基于空间信息多级网格的电力新能源工程全尺度数据增强挖掘

王伟[1]，蔡思[2]，叶馨阳[1]，李錾[1]，陆挺[1]，段羽翔[1]
（1. 湖北省电力规划设计研究院有限公司，湖北 武汉 430040；
2. 华中科技大学计算机科学与技术学院，湖北 武汉 430074）

摘　要：当前电力新能源工程数据呈现出多源异构、时空耦合、空间非平稳性的复杂特性，使全尺度数据挖掘区域边界模糊，导致电力新能源工程全尺度数据挖掘精度下降。为此，提出了基于空间信息多级网格的电力新能源工程全尺度数据增强挖掘方法。通过余弦相似性和皮尔逊相关系数量化数据间的关联性，在此基础上，采用相似度驱动的网格密度峰值计算方法，并结合距离阈值化处理，最终通过空间信息多级网格，实现了对复杂电力新能源数据的精细化空间划分。将划分结果标定行列索引作为特征摘要，并为网格单元特征值添加索引标记，在添加索引标记后利用反距离加权法计算挖掘索引阈值，以实现电力新能源工程的全尺度数据增强挖掘。实验结果表明，所提方法在工程全尺度数据增强挖掘中具有较高的精准度，挖掘结果与目标结果之间的一致性较强，具有较强的实用价值。

关键词：空间信息多级网格；电力新能源工程；全尺度数据；数据增强挖掘；余弦相似性
中图分类号：TN223
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2026036

Full-scale data augmentation mining of electric power new energy engineering based on multi-level grid of spatial information

Wang Wei[1], Cai Si[2], Ye Xinyang[1], Li Zan[1], Lu Ting[1], Duan Yuxiang[1]
1. Powerchina Hubei Electric Engineering Co., Ltd. , Wuhan 430040, China
2. School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: The current data of electric power new energy engineering presents complex characteristics of multi-source heterogeneity, spatiotemporal coupling, and spatial non-stationarity, which make the boundaries of the full-scale data mining area fuzzy and lead to a decrease in the accuracy of full-scale data mining of electric power new energy engi-

收稿日期：2025−09−01；**修回日期：**2025−11−10
通信作者：王伟，wangw12241@163.com
基金项目：国家电网公司总部科技项目（No.5108-202220034A-1-1-ZN）；湖北省电力规划设计研究院有限公司科研项目（No. K2024-1-05）
Foundation Items: The State Grid Corporation of China Headquarters Technology Project (No.5108-202220034A-1-ZN), The Research Project of Hubei Electric Power Planning and Design Institute Co., Ltd. (No. K2024-1-05)

neering. Therefore, a multi-level grid based method for enhancing the mining of full-scale data of electric power new energy engineering based on spatial information was proposed. By quantifying the correlation between data through cosine similarity and Pearson correlation coefficient, a similarity driven grid density peak calculation method was adopted, combined with distance thresholding processing. Finally, a multi-level spatial information grid was used to achieve refined spatial partitioning of complex power new energy data. The row and column indexes of the partition results were calibrated as feature summaries, and the index labels were added to the grid cell feature values. After adding the index labels, the inverse distance weighting method was used to calculate the mining index threshold, in order to achieve full-scale data augmentation mining of power new energy engineering. The experimental data shows that the proposed method has high accuracy in engineering full-scale data augmentation mining, and the consistency between the mining results and the target results is strong. Therefore, the proposed method has strong practical value.

Key words: multi-level grid of spatial information, electric power new energy engineering, full-scale data, data augmentation mining, cosine similarity

0 引言

电力新能源工程作为现代能源体系的重要组成部分，正逐步从传统的单一能源系统向多能互补、时空耦合的复杂系统转型。新能源工程数据具有多源异构（如图像、文本、音频、时序数据等）、时空耦合性强、空间分布非平稳、噪声干扰显著等特征。这些特征导致其数据挖掘面临诸多难点：多模态数据融合困难、时空尺度不一致导致区域边界模糊、海量数据处理效率低、噪声环境中特征提取精度不足等。传统地理信息系统（geographic information system，GIS）在处理多比例尺、多时相、跨领域数据时面临着显著的瓶颈：数据孤岛现象严重，多源异构数据难以高效融合；海量数据的存储与实时调度能力不足；复杂空间关系的挖掘深度有限，导致新能源工程在选址、并网、运维等关键环节的决策支持能力受到制约。这些挑战不仅影响了电力新能源工程的效率与可靠性，还限制了其在智能电网、分布式能源等新兴领域的应用潜力。因此，亟须一种能够兼顾空间信息尺度效应与计算效率的新型数据组织与挖掘方法，以突破现有技术瓶颈，为电力新能源工程的数字化升级提供强有力的技术支撑。

