

基于 2 阶段集成的多层网络社区发现算法

赵兴旺 张珧溥 梁吉业

(山西大学计算机与信息技术学院 太原 030006)

(计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)

(zhaoxw84@163.com)

Two-Stage Ensemble-Based Community Discovery Algorithm in Multilayer Networks

Zhao Xingwang, Zhang Yaopu, and Liang Jiye

(School of Computer and Information Technology, Shanxi University, Taiyuan 030006)

(Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan 030006)

Abstract Community discovery aims to uncover the community structure embedded in complex networks, which is one of the important tasks in complex network analysis. However, most of the existing community discovery methods are aimed at single-layer network data, and less research has been done on the widespread multi-layer networks in the real world. In order to solve the problem of community discovery in multi-layer networks, we propose a two-stage ensemble-based community discovery algorithm. The algorithm can improve the accuracy and interpretability of community discovery results. Firstly, after the base communities are obtained at each layer, the local ensemble is performed based on the base community structure information of each layer and the optimal base community structure of each other layer. Secondly, the stability of each community in the local community division of each layer is measured based on information entropy, and the accuracy of each local community division is evaluated through the results of other layer community division. Finally, a global weighted ensemble is carried out based on the importance of each community and the community structure to obtain the final community structure. A comparative analysis is carried out on artificial multi-layer networks and real multi-layer networks with existing multi-layer network community discovery algorithms. The experimental results show that our proposed algorithm is superior to the existing algorithms in terms of multi-layer modularity and normalized mutual information.

Key words multilayer network; community discovery; weighted ensemble; modularity; complex network analysis

摘要 社区发现旨在挖掘复杂网络蕴含的社区结构,是复杂网络分析的重要任务之一。然而,现有的社区发现方法大多针对单层网络数据,对现实世界中广泛存在的多层网络数据的研究较少。针对多层网络的社区发现问题,提出了一个基于 2 阶段集成的社区发现算法,以提高社区发现结果的准确性和可解释性。首先,在各层分别得到基社区划分;其次以各层社区划分结构信息为主并结合其他各层网络得到的基社区划分中最优的社区划分信息进行局部集成;再次,基于信息熵对各层局部社区划分中各个社区的稳定性进行度量,并通过其他层社区划分结果来对各个局部社区划分的准确性进行评价;最后,基于各个社区以及社区划分的重要性进行全局加权集成得到最终的社区划分结果。在人造多层网络和真实多层网络数据上与已有的多层网络社区发现算法进行了比较分析。实验结果表明,提出的算法在多层模块度、标准化

收稿日期: 2022-03-14; 修回日期: 2023-01-30

基金项目: 国家自然科学基金项目(62072293, U21A20473, 61976128, 62272285)

This work was supported by the National Natural Science Foundation of China (62072293, U21A20473, 61976128, 62272285).

通信作者: 梁吉业(lgy@sxu.edu.cn)

互信息等评价指标上优于已有算法。

关键词 多层网络；社区发现；加权集成；模块度；复杂网络分析

中图法分类号 TP391

社团结构是复杂网络的一个重要特征，即同一个社区内节点之间连接紧密而不同社区间节点之间连接稀疏^[1]。例如，社交网络的社区代表具有相似特征的人群，蛋白质交互网络中社区代表具有相似功能的生物组织模块，万维网中不同的社区代表不同功能的网页。社区发现是复杂网络分析挖掘中的重要任务之一，对于网络拓扑结构分析、功能分析和行为预测具有重要意义，已在社会学、生物信息学、交通系统等领域得到了广泛应用^[2-4]。例如，在社交网络中，社区发现有助于分析个体的行为模式、信息的传播方式和网络的变化趋势^[5]；在生物网络中，社区发现有助于分析属于共同功能模块的蛋白质之间的相互作用^[6]；在交通网络中，社区发现可以通过将城市道路网络划分为多个支线网络来实现对城市交通网络的区域控制，从而缓解城市交通拥堵问题^[7]。

针对不同的应用领域，研究者近年来已经开展了广泛研究，并提出了系列社区发现算法，主要包括：基于划分的方法^[8]、基于模块度的方法^[9-10]、基于谱的方法^[11]、基于动力学的方法^[12]、基于标签传播的方法^[13]等。已有算法大多适用于单层网络，然而现实世界中由多种类型节点及其连边关系组成的多层网络存在更为普遍^[14]。例如，社交网络中的个体之间可能存在不同类型的社交关系，如好友、关注、评论、转发等；在生物网络中，对于某些生物体来说，完整的蛋白质与蛋白质相互作用涉及数千种蛋白质分子之间多种不同的相互作用模式；在航空运输系统中，通过直飞航班对机场之间的连接建模，不同的商业航空公司可以被视为机场之间的不同连接模式。与单层网络相比，多层网络结构具有更加丰富的拓扑信息，有助于更为准确地探测网络中蕴含的社区结构。但是不同网络层蕴含的社区结构之间既存在一定的相关性又存在异质性，为社区发现任务带来了新的挑战。

近年来，一些学者针对多层网络的社区发现问题已经开展了有益探索，提出了相应的算法^[15-16]。其中，基于聚合的方法由于实现简单、具有良好的可扩展性等优点，得到了广泛关注。第1类基于聚合的方法为网络聚合，直接将多层网络合并为单层网络，然后利用传统社区发现算法实现社区划分。该类方法

虽然降低了社区发现后续研究的难度，但是在聚合过程中丢失了不同层网络结构的独特性，造成社区结果的不准确。第2类方法为基于集成学习的方法，首先在每层网络上分别利用传统社区发现算法进行社区划分，然后基于集成学习机制将每层的社区划分融合得到最终的社区结构。这类方法在不同网络层得到基社区划分后，往往各层社区划分分别进行融合或者不同层的社区划分进行统一融合，忽视了不同层社区之间的异质性；另一方面在融合过程中忽略了不同的基社区结构和社区划分之间的重要性，难以获得准确的社区结构^[17]。

针对上述问题，本文提出了一种基于2阶段集成的多层网络社区发现算法，旨在有效融合不同网络层获得的基社区划分之间的信息，提高社区发现结果的准确性和可解释性。在第1阶段，分别以各层网络生成的基社区划分为主，并结合其他各层网络得到的基社区划分中最优的社区划分结构信息进行局部集成；在第2阶段，首先对得到的局部社区划分及各个社区结构的重要性度量，然后进行全局加权集成得到最终的社区划分结果。

1 相关工作

1.1 基于扩展的方法

基于扩展的方法主要指将单层网络的社区发现方法直接扩展到多层网络，通常通过优化社区质量评估函数来实现。基于模块度函数优化是探测社区结构的经典方法，其核心思想是通过最大化模块度来获得社区划分结果。例如，Mucha等人^[18]将模块度扩展到了多层网络，开发了一种多层网络社区质量函数的广义框架用于检测多层网络的社区结构。该算法就是通过优化多层模块度扩展了传统的Louvain算法，从而得到一个多层模块度高的社区划分。Ma等人^[19]提出了一种用于多层网络社区检测的定量函数——多层模块化密度，解决了模块度的分辨率限制问题，为社区发现算法的设计提供了理论基础，并提出了一种半监督联合非负矩阵分解的多层网络社区发现算法S2-jNMF。该算法用贪婪搜索方法提取多层网络所有层中连接良好的子图作为先验信息，然

后将与多层网络相关的矩阵和先验信息联合分解为一个基矩阵和一个多系数矩阵用于社区发现。LART^[20]是一种基于局部自适应随机游走的多层网络社区发现算法,它首先对每一层网络运行不同的随机游走;然后利用每层转移概率获得节点之间的不同度量;最后,采用层次聚类方法生成社区划分,最终在多层模块度的最优值对应的水平上得到社区划分结果。Laishram 等人^[21]提出了一种新颖的基于采样的多层网络社区发现算法,使用来自探索成本低的层的信息来帮助探索成本高的层的信息。尽管该类方法通过对单层网络社区发现方法进行直接扩展可以对多层网络进行有效社区发现,但是存在着需要设置的参数较多、时间复杂度较高、易受噪声影响等缺点。

