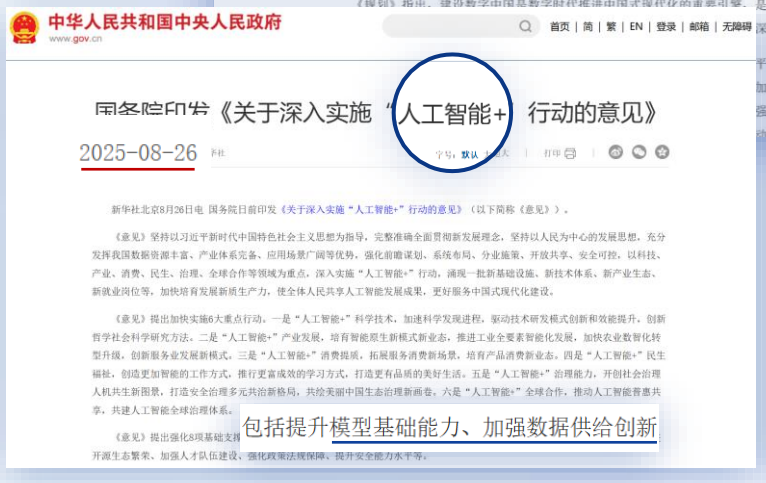
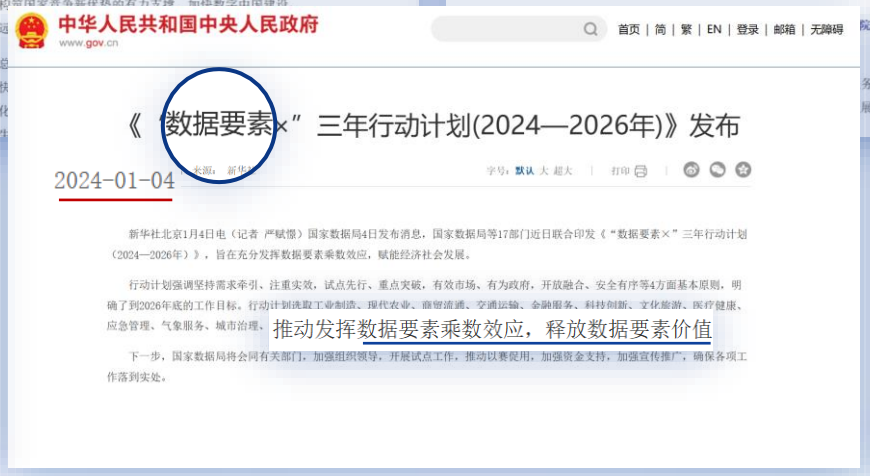
多源异质大数据联合分析

