基于异构网络的相关数据挖掘任务研究综述*

于诗睿 李爱花 林紫洛 唐小利

（中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100005）

〔摘要〕 详细阐述异构网络相关数据挖掘任务研究领域的基本问题、研究进展以及其在现实世界应用情况，对该领域未来研究热点提出展望，指出异构网络是情报学领域处理复杂大数据的一种有效工具。

〔关键词〕 异构网络；数据挖掘；表征学习；元路径；半结构化数据

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.04.005

A Review of Related Data Mining Tasks Based on Heterogeneous Networks *YU Shirui, LI Aihua, LIN Ziluo, TANG Xiaoli, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China*

〔**Abstract**〕 The paper elaborates the basic problems, research progress and its applications in the field of data mining tasks related to heterogeneous networks in detail, proposes the prospect of future research hotspots in this field, and points out that heterogeneous network is an effective tool for processing complex big data in the field of information science.

〔**Keywords**〕 heterogeneous network; data mining; representative learning; meta path; semi－structural data

1 引言

现实世界中各种事物和关系相互交织，图（又称为“网络”）作为连通数据结构的常用表现形式，能够通过对象及其之间的链接对实体和关系进行建模，是用于模拟对象交互的一种普遍方式[1]，成为当下信息基础设施建设的重要组成部分。此外随着信息技术的高速发展，产生了大量相互关联的结构化和半结构化数据。将这类复杂的交互成分建模为包含不同类型对象和链接的异构图，可实现对现实世界数据中丰富语义信息和结构信息的有效整合和全面反映。异构图已成为研究异构、多模态、多关系、多类型数据的强大模型[2]。基于异构图的分析也成为数据挖掘任务的新方向，并催生出推荐任务、节点分类和聚类、知识库完善等新型任务[3]。因此如何深入理解异构图中的数据，从中提取出潜在信息，开发有效算法促进异构图的应用，提升下游数据挖掘任务的效果，成为当下研究的热点问题。

本文全面总结近年来有关异构图数据挖掘的研究进展，并针对重点领域和研究前沿开展详细讨论。本文工作将有助于研究人员全面了解该领域，从而提升基于异构图的数据挖掘效果，同时也有助于该模型更好适用于现实世界应用，从而解决实际问题。

〔修回日期〕 2022－11－23

〔作者简介〕 于诗睿，硕士研究生，发表论文 1 篇；通信作者：唐小利，研究馆员，发表论文 93 篇。

〔基金项目〕 中国医学科学院医学与健康科技创新工程 2021 年重大协同创新项目“生物医学文献信息保障与集成服务平台”（项目编号：2021－I2M－1－033）。

2 异构网络表示学习

2.1 技术角度算法分类

近年来有关异构网络的研究越来越受到关注，其中异构图表示是一项非常重要的研究内容，为下游的网络分析和数据挖掘任务提供有力支撑[4]。从技术角度总结，异构网络表示学习常用算法可分为浅层模型和深层模型两类，见图 1。

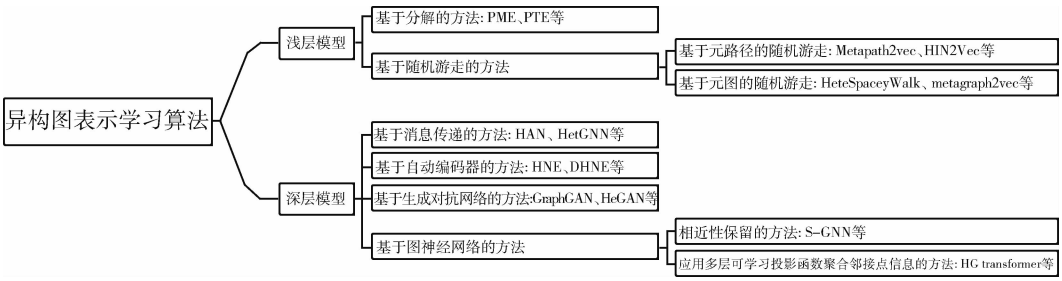

图 1 技术角度异构图表示学习算法分类

2.2 基于不同特点的算法分类

异构图表示学习的关键是保留图中节点嵌入的结构和属性，因此如何有效处理由多类型节点和关系融合导致的高阶复杂结构、将异质性属性表达的不同含义有效结合以及如何融入先进的领域知识促进异构图在真实世界的应用等，成为该领域算法改进亟待解决的问题。根据异构图表示学习使用的信息可将现有算法分为 5 类，见图 2。

