法，在保持视觉对比度的同时提升处理效率；邹黎敏等[13]结合熵权法与 t-SNE，对加权后的特征进行降维，在多个数据集上验证了其优于传统方法的降维效果。此外，均匀流形逼近与投影算法[14]作为一种基于黎曼几何与拓扑理论的非线性降维方法，因其在保持数据全局与局部结构方面的优越性能，近年来在机器学习领域受到广泛关注。周映宇等[15]采用 UMAP 替代传统 PCA，在公交车工况构建中实现了更低的信息损失与更高的表征效率；Chae 等[16]将 UMAP 与贝叶斯神经网络结合，提出多组分信息度量方法，显著提升了核电厂状态诊断的数据效率与分类精度；Tan 等[17]则在生物医学图像分割中引入 UMAP 表征采样，减少了标注数据需求并提升了分割稳定性；Mehrjardi 等[18]将 UMAP 与机器学习结合，实现了高性能的土壤类型划分；Anant 等[19]系统评估了 Aligned-UMAP 在纵向生物医学研究中的时序结构保持能力；黎耀康等[20]提出曼哈顿距离加权 UMAP 与改进宽度学习系统相结合的锂电池温度预测模型，在时空预测中表现出色；尹泽明等[21]则基于 UMAP 改进多域特征提取方法，在轴承故障诊断中实现 100%的识别准确率；张润等[22]对 UMAP 的理论基础与衍生方法进行了系统综述，为其在多领域应用提供了方法论参考。这些研究不仅展现了流形学习在复杂数据降维中的强大能力，也为特征工程与降维技术的融合创新提供了重要启示，特别是在处理具有多维度、非线性特性的综合评价数据时，如何科学地确定特征重要性并融入降维过程，成为提升分析效果的关键。

尽管上述研究在全球创新指数评估与流形学习方法应用中均取得了显著进展，但针对 GII 这类多维度综合评价数据的特点，现有研究在降维过程中较少考虑各指标对创新差异的实际贡献权重。特别是在机器学习领域，特征加权降维被认为是提升模型解释性和性能的重要手段，然而现有研究尚未系统探索将熵权法这一客观赋权方法与 UMAP 这一先进流形学习技术相结合，用于处理 GII 高维面板数据的非线性结构挖掘与特征重要性量化问题。特别是在面对 GII 多维度、指标非线性关联且重要程度不一的特性时，传统线性降维方法如 PCA 难以有效捕捉其内在流形结构，而直接使用流形学习方法如 UMAP 则可能忽视不同创新指标对整体差异的贡献差异，导致降维结果区分度不足。

基于以上背景，本研究提出一种融合特征权重的改进流形学习方法——熵权 UMAP，旨在更有效地揭示全球创新格局的内在结构与集群特征。研究以 2013~2022 年 118 个经济体的 GII 面板数据为基础，通过熵权法客观确定各创新指标的权重，并将其融入 UMAP 的距离度量中，构建加权降维模型；进而利用 K-Means 聚类对降维结果进行分组，并结合轮廓系数、Calinski-Harabasz 指数、改进 Davies-Bouldin 指数及簇内平方和等多种评估指标，以及 TOPSIS 综合评价方法，系统比较不同降维方案的聚类性能；最终基于最优聚类结果，结合现实发展背景对各类集群的形成机制与特征展开深入阐释。本研究不仅为 GII 等高维多指标体系的降维与聚类分析提供了新的方法框架，也为全球创新政策制定与效果评估提供了更稳健的数据分析工具。

2. 研究方法

2.1. 主成分分析

主成分分析(Principal Component Analysis, PCA)是一种线性降维方法，通过正交变换将原始相关特征转换为线性无关的主成分。该方法计算步骤流程如下：

(1) 对标准化数据矩阵 $X \in \mathbb{R}^{m\times n}$ 计算协方差矩阵， m 为样本数， n 为特征数。

$$C=\frac{1}{m-1}X^{\top}X$$

(2) 对协方差矩阵进行特征值分解：

$$Cv_i=\lambda_i v_i,\quad i=1,2,\cdots,n$$

(3) 根据累积方差贡献率选择主成分数 d：

$$\frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\geq\theta$$

(4) 构造投影矩阵并进行降维：

$$Z=XV_d$$

其中，$V_d=[v_1,v_2,\cdots,v_d]\in\mathbb{R}^{n\times d}$。

2.2. t 分布随机邻域嵌入

t 分布随机邻域嵌入(t-distributed Stochastic Neighbor Embedding, t-SNE)是一种适用于高维数据可视化的非线性降维算法，数学计算流程如下：

(1) 计算高维空间的条件概率分布：

$$p_{j|i}=\frac{\exp\left(-\left\|x_i-x_j\right\|^2/\left(2\sigma_i^2\right)\right)}{\sum_{k\neq i}\exp\left(-\left\|x_i-x_k\right\|^2/\left(2\sigma_i^2\right)\right)}$$

(2) 对称化得到联合概率分布：

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2m}$$

(3) 在低维空间中定义 t 分布概率：

$$q_{ij}=\frac{\left(1+\left\|y_i-y_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|y_k-y_l\right\|^2\right)^{-1}}$$

(4) 最小化 KL 散度损失函数：

$$\mathcal{L}_{\text{t-SNE}}=\mathrm{KL}\left(P\|Q\right)=\sum_{i\neq j}p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

(5) 梯度下降更新：

$$\frac{\partial\mathcal{L}}{\partial y_i}=4\sum_{j\neq i}\left(p_{ij}-q_{ij}\right)\left(y_i-y_j\right)\left(1+\left\|y_i-y_j\right\|^2\right)^{-1}$$