智 联 异 质 数 据 孤 岛 · 赋 能 精 准 分 析 决 策

作品类别 | 自然科学类学术论文
参赛成员 | 张云帆 陈欣禧 谭泽熙 甘国基
龙志全 高逸菲 蔡升宏 陈俊仰
指导老师 | 张逸群 姬玉柱 骆文君

大数据 已成为推动社会进步和经济发展的核心资源





充分开发数据资源，推进智能分析决策，成为重要命题

技术路线图

多源异质大数据联合分析

研究价值

✓ 丰富理论方法

统一距离度量 | 无监督联邦学习机制 | 持续联邦学习范式

✓ 疏通应用堵点

数据多源异质 | 真实环境标签匮乏 | 知识持续更新

创新架构

信息融合损耗 ↓

创新点一
异质特征数据距离新度量

自研统一度量+多层次耦合编码

数据增强

知识嗅探精度 ↑

创新点二
多源数据联邦聚类新策略

多粒度层次脱敏+簇分布自组织

持续更新

知识时效性 ↑

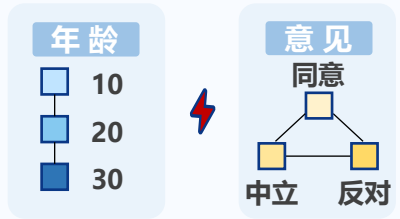
创新点三
动态环境持续学习新模型

异步权重更新+主动漂移感知

核心问题

信息融合损耗大

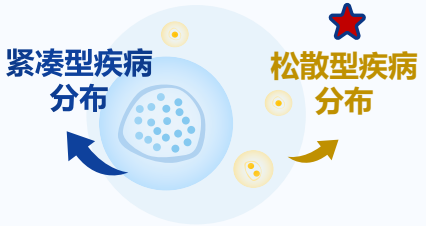
数据异质 → 对齐程度弱



+

联邦聚类精度低

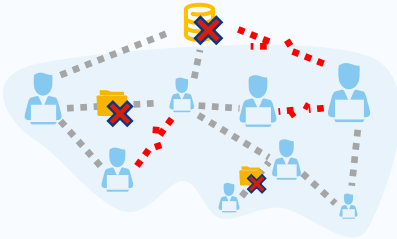
信息非完备 → 聚类依据匮乏



+

知识更新时效差

数据动态 → 更新链路复杂



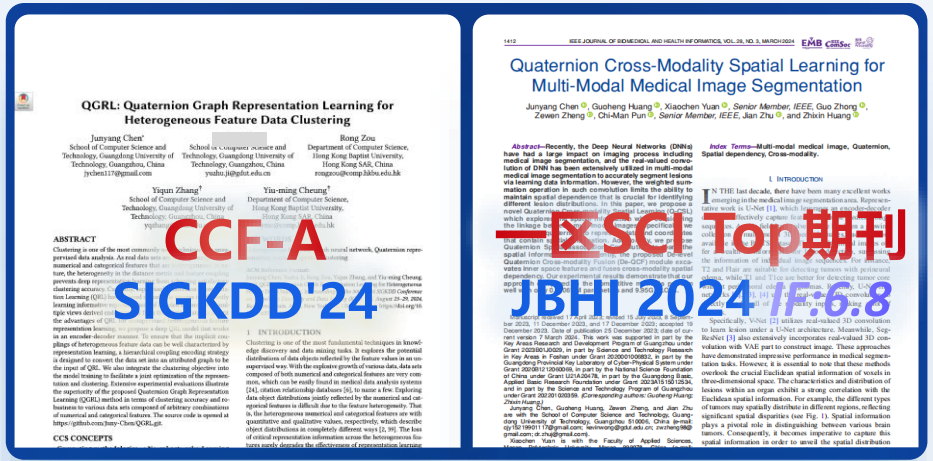
创新点一 异质特征数据距离新度量

痛点一 数据异质，对齐程度弱 ➡ 信息融合损耗大

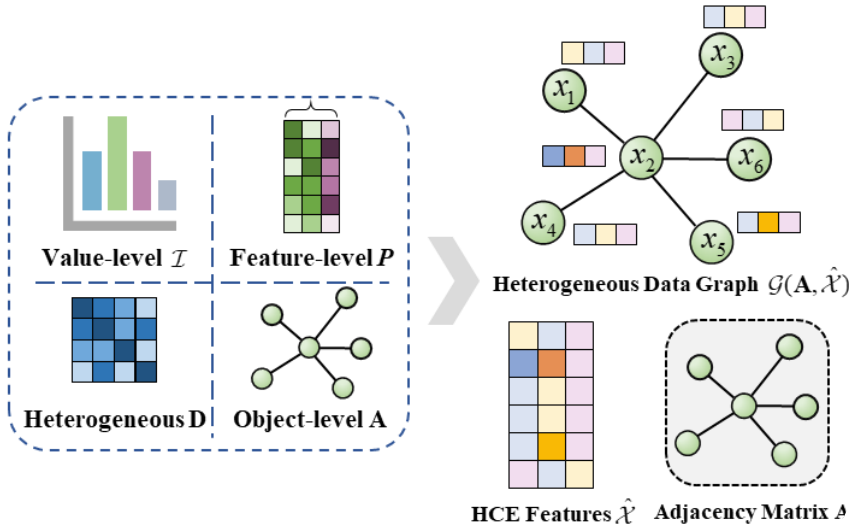
解决方法 ➡ 自研统一度量+多层次耦合编码

- 使用统一度量方法衡量异质指标间距离，弥合信息差
- 从“值-特征-异质-样本”四个层次进行编码，进一步提取隐式信息
- 通过端到端的机器学习方法实现下游任务自适应

解决成效 信息融合损失降低37.49%



层次耦合编码策略



统一度量，精准融合，减少信息流失

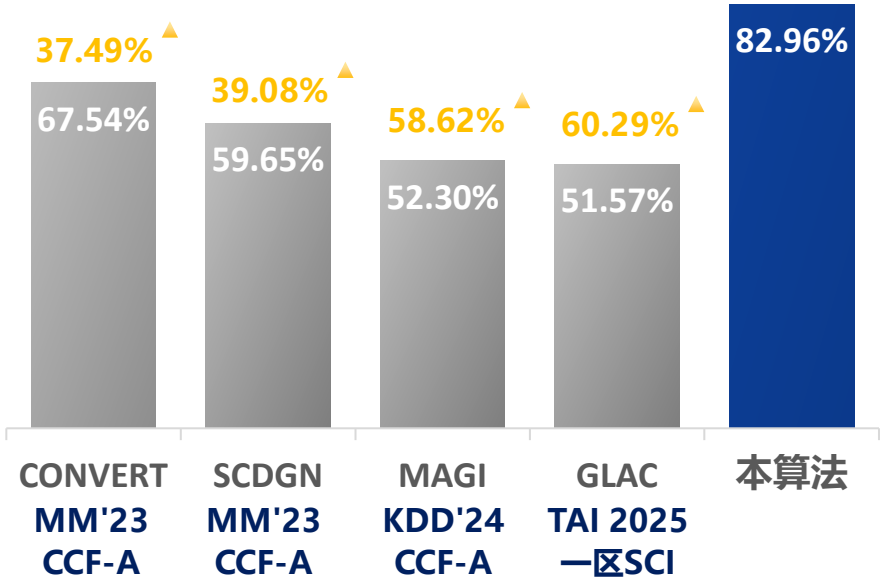
实验结果

当前领域先进算法： GLAC
(TAI 2025 一区SCI)