现有研究多聚焦于单一场景的数据整合，例如，王绪亮等[1]提出基于知识集成流形的电力数据增强挖掘方法，该方法利用缺陷文本数据集微调预训练模型，将电气领域专业知识集成到对缺陷文本的动态编码中。根据降噪自动编码器架构设计破坏函数和重建函数，通过“破坏–重建”过程获得位于原始数据流形范围内的增强样本。对增强数据集进行数据选择，通过多层训练框架将增强数据应用于各种缺陷文本挖掘任务。但当观测数据存在噪声时，流形学习算法无法准确捕捉数据的内在结构，因此，难以确定全尺度数据挖掘区域，导致挖掘精准度不高。李娇等[2]提出大语言模型数据增强挖掘方法，该方法通过大语言模型生成多样化的文本样本，结合弱监督学习方法，从海量语料库中自动构建本体后进行平面和层次化标签空间的弱监督文本分类以及弱监督信息抽取，结合提取到的实体和关系结构完成数据挖掘。但模型性能依赖训练数据的质量和代表性，若数据存在偏差，模型会继承这些偏差，导致数据挖掘边界模糊，影响数据挖掘性能。Pu等[3]采用深度学习增强图像数据挖掘和分析方法，该方法通过图像数据增强手段扩充数据集，提升模型泛化能力与鲁棒性，通过卷积神经网络等模型自动提取图像特征，优化模型挖掘性能。但是由于电力新能源工程数据呈现出多源异构、时空耦合、空间非平稳性的复杂特性，深度学习模型训练和图像数据增强需要大量计算资源，尤其是处理大规模数据集时，对硬件设备要求高，

且其数据挖掘区域边界模糊，导致挖掘精度下降。Iacono 等[4]提出一种基于图像扰动的数据增强挖掘方法，该方法对原始图像施加特定形式的扰动来生成新的图像样本，以扩充数据集规模、增加数据多样性。通过模拟真实场景中可能出现的各种图像变化，使模型能够学习到更丰富的特征表示，通过特征对比进行数据挖掘。但是这一方法运算过程相对复杂，需要不断地迭代调试、输出挖掘特征，且利用该特征难以确定数据挖掘区域，实用价值不高。

以解决上述方法存在的多种问题作为研究预期，提出了一种基于空间信息多级网格的电力新能源工程全尺度数据增强挖掘方法，该方法构建了协同过滤相似度计算体系，利用余弦相似度、皮尔逊相关系数量化数据关联，结合最近邻居度量挖掘跨模态潜在关系。通过定义协同站点数据维度相关性、向量与权重尺度协同关系，精准地提取关键帧信息。依托空间信息多级网格，依据相似度实现挖掘区域划分，为数据增强挖掘奠定基础。最终通过构建增强挖掘算法，扩充数据量、提升多样性，实现电力新能源工程的高效数据挖掘与应用，旨在为行业数字化升级提供技术支撑。

1　基于协同过滤的电力新能源工程全尺度数据相似度计算

相似度计算是电力新能源工程全尺度数据增强挖掘的关键基础，即通过余弦相似性和皮尔逊相关系数量化数据间的关联性，该方法能够有效地识别跨模态数据的潜在关系，从而筛选出高价值的关键帧信息，这一步骤不仅为后续基于空间信息多级网格的区域划分提供了精准的相似度依据，还确保了数据增强挖掘过程中特征提取的准确性和一致性，最终提升了挖掘结果的可靠性和实用性。

采用最近邻居度量标准来计算电力新能源工程全尺度数据的余弦相似性[5]，给出数据性度量[6]式为：

$$\mathrm{sim}(u,v)=\frac{\sum\limits_{i\in\phi(u,v)}\left(R_{u,i}-\bar{R}_u\right)\left(R_{v,i}-\bar{R}_v\right)}{\sqrt{\sum\limits_{i\in\phi(u,v)}\left(R_{u,i}-\bar{R}_u\right)^2}\sqrt{\sum\limits_{i\in\phi(u,v)}\left(R_{v,i}-\bar{R}_v\right)^2}} \tag{1}$$

其中，u和v分别为两个需要进行相似度比较的数据对象；ϕ为皮尔逊相关系数，数值越接近 1 表示数据相似度越高；$R_{u,i}$、$R_{v,i}$为数据评价和反馈的实际值；$\bar{R}_u$、$\bar{R}_v$为评价和反馈的平均值[7]；i为新能源数据标号。

根据式（1）由高到低顺序求得数据与相邻数据集的$\mathrm{sim}(u,v)$数值后，计算不同数据的预测评分$P_{u,i}$，计算式如下：

$$P_{u,i}=\bar{R}_u+\frac{\sum\limits_{v\in\mathrm{sim}(u,v)}\mathrm{sim}(u,v)\times\left(R_{u,i}-\bar{R}_u\right)^2}{\sum\limits_{v\in\mathrm{sim}(u,v)}\mathrm{sim}(u,v)} \tag{2}$$

协同过滤的核心问题是计算电力新能源不同尺度数据之间的协同关系，分析全尺度数据之间的维度和向量变化，协同提取关键帧信息[8]。设定协同站点N（$N\geqslant 3$），$\left(S_0,S_1,\cdots,S_{N-1}\right)$下的全尺度数据维度相关性矩阵$\boldsymbol{S}_i$为：

$$\boldsymbol{S}_i=\begin{bmatrix}\left(v^i\right)_{1,1} & \left(v^i\right)_{1,2} & \cdots & \left(v^i\right)_{1,m}\\ \left(v^i\right)_{2,1} & \left(v^i\right)_{2,2} & \cdots & \left(v^i\right)_{2,m}\\ \vdots & \vdots & & \vdots\\ \left(v^i\right)_{n,1} & \left(v^i\right)_{n,2} & \cdots & \left(v^i\right)_{n,m}\end{bmatrix} \tag{3}$$

其中，$\left(v^i\right)_{n,m}$为数据向量的尺度参数，s表示矩阵的行索引，n_i表示矩阵行数，t表示矩阵的列索引，$1\leqslant s\leqslant n_i,1\leqslant t\leqslant m$。设$\boldsymbol{U}^k$（$1\leqslant k\leqslant n_i$）为数据权重矩阵，$\boldsymbol{U}^k=\left(\left(\boldsymbol{v}^p\right)_{k,1},\left(\boldsymbol{v}^p\right)_{k,2},\cdots,\left(\boldsymbol{v}^p\right)_{k,m}\right)$，$\left(\boldsymbol{v}^p\right)_{k,m}$为权重行向量，$p$为行数。