2.3. 均匀流形近似与投影

均匀流形近似与投影(Uniform Manifold Approximation and Projection, UMAP)是基于流形学习和拓扑数据分析的降维方法，核心计算流程如下：

(1) 构建高维空间模糊拓扑：

$$p_{ij}=\exp\left(-\frac{\max\left(0,\left\|x_i-x_j\right\|-\rho_i\right)}{\sigma_i}\right)$$

其中，ρ_i 为到最近邻的距离，σ_i 通过困惑度参数确定。

(2) 定义低维空间相似性：

$$q_{ij}=\left(1+a\left\|y_i-y_j\right\|^{2b}\right)^{-1}$$

(3) 最小化交叉熵损失函数：

$$\mathcal{L}=\sum_{i,j}\left[p_{ij}\log\frac{p_{ij}}{q_{ij}}+\left(1-p_{ij}\right)\log\frac{1-p_{ij}}{1-q_{ij}}\right]$$

2.4. 熵权 UMAP

在标准 UMAP 基础上引入特征权重机制，用熵权法客观确定各创新指标重要性权重，将权重向量融入 UMAP 距离度量。这使算法降维时更敏感响应重要特征变化，增强对关键创新维度的保持能力，其数学计算流程如下：

(1) 熵权法计算特征权重，设标准化后特征矩阵为：$X=\left(x_{ij}\right)_{m\times n}$，其中，$m$ 为本数，n 为特征数。

1) 首先将各特征值按列归一化，得到该特征下每个样本的比重，即：

$$P_{ij}=\frac{x_{ij}}{\sum_{i=1}^{m}x_{ij}}$$

2) 计算特征熵值，根据信息熵定义，第 j 个特征的熵值计算公式为：

$$e_j=-\frac{1}{\ln m}\sum_{i=1}^{m}p_{ij}\ln p_{ij},\quad j=1,\cdots,n$$

3) 计算差异系数与权重：

$$g_j=1-e_j,\quad w_j=\frac{g_j}{\sum_{k=1}^{n}g_k},\quad j=1,\cdots,n$$

得到各特征的权重：

$$w=\left[w_1,w_2,\cdots,w_n\right]^{\mathrm{T}}$$

(2) 权重融入距离度量。在 UMAP 的距离计算中，采用加权欧氏距离：

$$d_{\text{weighted}}\left(X_i,X_j\right)=\sqrt{\sum_{k=1}^{n}w_k\cdot\left(x_{ik}-x_{jk}\right)^2}$$

其中，$X_i=\left[x_{i1},x_{i2},\cdots,x_{in}\right]^{\mathrm{T}}$ 为第 i 个样本的特征向量。

(3) 熵权 UMAP 降维：将熵权法确定的各指标的重要性权重应用于 UMAP 的所有距离计算步骤，剩余步骤与标准 UMAP 算法保持一致。

3. 全球创新指数可视化及聚类分析

3.1. 数据来源

本研究采用世界知识产权组织(WIPO)发布的全球创新指数数据集，时间跨度为 2013 年至 2022 年，

涵盖 118 个具有完整数据的国家及经济体。原始数据集包括七个核心创新支柱的年度指标值：制度建设、人力资本与研究、基础设施、市场成熟度、商业成熟度、知识与技术产出以及创意产出，每个支柱由多个二级指标构成[10]。

3.2. 降维可视化分析

图 1~4 展示了 PCA、t-SNE、UMAP 和熵权 UMAP 四种降维方法在二维空间的可视化结果。为直观识别散点图潜在簇结构及分布特征，在可视化图中添加辅助红线示意可能簇群边界。