知识融合精度

51.57% → 82.96%

本算法：QGRL
(SIGKDD'24 CCF-A)



异质信息**提取更充分**，分析决策**精度高**，发表CCF-A类成果

创新点二 多源数据联邦聚类新策略

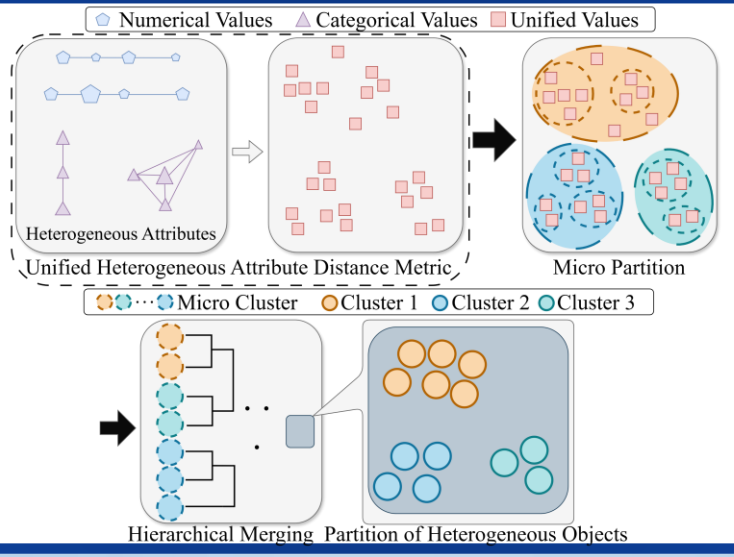
痛点二 信息非完备，聚类依据匮乏 ➡ 联邦聚类精度低

解决方法 ➡ 多粒度层次脱敏+簇分布自组织

- 微簇竞争遵循“胜者通吃”原则，不显著微簇被淘汰
- 显著微簇进行层次融合，得到数据的多层次分布结构
- 基于多层次结构形成不同粒度的样本群体，避免“漏网之簇”

解决成效 知识嗅探精度高达90.50%

微簇划分和分层合并策略



自动搜寻微簇，细查复杂分布，锁定关键少数

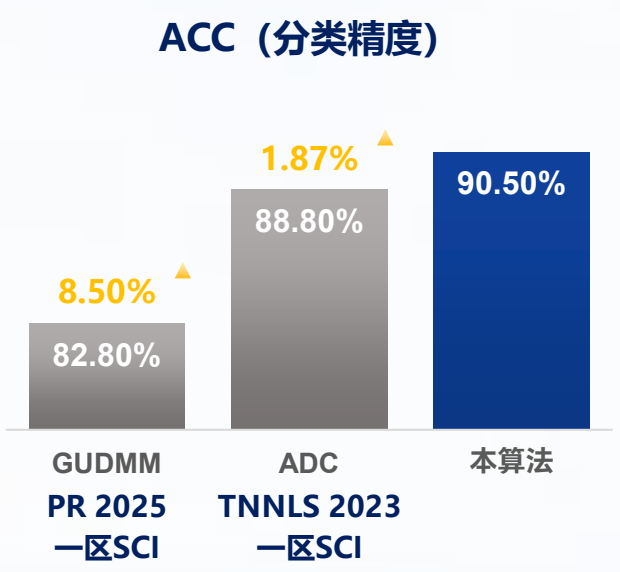
实验结果

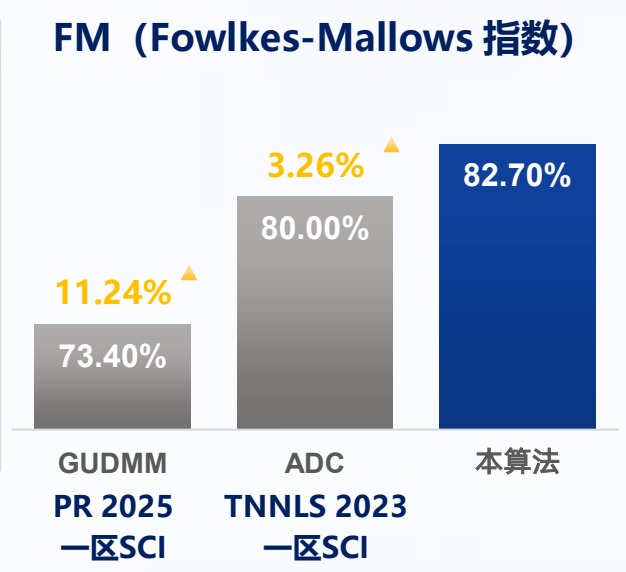
当前领域先进算法： GUDMM
(PR 2025 一区SCI)

