

基于SCL-CMM模型的超网络链路预测方法

任玉媛, 马宏, 刘树新, 王凯
(信息工程大学, 河南 郑州 450001)

摘要: 图为建模现实系统的内在交互提供了一种有效的方式, 但却无法显示捕获的多个实体之间广泛存在的高阶异质性, 超图则可以很好地突破低阶关系的限制。超网络的链路预测就是根据观测到的超图结构来预测未知的超链路, 因其可以充分地刻画复杂系统的关联模式而成为网络科学中的热点问题。现有的方法通常针对整个拓扑结构设计推理模型, 忽略了网络中隐含的聚集特性, 导致预测的超链路类别不全面。针对上述问题, 提出了基于超图谱聚类解析器的协调矩阵最小化 (coordination matrix minimization based on hyper graph spectral clustering parser, SCL-CMM) 模型的超网络链路预测方法。该方法将高阶超网络映射到具有一定语义的异质超图上, 然后利用谱聚类解析器来提取超链路的结构特征, 将原始超图重构为多个同质子图, 进而在子图的观测空间而不是整个网络的邻接空间推断潜在超链路的分布情况, 还原完整的超网络结构。该方法联合学习超网络的结构特征与集聚属性来建模各个子图的高阶非线性行为, 解决了异构超图链路预测类别单一、精度低的问题。在9个真实数据集上进行了大量的对比实验表明, 该方法在AUC (area under curve) 评分和召回率方面都显著优于现有方法。

关键词: 链路预测; 超图; 超网络; 拓扑结构; 聚类

中图分类号: TP301.6

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024039

Hypernetwork link prediction method based on the SCL-CMM model

REN Yuyuan, MA Hong, LIU Shuxin, WANG Kai
Information Engineering University, Zhengzhou 450001, China

Abstract: An effective method for the internal interaction within modeling reality systems is provided by graphs; however, they have been unable to effectively display and capture the high-order heterogeneity that widely exists between multiple entities. Hypergraphs have been recognized for their ability to surpass the limitations imposed by low-order relationships. Hypernetwork link prediction, which involves predicting unknown hyperlinks based on the observed hypergraph structure, has increasingly become a hot topic in network science due to its capacity to fully

收稿日期: 2023-01-04; 修回日期: 2023-06-03

通信作者: 任玉媛, 1697562142@qq.com

基金项目: 中原英才计划项目 (6212101510002)

Foundation Item: Zhongyuan Talent Program (6212101510002)

引用格式: 任玉媛, 马宏, 刘树新, 等. 基于SCL-CMM模型的超网络链路预测方法[J]. 网络与信息安全学报, 2024, 10(3): 52-65.

Citation Format: REN Y Y, MA H, LIU S X, et al. Hypernetwork link prediction method based on the SCL-CMM model[J]. Chinese Journal of Network and Information Security, 2024, 10(3): 52-65.

describe the association patterns of complex systems. Existing methods typically design reasoning models for the entire topology, often overlooking the implicit aggregation characteristics within the network, which leads to an incomplete prediction of hyperlink categories. To address these issues, a coordination matrix minimization model based on hypergraph spectral clustering parser (SCL-CMM) was proposed. Initially, higher-order hypernetworks were mapped into heterogeneous hypergraphs with certain semantics. Subsequently, the spectral clustering parser was employed to extract the structural features of hyperlinks. The original hypergraph was reconstructed into multiple homoprotonic graphs, and the distribution of potential hyperlinks was inferred within the observation space of the subgraph, rather than the entire adjacency space, in order to restore the complete hypernetwork structure. This method federated learned the structural characteristics and aggregation attributes of hypernetworks to model the high-order nonlinear behavior of each subgraph, thereby solving the problems of single category and low precision in heterogeneous hypergraphs link prediction. Extensive comparative experiments were conducted on nine real datasets, demonstrating that this method significantly outperformed existing methods in terms of AUC score and recall rate.