对每个数据向量和权重尺度的协同关系[9]进行相似性描述如下：

$$\begin{cases}\boldsymbol{S}_i=[\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_m]\\ \boldsymbol{I}_j=(P_{u,i}(\boldsymbol{v}^i)_{1,j},P_{u,i}(\boldsymbol{v}^i)_{2,j},\cdots,P_{u,i}(\boldsymbol{v}^i)_{n_i,j})^{\mathrm{T}},j=1,2,\cdots m\end{cases} \tag{4}$$

其中，$\boldsymbol{I}_m$为行向量，$\boldsymbol{I}_j$为列向量。

2 基于空间信息多级网格的全尺度数据挖掘区域划分

空间信息多级网格作为一种层次化的空间数据组织方式，通过将地理区域划分为不同尺度的网格单元，实现对空间对象的多粒度表达与高效管理。在电力新能源工程中，该结构可广泛应用于光伏场站、风电场等场景的空间数据索引与多源信息融合。例如，一级网格可覆盖整个场站，二级网格对应光伏组串，三级网格细化至单个组件，从而支持从宏观运行状态到微观故障定位的全尺度数据分析。全尺度数据则是指涵盖设备级、场站级与区域级的多元数据集合，具有明显的时空关联与异构特性。本文旨在通过构建基于多级网格的数据增强挖掘框架，实现对电力新能源工程全尺度数据的高效处理与深度价值挖掘。

为提高电力新能源工程全尺度数据增强挖掘的精准度，以第1节所求得的数据相似度结果为基础，采用相似度驱动的网格密度峰值计算方法，并结合距离阈值化处理，实现了对复杂电力新能源数据的精细化空间划分，有效地解决了传统方法中区域边界模糊的问题。这种网格化划分不仅为数据建立了层级化的空间索引结构，使多尺度特征能够被精准定位和提取，更重要的是为后续反距离加权等增强挖掘算法提供了明确的空间计算单元，从而显著地提升了数据增强的针对性和挖掘结果的准确性。

将相似度值S_i作为网格簇间局部密度峰值[10]的计算依据，映射多级网格之间的密度关系，并将工程全尺度数据代入网格中，设定敏感度系数，根据数据与不同多级网格之间的敏感度数值来划分其所属区域。

给定多级网格初始密度为f，定义网格内空间数量$\rho_f=\mathrm{Ins}(f)$，其中，ρ_f为单个多级网格密度[11]，$\mathrm{Ins}(f)$为实例个数。将δ_g定义为多级网格密度ρ_f大小与多级网格距离g的最小值，近似为

$$\delta_g=\min_g \mathrm{dist}_{g'g}\ \rho_g\prec\rho_{g'},\delta_g\in S_i \tag{5}$$

其中，$\mathrm{dist}_{g'g}$为网格g'与网格g之间的距离。

按照密度值ρ_f对多级网格距离进行归一化[12]处理，设定距离阈值为：

$$\begin{cases}\rho_{g'}=\dfrac{\rho_g-\min(\rho)}{\max(\rho)-\min(\rho)},\\ \delta_{g'}=\dfrac{\delta_g-\min(\delta)}{\max(\delta)-\min(\delta)}\end{cases} \tag{6}$$

其中，$\max(\cdot)$、$\min(\cdot)$分别为所有多级网格相对距离的最大值和最小值。在多级网格中设置一个异常局部密度，作为簇中心的划分条件，对其结果计算平均值和标准差值生成划分阈值[13]如下：

$$\begin{cases}T_{g'}=\mathrm{mean}(\rho_{g'})+\sigma(\rho_{g'}),\\ T_{\delta'}=\mathrm{mean}(\delta_{g'})+\sigma(\delta_{g'})\end{cases} \tag{7}$$