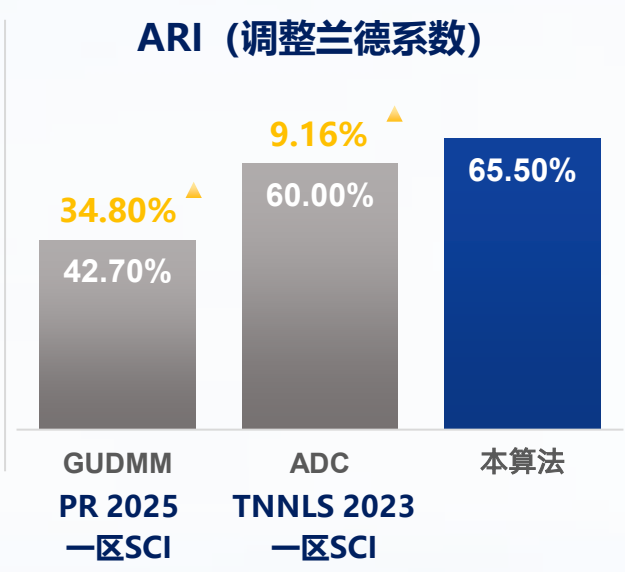
知识嗅探精度

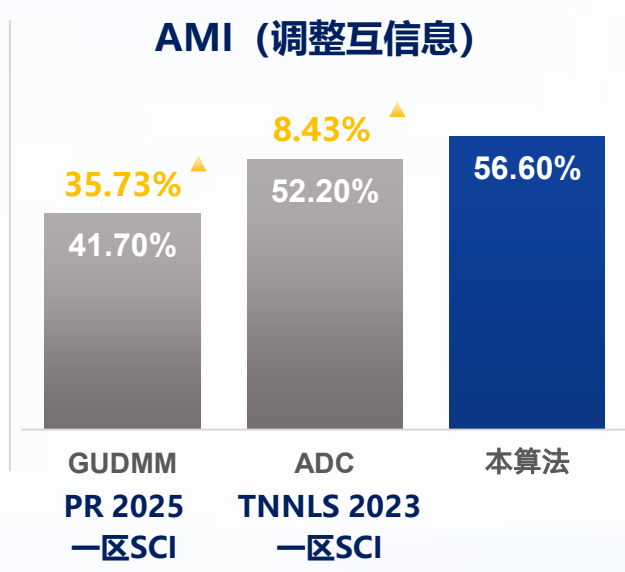
82.80% → 90.50%

本算法： MCDC
(ICDCS'24 CCF-B)