Keywords: link prediction, hypergraph, hypernetwork, topology, clustering

0 引言

从科学合作网络到关联网络,从通信网络到 Facebook、微博等社交网络,现实世界中的许多网络可以抽象为复杂网络,用节点表示实体,连边表示实体间的交互^[1]。复杂网络的链路预测是指根据网络已知的信息发现或预测未知或未来可能存在的链路^[2],在社会网络分析^[3-5]、推荐系统^[6-8]、生物信息学^[9-11]等领域具有广泛的实际应用价值,同时推动了网络演化与网络动力学等相关理论研究,受到了多个学科的关注。

传统的链路预测忽略了多个节点之间的交互,它们只关注成对的节点之间是否存在链路的可能性。然而,粗略地将关系压缩为二元关系会丢失许多有效信息。例如,代谢网络中,每个化学反应都关联着3种及其以上的物质,论文合著网络中,一篇论文往往由多个作者撰写。真实网络中的关系存在高阶交互性,普通的图不能描述这种错综复杂的结构。超图为建模这种高阶关系提供了有效方法,超图中的超边可以刻画由多个不同类型节点构成的超链路。

超链路预测是预测超图中缺失的高阶关系的问题^[12]。相对单一的复杂网络,高阶交互的超网络研究的结构更复杂,包含的信息更多元全面,但链路预测的本质没变,都是通过对边/超边的存在进行推断,以此实现预判网络演化趋势以及干预网络结构进而调整系统功能的目的^[13]。由于构成超链路的节点数量动态可变,现有的针对两个输入特征的链路预测方法不再适用于超网

络。同时基数可变问题使超图的链路预测变成了指数级的推理问题。对于有 n 个节点的超网络来说,潜在的超链路数为 $O(2^n)$ 。然而,现实世界不需要考虑所有超链路,针对具体情况可以过滤掉许多无意义的超边^[14]。因此,对于现实中有有限超链路来说,超链路预测是可行的问题。超图的链路预测越来越受到学术界的重视,在无向超图上取得了巨大进展。目前超链路预测方法主要是将问题转化成与超图相关的数学问题,包括超图上的超链路接近信息的排序问题^[15]、超图的不完整关联矩阵的补全问题^[16]、对偶超图的二元顶点分类问题^[17],同样强大的深度学习模型——图卷积网络^[12]和图自编码器^[18]也可以用于解决超图的链路预测问题。另外,Zhang等^[14]提出了协调矩阵最小化方法,使用期望最大化方法预测邻接空间中的超链路。尽管不同链路预测技术处理问题的方法不同,但都忽略了网络内部之间的差异性,不加区分地将整张图输入超链路预测模型中,这必然会导致低效的预测结果。例如,代谢网络中生物会进行不同类型的代谢反应,反应之间既会存在明显的性质差异,也会在条件、过程、作用等方面保持一定的相似性,整张网络会被这种浑然天成的现象分割成不同类别的集合,具有明显的集聚倾向,如果某个集合的规模较大或者相似性比较高,那么基于整个网络预测出来的缺失反应类型会倾向于该群体,而一些小集合缺失的反应会由于与当前网络的匹配度较低而被自动过滤。

针对上述挑战,本文提出基于 SCL-CMM

(coordination matrix minimization based on hypergraph spectral clustering paper) 模型的超网络链路预测方法。与其他基于整个超图的超链路预测方法^[18-21]不同, SCL-CMM 模型首先建模超网络为异质超图, 以保留网络的高阶语义和异质关系, 同时重构以超链路为样本点的图, 挖掘超链路的潜在特征, 实现超图的谱聚类, 从而捕获原超图中不同类别的子图的高阶信息; 然后, 在划分后的子图中建立独立的超链路预测器作为生成模型, 学习已知超边的深层逻辑, 从各自的候选超链路中筛选出与当前网络匹配度最高的超链路作为最终的预测结果。本文的创新点和主要贡献如下。