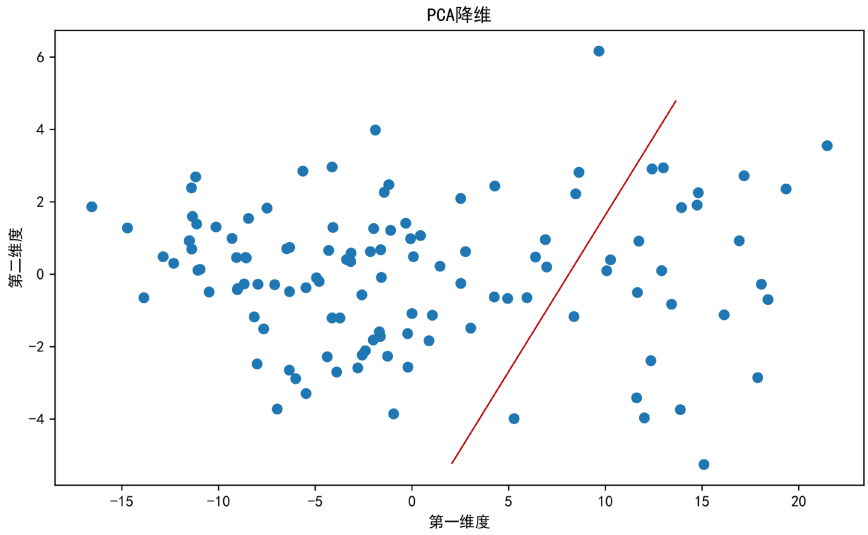

Figure 1. Distribution of PCA dimensionality reduction results
图 1. PCA 降维结果分布

PCA 作为线性降维方法，可视化结果呈典型散射分布，数据点在二维平面广泛散布，无明显簇状结构或密集区域，点间距分布均匀，无明显聚集中心或分组。

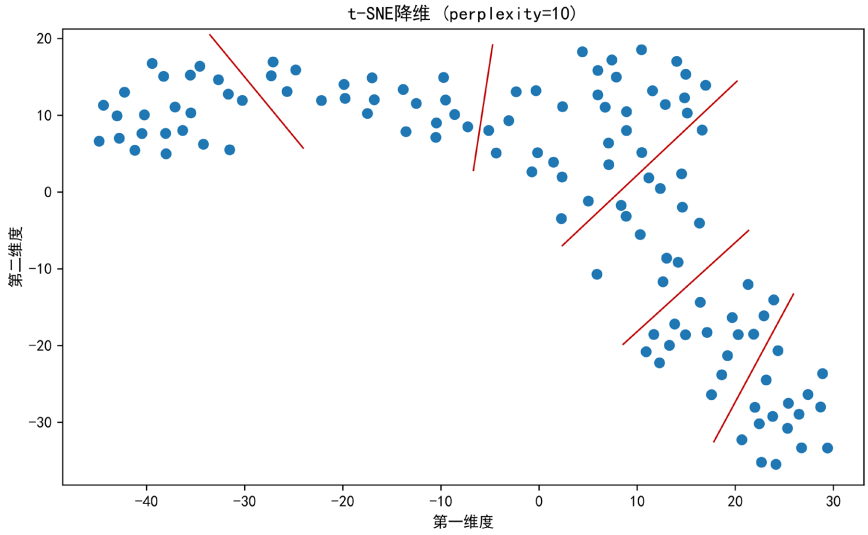

Figure 2. Distribution of t-SNE dimensionality reduction results
图 2. t-SNE 降维结果分布

t-SNE 降维结果有显著局部聚集特征，但整体分布分散，不同群体间有明显重叠，呈多中心、非均匀的复杂分布。数据点形成数个相对密集区域，该方法局部结构保持能力优于 PCA。但局部优化致全局结构失真，不同簇群相对距离不能准确反映原始空间关系，全局结构保持有局限。

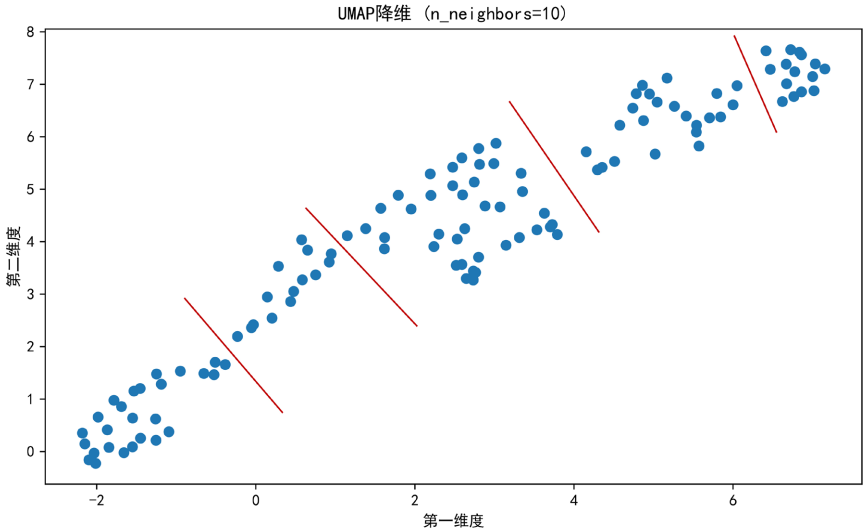

Figure 3. Distribution of UMAP dimensionality reduction results
图 3. UMAP 降维结果分布

UMAP 可视化结果呈现清晰多簇结构，能明显识别多个分离良好的簇群，簇内点分布紧凑，簇间有明显低密度间隙。为聚类分析提供理想基础；各簇群形状和大小有差异，反映不同创新群体在特征空间的非对称分布。

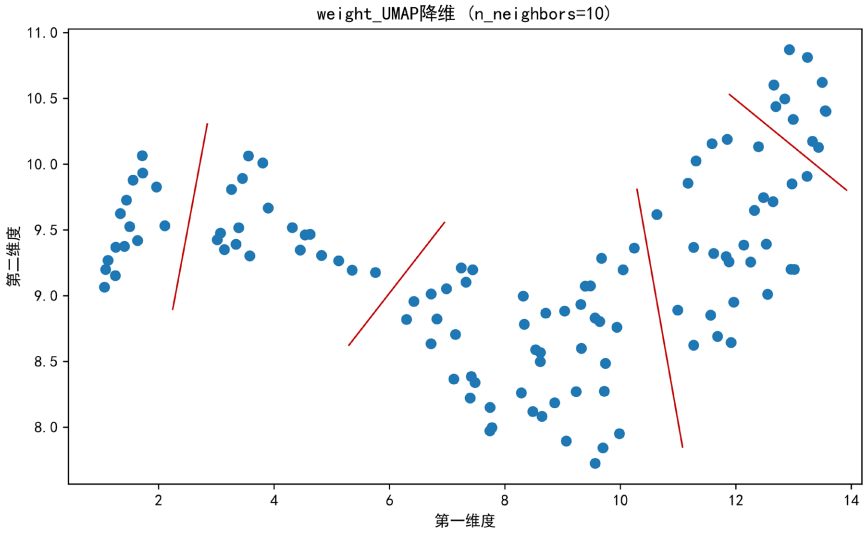

Figure 4. Distribution of entropy-weighted UMAP dimensionality reduction results
图 4. 熵权 UMAP 降维结果分布

熵权 UMAP 在普通 UMAP 基础上优化，可视化结果的簇结构清晰度最优。图中簇群边界更分明，簇内点分布更紧凑，簇间分离度提高。尤其通过熵权法引入的特征权重增强了重要创新指标区分能力，