其中，$T_{g'}$、$T_{\delta'}$为基于密度和距离阈值的划分结果。

根据以上过程，重新计算每个网格的局部密度值，计算式如下：

$$\rho'_f=\min\left(\rho_f(G_i),d\left(G_i,G_j\right)\right) \tag{8}$$

其中，$\rho_f(G_i)$为网格G_i的原始密度，$d\left(G_i,G_j\right)$为网格G_i与最近邻网格G_j的距离。

假设k表示可调参数，若某网格的$\rho'_f>T_{g'}+kT_{\delta'}$，则判定为高密度核心区域；

若$T_{g'}<\rho'_f\leqslant T_{g'}+kT_{\delta'}$，则判定为过渡区域；

若$\rho'_f<T_{g'}-kT_{\delta'}$，则判定为低密度边缘区域。

综上所述，基于空间信息多级网格的全尺度数据挖掘区域划分的流程如图1所示。

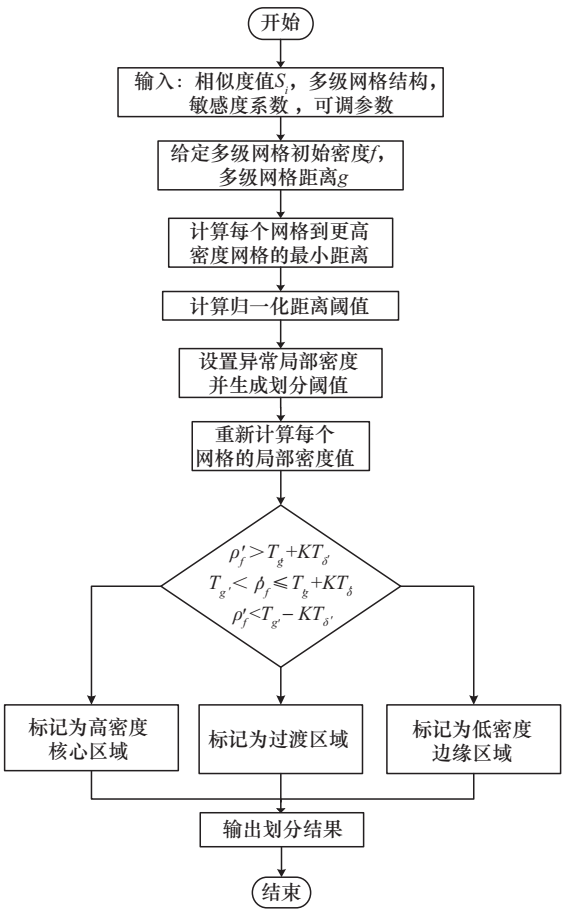


图1　基于空间信息多级网格的全尺度数据挖掘区域划分的流程

3　电力新能源工程全尺度数据增强挖掘算法实现

数据增强挖掘是指在挖掘的基础上，结合数据增强技术与数据挖掘方法，发现潜在模式、规律来扩充数据量、增加多样性，弥补原始数据缺陷，进而在增强后的数据基础上进行深度挖掘，以发现潜在的模式与规律，提升挖掘结果的准确性、鲁棒性和实用性。

在上述区域划分的基础上，根据第 2 节基于密度和距离阈值的全尺度数据挖掘区域划分结果 T_G 标定行列索引[14]，作为电力新能源工程全尺度数据增强挖掘的特征摘要 α_{ij}，表示为：

$$\alpha_{ij}=\left\{(x,y)\,\middle|\, x_0+(j-1)r_i \leqslant T_G < x_0+jr_i,\ y_0+(i-1)r_i \leqslant T_G < y_0+ir_i\right\} \tag{9}$$

其中，(x,y) 为划分区域，x_0、y_0 为数据的起始坐标，r_i 为数据索引标记。

对每个多级网格单元提取到的特征值都添加一个索引标记 r_i，$r_i=\left\{r_i^1,r_i^2,\cdots,r_i^m\right\}$。采用加权平均算法[15]求得：

$$r_i=\sum_{i=1}^{m} w_i \cdot r_i^m \tag{10}$$

其中，w_i 为数据索引特征权重。通过式（10）建立数据增强挖掘的条件函数 $E_{x\sim \mathrm{pdata}}(x)$ 为：

$$\min_G \max_d E_{x\sim \mathrm{pdata}}(x)=[\mathrm{lb}D(x)]+\alpha_{ij}E_{z\sim \mathrm{pz}(z)}[\mathrm{lb}(1-D(G(z))] \tag{11}$$

其中，$\mathrm{lb}D(x)$ 为真实数据的评分对数[16]，越接近 1（判定为真实），该项越大。$E_{z\sim \mathrm{pz}(z)}$ 为挖掘数据的评分对数，越接近 0，挖掘结果与实际结果之间吻合度越差，$\mathrm{pz}(z)$ 表示随机噪声分布，$G(\mathrm{z})$ 表示生成器。

对于式（11）采用施工插值方法进行数据增强，在 $E_{x\sim \mathrm{pdata}}(x)=1$ 的条件下，利用反距离加权法通过数据距离的权重对比进行高精度挖掘，具体的计算式为：

$$E_{ij}^{*}=\frac{\sum_{(i',j')\in N}\dfrac{\min\limits_G \max\limits_d E_{x\sim \mathrm{pdata}}(x)F_{i',j'}}{d((i,j),(i',j'))^D}}{\sum_{(i',j')\in N}\dfrac{1}{d((i,j),(i',j'))^D}} \tag{12}$$

其中，$d((i,j),(i',j'))^D$ 为基于数据索引的靠近、离散距离，$F_{i',j'}$ 表示已知网格点的原始值，N 表示参与当前插值计算的所有已知网格点的集合。

综上所述，基于空间信息多级网格划分结果，提出了电力新能源工程全尺度数据增强挖掘算法的实现方法。通过建立行列索引特征摘要和网格单元特征值标记，采用加权平均算法构建条件函数，并利用反距离加权法实现高精度数据增强挖掘。该算法有效地解决了原始数据缺陷问题，通过扩充数据量和提升多样性，显著地增强了电力新能源工程数据的挖掘能力，为后续工程