1) 提出基于 SCL-CMM 模型的超网络链路预测方法, 利用超图表示超链路高阶语义的同时捕捉了超链路的异质性与集聚的物理特征之间的关联, 联合聚类解析器和超链路预测器来生成每个子图的潜在超链路的拓扑结构, 保证了所有类别的超链路信息都得到合理且充分的利用, 突破了超链路预测统一建模导致的预测精度不高、结果覆盖不全的局限性。

2) 提出聚类解析器对超图进行子图划分的方法, 首先解构整个超网络, 构建以超边为样本点的图, 分析样本点的特征差异, 保证性质相近的超链路被划分到同一簇, 根据簇划分情况重构原超图为多个同质的子图。该方法不仅通过解重构过程合理地将谱聚类的思想迁移到超图中, 而且能够准确推断每个性质各异的子图, 不会出现合并小类的情况。

3) 在 6 个真实数据集上证明了所提方法的可行性和有效性, 与常见的基线方法相比, 在 AUC (area under curve)、召回率方面都取得了最好的效果, 并且在不同比例的训练集下进行实验, 所有的超链路预测技术融入聚类解析器后预测能力都有显著提高, 充分说明了超链路的聚类可以增强算法的预测能力。

1 问题描述

定义 1 给定点集 $V = \{v_1, v_2, \dots, v_m\}$, V 上的一个超图表示为 $H = E = \{E_1, E_2, \dots, E_n\}$ 。每个 E_i 都是由 V 中的部分节点构成的集合, $|E_i| \geq 2$, 且有 $\bigcup_{i=1}^n E_i = V$ 。超链路候选集 $C = \{E_1, E_2, \dots, E_n\}$

包含 H 中所有的链路以及不存在的超链路。

定义 2 超链路预测问题是一个二元组 (H, C) 。 H 表示不完整的超图, C 表示超链路候选集且 $H \subseteq C$ 。因此, 超链路预测的目标就是在 C 中寻找最匹配 H 的缺失链路集。

定义 3 用关联矩阵 $\mathbf{S}_{|V| \times |E|}$ 表示一个超图 H , 每一行代表一个节点, 每一列代表一条超边。矩阵中的每一个元素 S_{ij} 定义为

$$S_{ij} = S(v_i, E_j) = \begin{cases} 1, & v_i \in E_j \\ 0, & \text{其他} \end{cases} \quad (1)$$

定义 4 用邻接矩阵 $\mathbf{A}_{|V| \times |V|}$ 表示节点之间的关系, 可以通过 $\mathbf{A} = \mathbf{S}\mathbf{S}^T$ 计算得到。矩阵中的元素 A_{ij} 表示第 i 个节点和第 j 个节点共同出现在超边的次数。

定义 5 超边的高斯距离决定了链路之间的相似性, 高斯距离越大, 它们越相似。超边的相似度矩阵 $\mathbf{Sim}_{|E| \times |E|}$ 中每个元素的值定义如下。

$$\mathbf{Sim}_{ij} = \text{distance}(E_i, E_j) = \exp \left(- \frac{\|x_{E_i} - x_{E_j}\|^2}{2\sigma^2} \right) \quad (2)$$

定义 6 采用 KNN (k-nearest neighbor) 算法遍历所有链路, 找到每条链路距离最近的 k 个邻居, 划分为每条链路的邻域。只有当超链路 E_i 和 E_j 彼此都在对方的邻域中, 它们才具有邻接关系, 邻接值 w_{ij} 为两个链路之间的相似度, 否则 $w_{ij} = 0$ 。权重矩阵 $\mathbf{W}$ 的每个元素定义如下。

$$W_{ij} = W_{ji} = \begin{cases} \exp \left(- \frac{\|x_{E_i} - x_{E_j}\|^2}{2\sigma^2} \right), & E_i \in \text{KNN}(E_j) \text{ 且 } E_j \in \text{KNN}(E_i) \\ 0, & E_i \notin \text{KNN}(E_j) \text{ 或 } E_j \notin \text{KNN}(E_i) \end{cases} \quad (3)$$