分布检测更细致，知识粒度自适应，获IEEE最佳论文奖

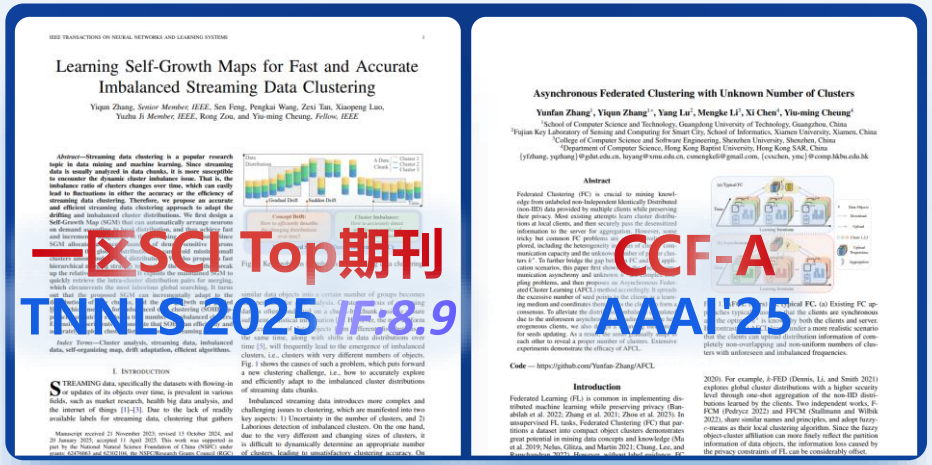
创新点三 动态环境持续学习新模型

痛点三 数据动态，更新链路复杂 ➡ 知识更新时效差

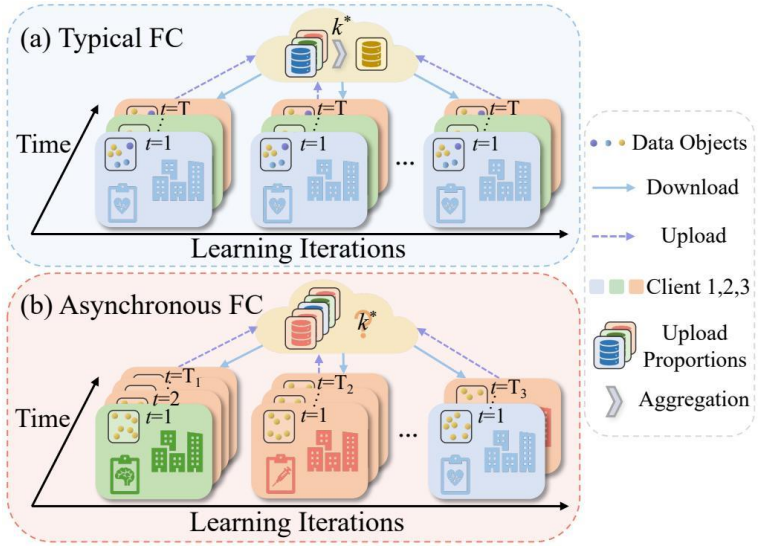
解决方法 ➡ 异步权重更新+主动漂移感知

- 对各节点群体检测数据进行知识概括，提取关键信息
- 对各节点数据实施脱敏处理以确保隐私安全，并通过跨节点的知识互补与簇分布推理，实现整合分析
- 动态接收并更新各节点数据

解决成效 知识更新时效提升高达74.28%



自组织联邦学习+自适应异步更新



异步联邦新架构，打通隐私壁垒，提升分析效能

实验结果

当前领域先进算法：FedSC
(AAAI'25 CCF-A)

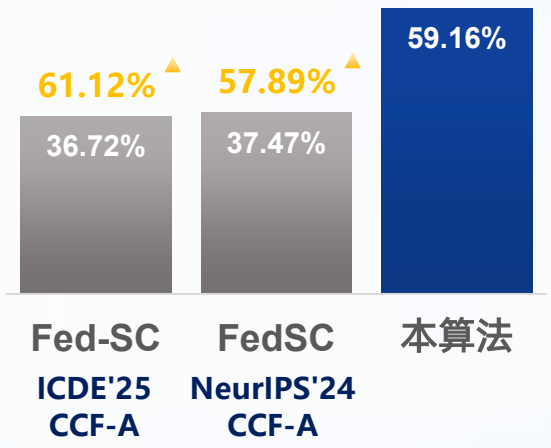
全局知识嗅探精度

82.52% → 89.67%

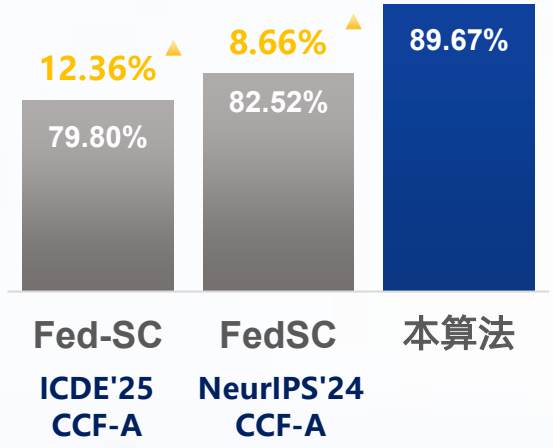
本算法：AFCL
(AAAI'25 CCF-A)