有效分离降维空间中可能重叠的群体。

基于散点图观察，数据在多种降维方法下有清晰聚集结构。从不同可视化结果看，数据可能存在 2~6 个明显簇群，部分视图有 2~3 个主簇，细致分布中可辨识 4~6 个子簇。因此可初步确定聚类簇数 k 为 2、3、4、5、6，后续结合评估方法进一步确定最优簇数。

3.3. 聚类分析

3.3.1. 聚类结果评估指标体系构建

为系统评估不同降维方法对聚类效果的提升作用，本研究构建了一个多维度、互补性的评估体系，包含四个核心指标及一个综合决策方法。

(1) 轮廓系数(Silhouette Coefficient)：该指标同时考虑了簇内紧密度与簇间分离度，通过计算单个样本与同簇及其他簇样本的距离关系来评估其归属的合理性。计算公式为：

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

其中 a_i 为样本 i 到同簇其他样本的平均距离，b_i 为样本 i 到最近其他簇样本的平均距离，其取值范围为 $[-1,1]$。

(2) Calinski-Harabasz 指数：该指标通过衡量簇间离散度与簇内离散度的比值来评估聚类有效性，尤其适用于评估样本分布呈凸形或类球形簇的聚类结果。计算公式如下：

$$\mathrm{CH} = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(n-k)}$$

式中 B_k 为簇间离散度矩阵，W_k 为簇内离散度矩阵，k 为聚类数，n 为样本总数。分子反映簇间方差，分母反映簇内方差，该比值越大表示聚类结构分离度越好、簇内集中度越高。

(3) Davies-Bouldin 指数：该指标通过量化簇内紧密度与簇间分离度的平衡关系来评估聚类质量。原始计算公式为：

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{\bar{d}_i + \bar{d}_j}{d(c_i, c_j)}\right)$$

其中 $\bar{d}_i$ 为簇 i 内所有样本到簇中心的平均距离，$d(c_i, c_j)$ 为簇中心 c_i 与 c_j 间的距离。为保持评估体系的一致性，本研究采用 $1-\mathrm{DB}$ 的形式将其转化为正向指标，值越大代表聚类质量越高。

(4) 簇内平方和(SSE)：反映聚类结果紧凑性的基础指标，SSE 值越小，表示簇内样本越接近其中心，即聚类结果的紧凑性越好。计算公式为：

$$\mathrm{SSE} = \sum_{i=1}^{k}\sum_{x \in C_i} \|x - \mu_i\|^2$$

其中 C_i 为第 i 个簇，μ_i 为该簇的质心。

3.3.2. 各降维方法下聚类性能的综合评估

在完成多种降维方法的对比实验后，我们得到了如表 1 所示的综合评估结果。该表系统整理了以下五种方法在不同聚类数配置下的性能表现，涵盖了多个关键评价指标。

从综合 Topsis_score 排名看，熵权 UMAP 降维方法在聚类数为 5 时表现最优，综合得分 0.902466 排第一，聚类数为 6 和 3 的熵权 UMAP 方法分列第二、第三，表明熵权 UMAP 在多种聚类设置下综合性能良好。相比之下，原始数据各聚类数下综合得分普遍低、排名靠后，说明直接对高维原始数据聚类效

果有限，降维可提升聚类质量。

Table 1. Comprehensive evaluation results of clustering for various dimensionality reduction methods
表 1. 各降维方法聚类综合评估结果

序号	方法	聚类数	轮廓系数	CH 指数	1-DB 指数	SSE	Topsis_score	Ranks
1	原始数据	2	0.522815	193.7	0.312599	4419.77	0.598836	19
2	原始数据	3	0.363831	168.465	0.0552189	3002.68	0.484178	21
3	原始数据	4	0.296128	140.282	−0.0766611	2515.11	0.429802	23
4	原始数据	5	0.217944	116.312	−0.356596	2305.93	0.364347	24
5	原始数据	6	0.189108	103.242	−0.438848	2103.76	0.351035	25
6	PCA	2	0.619681	294.177	0.475944	2907.92	0.698189	10
7	PCA	3	0.513124	324.258	0.388377	1548.73	0.673912	16
8	PCA	4	0.470497	318.951	0.306562	1094.64	0.641434	17
9	PCA	5	0.421753	297.66	0.259634	891.284	0.602397	18
10	PCA	6	0.394101	291.029	0.185317	734.86	0.573417	20
11	t-SNE	2	0.531482	194.003	0.331436	32876.5	0.45234	22
12	t-SNE	3	0.584222	363.887	0.462681	11988.9	0.675574	15
13	t-SNE	4	0.537749	377.114	0.391087	8042.85	0.679916	12
14	t-SNE	5	0.49503	394.836	0.361562	5866.56	0.676109	14
15	t-SNE	6	0.493156	415.25	0.313544	4496.92	0.682669	11
16	UMAP	2	0.582435	262.284	0.447715	478.119	0.679032	13
17	UMAP	3	0.617255	492.437	0.512385	163.024	0.81417	7
18	UMAP	4	0.571594	537.266	0.446043	102.994	0.821492	6
19	UMAP	5	0.545316	573.407	0.429373	73.2092	0.826297	5
20	UMAP	6	0.522332	577.424	0.319256	58.2264	0.791999	8
21	weight_umap	2	0.611393	299.784	0.49641	511.538	0.707682	9
22	weight_umap	3	0.632673	545.311	0.525614	174.86	0.849254	3
23	weight_umap	4	0.56973	549.485	0.454357	118.574	0.828884	4
24	weight_umap	5	0.585141	654.938	0.484045	75.8023	0.902466	1
25	weight_umap	6	0.535873	762.369	0.398842	52.3251	0.878065	2