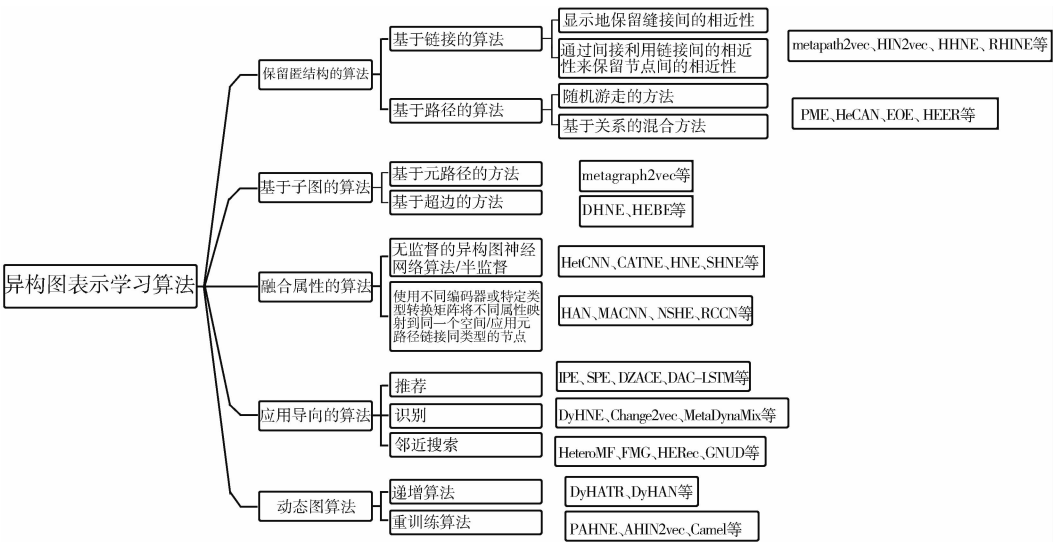

图 2 基于不同特点的异构图表示学习算法分类

3 数据挖掘任务

3.1 概述

异构网络为数据挖掘任务提供了新的范式，是情报学领域进行大数据分析的有效工具。本文调研了 Web of Science 数据库中相关文献，发现目前数据挖掘任务主要可分为聚类、分类、链路预测、排序、推荐、信息融合、相似度测量 7 类，见图 3。本文选择 3 种典型任务进行总结。

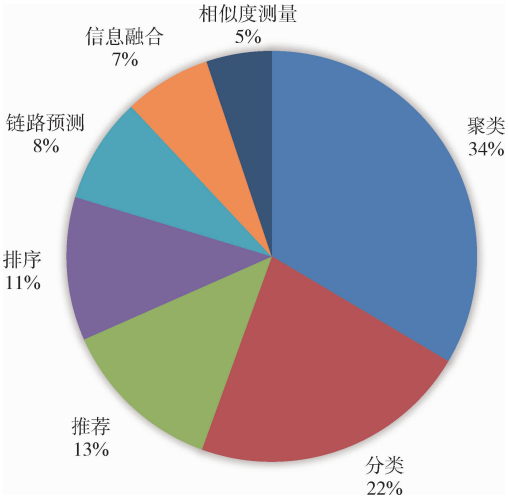

图 3 基于异构网络的数据挖掘任务研究分布情况

3.2 聚类

聚类分析是将一组数据对象划分为一组集群，并使每个集群中的对象彼此相似，但又与其他集群中的对象不同的过程。异构图中集成的多类型对象和链接给聚类任务带来巨大挑战。根据集成信息或任务类型，异构图聚类分析可分为以下 4 类：属性信息集成的聚类分析，文本信息集成的聚类分析，用户指南信息集成的聚类分析，与排序任务、社区检测等其他数据挖掘任务集成的聚类分析等。近年来元路径机制为异构图聚类提供了新的方法，但还存在许多问题亟待解决：一是现有模型无法有效集成多路径图的聚类结果，聚类质量不高；二是现有的相似度聚类方法顶点分配和聚类目标的粒度过于粗糙，无法反映现实情况；三是仅考虑顶点同质性聚类，没有考虑边聚类，聚类结果不准确；四是缺乏将顶点聚类和边聚类技术有效结合的模型[5]。其中较典型的研究内容，见表 1。