考虑更具挑战性场景： ✦异步通讯（收发不同步） ✦先验知识不可知（真实类别数未知）

在数据集Breast Cancer上对比



在数据集Live上对比



分析鲁棒性强，决策精度再提升，发表CCF-A类成果

科研成果

1篇 CCF-B类

2篇 CCF-A类

4篇 一区SCI

1篇 IEEE最佳论文奖

团队取得论文成果17篇 其中10篇为本科生一作

另有高水平论文在投6篇

学术认可

查新项目名称

多源异质大数据联合分析

查新结论

国内外文献检索中未见相同文献报道
本查新项目具有新颖性

报告编号：202521C0725459

科 技 查 新 报 告

项目名称：多源异质大数据联合分析

委 托 人：广东工业大学

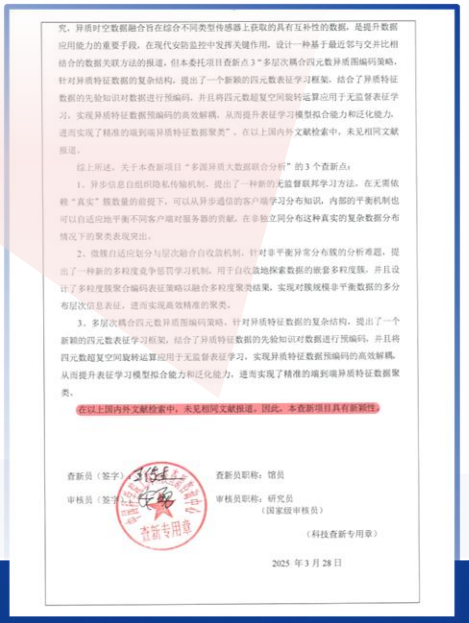
委托日期：二〇二五年三月十八日

查新机构：中国科学院上海科技查新咨询中心
（科技查新专用章）

查新完成日期：二〇二五年三月二十八日

中 华 人 民 共 和 国 科 学 技 术 部

二〇一六年制



在以上国内外文献检索中，未见相同文献报道。因此，本查新项目具有新颖性。

查新员（签字）：

审核员（签字）：

查新员职称：馆员

审核员职称：研究员
（国家级审核员）

（科技查新专用章）

2025 年 3 月 28 日

论文引用成果

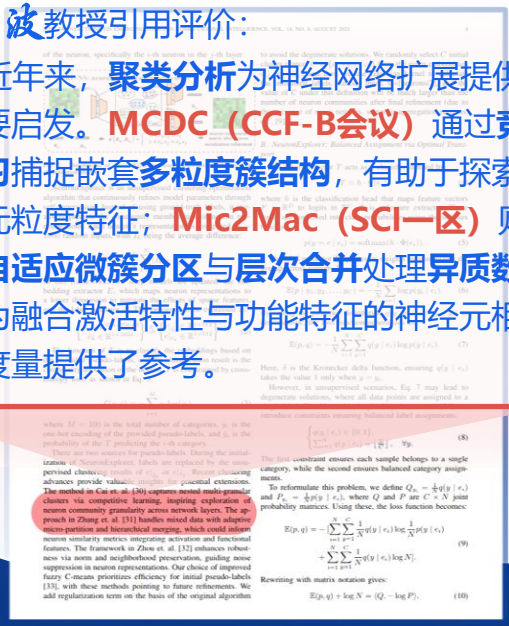
均发表于2024-2025年
累计被引次数100+

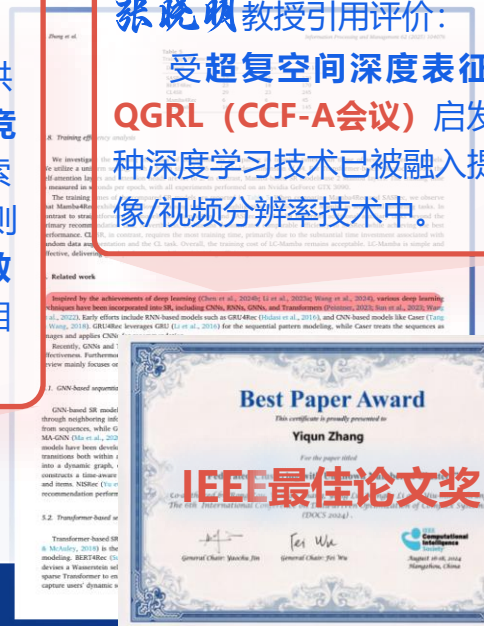
获知名学者多次正面引用与评价

- **教育部长江学者 香港浸会大学讲席教授** 张晓明
- **国家杰青 西安电子科技大学校长** 高新波

高新波教授引用评价：

近年来，**聚类分析**为神经网络扩展提供了重要启发。**MCDC（CCF-B会议）**通过**竞争学习**捕捉嵌套**多粒度簇结构**，有助于探索神经元粒度特征；**Mic2Mac（SCI一区）**则基于**自适应微簇分区**与**层次合并**处理**异质数据**，为融合激活特性与功能特征的神经元相似性度量提供了参考。





张晓明教授引用评价：

受**超复空间深度表征学习QGRL（CCF-A会议）**启发，各种深度学习技术已被融入提升图像/视频分辨率技术中。

已录用论文成果列表

No.	名称	
1	Yiqun Zhang,..., Zexi Tan, Xiaopeng Luo, Yuzhu Ji*, "Learning SelfGrowth Maps for Fast and Accurate Imbalanced Streaming Data Clustering", *TNNLS*, 2025. 探索动态数据分布自适应的高精度高效率非平衡数据聚类算法	一区SCI IF: 8.9
2	Yuzhu Ji, Wei Hu, Guoji Gan, Zhiquan Long, Yiqun Zhang*, An Zeng*, "SMACNet: A Unified Framework for One-shot Talking Head Synthesis via Subtle Motion and Appearance Compensation", *TCE*, 2025. 提出能够捕捉动态数据中细微变化的深度表征与生成模型	一区SCI IF: 10.9
3	Junyang Chen,..., and Zhixin Huang, "Quaternion Cross-Modality Spatial Learning for Multi-Modal Medical Image Segmentation", *JBHI*, 2024. 面向医学影像分割应用的距离度量空间学习深度模型	一区SCI IF: 6.8
4	Rong Zou, Yunfan Zhang, ..., Zexi Tan, Yiqun Zhang* and Yiu-ming Cheung*, "SDENK: Unbiased Subspace Density-k-Clustering", *Neurocomputing*, 2025. 密度分布学习与子空间学习融合互补的无偏聚类算法	一区SCI IF: 6.5
5	Yunfan Zhang, Rong Zou, Yiqun Zhang*,..., "Adaptive Micro Partition and Hierarchical Merging for Accurate Mixed Data Clustering", *CAIS*, 2024. 提出数据微簇划分机制，引导高精度异质数据聚类	二区SCI IF: 4.6
6	Chuyao Zhang, Xinxi Chen, Zexi Tan,...,Yuzhu Ji,Yiqun Zhang*, "Towards Clustering of Incomplete Mixed-Attribute Data", *Expert Systems*, 2025. 对异质特征数据进行缺失值补全以增强聚类精度	二区SCI IF: 2.3
...		
8	Rong Zou, Yunfan Zhang, Yiqun Zhang,..., Yiu-ming Cheung*, "Federated Clustering with Unknown Number of Clusters", *IEEE DOCS*'24. 针对普遍的真实簇数未知情形提出簇分布自探索联邦聚类算法	EI IEEE 最佳论文奖