从不同降维方法对比结果看，UMAP 与熵权 UMAP 整体优于 PCA 和 t-SNE。聚类数为 3 至 5 时，UMAP 系列方法在轮廓系数和 1-DB 指数上表现突出，聚类结果有合理紧密度与分离度，CH 指数与 SSE 的优良表现验证了簇结构内部一致性。不过，CH 指数与 SSE 基于方差计算，对 K-Means 生成的球形簇结构有天然偏好。为全面评估聚类质量，本研究采用轮廓系数等对簇形状假设少的指标，形成互补评估体系。

值得注意的是，随聚类数增加，各方法轮廓系数普遍下降，这符合聚类分析规律，过细划分会降低

样本归属明确性。熵权 UMAP 在聚类数为 6 时 CH 指数仍较高(762.369)，反映其在高聚类数目下能产生方差分离度良好的簇结构，但要结合轮廓系数(0.585141)等指标综合判断，避免过度聚类。综上，本实验最优聚类方案是采用熵权 UMAP 降维并设定聚类数为 5，该方案在多项评估指标中表现均衡突出，尤其在轮廓系数上表现稳定，增强了结果可信度。

3.3.3. 熵权 UMAP 聚类结果分析

基于熵权 UMAP 降维与 K-Means 聚类算法得到的 5 类划分结果，结合全球创新发展的现实格局、经济发展阶段、区域地理特征及政策环境差异，对各聚类的形成逻辑与实践意义展开深度解析，具体如下表 2 所示。

Table 2. Statistical summary table of entropy-weighted UMAP-K5 clustering results
表 2. 熵权 UMAP-K5 聚类结果统计汇总表

聚类编号	国家数量	占比(%)	主要区域分布	核心特征
1	15	12.7	北美(3 个)、欧洲(8 个)、亚洲(4 个)	引领型创新强国经济体
2	18	15.3	欧洲(12 个)、大洋洲(2 个)、亚洲(4 个)	均衡型高成熟度经济体
3	18	15.3	欧洲(11 个)、亚洲(5 个)、美洲(2 个)	稳健型中坚力量经济体
4	29	24.6	美洲(11 个)、亚洲(9 个)、欧洲(6 个)	成长型潜力经济体
5	38	32.2	非洲(21 个)、亚洲(12 个)、美洲(5 个)	追赶型新兴经济体

聚类 1：引领型创新强国经济体。以美、德、日、中等 15 国为核心，构成全球创新第一梯队。研发投入高，人力资本雄厚，具备从基础研究到产业应用的完整生态，且制度环境成熟，知识产权保护有力，政府与市场协同高效。

聚类 2：均衡型高成熟度经济体。包括澳、加、奥等 18 个高收入发达国家。创新体系完善，科研基础扎实，以稳健渐进式创新为主，聚焦清洁能源、医疗健康等领域，注重社会福利与可持续，较少追求颠覆性突破。

聚类 3：稳健型中坚力量经济体。含波、土、阿联酋等 18 个中等偏上收入国家。创新路径体现“借力”与“聚焦”：融入区域一体化获取外部资源，在汽车、IT 外包等细分领域构建优势，政府通过科技园区、税收优惠等政策关键驱动。

聚类 4：成长型潜力经济体。集合巴、沙、智、白俄等 29 国，处于创新转型期。三类路径：资源国(如沙特)以能源财富投资未来科技；新兴市场(如巴西)在优势产业有亮点但易受宏观波动影响；后计划经济体(如塞尔维亚)正融入开放型创新网络。

聚类 5：追赶型新兴经济体。涵盖赞、坦、尼等 38 个亚非发展中国家，规模最大。面临基础设施薄弱、研发投入低、人力资本不足等制约，经济依赖初级产业，创新需求弱，制度环境不完善，体现全球创新鸿沟与突破低水平均衡的挑战。

3.3.4. 聚类结果地理可视化对比

为进一步深入评估降维方法对聚类结果的影响，本研究将熵权 UMAP 方法与 Eufrazio 和 Costa [10] 采用的 PCA 方法处理 GII 面板数据进行了地理可视化对比，并以中国为例展开具体分析。

在图 5 基于 PCA 的聚类结果中，中国与比利时、保加利亚等国被归为一类。该聚类方式受全局方差结构主导，划分结果地理分布分散、发展阶段混合，未反映中国独特性，难识别其与相似经济体的内在关联。相比之下，图 6 基于熵权 UMAP 的聚类结果将中国与日本、韩国等创新活跃型经济体聚为一类，区分了与传统工业国及资源依赖型经济体的边界。该方法敏感捕捉局部数据结构，识别中国多维特征，

归入符合其“系统竞争力”属性的类别，有更优的经济解释力和类别区分度。

综合来看，熵权 UMAP 保留高维数据局部邻近关系，能更精细识别国家间本质相似性，聚类结果符合经济学直觉，有政策启示意义。相较于 PCA 过度依赖全局方差导致的类别混杂问题，熵权 UMAP 能依据多维相似性合理细致分类，适用于地理邻近但发展路径分化、或地理分散但结构趋同的经济体分析。这一发现强化了熵权 UMAP 在复杂社会经济数据挖掘中的价值，展现更强特征识别能力和分析鲁棒性，为发展路径比较、创新政策评估及国家分类与区域研究提供了可靠方法与新视角。