1.2 基于聚合的方法

为了降低社区检测的难度,基于多层网络的聚合结构进行社区发现的相关研究具有实现简单、可扩展性强的特点。这些研究主要包括基于网络聚合和基于集成学习 2 类思路。

基于网络聚合的方法是将一个多层网络压缩成单个网络,然后采用单层网络的社区发现算法对其进行社区检测。最简单的策略就是对邻接矩阵进行加权,其中 2 个节点之间的权值为它们在多层网络中相连的层数。Berlingerio 等人^[22]对简单的加权策略进行了修改并提出了基于共同邻居数的邻接矩阵的加权策略,这种策略不再考虑 2 个节点之间的直接联系,而是关注它们的邻居,共同邻居数越多则越可能在同一个社区。Zhu 等人^[23]进一步考虑了每层网络的重要性,将单层网络之间的相关性与单层网络的重要性密切联系起来,若某一单层网络与其他的单层网络之间的相关性更大,则表明其重要性更强。同时判断 2 个节点之间是否存在路径来衡量节点相似度,然后利用单层网络重要性和节点相似度构造统一矩阵,最后再用单层网络社区发现算法得到最终的社区划分结果。基于网络聚合方法在聚合过程中丢失了不同层网络结构的独特性,造成社区结果的不准确。

基于集成学习的方法首先利用传统算法对多层网络的各个层进行社区发现,然后基于集成学习机制将各层得到的社区划分结果进行融合得到最终的社区划分^[17]。Tang 等人^[24]提出了主模块化最大化方法,首先通过模块化最大化从多层网络的每个层中提取结构特征,然后对这些特征进行跨层次分析,采用 PCA 技术找到低维嵌入,使从所有层中提取的特征彼此高度相关,最后通过 k-means 聚类算法找到社

区划分。Berlingerio 等人^[25]提出了一种基于频繁模式挖掘的多层网络的社区发现算法,能够从单层社区成员中提取频繁的封闭项集来提取多层社区。首先用单层社区发现算法获得每层网络上的社区,然后用频繁封闭项集算法将至少在最少数量的层上属于同一社区的节点分配到同一社区。Tagarelli 等人^[17]提出了一种新的基于模块化优化的集成的多层社区发现算法,此方法在寻找共识社区结构时,不仅捕获了节点的原型社区成员,而且能够保存多层网络的拓扑信息,在获得共识解的过程中将多层网络的拓扑上界和拓扑下界作为搜索空间,共识函数也是通过基于模块化优化的方法,同时考虑了社区内和社区间的连通性。这类方法在得到不同层的社区划分后,往往各层社区划分分别进行融合或者不同层的社区划分进行统一融合。由于每一层网络中节点之间的交互方式以及社区结构的不同,进行简单集成容易忽视不同层社区之间的异质性;而且在融合过程中忽略了不同的基社区结构和社区划分之间的重要性,难以获得准确的社区结果。

2 基于集成的多层网络社区发现算法

本节首先对基于集成学习的多层网络社区发现问题进行了描述,接着对本文提出的基于 2 阶段集成的多层网络社区发现算法进行了介绍。

2.1 问题描述

一个多层网络可以定义为

$$\mathcal{G}_L = (G, \mathcal{A}), \quad (1)$$

其中 $G = \{G_\alpha | \alpha \in \{1, 2, \dots, L\}\}$ 表示 L 层网络构成的集合, $G_\alpha = (V_\alpha, E_\alpha)$ 为多层网络中的第 α 层, $V_\alpha \subseteq V = \{v_1, v_2, \dots, v_N\}$, V_α 表示第 α 层的节点集, V 为多层网络的所有节点的集合, $E_\alpha \subseteq V_\alpha \times V_\alpha$ 表示第 α 层内节点之间连接的边集; \mathcal{A} 表示任意 2 个不同层的节点之间的连接,可以表示为

$$\mathcal{A} = \{E_{\alpha\beta} \subseteq V_\alpha \times V_\beta | \alpha, \beta \in \{1, 2, \dots, L\}, \alpha \neq \beta\}, \quad (2)$$

即 E_α 中的元素表示层内顶点间连接, $E_{\alpha\beta}$ 表示层间顶点间的连接。本文主要讨论的对象为所有层共享相同节点集的多层网络,即在给定的多层网络 $\mathcal{G}_L = (G, \mathcal{A})$ 中,对于每一层 $G_\alpha = (V_\alpha, E_\alpha) \in \mathcal{G}$ 都有 $V_\alpha = V (V = \{v_1, v_2, \dots, v_N\})$,但是每一层内的边集不同,即 $G_\alpha = (V, E_\alpha)$ 。另外,对于层间连接

$$\mathcal{A} = \{E_{\alpha\beta} \subseteq V_\alpha \times V_\beta | \alpha, \beta \in \{1, 2, \dots, L\}, \alpha \neq \beta, V_\alpha = V_\beta\}, \quad (3)$$

即所有层的节点是对齐的。如图 1 为一个多层网络示

意图，由3层网络组成，每层均包括6个节点，每层内的连边表示节点间的一种关系，层间的虚线表示不同层的节点是对齐的。

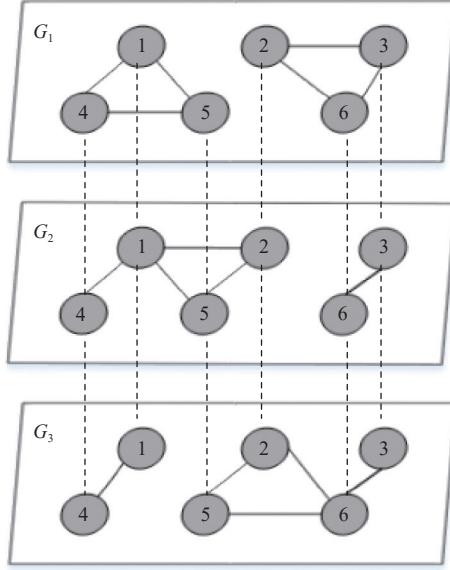


Fig. 1 An example of multi-layer network

图1 多层网络示例

给定一个 L 层的多层网络 $\mathcal{G}_L = (G, \mathcal{A})$ ，针对每层网络 $G_\alpha = (V, E_\alpha)$ 分别多次利用传统单层网络社区发现算法得到该层的社区结构为 $B_\alpha = \{B_\alpha^1, B_\alpha^2, \dots, B_\alpha^{M_\alpha}\}$ ，其中 $B_\alpha^{M_\alpha}$ 表示第 α 层网络共生成 M_α 个基社区划分，则多层网络 \mathcal{G}_L 的基社区结构表示为 $B = \{B_1, \dots, B_\alpha, \dots, B_L\}$ ，多层网络总的社区结构划分方案个数为 $M = \sum_{\alpha=1}^L M_\alpha$ 。基于集成的社区发现算法的目的是基于各层网络产生的基社区结构 $B = \{B_1, \dots, B_\alpha, \dots, B_L\}$ ，通过集成函数获得给定多层网络 $\mathcal{G}_L = (G, \mathcal{A})$ 的最终社区划分结果 $C^* = \{C_1, C_2, \dots, C_k\}$ ，且 $\cup_{i=1}^k C_i = V, C_i \cap C_j = \emptyset$ ，其中 k 表示最终社区个数。

2.2 基于2阶段集成的多层网络社区发现算法

本文算法的主要步骤如图2所示，主要包括基社区划分生成、局部集成、全局加权集成3个步骤。具体地：1)基社区划分生成是指在多层网络的每一层分别多次利用相同传统单层网络社区发现算法产生一组基社区划分；2)分别基于单层网络生成的基社区划分和其他各层网络得到的基社区划分中最优的社区划分信息对各层进行局部集成得到局部社区划分结果；3)分别对局部集成得到的社区划分及各个社区的重要性进行度量并进行全局加权集成得到最终社区结果。