No.	名称	
9	Yunfan Zhang, Yiqun Zhang* , Yang Lu, Mengke Li, Yiu-ming Cheung, "Asynchronous Federated Clustering with Unknown Number of Clusters", *AAAI'25*. 本地节点异步传输且合理簇数未知情形下的联邦聚类范式	CCF-A
10	Junyang Chen, Yuzhu Ji , Rong Zou, Yiqun Zhang, Yiu-ming Cheung, "QGRL: Quaternion Graph Representation Learning for Heterogeneous Feature Data Clustering", *SIGKDD*'24. 引入超复空间深度表征学习机制突破异质特征数据聚类精度瓶颈	CCF-A
11	Shenghong Cai, Yiqun Zhang*, ..., Hong Jia, Peng Liu, "Robust Categorical Data Clustering Guided by Multi-Granular Competitive Learning", *ICDCS*'24. 拓展多粒度竞争学习理论提升聚类分析鲁棒性	CCF-B
12	Yunfan Zhang, ..., Yiqun Zhang*, Yiu-ming Cheung, "Towards Unbiased Minimal Cluster Analysis of Categorical-and-Numerical Attribute Data", *ICPR*'24. 提出异质特征数据最小簇理论实现低归纳偏执聚类分析	CCF-C
13	Pengkai Wang†, Yunfan Zhang†, Yiqun Zhang*, ..., Yiu-ming Cheung, "Clustering by Learning the Ordinal Relationships of Qualitative Attribute Values", *IJCNN*'24. 进行异质特征数据取值顺序关系学习以增强聚类精度	CCF-C
14	Haoyi Xiao, Xinxi Chen, Xiaopeng Luo* , ..., Wei Ai, "MACL: Metric and Attribute Space Co-Learning for Qualitative Data Clustering" , *ICIC*'25. 提出度量空间和属性子空间协同学习机制增强聚类精度	CCF-C
...		
17	Zexi Tan, Tao Xie, Binbin Sun, Xiang Zhang, Yiqun Zhang* and Yiu-ming Cheung, "MEET-Sepsis: Multi-Endogenous-View Enhanced Time-Series Representation Learning for Early Sepsis Prediction", *PRICAI*'25. 面向脓毒症早期预测的多源内生视图增强时序表征学习	CCF-C

注： 团队本科生成员 指导老师 通讯作者*

专家推荐

沈卫明

加拿大工程院院士
华中科技大学教授
福耀科技大学讲席教授

推荐理由摘选

"对无监督联邦学习和异质数据分析的研究具有前沿性和前瞻性，项目所提出的方案为解决该领域的关键基础性技术难题提供了创新思路。"

蔡宏民

国家杰出青年科学基金获得者
华南理工大学教授

推荐理由摘选

"成果构成了联邦异质数据分析方法体系，理论性和科学性较强，研究具有前瞻性，拓展了数据科学和机器学习研究领域。在许多重要行业有良好应用前景，有望提升行业应用中的数据分析效能。"

团队成员

张云帆（负责人）

- 录用CCF-A会议一篇（一作）
- 录用一区SCI期刊一篇（一作）
- 录用一区SCI期刊一篇（第一本科生作者）
- 录用四区SCI期刊一篇（三作）
- 录用CCF-C会议两篇（一作）
- 录用IEEE会议一篇（获最佳论文奖）
- 获得本科生国家奖学金
- 已被香港高校直博全奖录取

谭泽熙

- 录用一区SCITop期刊一篇（第一本科生作者）
- 录用一区SCI期刊一篇（三作）
- 录用二区SCI/CCF-C期刊一篇（三作）
- 录用四区SCI期刊一篇（第一本科生作者）
- 录用CCF-C会议论文一篇（一作）
- 录用EI会议论文一篇（一作）
- 国家级大创项目第一负责人
- 计算机设计大赛国赛一等奖第一负责人
- 广东省级科技创新项目第一负责人