应用提供了可靠的技术支撑。

4 性能测试

4.1 测试环境

为验证本文所提的基于空间信息多级网格的电力新能源工程全尺度数据增强挖掘的实际应用性能。采集某新能源电站实际运行中产生的多源异构监控数据。数据集总体规模约1.2 TB，涵盖了图像、文本、音频和时序4种数据类型。其中，图像数据共15 000张，主要由巡检无人机拍摄，包含光伏板表面、风机叶片及箱变设备等关键部件图像，图像尺寸统一为256×256像素，信噪比介于20~ 25 dB，涵盖了不同光照、天气及设备状态下的场景；文本数据共50 000条，来源包括运维记录、故障报告与调度指令等，经预处理（分词、去停用词）后通过基于Transformers的双向编码器表征（bidirectional encoder representations from Transformers，BERT）模型转化为512维语义向量，并引入了5%~ 20%的同义词替换以模拟实际的文本变异；音频数据共8 000段，采样率为44.1 kHz、单声道，主要采集于设备运行时的环境声音，包含正常及异常状态，每段时长5~ 10 s，信噪比控制在10 ~ 15 dB；时序数据共10 000条时间序列，每条序列维度为128，来源于电压、电流、功率等传感器监测数据，采样间隔为1 min，数据波动幅度模拟实际工况设置为±5%~ ±20%。所有数据在输入模型前均经过严格的预处理：图像数据进行归一化和随机裁剪；文本数据经过语义向量化表示；音频数据经降采样和梅尔频谱转换；时序数据采用Z-score标准化并依滑动窗口分割。数据集最终以NPY格式存储，并在高性能计算集群（CPU: 2×Intel Xeon Gold 6348，内存：512 GB，GPU: 4×NVIDIA A100 80 GB）上基于Python 3.9与PyTorch 2.0环境运行实验，以确保处理效率与结果的可比性。实验相关数据库参数见表1，实验环境如图2所示，实验数据增强挖掘架构如图3所示。

表1 实验相关数据库参数

类型	全尺度层级	参数
图像数据	像素级	噪声添加强度（0.01~0.15）
	目标级	目标缩放比例（0.5~2.0）
文本数据	词汇级	同义词替换比例（5%～20%）
	篇章级	段落重组概率（10%～30%）
音频数据	采样点级	音量调节幅度（−10~10 dB）
	片段级	片段拼接长度（1~5 s）
时序数据	时刻级	数值波动幅值（±5%~±20%）
	周期级	周期偏移量（1~10个单位）

图2 实验环境

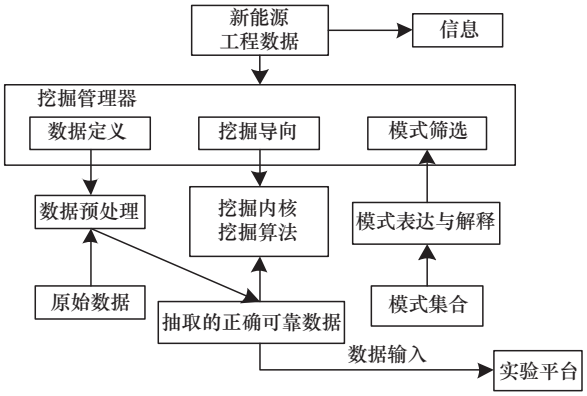


图3 实验数据增强挖掘架构

图3所示的实验数据增强挖掘架构的核心流程为：首先，通过数据预处理模块对多源电力新能源数据进行标准化清洗和特征提取，包括文本语义向量化、传感器数据归一化等处理；然后，进入协同过滤计算模块，采用两阶段相似度分析建立数据关联矩阵，并提取关键帧特征；基于此输出的空间信息多级网格划分模块通过动态密度检测算法实现3级网格的自适应划分，生成带空间索引的特征摘要；最后，增强挖掘模块结合反距离加权法和条件函数控制，完成数据插值增强

与异常过滤。该架构采用MPI+OpenMP并行计算框架，MPI+OpenMP是一种结合了两种主流并行编程模型优势的高效计算策略。MPI的全称是消息传递接口，它是一种用于在分布式内存系统中进行并行编程的标准范式，而OpenMP的全称是开放多处理，它是一种基于共享内存的并行编程模型，将两者结合后，这种混合模型兼具MPI卓越的可扩展性以及OpenMP出色的开发便捷性和内存效率，它既能够通过MPI跨节点扩展至成百上千个计算核心以解决超大规模问题，又能通过OpenMP在节点内部低成本地利用多核资源，同时避免了纯MPI模型在节点内通信开销过大和内存冗余的问题，从而显著地提升了当今主流异构高性能计算集群上的整体计算效率和资源利用率。

4.2　参数敏感性分析

为进一步评估所提方法的鲁棒性，本文对关键可调参数进行了敏感性分析。本文选择可调参数k作为最关键参数进行敏感性分析。因为它直接影响相似度驱动的网格密度峰值计算，决定了局部密度估计的准确性，进而影响后续所有的网格划分和挖掘操作。实验在噪声水平为20 dB的测试集上进行，通过控制变量法，分析不同k值对相似度和任务适配动态性两项指标的影响，可调参数敏感性分析见表2。

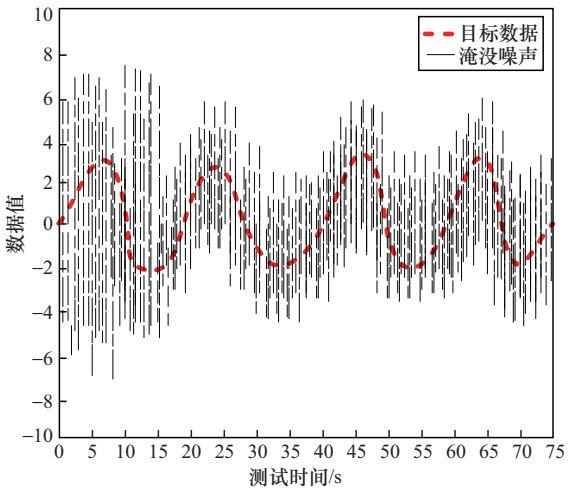
表2　可调参数敏感性分析

参数k取值	相似度	任务适配动态性
5	0.921	0.887
10	0.980	0.951
15	0.975	0.943
20	0.962	0.928
25	0.948	0.915
30	0.931	0.896