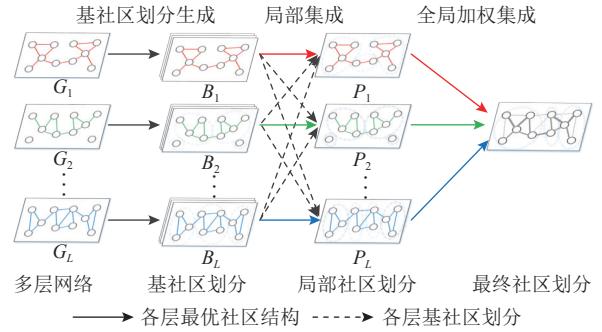


Fig. 2 Illustration of our proposed algorithm process

图2 本文算法流程示意图

2.2.1 基社区划分生成

该步骤的目的是为不同层的网络分别生成多组准确性高、差异性强的基社区划分。给定一个 L 层的多层网络 $\mathcal{G}_L = (G, \mathcal{A})$ ，在每层网络 $G_\alpha = (V, E_\alpha)$ 分别多次利用传统单层网络社区发现算法获得该层的社区划分 $B_\alpha = \{B_\alpha^1, B_\alpha^2, \dots, B_\alpha^{M_\alpha}\}$ ，其中 $B_\alpha^{M_\alpha}$ 表示第 α 层网络共生成 M_α 个基社区划分。为了生成准确性高、多样性强的基社区划分，本文采用经典的LM算法^[10]对每层网络分别进行社区发现。

2.2.2 局部集成

由于不同层产生的基社区划分之间既存在异质性，又存在一定的相关性。因此，在局部社区划分集成阶段各层结合其他各层的最优社区结构信息进行集成。本文中将各个层中最优的社区划分结果加入其他层，以此达到提升局部模型性能的目的。模块度是社区划分质量优劣的评价指标^[26]。模块度反映了网络社区结构内部连接的强弱，指的是社区内的边占所有边的比例减去在同样的社区结构下随机放置社区内部的边占所有边的比例的期望值，定义为

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(i, j), \quad (4)$$

其中 m 表示网络的总边数， A 表示网络的邻接矩阵， k_i 和 k_j 分别代表节点 v_i 和 v_j 的度，当节点 v_i 和 v_j 属于同一个社区时， $\delta(i, j) = 1$ ，否则， $\delta(i, j) = 0$ 。模块度越高，则社区划分质量越高。针对第 α 层网络获得的基社区划分 $B_\alpha = \{B_\alpha^1, B_\alpha^2, \dots, B_\alpha^{M_\alpha}\}$ ，可以通过计算各个划分下的模块度指标来选出第 α 层中最优的基社区划分。因此，在进行局部集成时，每一层的基社区划分由该层产生的 M_α 个基社区和其他 $L-1$ 层分别选出的最优社区划分组成。基于基社区划分信息，从局部角度每层分别构造节点之间的共协矩阵 A_α （ $1 \leq \alpha \leq L$ ），记为

$$A_\alpha(i, j) = \frac{n_{ij}}{M_\alpha + L - 1}, \quad (5)$$

其中 n_{ij} 表示在第 α 层中节点 v_i 和 v_j 在参与集成的基社区划分中被分配到同一个社区的数量, $M_\alpha + L - 1$ 是参与集成的社区划分的数量。基于共协矩阵, 利用层次聚类即可得到每层的局部社区划分结果 P_α ($1 \leq \alpha \leq L$), 则多层网络 G_L 的局部社区划分结果记为 $\mathcal{E} = \{P_1, P_2, \dots, P_L\}$ 。

2.2.3 全局加权集成

在集成过程中, 不同社区划分以及社区的优劣均会对最终社区的检测产生不同的影响。本节通过社区划分和社区质量进行评价来对集成过程进行加权。局部社区划分结果 E 中, 每层网络的 P_α 的优劣通过该划分方案运用到其他层网络上的模块度来衡量, 定义为

$$Q(P_\alpha, G_\beta) = \frac{1}{2m_\beta} \sum_{i,j} \left(A_{ij}^\beta - \frac{k_i^\beta k_j^\beta}{2m_\beta} \right) \delta_\alpha(i, j), \quad (6)$$

其中 m_β 表示第 β 层网络的总边数, A^β 表示第 β 层网络的邻接矩阵, k_i^β 和 k_j^β 分别代表节点 v_i 和 v_j 在第 β 层网络中的度, 当节点 v_i 和 v_j 在 P_α 属于同一个社区时, $\delta_\alpha(i, j) = 1$, 否则, $\delta_\alpha(i, j) = 0$ 。如果 $Q(P_\alpha, G_\beta)$ 越高, 则表示相对于第 β 层 P_α 的质量越优。因此, 在全局集成中第 α 层获得的 P_α 的权重定义为

$$W(P_\alpha) = \frac{1}{L-1} \sum_{\beta=1, \alpha \neq \beta}^L Q(P_\alpha, G_\beta). \quad (7)$$

然后, 对 $\mathcal{E} = \{P_1, P_2, \dots, P_L\}$ 中各层 P_α ($1 \leq \alpha \leq L$)的权重进行归一化, 则有

$$\overline{W}(P_\alpha) = \frac{W(P_\alpha)}{\sum_{\gamma=1}^L W(P_\gamma)}. \quad (8)$$

信息熵是信息论中用于度量系统不确定性的一种重要度量方法。如果一个社区结构相对其他划分越稳定, 则表明该社区的不确定性越小、可靠性越强, 该社区在集成过程中的权重越大。

局部社区划分结果 $\mathcal{E} = \{P_1, P_2, \dots, P_L\}$ 中第 α 层的社区划分记为 $P_\alpha = \{C_1^\alpha, C_2^\alpha, \dots, C_{n^\alpha}^\alpha\}$, 其中 C_i^α 表示 P_α 中的第 i 个社区, n^α 表示 P_α 中社区的数量。因此, \mathcal{E} 中所有社区可以表示为

$$C = \{C_1, C_2, \dots, C_{n_c}\}, \quad (9)$$

其中 C_i 表示第 i 个社区, n_c 表示 \mathcal{E} 中所有社区的总个数。

给定一个社区 $C_i \in C$ 和网络层 $G_\alpha \in \mathcal{E}$, 如果社区 C_i 不属于 P_α , 那么 C_i 中的节点可能属于 P_α 中的多个社区, C_i 相对于 P_α 的不确定性可以通过考虑 C_i 中的对象在 P_α 中各个社区中的分布来进行计算。那么, 社区 C_i 关于第 α 层网络 G_α 的 P_α 的不确定性可表示为

$$H^\alpha(C_i) = - \sum_{j=1}^{n^\alpha} p(C_i, C_j^\alpha) \ln p(C_i, C_j^\alpha), \quad (10)$$

其中

$$p(C_i, C_j^\alpha) = \frac{|C_i \cap C_j^\alpha|}{|C_i|}, \quad (11)$$

其中 n^α 表示社区划分 P_α 中的社区数量, C_j^α 表示 P_α 中的第 j 个社区, $|C_i \cap C_j^\alpha|$ 表示2个社区中共有节点的个数, $|C_i|$ 表示 C_i 中节点的个数。 $p(C_i, C_j^\alpha) \in [0, 1]$, 所以 $H^\alpha(C_i) \in [0, +\infty]$ 。

因此, C_i 关于多层网络 G_L 的局部社区划分 \mathcal{E} 的不确定性可以通过 C_i 关于 \mathcal{E} 中每层网络的社区划分结构的不确定性的和来表示, 即

$$H^\mathcal{E}(C_i) = \sum_{\alpha=1}^L H^\alpha(C_i). \quad (12)$$

基于信息熵来衡量每个社区关于局部社区划分 \mathcal{E} 的不确定性, 那么进一步可通过归一化来衡量 E 中每个社区的权重, 即

$$W(C_i) = \frac{e^{-H^\mathcal{E}(C_i)}}{\sum_{j=1}^{n_c} e^{-H^\mathcal{E}(C_j)}}. \quad (13)$$

由于 $H^\mathcal{E}(C_i) \in [0, +\infty]$, 所以 $W(C_i) \in [0, 1]$, 即一个社区的不确定越小, 则该社区的可靠性就越高, 在集成过程中的权重越大。