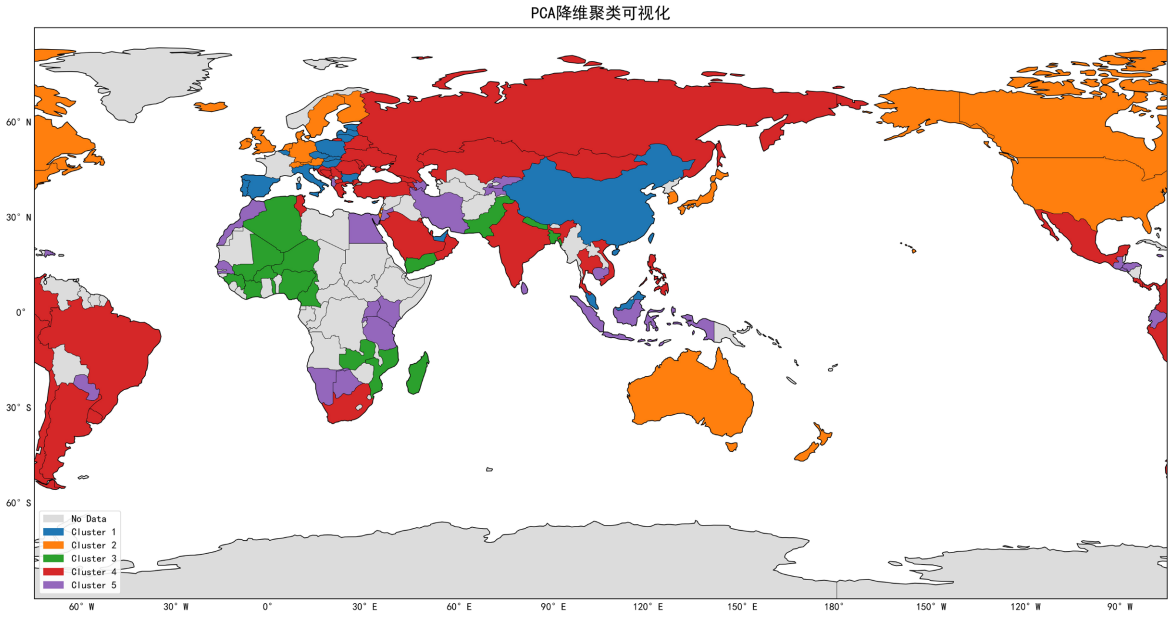

Figure 5. PCA clustering results
图 5. PCA 聚类结果

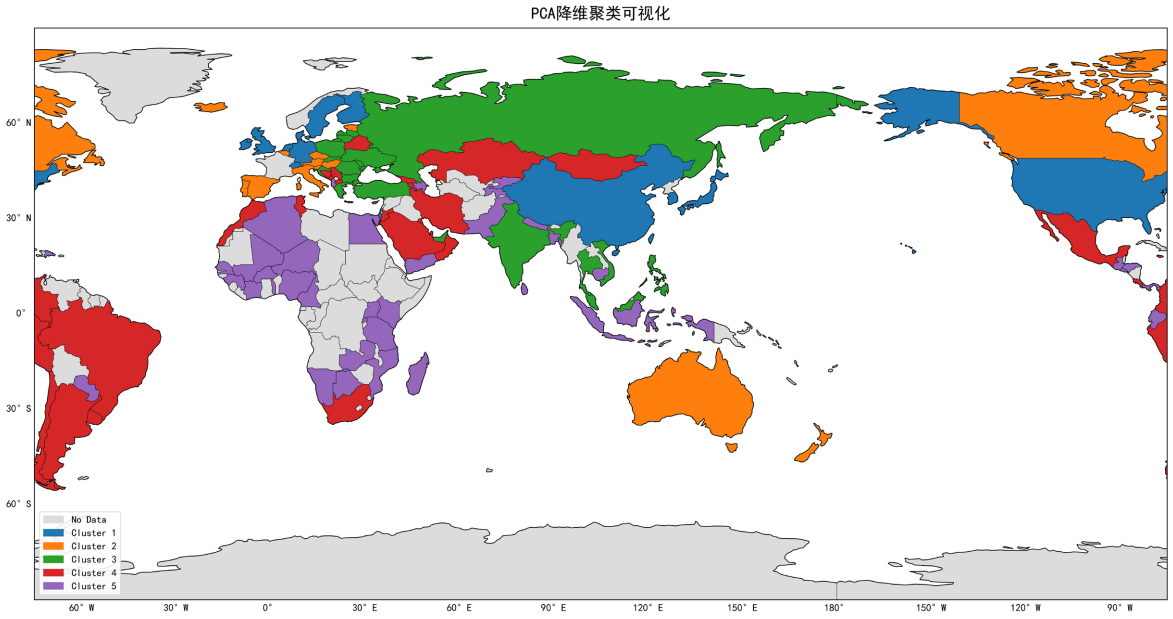

Figure 6. Entropy-weighted UMAP clustering results
图 6. 熵权 UMAP 聚类结果

3.4. 特征权重分析及其与聚类结果的关联

3.4.1. 特征权重分析

本研究采用熵权法对全球创新指数(GII)数据集中的特征变量进行客观赋权。计算前对所有数据进行标准化处理以消除量纲影响，最终确保所有权重之和为 1。

权重分析结果显示，各特征权重分布较为均衡，无单一主导变量。权重最大值为 0.0313，最小值为 0.0030，均值为 0.01，标准差为 0.0059。从区间看，58%的特征权重低于 0.01，42%的特征权重介于 0.01 至 0.05 之间。权重排序清晰地识别出驱动聚类划分的核心特征。排名前五的特征均属于创新产出维度，其具体权重与占比如下表 3 所示。