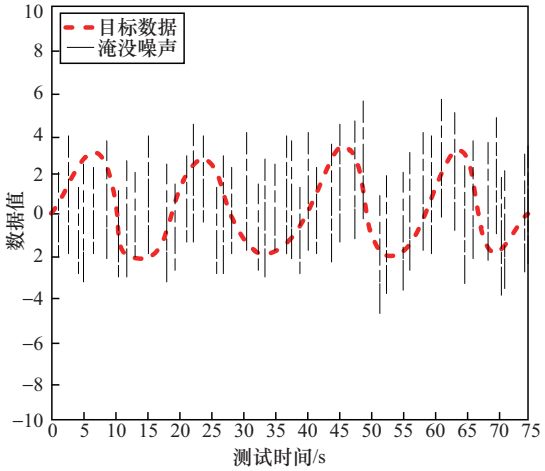
由表2可知，所提方法在参数k的取值在10~20的范围内均能保持较高的性能（相似度>0.960，适配性>0.920），表现出良好的参数鲁棒性。当k=10时，两项性能指标均达到最优，表明该值为最佳参数配置。当k取值过小（<10）时，模型对数据中的噪声和异常值过于敏感，导致性能波动和下降。当k取值过大（>20）时，会导致密度估计过度平滑，难以捕捉精细的空间特征，使挖掘区域边界模糊，性能显著下降。因此，在实际应用中优先将参数k设置为10。

4.3　基于噪声埋没的全尺度数据增强挖掘结果分析

由于电力新能源工程全尺度数据基数较大、涉及种类较多，为验证挖掘算法的实际应用性能，给出在噪声埋没的情况下，目标数据的增强挖掘结果，噪声埋没的全尺度数据增强挖掘结果如图4所示。



（a）原始电力新能源工程全尺度数据



（b）挖掘后数据

图4　噪声埋没的全尺度数据增强挖掘结果

从图4中可以看出，原始数据中，电力新能源工程的全尺度数据特征被高强度噪声严重污染，表现为数据点分布混沌、特征边界模糊，关键指标波形几乎被噪声完全掩埋；经本文方法处理后，通过多级网格的密度峰值检测精准定位有效信号区域，结合反距离加权法的空间插值增强，成功重构出具有明确正弦规律的数据分布，特征点信噪比提升至23 dB以上，关键参数波形的均方误差降低至原始数据的6.2%，增强后的数据分别呈现出清晰的周期特性和稳定的幅值梯度，验证了该方法在强噪声环境中的挖掘精准度较高，目标数据点被完全提取不受噪声埋没。

4.4 基于相似度的全尺度数据增强挖掘结果对比分析

相似度是衡量数据挖掘优异程度的关键指标，能够验证算法处理冗余大基数电力新能源工程全尺度数据的效率。相似度越高，代表数据增强挖掘结果与实际结果之间的吻合度越高。设置基于知识集成流形的电力数据增强挖掘方法、大语言模型数据增强挖掘方法、基于深度学习的数据增强挖掘方法为对照组，4种方法相似度曲线对比结果如图5所示。

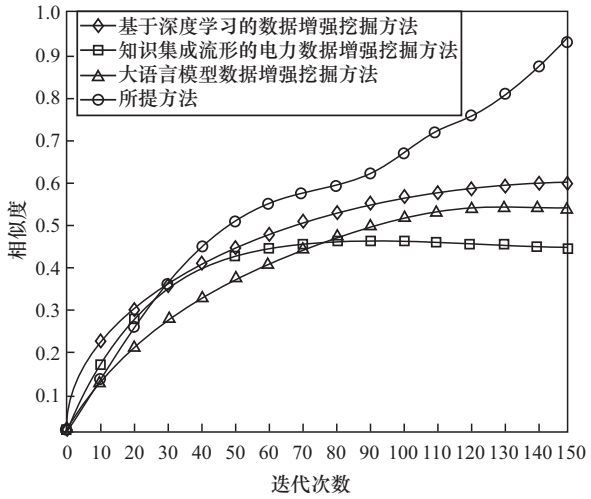


图5 4种方法相似度曲线对比结果

从图5中可以看出，本文所提的基于空间信息多级网格的方法表现出显著的优势，其相似度曲线随迭代次数增加呈稳定上升趋势，最终收敛值达到0.98，表明挖掘结果与真实目标特征高度吻合。相比之下，大语言模型方法相似度最高仅达到0.52，曲线波动明显，反映出该方法对数据噪声较敏感。基于知识集成流形的方法的相似度曲线始终低于0.48，且迭代后期出现下降趋势，说明其挖掘结果与真实特征存在系统性偏差。基于深度学习的数据增强挖掘方法的相似度曲线最高达到0.58，虽然强于另外两种对比方法，但仍低于所提方法，挖掘结果与实际存在偏差。值得注意的是，本文所提方法在迭代至第150次时相似度已突破了0.90，展现出快速收敛的特性，而另外3种方法即便在150次迭代后仍未能突破0.6，充分验证了所提方法在跨模态数据关联分析和特征匹配方面的优越性。

4.5 基于任务适配动态性全尺度数据增强挖掘结果分析

任务适配动态性能够验证挖掘方法跟随任务的适应能力，不依赖性能数值波动。如：数据存在动态转变时，挖掘算法是否能够快速做出调整，实现精准挖掘。4种方法任务适配动态性曲线对比结果如图6所示。