基于式(8)(13)给出的社区划分以及社区的权重, 从全局角度构造节点之间的加权共协矩阵, 表示为

$$D = (\tilde{a}_{ij})_{N \times N}, \quad (14)$$

其中

$$\tilde{a}_{ij} = \frac{1}{L} \times \sum_{\alpha=1}^L \overline{W}(P_\alpha) \times w_i^\alpha \times \delta_{ij}^\alpha, \quad (15)$$

$$w_i^\alpha = W(\text{Cls}^\alpha(v_i)), \quad (16)$$

$$\delta_{ij}^\alpha = \begin{cases} 1, & \text{if } \text{Cls}^\alpha(v_i) = \text{Cls}^\alpha(v_j), \\ 0, & \text{otherwise,} \end{cases} \quad (17)$$

其中 $\text{Cls}^\alpha(v_i)$ 表示节点 v_i 在 P_α 中所属社区的标签。

基于加权共协矩阵, 通过层次聚类获得多层网络 G_L 的最终社区划分结果 $C^* = \{C_1, C_2, \dots, C_k\}$ 。

基于以上对各阶段集成过程的描述, 本文提出的基于2阶段集成的多层网络社区发现算法描述如算法1。

算法1. 基于2阶段集成的多层网络社区发现算法。

输入: 多层网络 $G_L = (G, \mathcal{A})$, 社区划分个数 k , 各层网络社区基划分次数 M ;

输出: 多层网络 G_L 的最终社区划分结果 $C^* =$

$\{C_1, C_2, \dots, C_k\}$.

- ① 在每层网络上利用传统的 LM 社区发现算法
获得基社区划分结果, 记为 $B = \{B_1, B_2, \dots, B_L\}$;
- ② for $B_\alpha \in B$
- ③ $B_\alpha^b \leftarrow B_\alpha$; /*根据式(4)计算得到各层最好的社
区划分*/
- ④ $\{B_1, \dots, B_{\alpha-1}, B_{\alpha+1}, \dots, B_L\} \leftarrow B_\alpha^b$; /*将 B_α^b 添加到
 B 中除 B_α 外的其他集合中*/
- ⑤ end for
- ⑥ for $G_\alpha \in G$
- ⑦ $A_\alpha \leftarrow B_\alpha$; /*根据式(5)计算共协矩阵*/
- ⑧ $P_\alpha \leftarrow A_\alpha$; /*利用层次聚类获得局部社区划分*/
- ⑨ end for
- ⑩ for $P_\alpha \in \mathcal{E}$
- ⑪ $W(P_\alpha) \leftarrow P_\alpha$; /*根据式(7)计算局部社区划
分的权重*/
- ⑫ end for
- ⑬ $\bar{W}(P_\alpha) \leftarrow W(P_\alpha)$; /*根据式(8)对权重进行归一化*/
- ⑭ for $C_i \in \mathcal{E}$
- ⑮ $H^e(C_i) \leftarrow C_i$; /*根据式(12)计算社区不确定*/*
- ⑯ end for
- ⑰ $W(C_i) \leftarrow H^e(C_i)$; /*根据式(13)计算各个社区权
重*/
- ⑱ 根据式(14)构造加权共协矩阵, 并利用层次
聚类得到最终社区结构 $C^* = \{C_1, C_2, \dots, C_k\}$.

2.3 时间复杂度分析

本文提出的基于 2 阶段集成的多层网络社区发现算法主要包括 3 个步骤: 1) 基社区划分生成; 2) 局部集成; 3) 全局加权集成.

在步骤 1 中采用 LM 算法^[10], 假设一个单层网络的边的数量为 e , 则其时间复杂度为 $O(e)$, 所以步骤 1 的时间复杂度为 $O(LMe)$. 步骤 2 中主要包括构造共协矩阵, 因为 B_α 中的社区划分数量为 $M_\alpha + L - 1$, 因此其时间复杂度为 $O((M_\alpha + L - 1)N^2)$; 此外, 利用层次聚类算法获得局部社区划分结果的时间复杂度为 $O(N^2)$, 所以步骤 2 的时间复杂度为 $O(L(M_\alpha + L - 1)N^2 + LN^2)$. 步骤 3 通过信息熵来计算各个社区的稳定性, 假设步骤 2 得到的总社区数量为 n_C , 则计算社区不稳定性的
时间复杂度为 $O(n_C)$, 基于共协矩阵利用层次聚类计算最终社区划分结果的时间复杂度为 $O(N^2)$, 所以步骤 3 的时间复杂度为 $O(n_C + N^2)$. 因此, 算法的总时间复杂度为 $O(LMe + L(M_\alpha + L - 1)N^2 + LN^2 + n_C + N^2)$. 由于 L, M, n_C 均小于节点个数 N , 所以算法的总时间复杂度为 $O(N^2)$.

3 实验分析

为了验证本文算法的有效性, 与已有的多层网络社区发现算法在人工合成多层网络和 12 个真实多层网络数据集上进行了实验比较分析.

3.1 数据集

3.1.1 人工合成多层网络

为了生成人工合成多层网络数据, 本文采用了 Bazzi 等人^[27] 提出的算法. 算法参数设置为: 网络的层数 $L = 3, 5, 7$, 每层网络的规模 $N = 1000$ 或 $N = 5000$, 社区个数分别为 10 和 50, 其中每个节点都出现在每一层中, 即在这些网络中分别总共有 3000, 5000, 7000, 15000, 25000, 35000 个节点, 节点的度设置为其算法默认值, 即最小度为 3, 最大度为 150. 混合参数 $\mu \in [0, 1]$, μ 的值越接近 1, 生成网络的随机性越强. 参数 $\tilde{p} \in [0, 1]$ 表示层间相关系数, 当 $\tilde{p} = 0$ 时, 表示产生的多层网络各层之间社区划分是相互独立; 当 $\tilde{p} = 1$ 时, 所产生的多层网络各层之间的社区划分是相同的. 由于本文算法为通过集成策略产生统一的社区划分, 因此选择参数 $\tilde{p} = 1$.

3.1.2 真实多层网络

实验中采用的真实多层网络数据集的基本情况如表 1 所示. 所有真实网络数据的社区划分未知.

Table 1 Real Multilayer Networks

表 1 真实多层网络

| 数据集 | 层数 | 顶点数 | 边数 |
|-------------|----|-------|-------|
| AUCS | 5 | 61 | 620 |
| EUAir | 37 | 450 | 3 588 |
| Pierreauger | 16 | 514 | 7 153 |
| CKM | 3 | 246 | 1 551 |
| Kapferer | 4 | 39 | 1 018 |
| Gallus | 6 | 313 | 388 |
| Herpes4 | 4 | 216 | 259 |
| Hiv1 | 5 | 1 005 | 1 355 |
| Plasmodium | 3 | 1 023 | 2 521 |
| Xenopus | 5 | 461 | 620 |
| Bos | 4 | 321 | 325 |
| Candida | 7 | 367 | 397 |

AUCS 数据集描述某大学研究部门的 61 名员工节点(包括教授、博士后、博士生和行政人员)之间的 Facebook 关系、午餐关系、合作关系、休闲关系和

工作关系^[28]. EUAir 数据集是欧洲航空交通网络, 每层网络对应在欧洲运营的不同航空公司^[29]. Pierreauger 数据集表示天文台科学家之间的协作关系, 每一层代表一种协作任务^[30]. CKM 数据集由美国 4 个城镇的 246 名医生组成, 主要讨论网络关系对医生用药决策的影响, 包括通常找谁获取用药建议、经常与谁讨论和与谁关系最好 3 层网络^[30]. Kapferer 网络中节点为裁缝店的工人, 关系为他们之间的工作和友谊互动^[31].

另外, 在包含相互作用数据集的生物通用存储库中选取了 Gallus^[32], Herpes^[33], Hiv1^[33], Plasmodium^[32], Xenopus^[33], Bos^[32], Candida^[32] 共 7 个多层次网络数据集. 这些网络代表蛋白质之间的相互作用, 不同的层对应于不同性质的相互作用, 即直接作用、物理关联、共定位、关联、抑制、增强和合成遗传相互作用. 每种生物体所对应的层数为 3~7 不等.