定义 7 每条超边的度 d_i 定义为与 E_i 相邻的超边的邻接权重之和, 表示为 $d_i = \sum_{j=1}^n w_{ij}$ 。根据超边的度的定义, 得到超边的度矩阵 $\mathbf{D}_{|E| \times |E|}$, 它是一个对角矩阵, 对角元素为样本点的度, 其余全为 0, 定义如下。

$$D_{ij} = \begin{cases} \sum_j w_{ij}, & i = j \\ 0, & i \neq j \end{cases} \quad (4)$$

定义 8 谱聚类中的拉普拉斯矩阵定义为

$$L = D - W \quad (5)$$

引理 1 拉普拉斯矩阵 L 是对称半正定矩阵，因此它的特征值都是实数。

引理 2 对于任意的实向量 $f \in \mathbb{R}^n$ 都存在

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^n n w_{ij} (f_i - f_j)^2 \quad (6)$$

2 本文方法

2.1 总体架构

本文提出的 SCL-CMM 模型联合学习超网络的异质性和高阶交互性，实现更准确和全面的预测。如图 1 所示，SCL-CMM 由聚类解析器作为推理模型来推测所需的 k 个子图以及 k 个超链路预测器，其作为生成模型来重建不完整超图中潜在超链路的拓扑组成。具体来说，超图的聚类解析器包括 3 个过程：首先解构输入的超图的关联矩阵，将其重新组合成以超边 E_i 为顶点的图；然后学习超边的结构特征进行聚类；最终根据聚类结果 $C_1, C_2, \dots, C_k$ 重构原超图为 k 个类别的子图 $S_1, S_2, \dots, S_k$ 。超链路预测器基于 k 个子图的拓扑结构 $S_1, S_2, \dots, S_k$ ，利用期望最大化算法（expectation maximization algorithm, EM）优化参数 Λ 和 Z ，分别从各自的候选超链路中筛选最适合不完整超图 S_i 的超链路 ΔS_i ，构建完整超图的拓扑，得到缺失矩阵 ΔS 。

2.2 超图谱聚类解析

本文的研究对象是超图中的超链路，聚类的目标是将超图根据超链路的相似性划分 k 个子图。超链路的基数可变，并且维度一般比较大，直接采用传统的聚类算法处理大量的超链路不仅效率低，而且在高维数据上不能取得好的聚类效果。因此，本文从无向图的谱聚类理论出发，开发了适用于超图的谱聚类解析器。在图论中，聚类就相当于图的分割，重点在于找到合理的切割方式，保证子图内部尽可能相似，子图之间尽可能相异。以此为原则，本文生成以超边为样本点的图，计算超边的相似性来进行聚类，根据簇划分结果重构原超图 H 为 k 个子图，每个子图都是一组最为相似的超链路构成的集合。下面通过数学公式定义最佳的切割方式。

2.2.1 损失函数

假设 $H_1, H_2, \dots, H_k$ 为切割后的 k 个子图，它们满足如下条件：① $H_1 \cup H_2 \cup \dots \cup H_k = H$ ；② $H_1 \cap H_2 \cap \dots \cap H_k = \emptyset$ 。由定义 6 可以计算每条超边的权重矩阵，进而得到每个子图内部的权重以及子图之间的权重。为了得到相似性最高的 k 个区域，就要切断权重较小的边，同时为了避免局部最小化的情况出现，需要限制子图内部的权重，由此得到目标函数如下。