表 1 聚类任务中的代表性研究

代表性研究	作者	研究内容
集成文本信息的聚类分析	Wang Q 等[6]	提出一个用于主题挖掘和多对象聚类的同一主题模型
基于元路径的聚类分析	Zhou Y 等[7]	提出一种基于元路径图的聚类框架 VERufiCuis TER，将元路径顶点中心聚类与元路径边缘中心聚类相结合以提高异构图聚类任务的质量
动态聚类分析	Kumar D 等[8]	改进 GraphIVE 性能，提出一种基于顶点切割的异构图聚类处理框架 GraphSteal，并能实现对图的动态重聚类
异常值检测	Gupta M 等[9]	提出一种基于联合非负矩阵分解的异常值感知方法来发现流行社区分布模式

3.3 分类

分类任务是通过构建模型或分类器来预测类别标签的数据分析任务。其中基于图的分类任务需要考虑对象之间存在的链接及之间的相关性。异构图分类问题的研究具有独特特征。首先，异构图中包含的对象类型不同，需要同时对多种类型的对象进行分类。其次，异构图中对象的标签是由不同类型对象和不同类型链接的共同影响所决定的。目前有许多工作将同构图的方法扩展到异构图中，如归纳分类方法、多标签分类方法、同构图标签传播方法等。元路径作为异构图的一种独特特性也被广泛用于异构图的分类任务，常用于特征生成。考虑到元路径的缺陷还提出基于元图的方法。与聚类问题类似，分类任务也常与其他数据挖掘任务结合研究，见表 2。

表 2 分类任务中的代表性研究

代表性研究	作者	研究内容
转导分类方法	Jacob Y 等[10]	提出一种通过计算空间中节点潜在表示来标记不同类型节点的方法
归纳分类方法	Rossi R G 等[11]	使用二分异构网络表示文本文档集合，并提出 IMBHN 归纳分类模型，为文本术语分配权重

续表 2

代表性研究	作者	研究内容
多标签分类方法	Zhou Y 等[12]	一种以活动 - 边缘为中心的多标签分类框架，用于分析具有 3 个独特特征的异构信息网络
同构图标签传播方法	Hwang T H 等[13]	改进原有同构图标签传播算法，提出 MINProp 算法，是一个在异构子网间传播信息的简单的正则化框架
基于元路径方法	Luo C 等[14]	提出一种转导分类方法 HetPathMine，其中使用元路径用于该方法中的特征生成
基于元图方法	Zhang J 等[15]	提出一种基于加权元图的异构信息网络分类框架，能够挖掘出不同节点潜在的结构特征，并增强节点间的语义关系

3.4 链路预测

链路预测的目标是利用网络中的可用信息来检测缺失的链路或预测未来可能形成的关系。异构网络中链路预测的重点是捕捉不同类型链接间的复杂关系，并利用互补的预测信息，预测多种类型的链接[16]。异构网络的链路预测方法通常分为以下 3 类：第 1 类是将异构网络简化为同构网络，通过探索不同节点和关系的类型来提取目标节点对间的相似特征，但此类方法不能推广到其他复杂的异构网络中；第 2 类是使用两种元结构，即网络模式和元路径，探索节点相似性；第 3 类使用概率模型进行链路预测。此外有监督链路预测也是一种提取异质特征的有效方法，可以与元结构方法结合使用，可提升提取结构特征和语义特征的能力。上述方法主要集中在对单个异构网络和静态异构网络链路预测的研究。还有一些对跨多个对齐异构网络的链路预测问题和动态链路预测问题的研究方法也非常重要，见表 3。

表 3 链路预测任务中的代表性研究

代表性研究	作者	研究内容
将异构网络简化为同构网络	Bao Z F 等[17]	提出一种针对社交网络的链路预测方法 sonLP，使用主成分分析来识别对链路预测重要的特征
基于元路径的链路预测	Hadi H 等[18]	将元路径中包含的不同节点和边的异质特征与有监督链路预测相结合，提出一种基于核的单类异构网络链路预测方法
基于概率模型的链路预测	Yang Y 等[19]	使用主题因子图模型（topical factor graph model，TFGM）定义一个潜在主题层来链接多个网络，并设计半监督学习模型来挖掘异构网络间的竞争关系
融合多源信息的链路预测	Jiang T 等[20]	提出一种基于自适应重要性抽样的多源异构信息融合图表示学习模型，该模型融合了图结构和文本描述信息，能够为实体生成丰富的属性嵌入
跨多网络的链路预测	Zhang Y T 等[21]	提出一个基于能量的模型 COSNET，同时也考虑多个网络间的局部和全局一致性
动态链路预测	Aggarwal C C 等[22]	开发一种两级方案用于解决时间和异构信息网络中的动态链接推断问题，该方法可以通过结合拓扑和类型信息做出有效的宏观和微观决策