3.2 评价指标

当数据集的真实社区结果未知时, 采用 Mucha 等人^[18] 提出的多层次模块度 Q_M 来评价多层次网络的社区质量, 定义为

$$Q_M = \frac{1}{2\mu} \sum_{i,j,\alpha,\beta} \left\{ \left(A_{ija} - \gamma \frac{k_{ia}k_{ja}}{2m_\alpha} \right) \delta_{\alpha\beta} + \delta_{ij} C_{j\alpha\beta} \right\} \delta(g_{ia}, g_{j\beta}), \quad (18)$$

其中 μ 表示多层次网络中的连接的数量, 分辨率 γ 以控制每层网络的社区规模和数量. A_{ija} 表示网络第 α 层的邻接矩阵节点 v_i 和节点 v_j 之间的连接关系, $\delta_{\alpha\beta}$ 在第 α 层和第 β 层为同一层时为 1, 否则为 0; δ_{ij} 在节点 v_i 和节点 v_j 为同层节点时为 1, 否则为 0. $C_{j\alpha\beta}$ 表示 α 层和 β 层之间与节点 v_j 的层间连边数, $k_{ia} = \sum_{i,j} A_{ija}$ 表示节点 v_i 在 α 层上的度, $m_\alpha = \sum_i k_{ia}$ 表示 α 层的总边数. 如果节点 v_i 和节点 v_j 的社区分布相同, 则 $\delta(g_{ia}, g_{j\beta}) = 1$, 否则为 0.

对于有真实社区划分的数据集, 利用标准化互信息 (NMI)^[34] 来评价社区划分结果的质量, NMI 定义为

$$NMI(X, Y) = \frac{-2 \sum_{i=1}^{c_X} \sum_{j=1}^{c_Y} N_{ij} \text{lb} \frac{N_{ij}\tilde{N}}{N_i N_j}}{\sum_{i=1}^{c_X} N_i \text{lb} \frac{N_i}{N} + \sum_{j=1}^{c_Y} N_j \text{lb} \frac{N_j}{N}}, \quad (19)$$

其中 X 和 Y 分别表示数据集的真实社区划分结果和算法得到的社区划分结果, c_X 和 c_Y 分别表示 X 和 Y 中的社区数量, \tilde{N} 是一个混淆矩阵, 其元素 N_{ij} 表示在 X 的社区 i 与 Y 的社区 j 中共同出现的节点数量, N_i 表

示 \tilde{N} 的第 i 行的元素的和, N_j 表示 \tilde{N} 的第 j 列的元素的和, N 表示节点数量. $NMI(X, Y)$ 的范围为 [0,1], 如果 X 和 Y 完全相同, 则 $NMI(X, Y) = 1$; 如果 X 和 Y 完全不同, 则 $NMI(X, Y) = 0$.

3.3 对比方法及参数设置

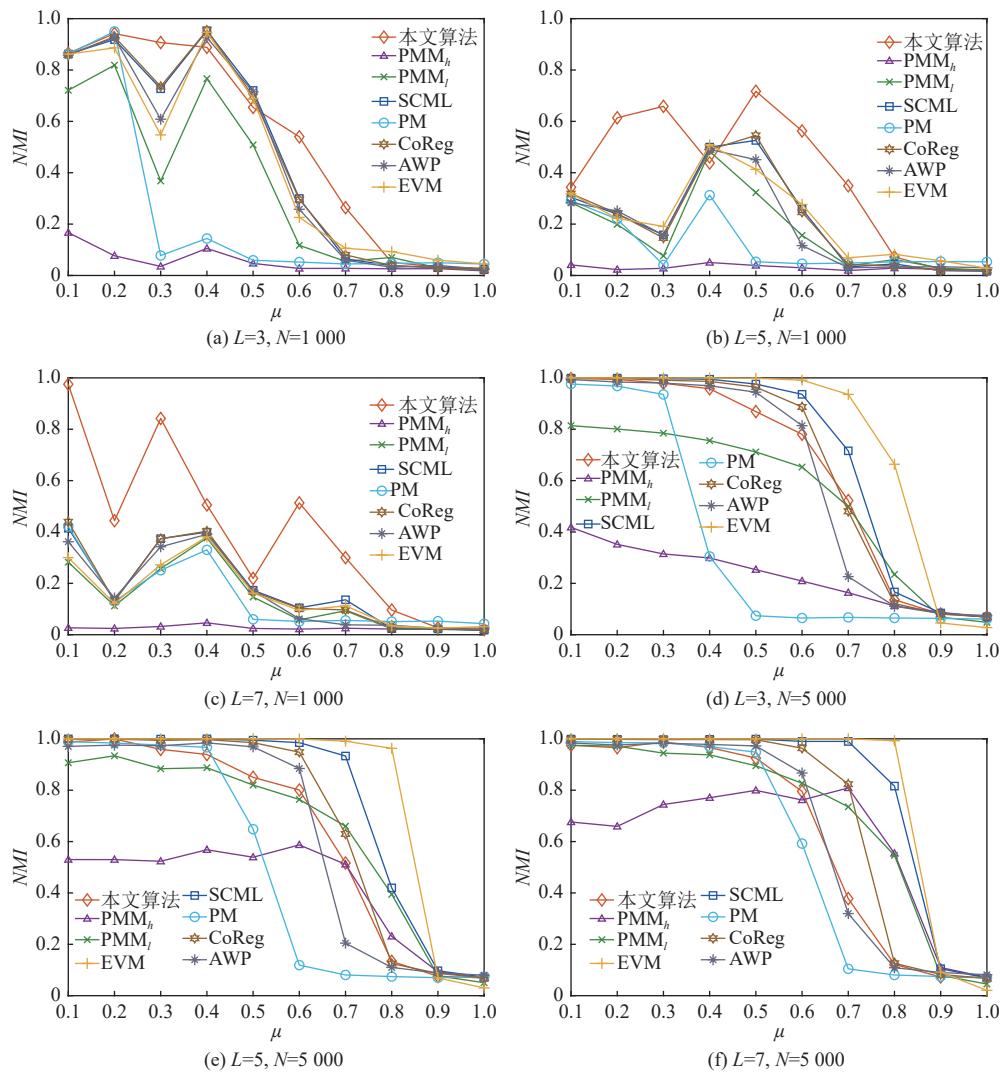
本文算法与已有多层网络社区发现基线方法进行了比较, 代表性方法包括 PMM^[24], SCML^[35], PM^[36], CoReg^[37], AWP^[38], EVM^[39].

根据已有文献实验分析中的建议, 对基线方法的参数进行了设置. 其中, 本文方法与基线方法 PMM, SCML, PM, CoReg, AWP 共同的参数为社区个数 k , 为了比较的公平性, 对于社区已知的人造网络个数 k 均设置为真实的社区个数, 对于社区未知的真实网络将社区个数 k 设置为 $k \in [2, \sqrt{N}]$, 其中 N 表示节点个数, 各个算法分别在不同的 k 值下运行, 并从结果中选取最优值进行比较. PMM 算法中结构特征的数量可以是 $1 \sim (N-2)$ (N 为节点总数) 之间的任意数字, 根据建议, 对于人工多层次网络和真实多层次网络均设置了 2 个值, 人工多层次网络比较低的常数为 50, 真实多层次网络比较低的常数为 10, 用 PMM_l 来表示; 另一个是根据节点总数来设定的值, 2 类数据集上都设置为 $N/2$, 用 PMM_h 来表示, k-means 算法执行的次数设置为 10. SCML 算法中正则化参数 $\lambda = 0.5$; PM 算法中参数 $p = -10$; CoReg 算法正则化参数 $\lambda = 0.01$; EVM 算法正则化参数 $\gamma = 0.5$. 本文算法中各层网络生成的基社区划分个数 $M = 10$.