Table 3. Top 5 Core features by weight ranking
表 3. 权重排名前五的核心特征

排名	特征类别	特征名称	权重值	权重占比
1	创意产出类	2022_Creative outputs index	0.0313	3.13%
2	创新指数类	2022_Innovation Output Sub-Index	0.0250	2.50%
3	知识技术产出类	2022_Knowledge and technology outputs index	0.0240	2.40%
4	知识技术产出类	2021_Knowledge and technology outputs index	0.0224	2.24%
5	知识技术产出类	2020_Knowledge and technology outputs index	0.0221	2.21%

如表所示，前五个核心特征的累计权重已达 12.53%，且在结果中显示，前二十个特征累计权重接近 40%，构成了影响聚类决策的“核心特征集合”。进一步分析发现，创新产出类特征的总权重显著高于机构环境、人力资本等基础支撑类特征。同时，同类型特征中 2022 年数据的权重普遍高于 2021 与 2020 年，这初步提示，近期的创新产出表现可能在区分国家创新层级时具有更强的时效性与解释力。

3.4.2. 特征权重在聚类中的作用

为验证熵权法所赋予权重的实际意义，进一步比较了高、低权重特征在不同聚类中的表现差异。下表 4 展示了权重排名前五的特征与三项低权重特征在各聚类中的均值分布。

Table 4. Comparison of means and differences of high- and low-weight features in each cluster (after standardization)
表 4. 高、低权重特征在各聚类中的均值与差异对比(标准化后)

特征类别	特征名称	特征权重	聚类 1 均值	聚类 2 均值	聚类 3 均值	聚类 4 均值	聚类 5 均值	差异倍数
高权重特征	2022_Creative outputs index	0.0313	0.82	0.65	0.48	0.33	0.21	3.90
	2022_Innovation Output Sub-Index	0.0250	0.79	0.61	0.45	0.30	0.22	3.59
	2022_Knowledge and technology outputs index	0.0240	0.76	0.58	0.43	0.29	0.24	3.17
	2021_Knowledge and technology outputs index	0.0224	0.74	0.56	0.41	0.28	0.23	3.22
	2020_Knowledge and technology outputs index	0.0221	0.72	0.54	0.39	0.27	0.22	3.27
低权重特征	2019_Infrastructure index	0.0032	0.51	0.47	0.43	0.39	0.39	1.31
	2018_Market environment index	0.0035	0.49	0.45	0.42	0.38	0.38	1.29
	2017_Institution index	0.0038	0.48	0.44	0.41	0.37	0.37	1.30

数据清晰显示，高权重特征如创意产出与知识技术产出指数在不同聚类间有显著梯度差异，顶尖引

领集群(聚类 1)与转型新兴集群(聚类 5)均值差异倍数达 3.17 至 3.90。例如，“2022 年创意产出指数”在聚类 1 均值为 0.82，在聚类 5 仅为 0.21，呈单调递减趋势。相比之下，低权重特征如基础设施、市场环境、制度指数在各聚类分布区间狭窄(0.37~0.51)，差异倍数仅 1.3 左右，无法刻画不同创新层级国家的结构性差距。这证实熵权法识别的高权重特征是驱动聚类形成的区分依据，特征权重与实际区分效力一致，验证了熵权法的合理性与有效性。因此，本部分分析验证了熵权 UMAP 模型中特征权重的可靠性，还揭示全球创新格局分化关键维度：国家层级差异主要体现为近期创新产出效能，尤其是创意与知识技术成果转化的系统性差距，而非基础设施或制度环境投入。这为理解创新集群形成机制提供依据，也为后续政策分析明确了核心关注维度。

4. 总结

本研究针对全球创新指数(GII)高维、非线性、特征权重差异显著的特点，提出了一种融合特征权重的改进流形学习方法——熵权 UMAP。通过对 2013~2022 年 118 个经济体的 GII 数据进行标准化预处理，引入熵权法对各创新指标进行客观赋权，并将其融入 UMAP 的距离度量中，构建了能够反映特征重要性差异的加权降维模型。在此基础上，采用 K-Means 聚类方法对降维后的数据进行分组，并综合利用轮廓系数、Calinski-Harabasz 等指数以及簇内平方和等评估指标，结合 TOPSIS 决策模型对不同降维方案下的聚类效果进行全面量化评估。

实验结果表明，与其他降维方法相比，熵权 UMAP 能够更有效地揭示全球创新数据的内在流形结构，其降维结果在可视化中呈现出更清晰的簇群边界和更紧凑的簇内分布。特别是在聚类数为 5 时，熵权 UMAP 方案的综合评估得分最高(TOPSIS 评分为 0.902466)，各项聚类指标均表现优异，说明该方法在保持局部相似性的同时，也更好地维持了全局拓扑结构，显著提升了聚类的解释性和稳健性。

基于熵权 UMAP 与 K-Means 的聚类结果，本研究将 118 个经济体划分为五个具有明显特征差异的创新集群：全球顶尖创新引领集群、中低创新水平多元发展集群、中等创新水平区域代表集群、较高创新水平稳健发展集群以及中创新水平转型与新兴集群。这一划分不仅与全球创新发展的现实格局高度吻合，也从数据层面印证了不同国家群体在创新基础、产出效能与发展路径上的系统性差异。

进一步的特征权重分析表明，影响聚类结构的关键特征主要集中于近期的创新产出类指标(如创意产出、知识与技术产出)，其权重显著高于基础设施、制度环境等支撑性指标。高权重特征在各聚类间表现出显著的梯度差异，验证了熵权法赋权的合理性，也说明创新活动的最终产出效能是区分不同创新梯队的关键依据。因此，本研究不仅在方法层面提供了一种适用于多指标综合评价体系的加权流形学习框架，也在实践层面为政策制定者提供了明确的方向：应重点关注创新产出效能的提升，而非仅仅依赖资源投入，从而为各国特别是创新追赶型经济体优化创新政策、精准评估成效提供了理论依据与数据支持。

基金项目

何飞雪的研究受到重庆工商大学研究生创新型科研项目资助(项目编号：yjscxx2025-269-23)。

参考文献

[1] 《2025 年全球创新指数》勾勒“强技术、慢落地”的全球创新脉搏[J]. 科技导报, 2025, 43(18): 6.