4 现实世界应用

异构网络包含的数据规模巨大，能够表达丰富的语义信息，广泛应用于引文网络、生物学网络、商业网络、媒体网络、社会网络等多种领域。其中生物学网络和电商网络最为典型。

4.1 生物学网络

生物学系统具有高度复杂的特点，构建异构网络可以形成统一框架来有效处理这类复杂数据，主要具有以下两项优势：首先，异构网络通过整合先

验知识提升预测的可信度并发现潜在知识，从而消除原始数据中的假阳性结果和噪音；其次，异构网络可以通过内部间接关联将不同生物学领域的数据链接起来。对生物学异构网络进行分析的主要方法有基于网络的链路预测，利用节点间的相似性搜索算法如 Katz 测度、随机游走（random walk，RW）的转移矩阵等来获取局部或全局网络的拓扑结构；还有一些方法基于元路径来定义具有不同语义类型的路径，然后计算路径间的相似度，如基于元路径的相似度算法（meta - path based similarity，PathSim）、异构网络的相似性度量（heterogeneous network based similarity，HeteSim）等算法。其中异构生物学网络的应用主要分为药物重定位、基因 - 表型关系识别、非编码 RNA 功能注释、人类微生物 - 疾病关联 4 个细分领域。分析过程中面临的主要挑战和未来研究热点包括：对生物学数据中噪声、缺失值的处理，面对组学数据不平衡时对关键数据的筛选，如何有效处理生物学数据中的多对多关系，开发能够有效处理生物学数据噪声多、数据稀疏等面向不同领域更为专指问题的相关模型算法，为适应精准医学的发展趋势开发新型高分辨率测量技术，有效利用异构网络解决生物学领域单细胞组学和人类细胞图谱的相关问题，开发更具普适性的标准框架实现社区驱动的知识共享。异构网络自身具有的独特特性能很好地处理具有噪声多、稀疏和复杂等特点的生物学数据，成为解决目前生物学领域棘手难题的强大工具[23]。

4.2 电商网络

随着互联网的快速普及，一些大型电商平台发展迅速。电商平台中涉及大量异构对象和交互信息，数据量庞大且复杂，使异构网络成为服务于用户商品推荐、意向推荐、用户画像和欺诈检测等各项任务的关键手段。其中推荐是电商平台的一项重要服务，异构网络可以用来对用户、商品和辅助信息进行建模，实现对用户的商品推荐。意图推荐是根据用户的历史行为实现对用户意愿的自动推荐，而无需任何输入信息。如 Fan S 等[24]提出将用户意图表示为搜索框中的默认查询，从而将意向推荐问题转化为查询推荐问题。用户画像任务在电商平台的个性化服务提供中具有重要地位，通过将用户丰富的交互信息建模为异构图从而丰富用户特征。随着电商平台的发展，系统中出现了许多欺诈者，以不正当手段从交易中牟利。由于欺诈者行为模式具有异质性，可以通过异构图检测这些恶意账户。

5 讨论和展望

5.1 解决复杂网络构建问题

实际应用中的异构网络非常复杂。首先，网络中的对象可能与现实世界中的实体不完全对应，网络中一个对象可指代多个实体或不同对象可指代同一实体。其次，对象间的关系可能没有明确给出或不完整。最后，对象和链接可能不可靠。因此需要对网络中的数据进行清洗和整合从而构建高质量的网络。此外网络中存在一些如文本数据、多媒体数据等非结构化数据，使异构网络的构建更具挑战性。应考虑将信息抽取、自然语言处理和其他技术相结合，实现对高质量异构网络的构建，为后续数据挖掘任务奠定基础。

5.2 开发更强大的数据挖掘算法

开发基于异构网络的相关算法要充分考虑异构网络的两个重要特征，即结构复杂性和语义丰富性。目前开发出的算法存在一定局限性。现实世界网络中的数据通常更加复杂不规则，网络的链接和对象上提供的属性信息未被充分利用。还需要构建动态网络考虑时间因素的影响；将不同网络中的同一对象进行对齐，有效融合不同网络的信息；设计更强大、更灵活的异构网络算法。