3.4 实验结果分析

3.4.1 人工网络实验结果分析

在不同的参数下生成了 60 个不同的人工合成网络, 针对各个网络, 对比算法均在不同参数设置下运行 10 次, 然后取结果的平均值进行比较. 各个算法在不同人工网络的 NMI 值随网络参数 μ 的变化如图 3 所示. 随着 μ 的逐渐增大, 不同算法的性能均出现了下降趋势. 这是因为随着 μ 的增大网络的随机性变强, 网络蕴含的社区结构复杂导致社区发现算法识别正确社区的难度加大. 当多层次网络的节点数量 $N = 1000$ 时, PMM_h 产生了较差的结果, 且在所有网络中, PMM_l 和 PMM_h 产生的结果有着较大的差异, 说明不同的结构特征数量对 PMM 产生了较大的影响, 而其取值为 $1 \sim (N-2)$, 特别对于规模较大的网络, 选取合适的结构特征数量不容易. 其他算法的结果都随着 μ 的变化出现了较大的波动, 从图 3 可以看出本文算法在大部分情况下优于其他算法. 当多层次网络的节点数量 $N = 5000$ 时, 基本上所有算法的结果随着 μ

Fig. 3 Comparison results of NMI on synthetic networks图3 不同算法在合成网络上的 NMI 比较结果

的增大而减小。例如，PM 算法的结果在 $L=3, 5, 7$ 的情况下分别在 $\mu=0.4, 0.5, 0.6$ 的时候明显下降，PMM 算法对于不同的提取的结构特征数量值，结果产生了较大的差异；AWP 算法的结果在 3 种不同层数的网络中 $\mu<0.7$ 时均表现较好，但在 $\mu=0.7$ 出现了明显下降；SCML 算法表现较好，这是因为 SCML 算法通过将包含在单个图层中的信息转换为 Grassmann 流形上的子空间并利用 Grassmann 流形上的距离分析找到一个具有代表性的子空间，最后在这个代表性子空间上进行聚类。但是，SCML 的聚类结果容易受到与代表性子空间距离更近的单个图层子空间上聚类结果的影响，当所有单个图层都包含丰富的信息时，SCML 将产生更好的结果，若与代表性子空间距离更近的单个图层上的聚类质量较差时也会导致 SCML 产生较差的结果。下面我们将选取 $L=3, N=5000, \mu=0.5$ 和 $L=3, N=1000, \mu=0.6$ 这 2 种不同情况的人工

网络来说明 SCML 算法为何在 $N=5000$ 时取得较好的结果。如表 2 所示，在 $L=3, N=5000, \mu=0.5$ 上代表性子空间与第 2 层的子空间距离最近，且从表 2 中 SCML 算法在单层网络上得到的结果和 NMI 值，可以看出各层网络均包含较为丰富的信息，所以 SCML 得到了更好的结果。如表 3 所示，在 $L=3, N=1000, \mu=0.6$ 上代表性子空间与第 2 层的子空间距离最近，与表 2 不同的是该网络的第 2 层得到了低质量的聚类结果，SCML 产生了较差的结果。说明当多层网络的各层网络中都包含较为丰富的信息时，SCML 产生更好的结果，而当各层网络中存在信息较少且该层网络的子空间与 SCML 计算所得的代表性子空间距离即使较近时，SCML 也会忽略信息较多的层，从而导致最终产生较差的结果。基于以上 2 个人工网络对本文提出的方法进行分析，在 $L=3, N=5000, \mu=0.5$ 上，本文算法的 NMI 分别为 0.8021, 0.8131, 0.8023。可

以看出,该网络中各层都包含了较为丰富的信息,通过式(14)进行全局加权集成后,得到的 $NMI = 0.8501$;在 $L = 3, N = 1000, \mu = 0.6$ 上,本文算法的 NMI 分别为 0.5440, 0.2247, 0.0247. 可以看出该网络中各层之间的差异性较大.

Table 2 Result Analysis of SCML on Network with $L=3, N=5\,000, \mu=0.5$

表 2 SCML 在 $L=3, N=5\,000, \mu=0.5$ 网络上的结果分析

| 网络 | 代表性子空间与各层子空间的距离 | NMI |
|-----------------|-----------------|--------|
| G_1 | 0.4316 | 0.6944 |
| G_2 | 0.4072 | 0.7033 |
| G_3 | 4.2899 | 0.6947 |
| \mathcal{G}_3 | | 0.9760 |

Table 3 Result Analysis of SCML on Network with $L=3, N=1\,000, \mu=0.6$

表 3 SCML 在 $L=3, N=1\,000, \mu=0.6$ 网络上的结果分析

| 网络 | 代表性子空间与各层子空间的距离 | NMI |
|-----------------|-----------------|--------|
| G_1 | 2.4602 | 0.4528 |
| G_2 | 2.0946 | 0.192 |
| G_3 | 2.5552 | 0.0192 |
| \mathcal{G}_3 | | 0.2996 |

通过式(14)进行全局加权集成后,得到的最终结果的 $NMI=0.5489$. 通过对比表 2、表 3 与本文算法得到的 NMI 值可以看出 $L=3, N=5\,000, \mu=0.5$ 中各层网络包含信息较为平均,而 $L=3, N=1\,000, \mu=0.6$ 中各层网络包含信息的差异较大,所以当 $N=5\,000$ 时,由于

单层网络之间差异较小,且随着 μ 的增大,各层网络社区结构难以识别,导致各层网络上得到的基社区划分较差,从而导致经过集成产生的社区质量也较低;当 $\mu < 0.7$ 时,本文算法在 3 种不同层数的网络上的实验结果变化较小,说明本文算法的鲁棒性较强. CoReg 和 EVM 同样在 $N=5\,000$, 即单层网络差异较小的人工网络上表现出了良好的性能,而在单层网络差异较大的人工网络上表现较差.

3.4.2 真实网络实验结果分析

不同算法在真实网络的实验结果如表 4 所示,其中每个算法在不同参数设置下均运行 10 次,从不同社区个数得到的结果中选取最优模块度值进行比较. 对不同算法在每个数据集上社区发现结果的最优结果和次优结果进行加粗标注. 从表 4 可以看出,本文算法在除 AUCS 之外的 11 个数据集上都取得了最优结果或次优的结果,其中在 EUAir, Gallus, Herpes4, Hiv1, Plasmodium, Xenopus, Bos 共 7 个数据集上本文算法的结果明显优于次优的结果. 总之,本文算法在真实数据集上表现出了较为良好的性能.

3.4.3 鲁棒性分析

为了验证本文算法的有效性受基社区划分数 M 变化的影响,对采用 $\mu = 0.3$ 时不同规模的 6 个人工多层网络和部分层数较少的真实多层网络进行了鲁棒性测试. 其中, $M \in \{10, 20, 30, 40, 50\}$, 对于不同的 M 值,取本文算法在每个网络上执行 10 次的结果的平均值进行比较,实验结果如图 4 和图 5 所示.