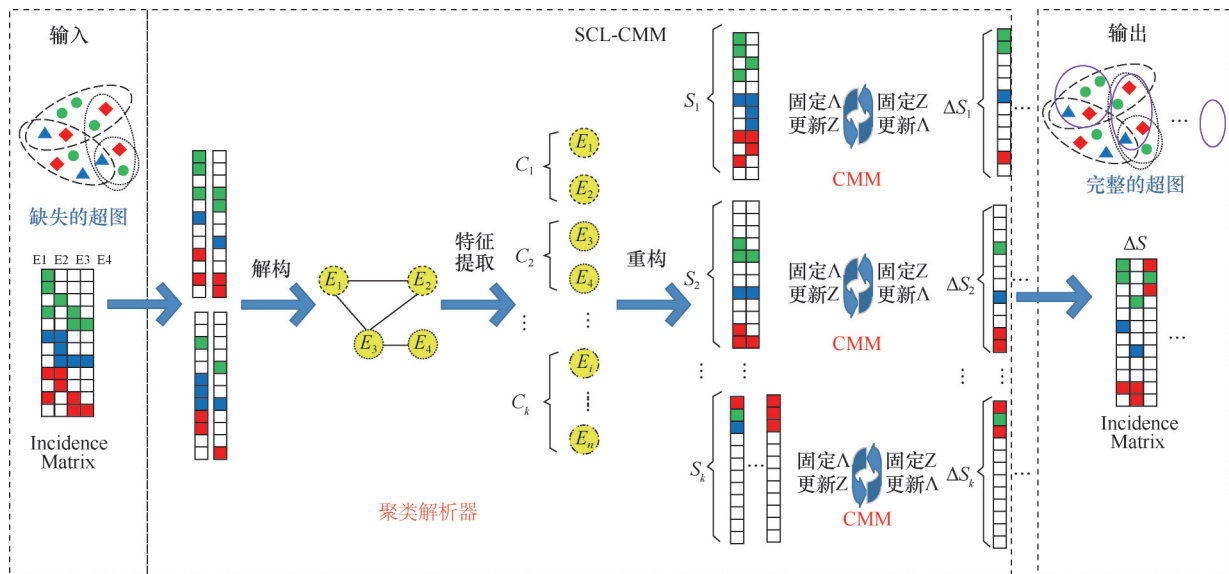


图 1 基于超图谱聚类解析器的协调矩阵最小化模型的框架

Figure 1 The framework of the hypergraph spectral clustering parser-based coordination matrix minimization model

$$Hcut(H_1, H_2, \dots, H_k) = \frac{\text{cut}(H_1, H_2, \dots, H_k)}{\sum_{i=1}^k \text{vol}(H_i)} = \frac{1}{2} \sum_{i=1}^k \frac{W(H_i, \overline{H_i})}{\text{vol}(H_i)} \quad (7)$$

其中, $\text{cut}(H_1, H_2, \dots, H_k) = \frac{1}{2} \sum_{i=1}^k W(H_i, \overline{H_i})$ 表示子图之间的权重, $\text{vol}(H_i) = \sum_{E_p \in H_i, E_q \in H_i} w_{pq} = \sum_{j=1}^k W(H_i, H_j) - W(H_i, \overline{H_i})$ 表示子图内部的权重, 通过最小化二者之间的比值来寻找聚类效果最好的 k 个子图。

为了最小化目标函数, 本文引入 k 个子图 $\{H_1, H_2, \dots, H_k\}$ 的指示变量矩阵 $\mathbf{R}_{n \times k} = \{r_1, r_2, \dots, r_k\}$, 用来表示超图中的超边和子图之间的关系。每个指示向量 $\mathbf{r}_j$ 对应一个子图 H_j , 是一个 n 维列向量, 每一维 r_{ji} 代表子图 H_j 对超边 E_i 的指示结果, 表示如下。

$$r_{ji} = \begin{cases} 0, & E_i \notin H_j \\ \frac{1}{\sqrt{|\text{vol}(H_j)|}}, & E_i \in H_j \end{cases} \quad (8)$$

引理 3 每条超边只能属于一个子图, 所以 k 个子图的指示向量之间相互正交, 即 $\mathbf{r}_i^* \mathbf{r}_j = 0 (i \neq j)$, 且指示向量和损失函数之间存在以下性质。

- (i) $\mathbf{r}_i^T \mathbf{D} \mathbf{r}_i = 1$,
- (ii) $\mathbf{r}_i^T \mathbf{L} \mathbf{r}_i = Hcut(H_i, \overline{H_i})$ 。