[2] 孙玛媛, 习怡衡, 王海燕. 从全球创新指数看中国创新能力——基于国家创新体系视角[J]. 经济体制改革, 2025(1): 164-173.

[3] Nasir, M.H. and Zhang, S. (2024) Evaluating Innovative Factors of the Global Innovation Index: A Panel Data Approach. *Innovation and Green Development*, **3**, Article ID: 100096. https://doi.org/10.1016/j.igd.2023.100096

[4] Ma, X., Hao, Y., Li, X., Liu, J. and Qi, J. (2023) Evaluating Global Intelligence Innovation: An Index Based on Machine Learning Methods. *Technological Forecasting and Social Change*, **194**, Article ID: 122736.

https://doi.org/10.1016/j.techfore.2023.122736

[5] Brás, G.R. (2023) Pillars of the Global Innovation Index by Income Level of Economies: Longitudinal Data (2011-2022) for Researchers' Use. *Data in Brief*, **46**, Article ID: 108818. https://doi.org/10.1016/j.dib.2022.108818

[6] Huarng, K. and Yu, T.H. (2022) Analysis of Global Innovation Index by Structural Qualitative Association. *Technological Forecasting and Social Change*, **182**, Article ID: 121850. https://doi.org/10.1016/j.techfore.2022.121850

[7] Yu, T.H., Huarng, K. and Huang, D. (2021) Causal Complexity Analysis of the Global Innovation Index. *Journal of Business Research*, **137**, 39-45. https://doi.org/10.1016/j.jbusres.2021.08.013

[8] Crespo, N.F. and Crespo, C.F. (2016) Global Innovation Index: Moving beyond the Absolute Value of Ranking with a Fuzzy-Set Analysis. *Journal of Business Research*, **69**, 5265-5271. https://doi.org/10.1016/j.jbusres.2016.04.123

[9] El, B.R. and Maymoni, L. (2022) How Can Lower-Income Countries Integrate in the Innovation-Led Global Economy? *International Journal of Innovation Studies*, **6**, 153-165.

[10] Eufrazio, E. and Costa, H. (2025) Comprehensive Dataset of Global Innovation Index Panel Data (2013-2022): Clustering with K-Means and Principal Component Analysis. *Data in Brief*, **63**, Article ID: 112194. https://doi.org/10.1016/j.dib.2025.112194

[11] Allaoui, M., Belhaouari, S.B., Hedjam, R., Bouanane, K. and Kherfi, M.L. (2025) t-SNE-PSO: Optimizing t-SNE Using Particle Swarm Optimization. *Expert Systems with Applications*, **269**, Article ID: 126398. https://doi.org/10.1016/j.eswa.2025.126398

[12] 谢斌, 徐燕, 王冠超, 等. t-SNE 最大化的自适应彩色图像灰度化方法[J]. 中国图象图形学报, 2024, 29(8): 2333-2349.

[13] 邹黎敏, 唐永欣. 基于机器学习的我国天然气进口量预测及其运输安全评价[J]. 工业技术经济, 2025, 44(2): 108-118.

[14] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[15] 周映宇, 何玲, 陈家兑, 等. 基于 UMAP-KNN 的公交车工况构建方法[J]. 机械设计与制造, 2025: 1-7.

[16] Chae, Y.H., Koo, S.R., Choi, J. and Kim, J. (2026) Enhanced Learning for Nuclear Power Plant Condition Diagnoses Using Information Metric Based on Bayesian Neural Networks and UMAP. *Nuclear Engineering and Technology*, **58**, Article ID: 103886. https://doi.org/10.1016/j.net.2025.103886

[17] Tan, H.S., Wang, K. and Mcbeth, R. (2024) Exploring UMAP in Hybrid Models of Entropy-Based and Representativeness Sampling for Active Learning in Biomedical Segmentation. *Computers in Biology and Medicine*, **176**, Article ID: 108605. https://doi.org/10.1016/j.compbiomed.2024.108605

[18] Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kebonye, N.M., Kakhani, N., Ghebleh-Goydaragh, M., Heung, B., *et al.* (2024) High-Performance Soil Class Delineation via UMAP Coupled with Machine Learning in Kurdistan Province, Iran. *Geoderma Regional*, **36**, e00754. https://doi.org/10.1016/j.geodrs.2024.e00754

[19] Dadu, A., Satone, V.K., Kaur, R., Koretsky, M.J., Iwaki, H., Qi, Y.A., *et al.* (2023) Application of Aligned-UMAP to Longitudinal Biomedical Studies. *Patterns*, **4**, Article ID: 100741. https://doi.org/10.1016/j.patter.2023.100741

[20] 黎耀康, 杨海东, 徐康康, 等. 基于加权 UMAP 和改进 BLS 的锂电池温度预测[J]. 储能科学与技术, 2024, 13(9): 3006-3015.

[21] 尹泽明, 王彩年, 王智, 等. 基于 UMAP 改进的多域特征提取方法及轴承故障诊断[J]. 组合机床与自动化加工技术, 2024(1): 160-163.

[22] 张润, 李晓斌, 徐亚敏. 一致流形逼近与投影算法综述[J/OL]. 计算机科学, 2025: 1-16. https://link.cnki.net/urlid/50.1075.tp.20250707.1434.026, 2025-07-08.