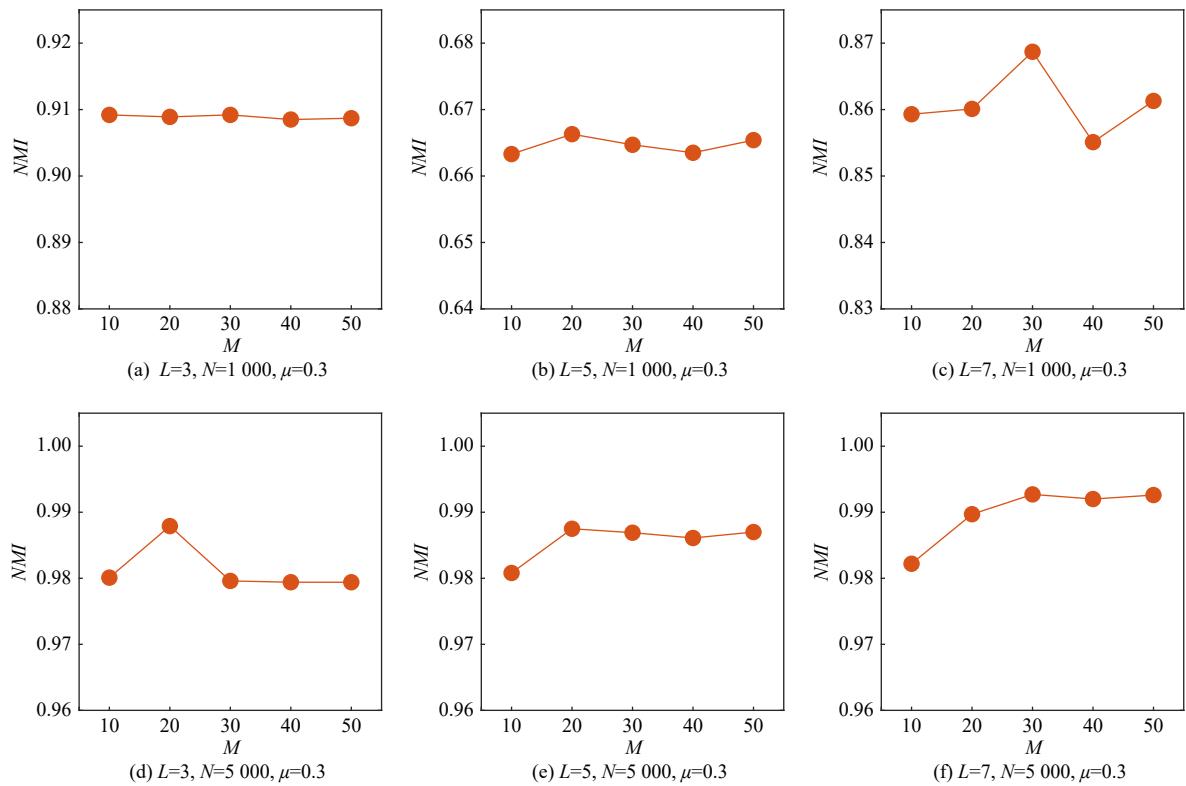
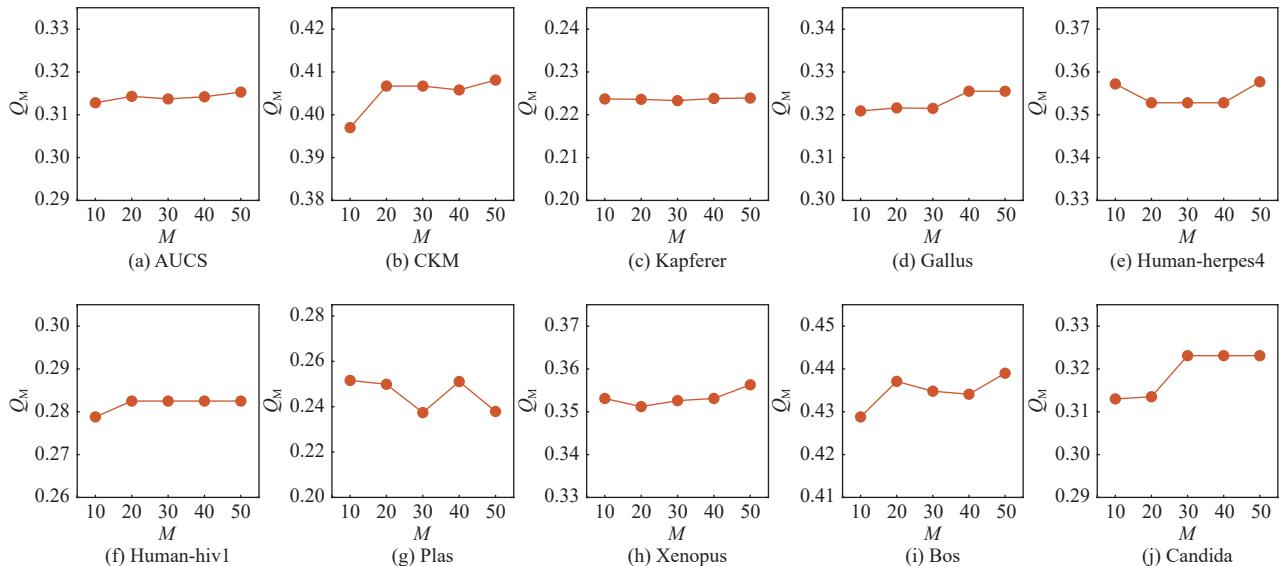
在人工多层网络数据集中,采用 NMI 指标对社

Table 4 Comparison of Experimental Results on Real Networks in Terms of Q_M

表 4 真实网络实验结果的 Q_M 比较

| 数据集 | PMM _h | PMM _l | SCML | PM | CoReg | AWP | EVM | 本文算法 |
|-------------|------------------|------------------|---------------|---------------|---------------|---------------|--------|---------------|
| AUCS | 0.2668 | 0.2922 | 0.3193 | 0.3103 | 0.3174 | 0.3095 | 0.3079 | 0.3143 |
| EUAir | 0.1214 | 0.1272 | 0.1776 | 0.2403 | 0.1681 | 0.2209 | 0.1646 | 0.2439 |
| Pierreauger | 0.0897 | 0.1220 | 0.1224 | 0.1257 | 0.1314 | 0.1056 | 0.1183 | 0.1278 |
| CKM | 0.2964 | 0.3508 | 0.4191 | 0.3274 | 0.3269 | 0.2809 | 0.3316 | 0.4067 |
| Kapferer | 0.2013 | 0.2212 | 0.2146 | 0.2417 | 0.2196 | 0.2008 | 0.1817 | 0.2236 |
| Gallus | 0.1117 | 0.1580 | 0.1772 | 0.2488 | 0.3026 | 0.2379 | 0.2488 | 0.3059 |
| Herpes4 | 0.1254 | 0.2157 | 0.0938 | 0.2236 | 0.3393 | 0.2522 | 0.1171 | 0.3572 |
| Hiv1 | 0.0813 | 0.2174 | 0.2295 | 0.2510 | 0.2170 | 0.1715 | 0.1623 | 0.2825 |
| Plasmodium | 0.0703 | 0.1484 | 0.2488 | 0.2495 | 0.2293 | 0.1663 | 0.1102 | 0.2516 |
| Xenopus | 0.1642 | 0.2179 | 0.2112 | 0.2559 | 0.2852 | 0.2874 | 0.2544 | 0.3136 |
| Bos | 0.1283 | 0.1979 | 0.1478 | 0.2493 | 0.3596 | 0.3224 | 0.3362 | 0.4173 |
| Candida | 0.0193 | 0.1130 | 0.1867 | 0.2643 | 0.3484 | 0.2758 | 0.2111 | 0.3135 |

注: 加粗数字表示最优和次优结果.

Fig. 4 NMI of the proposed algorithm changes with varying ensemble number M 图4 本文算法的 NMI 随基社区划分数 M 的变化Fig. 5 Q_M of the proposed algorithm changes with varying ensemble number M 图5 本文算法的 Q_M 随基社区划分数 M 的变化

区划分结果进行评价,统计在不同 M 下 NMI 的变化情况。由图4可知,当 $L=7, N=1\,000, \mu=0.3$ 时本文算法的 NMI 值出现了最大的波动(0.013),其次当 $L=7, N=5\,000, \mu=0.3$ 时 NMI 值为 0.01,在其他人工网络数据集上 NMI 值的波动范围均在 0.01 以内。在真实多层网络数据集中,通过多层模块度指标 Q_M 对各个

结果的变化范围进行分析。在 Plas 数据集上,本文算法的结果在不同的 M 值下 Q_M 值的波动为 0.015,在 CKM 和 Candida 数据上的 Q_M 值的波动范围均为 0.011,而在其他数据集上不同的 M 值下 Q_M 值变化范围均小于 0.01。结果表明,本文算法的性能受基各层社区划分个数 M 的影响较小,具有良好的鲁棒性。

4 总 结

近年来,有效融合多层网络信息并对其进行社区发现成为了一个重要的研究内容。为此,本文提出了一种基于2阶段集成的多层网络社区发现算法。首先,第1阶段在各层进行局部集成时结合了其他层的最优社区划分信息,有效利用了不同网络层社区结构的相关性;其次,第2阶段中基于其他网络层的社区结构信息对各层融合得到的社区划分及社区结构的重要性进行评价加权,再进行全局加权集成,综合考虑了各层局部社区划分结果对集成的影响;最后,通过在人工生成多层网络和真实多层网络上进行实验分析,对本文算法的有效性和鲁棒性进行了验证。然而,本文提出的算法面临时间复杂度较高的问题,在未来的工作中将考虑如何提高算法的计算效率来应对大规模多层网络社区发现面临的挑战。