陈俊仰

- 录用一区SCI期刊一篇（一作）
- 录用CCF-A会议一篇（一作）
- 在投CCF-A会议一篇（一作）
- 在投一区SCI期刊一篇（二作）
- 保研至清华大学

陈欣禧

- 录用二区SCI/CCF-C期刊一篇（第一本科生作者）
- 录用CCF-C会议一篇（二作）
- 二审一区SCITop期刊一篇（二作，指导老师一作）
- 在投一区SCI期刊一篇（一作）
- 国家级大创项目第一负责人

蔡升宏

- 录用CCF-B会议一篇（一作）
- 在投CCF-A会议一篇（一作）
- 在投一区SCITop期刊一篇（二作）
- 国家级大创项目第一负责人
- 多所港澳高校直博全奖条件录取

甘国基

- 录用一区SCI期刊一篇（第一本科生作者）
- 在投CCF-B会议一篇
- 实审&授权国家发明专利各一项
- 保研至广东工业大学

龙志全

- 录用一区SCI期刊一篇（四作）
- 在投CCF-B会议一篇
- 实审&授权国家发明专利各一项

应用价值

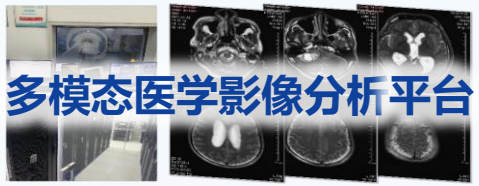
现实应用价值

✓ 与各三甲医院展开广泛合作与应用部署
（如广州医科大学附属第三医院、深圳市妇幼保健院）

✓ 成果价值广受合作单位认可

预处理 + 预分析
医院本地数据

提取 + 融合
复杂临床信息



联合分析
甲方医院-乙方企业

应用拓展

通用分析工具
- 数据约简
- 簇数自动判定
- 异常检测
- 数据可视化

医学领域应用
- 辅助诊断
- 疾控与健康数据分析

国防军工领域应用
- 网络攻击检测
- 情报分析与整合

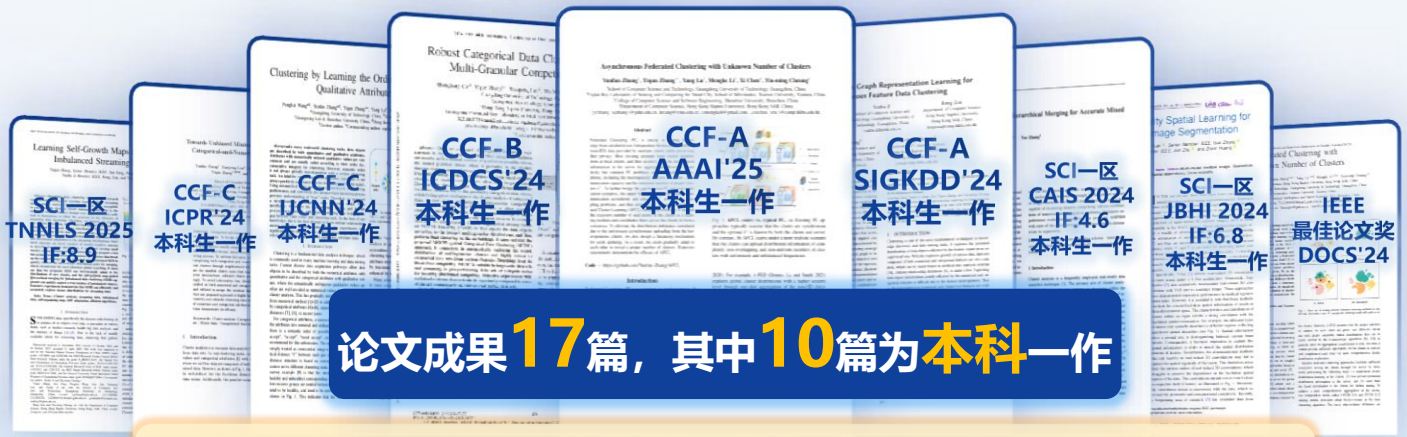
金融领域应用
- 信用风险评估
- 实时欺诈检测

显著实际意义 & 广泛应用价值

多源异质大数据联合分析

智联异质数据孤岛·赋能精准分析决策

敬请各位
评委老师指导



论文成果17篇，其中10篇为本科一作

团队已有成员保送至清华大学、全奖直博至香港高校

科学性

面向应用需求和技术现状解决系列关键科学问题：

- 率先统一异质特征数据距离度量
- 创新自组织联邦聚类策略
- 提出原创持续联邦学习模型

先进性

被知名学者正面评价引用，查新报告证实新颖性：

- 异质数据基础测度：信息融合损失降低高达37.49%
- 多源数据联邦聚类：知识嗅探精度高达90.05%
- 动态环境持续学习：知识更新时效提升高达74.27%

拓展性

- 为疾控健康数据分析、信用风险评估等诸多领域提供通用方法和技术支持