元路径和元图是异构网络的典型特征，在语义获取和特征选择方面表现出强大性能，但存在一定缺陷。一是无法捕获更细微的语义信息；二是该类方法不考虑链接上的属性值。此外该类方法在语义获取方面还面临两点挑战。一是考虑如何从复杂网络中自动提取具有最佳解释性的元路径；二是元路径权重的确定，能够体现网络中路径的重要性，表达不同类别的语义信息。未来要扩展原有算法，或

设计能够获取更详细语义的工具，实现对更复杂网络的分析。

5.3 基于异构网络的深度图学习算法

近年来，基于异构图的神经网络算法开始受到广泛关注。但目前对异构图的深度神经网络算法理论分析仍有不足。此外新技术开发也是研究方向。其中，一个重要方向是自我监督学习，即利用预先标注任务训练神经网络，减少对人工标签的依赖，有效解决实际情况中标签不足的问题，在异构图中性能显著，有待进一步探索；另一个方向是对异构图神经网络的预训练。目前针对异构图的神经网络方法缺乏迁移能力，不仅耗时且需要大量标签，因此考虑使用具有较强泛化能力的预训练异构神经网络，实现使用少量标签进行微调。

6 语结

由于异构网络能够包含丰富的结构和语义信息，是情报学领域处理复杂大数据的一种有效工具。本文对基于异构网络的相关数据挖掘任务进行系统梳理，介绍了有关方面的最新进展，并指出未来发展方向，为研究人员提供一些新的视角。

参考文献

1 XIA F, SUN K, YU S, et al. Graph learning: a survey [J]. IEEE transactions on artificial intelligence, 2021, 2 (2): 109 – 127.

2 YANG C, XIAO Y, ZHANG Y, et al. Heterogeneous network representation learning: aunified framework with survey and benchmark [J]. IEEE transactions on knowledge and data engineering, 2022, 34 (10): 4854 – 4873.

3 XIE Y, YU B, LV S, et al. A survey on heterogeneous network representation learning [J]. Pattern recognition, 2021, 116 (3): 107936 – 107950.

4 周丽华，王家龙，王丽珍，等. 异质信息网络表征学习综述 [J]. 计算机学报，2022，45 (1)：160 – 189.

5 DONG Y X, HU Z N, WANG K S, et al. Heterogeneous network representation learning [C]. online: 29th International Joint Conference on Artificial Intelligence, 2021.

6 WANG Q, PENG Z H, WANG S Z, et al. cluTM: content and link integrated topic model on heterogeneous information networks [C]. Qingdao: 16th International Conference on Web – Age Information Management (WAIM), 2015.

7 ZHOU Y, LIU L, BUTTLER D, et al. Integrating vertex – centric clustering with edge – centric clustering for meta path graph analysis [C]. Sydney: 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2015.

8 KUMAR D, RAJ A, DHARANIPRAGADA J. GraphSteal: dynamic re – partitioning for efficient graph processing in heterogeneous clusters [C]. Honolulu: 10th IEEE International Conference on Cloud Computing (CLOUD), 2017.

9 GUPTA M, GAO J, HAN J. Community distribution outlier detection in heterogeneous information networks [C]. Berlin: Machine Learning and Knowledge Discovery in Databases, 2013.

10 JACOB Y, DENOYER L, GALLINARI P, et al. Learning latent representations of nodes for classifying in heterogeneous social networks [C]. New York: 7th ACM International Conference on Web Search and Data Mining (WSDM), 2014.

11 ROSSI R G, FALEIROS T D, LOPES A D, et al. Inductive model generation for text categorization using a bipartite heterogeneous network [C]. Brussels: 12th IEEE International Conference on Data Mining (ICDM), 2012.

12 ZHOU Y, LIU L. Activity – edge centric multi – label classification for mining heterogeneous information networks [C]. New York: 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2014.

13 Hwang T H, Kuang R. A heterogeneous label propagation algorithm for disease gene discovery [C]. Columbus: 10th SIAM International Conference on Data Mining (SDM), 2010.

14 LUO C, GUAN R, WANG Z, et al. HetPathMine: anovel transductiveclassification algorithm on heterogeneous information networks [C]. Amsterdam: Advances in Information Retrieval, 2014.