作者贡献声明:赵兴旺提出论文算法思想、设计实验方案并修订论文;张珧溥完成实验并撰写论文;梁吉业提出算法的指导意见和审核论文。

参 考 文 献

- [1] Fortunato S, Hric D. Community detection in networks: A user guide[J]. *Physics Reports*, 2016, 659: 1–44
- [2] Girvan M, Newman M E J. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Sciences*, 2002, 99(12): 7821–7826
- [3] Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete interactome[J]. *Science*, 2015, 347(6224): 1257601
- [4] Wang Xiaoyang, Zhang Ying, Zhang Wenjie, et al. Efficient distance-aware influence maximization in geo-social networks[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 29(3): 599–612
- [5] Zhang Fan, Zhang Wenjie, Zhang Ying, et al. OLAK: An efficient algorithm to prevent unraveling in social networks[J]. *Proceedings of the VLDB Endowment*, 2017, 10(6): 649–660
- [6] Rahman M S, Ngom A. A fast agglomerative community detection method for protein complex discovery in protein interaction networks[C/OL] //Proc of the 8th IAPR Int Conf on Pattern Recognition in Bioinformatics. Berlin: Springer, 2013[2022-03-14].https://link.springer.com/chapter/10.1007/978-3-642-39159-0_1
- [7] Guo Lu, Cui Ying, Liang Haili, et al. Spectral bisection community detection method for urban road networks[C] //Proc of the 40th Chinese Control Conf (CCC). Piscataway, NJ: IEEE, 2021: 806–811
- [8] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. *Proceedings of the National Academy of Sciences*, 2004, 101(9): 2658–2663
- [9] Newman M E J. Fast algorithm for detecting community structure in networks[J]. *Physical Review E*, 2004, 69(6): 066133
- [10] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 2008(10): P10008
- [11] Newman M E J. Spectral methods for community detection and graph partitioning[J]. *Physical Review E*, 2013, 88(4): 042822
- [12] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. *Proceedings of the National Academy of Sciences*, 2008, 105(4): 1118–1123
- [13] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. *Physical Review E*, 2007, 76(3): 036106
- [14] Wu Zongning, Di Zengru, Fan Ying. The structure and function of multilayer networks: Progress and prospects[J]. *Journal of University of Electronic Science and Technology of China*, 2021, 50(1): 106–120 (in Chinese)
(吴宗柠, 狄增如, 樊瑛. 多层网络的结构与功能研究进展[J]. 电子科技大学学报, 2021, 50(1): 106–120)
- [15] Chen Kejia, Chen Liming, Wu Tong. Survey on community detection in multi-layer networks[J]. *Journal of Frontiers of Computer Science and Technology*, 2020, 14(11): 1801–1812 (in Chinese)
(陈可佳, 陈利明, 吴桐. 多层网络社区发现研究综述[J]. 计算机科学与探索, 2020, 14(11): 1801–1812)
- [16] Huang Xinyu, Chen Dongming, Ren Tao, et al. A survey of community detection methods in multilayer networks[J]. *Data Mining and Knowledge Discovery*, 2021, 35(1): 1–45
- [17] Tagarelli A, Amelio A, Gullo F. Ensemble-based community detection in multilayer networks[J]. *Data Mining and Knowledge Discovery*, 2017, 31(5): 1506–1543
- [18] Mucha P J, Richardson T, Macon K, et al. Community structure in time-dependent, multiscale, and multiplex networks[J]. *Science*, 2010, 328(5980): 876–878
- [19] Ma Xiaoke, Dong Di, Wang Quan. Community detection in multi-layer networks using joint nonnegative matrix factorization[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2018, 31(2): 273–286
- [20] Kuncheva Z, Montana G. Community detection in multiplex networks using locally adaptive random walks [C] //Proc of the 2015 IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining. New York: ACM, 2015: 1308–1315
- [21] Laishram R, Wendt J D, Soundarajan S. MCS+: An efficient algorithm for crawling the community structure in multiplex networks[J]. *ACM Transactions on Knowledge Discovery from Data*, 2021, 16(1): 1–32
- [22] Berlingero M, Coscia M, Giannotti F. Finding and characterizing communities in multidimensional networks[C] //Proc of the 2011 Int Conf on Advances in Social Networks Analysis & Mining.

- Piscataway, NJ: IEEE, 2011: 490–494
- [23] Zhu Guangyao, Li Kan. A unified model for community detection of multiplex networks[C] //Proc of the 15th Int Conf on Web Information Systems Engineering. Berlin: Springer, 2014: 31–46
- [24] Tang Lei, Wang Xufei, Liu Huan. Uncovering groups via heterogeneous interaction analysis[C] //Proc of the 9th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2009: 503–512
- [25] Berlingerio M, Pinelli F, Calabrese F. Abacus: Frequent pattern mining-based community discovery in multidimensional networks[J]. *Data Mining and Knowledge Discovery*, 2013, 27(3): 294–320
- [26] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. *Physical Review E*, 2004, 69(2): 026113
- [27] Bazzi M, Jeub L G S, Arenas A, et al. A framework for the construction of generative models for mesoscale structure in multilayer networks[J]. *Physical Review Research*, 2020, 2(2): 023100
- [28] Magnani M, Micenková B, Rossi L. Combinatorial analysis of multiple networks [J]. arXiv preprint, arXiv: 1303.4986, 2013
- [29] Cardillo A, Gómez-Gardenes J, Zanin M, et al. Emergence of network features from multiplexity[J]. *Scientific Reports*, 2013, 3(1): 1–6
- [30] Coleman J, Katz E, Menzel H. The diffusion of an innovation among physicians[J]. *Sociometry*, 1957, 20(4): 253–270
- [31] Kapferer B. Strategy and Transaction in an African Factory: African Workers and Indian Management in a Zambian Town [M]. Manchester: Manchester University Press, 1972
- [32] De Domenico M, Nicosia V, Arenas A, et al. Structural reducibility of multilayer networks[J]. *Nature Communications*, 2015, 6(1): 1–9
- [33] De Domenico M, Porter M A, Arenas A. MuxViz: A tool for multilayer analysis and visualization of networks[J]. *Journal of Complex Networks*, 2015, 3(2): 159–176
- [34] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification[J]. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, 2005(9): P09008
- [35] Dong Xiaowen, Frossard P, Vandergheynst P, et al. Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds[J]. *IEEE Transactions on Signal Processing*, 2013, 62(4): 905–918
- [36] Mercado P, Gautier A, Tudisco F, et al. The power mean Laplacian for multilayer graph clustering[C] //Proc of the 21st Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2018: 1828–1838
- [37] Kumar A, Rai P, Daume H. Co-regularized multi-view spectral clustering[C] //Proc of the 25th Int Conf on Neural Information Processing Systems. La Jolla, CA: NIPS, 2011: 1413–1421
- [38] Nie Feiping, Tian Lai, Li Xuelong. Multiview clustering via adaptively weighted procrustes[C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 2022–2030
- [39] Venturini S, Cristofari A, Rinaldi F, et al. Louvain-like methods for community detection in multiplex networks[J]. arXiv preprint, arXiv: 2106.13543, 2021



Zhao Xingwang, born in 1984. PhD, associate professor, PhD supervisor. Member of CCF. His main research interests include data mining and machine learning.

赵兴旺, 1984 年生. 博士, 副教授, 博士生导师. CCF 会员. 主要研究方向为数据挖掘、机器学习.



Zhang Yaopu, born in 1997. Master. Student member of CCF. His main research interest includes network data mining.

张姚溥, 1997 年生. 硕士. CCF 学生会员. 主要研究方向为网络数据挖掘.



Liang Jiye, born in 1962. PhD, professor, PhD supervisor. Fellow of CCF. His main research interests include artificial intelligence, machine learning, and data mining.

梁吉业, 1962 年生. 博士, 教授, 博士生导师. CCF 会士. 主要研究方向为人工智能、机器学习、数据挖掘.