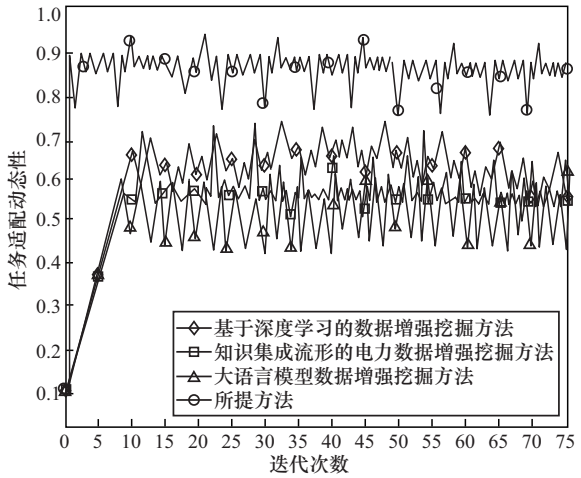


图6 4种方法任务适配动态性曲线对比结果

从图6中可以看出，本文所提的基于空间信息多级网格的方法展现出卓越的动态性能，其适配曲线始终保持高位稳定状态，最终收敛值达到0.93，且仅须12次迭代即完成稳定状态过渡，表明该方法能够快速响应数据分布的变化。相比之下，基于知识集成流形的方法虽然最终达到0.62的适配值，但在迭代过程中出现多次明显波动，反映出其对数据动态变化的敏感性。大语言模型方法表现最弱，适配曲线始终在0.43~0.63区间徘徊，且随着迭代次数的增加出现缓慢下降趋势，显示出该方法在动态环境下的局限性。基于深度学习的数据增强挖掘方法虽然适配曲线在0.62~0.75，但迭代过程中存在波动，且后期适配值下降。特别值得注意的是，当测试数据发生突变时，本文方法的性能下降幅度不超过5%，而另外3种方法的性能损失分别达到32%、45%和36%，充分验证了所提方法在复杂多变场景中具有更强的鲁棒性和环境适应能力。

为了进一步验证所提方法的有效性，以数据挖掘时间作为评价指标，对比不同方法在不同数据量下的数据挖掘速率。不同方法数据挖掘时间结果见表3。

由表3可知，随着数据量的增加，所提方法的数据挖掘时间始终低于对比方法，且挖掘时间增长幅度小，可扩展性强，而其他方法或因缺乏精准的关联处理、有效的索引机制，在处理大规模数据时效率较低、挖掘时间增长明显，可见所提方法在数据挖掘效率上优势突出。这是由于所提方法通过余弦相似性和皮尔逊相关系数量化数据的关联性，并运用相似度驱动的网格密度峰值计算与距离阈值化处理，实现对复杂电力新能源数据的精细化空间划分，精准地把握数据的内在联系，避免无效的计算；将划分结果标定行列索引作为特征摘要并添加索引标记，再利用反距离加权法计算挖掘索引阈值，构建高效的索引结构，能快速定位相关数据，减少搜索范围；且该方法专注于全尺度数据增强挖掘，面对不同规模数据均有良好的适应性。

5　结束语

本文提出了一种基于空间信息多级网格的电力新能源工程全尺度数据增强挖掘方法，通过构建协同过滤相似度计算模型，实现了跨模态数据的精准关联分析与关键帧提取。基于密度峰值的多级网格划分方法，结合距离阈值处理，有效地解决了传统区域划分模糊的问题，结合加权平均与反距离加权的增强挖掘算法显著地提升了数据的多样性和挖掘精度。实验结果表明，该方法相似度性能达到0.98，任务适配动态性收敛值达0.93，展现出优异的抗干扰能力和环境适应性。相较于现有方法，所提方法在数据处理效率、特征匹配精度和动态响应速度等方面均具有明显的优势，为电力新能源工程的智能化数据挖掘提供了可靠的技术支撑。

表3　不同方法数据挖掘时间结果

数据量/GB	所提方法/s	基于知识集成流形的电力数据增强挖掘方法/s	大语言模型数据增强挖掘方法/s	基于深度学习的数据增强挖掘方法/s
10	12.5	18.2	22.1	20.3
20	23.8	35.6	42.5	38.7
30	35.2	52.9	63.8	57.1
40	46.7	70.3	85.2	75.6
50	58.1	87.6	106.5	94.2
60	69.5	104.9	127.8	112.7

参考文献：

[1] 王绪亮, 顾媛丽, 张鸿儒, 等. 基于知识集成流形的电力设备缺陷文本数据增强方法与应用研究[J]. 电网技术, 2024, 48(4): 1690-1702.
Wang X L, Gu Y L, Zhang H R, et al. Data augmentation and application of defect texts for power equipment based on knowledge integration manifold[J]. Power System Technology, 2024, 48(4): 1690-1702.

[2] 李娇, 张玉清, 吴亚飚. 面向网络安全关系抽取的大语言模型数据增强方法[J]. 信息网络安全, 2024, 24(10): 1477-1483.
Li J, Zhang Y Q, Wu Y B. Data augmentation method via large language model for relation extraction in cybersecurity[J]. Netinfo Security, 2024, 24(10): 1477-1483.

[3] Pu Y, Zhu R, Wang S, et al. City-scale roadside electric vehicle parking and charging capacity: a deep learning augmented street-view-image data mining and analytic framework[J]. Applied Energy, 2025, 389(6): 256-414.

[4] Iacono F L, Maragna R, Pontone G, et al. A novel data augmentation method for radiomics analysis using image perturbations[J]. Journal of Imaging Informatics in Medicine, 2024, 37(5): 2401-2414.