15 ZHANG J, LI T, JIANG Z, et al. A noval weighted meta graph method for classification in heterogeneous information networks [J]. Applied sciences, 2020, 10 (5): 1603.

16 CHEN Z Y, FAN Z P, SUN M H. Tensorial graph learning for link prediction in generalized heterogeneous networks [J]. European journal of operational research, 2021, 290 (1): 219 – 234.

17 BAO Z F, ZENG Y, TAY Y C. sonLP: social network link prediction by principal component regression [C]. Niagara Falls: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2013.

18 SHAKIBIAN H, CHARKARI N M, JALILI S. Multi-kernel one class link prediction in heterogeneous complex networks [J]. Applied intelligence, 2018, 48 (10): 3411-3428.

19 YANG Y, TANG J, LI J Z. Learning to infer competitive relationships in heterogeneous networks [J]. Acmtransactions on knowledge discovery from data, 2018, 12 (1): 12.

20 JIANG T, WANG H, LUO X, et al. MIFAS: multi-source heterogeneous information fusion with adaptive importance sampling for link prediction [J]. Expert systems, 2022, 39 (4): 1-13.

21 ZHANG Y T, TANG J, YANG Z L, et al. COSNET: connecting heterogeneous social networks with local and global consistency [C]. Sydney: 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2015.

22 AGGARWAL C C, XIE Y, YU P S. A framework for dynamic link prediction in heterogeneous networks [J]. Statistical analysis and data mining, 2014, 7 (1): 14-33.

23 TSUYUZAKI K, NIKAIDO I. Biological systems as heterogeneous information networks: amini-review and perspectives [EB/OL]. [2021-12-24]. https://www.semanticscholar.org/paper/Biological-Systems-as-Heterogeneous-Information-A-Tsuyuzaki-Nikaido/0ed218b3711004069bf5bddb0c44fca319e6de0a.

24 FAN S, ZHU J, HAN X, et al. Metapathguided heterogeneous graph neural network for intent recommendation [C]. Anchorage: 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 2019.

2023 年《医学信息学杂志》编辑出版重点选题计划

2023 年本刊将继续以“学术性、前瞻性、实践性”为特色，及时追踪并深入报道国内外医学信息学领域前沿热点，反映学科研究动态，展示学科研究与应用成果，引领学科发展方向。2023 年度编辑出版重点选题包括但不限于以下方向：

1 党的二十大精神引领下的医学信息学研究领域新使命、新格局；2 数字中国、健康中国建设过程中的数字技术与系统思维；3 健康中国战略背景下医药卫生信息化发展规划政策解读；4 人口健康数据助力现代医学发展及健康中国建设研究；5 新一代信息技术在卫生健康行业的重点应用；6 智能算法、算力平台建设及其医学应用；7 生成式人工智能、人工智能工程化在医疗健康领域的应用探索；8 智慧医疗、智慧健康、智慧养老服务体系构建；9 虚拟生理人体建模与仿真关键技术实现；10 人工智能辅助药物发现及药品供应保障智慧监测；11 数智赋能的重大突发公共卫生事件预测；12 数据与知识驱动的临床辅助决策支持系统；13 健康数据生态系统治理体系研究；14 元宇宙和数字孪生在健康医疗领域的创新应用；15 真实世界数据研究方法、案例及其对医疗卫生决策的助推作用；16 医学小数据与暗数据价值评估；17 面向多模态医疗健康数据的知识组织、知识发现方法；18 可计算医学知识、元知识与医学知识图谱；19 开放医学数据中敏感数据识别与隐私计算；20 大规模人群队列研究及其共享平台建设；21 居民健康指数构建与精准画像研究；22 全民健康信息平台建设、应用、共享与评价；23 医疗卫生信息系统互联互通及其相关标准完善、应用推广与服务管理；24 数字健康融合发展创新体系建设；25“互联网+医疗健康”关键技术与新业态；26 基于主动健康理念的用户健康信息行为研究；27 网络伪健康知识评估机制与虚假信息治理；28 数字时代的心理健康；29 智慧医学图书馆建设管理、理念创新及智慧馆员培养；30 数字乡村背景下农村医疗健康信息协同与智慧服务；31 数字信息生态与智慧养老视域下健康信息服务适老化发展模式；32 医学文献资源精细组织与精准服务模式；33 医学信息学及其分支学科建设与创新发展；34“新医科”背景下医学信息学高层次、复合型人才培养。

《医学信息学杂志》编辑部