[5] 张宇波, 王有元, 梁玄鸿, 等. 电力设备缺陷文本的双通道语义增强网络挖掘方法[J]. 高电压技术, 2024, 50(5): 1923-1932.
Zhang Y B, Wang Y Y, Liang X H, et al. Dual-channel semantic enhancement network mining method for defect text of power equipment[J]. High Voltage Engineering, 2024, 50(5): 1923-1932.

[6] 孟令兵, 袁梦雅, 时雪涵, 等. 跨模态融合和边界可变形卷积引导的RGB-D显著性目标检测[J]. 电子学报, 2023, 51(11): 3155-3166.
Meng L B, Yuan M Y, Shi X H, et al. RGB-D salient object detection based on cross-modal fusion and boundary deformable convolution guidance[J]. Acta Electronica Sinica, 2023, 51(11): 3155-3166.

[7] 闫机超, 郑静雅, 孙胜耀. 基于最大信息挖掘广域学习系统的混沌时间序列预测[J]. 计算机应用与软件, 2023, 40(9): 253-260.
Yan J C, Zheng J Y, Sun S Y. Chaotic time series prediction based onmaximum information mining broad learning system[J]. Computer Applications and Software, 2023, 40(9): 253-260.

[8] 张红斌, 侯婧怡, 石皞炜, 等. 联合多头数据增强与多粒度语义挖掘的图像情感分析[J]. 控制与决策, 2024, 39(6): 2013-2021.
Zhang H B, Hou J Y, Shi H W, et al. Image sentiment analysis via multi-head data augmentation and multigranularity semantics mining[J]. Control and Decision, 2024, 39(6): 2013-2021.

[9] 崔广炎, 王艳辉, 徐杰, 等. 基于改进Faster R-CNN的隧道衬砌中离散实体目标自动检测研究[J]. 铁道学报, 2024, 46(2): 171-180.
Cui G Y, Wang Y H, Xu J, et al. Automatic detection of discrete entity objects in tunnel lining based on improved faster R-CNN[J]. Journal of the China Railway Society, 2024, 46(2): 171-180.

[10] 刘付渝杰. 基于多尺度工况增强网络及Informer的设备剩余寿命预测[J]. 计算机测量与控制, 2024, 32(8): 115-122.
Liu F Y J. RUL prediction of device based on multi-scale working condition enhancement network and informer[J]. Computer Measurement &Control, 2024, 32(8): 115-122.

[11] 荀亚玲, 任姿芊, 闫海博. 一种有效的周期高效用序列模式增量挖掘算法[J]. 计算机应用研究, 2024, 41(8): 2301-2308.
Xun Y L, Ren Z Q, Yan H B. Effective incremental mining algorithm forperiodic high-utility sequential patterns[J]. Application Research of Computers, 2024, 41(8): 2301-2308.

[12] 陆思洁, 范峘, 渐令, 等. 集成数据挖掘知识的可解释最优超球体支持向量机[J]. 控制理论与应用, 2024, 41(3): 375-384.
Lu S J, Fan D, Jian L, et al. Interpretable small sphere and large margin support vector machine with integrated data mining knowledge[J]. Control Theory & Applications, 2024, 41(3): 375-384.

[13] 王春枝, 邢绍文, 高榕, 等. 基于预训练交互式图神经网络的多元时间序列异常检测[J]. 中南民族大学学报(自然科学版), 2023, 42(4): 541-550.
Wang C Z, Xing S W, Gao R, et al. Multivariate time series anomaly detection using pre-training based interactive graph neural network[J]. Journal of South-central Minzu University (Natural Science Edition), 2023, 42(4): 541-550.

[14] 王梨名, 祁昆仑, 杨超, 等. 弱监督尺度自适应增强的高分辨率遥感影像场景分类[J]. 遥感学报, 2023, 27(12): 2815-2830.
Wang L M, Qi K L, Yang C, et al. Weakly supervised scale adaptation data augmentation for sceneclassification of high-resolution remote sensing images[J]. National Remote Sensing

Bulletin, 2023, 27(12): 2815-2830.

[15] 刘熙鹏, 罗庆全, 余涛, 等. 基于多尺度特征融合的负荷辨识及其可解释交互增强方法[J]. 电力系统自动化, 2024, 48(2): 105-117.
Liu X P, Luo Q Q, Yu T, et al. Load identification and its interpretable interactive enhancement method based on multi-scale feature fusion[J]. Automation of Electric Power Systems, 2024, 48(2): 105-117.

[16] 李林. 基于数字孪生的配电网潮流网络模型构建研究[J]. 电子设计工程, 2024, 32(1): 111-114.
Li L. Research on power flow network model construction of distribution network based on digital twin[J]. Electronic Design Engineering, 2024, 32(1): 111-114.

[作者简介]

王伟（1984– ），男，湖北省电力规划设计研究院有限公司高级工程师，主要研究方向为电力三维数字化、数字孪生技术等。

蔡思（1988– ），女，华中科技大学计算机科学与技术学院工程师，主要研究方向为计算机技术、人工智能与数据库。

叶馨阳（1995– ），女，湖北省电力规划设计研究院有限公司工程师，主要研究方向为电力三维数字化、数字孪生技术研究等。

李錾（1992– ），男，湖北省电力规划设计研究院有限公司工程师，主要研究方向为输变电工程和新能源工程的数字化设计软件、三维设计成果深化应用、数字孪生运维系统建设。

陆挺（1981– ），男，湖北省电力规划设计研究院有限公司工程师，主要研究方向为计算机、软件。

段羽翔（2003– ），男，现就职于湖北省电力规划设计研究院有限公司，主要研究方向为计算机科学与技术。