证明

(i) 对公式 $\mathbf{r}_i^T \mathbf{D} \mathbf{r}_i = 1$ 的证明。

$$\mathbf{r}_i^T \mathbf{D} \mathbf{r}_i = \sum_{j=1}^n r_{ij}^2 D_{jj} = \frac{1}{\text{vol}(H_i)} \sum_{E_j \in H_i} w_{E_j} = \frac{1}{\text{vol}(H_i)} \text{vol}(H_i) = 1 \quad (9)$$

(ii) 对公式 $\mathbf{r}_i^T \mathbf{L} \mathbf{r}_i = Hcut(H_i, \overline{H_i})$ 的证明。

$$\begin{aligned} \mathbf{r}_i^T \mathbf{L} \mathbf{r}_i &= \mathbf{r}_i^T \mathbf{D} \mathbf{r}_i - \mathbf{r}_i^T \mathbf{W} \mathbf{r}_i = \\ &= \frac{1}{2} \sum_{p=1}^n \sum_{q=1}^n w_{pq} (r_{ip} - r_{iq})^2 = \\ &= \frac{1}{2} \left[\sum_{E_p \in H_i, E_q \in \overline{H_i}} w_{pq} \left(\frac{1}{\sqrt{|\text{vol}(H_i)|}} - 0 \right)^2 + \right. \end{aligned}$$

$$\begin{aligned} &\left. \sum_{E_p \in \overline{H_i}, E_q \in H_i} w_{pq} \left(0 - \frac{1}{\sqrt{|\text{vol}(H_i)|}} \right)^2 \right] = \\ &= \frac{1}{2} \left[\sum_{E_p \in H_i, E_q \in \overline{H_i}} w_{pq} \frac{1}{\text{vol}(H_i)} + \sum_{E_p \in \overline{H_i}, E_q \in H_i} w_{pq} \frac{1}{\text{vol}(H_i)} \right] = \\ &= \frac{1}{2} \left[\text{cut}(H_i - \overline{H_i}) \frac{1}{\text{vol}(H_i)} + \text{cut}(\overline{H_i} - H_i) \frac{1}{\text{vol}(H_i)} \right] = \\ &= \frac{\text{cut}(H_i - \overline{H_i})}{\text{vol}(H_i)} = Hcut(H_i, \overline{H_i}) \quad (10) \end{aligned}$$

由式(9)、式(10)可将目标函数进一步缩放到以下形式。

$$Hcut(H_1, H_2, \dots, H_k) = \sum_{i=1}^k \mathbf{r}_i^T \mathbf{L} \mathbf{r}_i = \sum_{i=1}^k (\mathbf{R}^T \mathbf{L} \mathbf{R})_{ii} = \text{tr}(\mathbf{R}^T \mathbf{L} \mathbf{R}) \quad (11)$$

此时优化目标为在 $Hcut(H_1, H_2, \dots, H_k)$ 取得最小值的情况下的 k 个子图的指示变量矩阵 $\mathbf{R}$ 。毫无疑问, 一旦获得了指示向量矩阵, 即可得到每个子图分配的超边情况, 但要想找到一个代价最小的目标, 最坏的情况需要遍历所有子图的指示向量矩阵, 每个子图的指示向量有 n 个未知解, 每个解有 2 种可能性, 因此求解 k 个子图的指示向量矩阵 $\mathbf{R}$ 的问题规模就相当于复杂度 $O(k2^n)$, 是一个 NP 难问题, 在有限的时间内很难用精确算法求解, 需继续转换优化目标。

对 $\mathbf{R}$ 执行归一化操作, 令 $\mathbf{R} = \mathbf{D}^{-1/2} \mathbf{F}$, 将其代入上式得:

$$\begin{aligned} \mathbf{R}^T \mathbf{D} \mathbf{R} &= (\mathbf{D}^{-1/2} \mathbf{F})^T \mathbf{D} \mathbf{D}^{-1/2} \mathbf{F} = \\ \mathbf{F}^T \mathbf{D}^{-1/2} \mathbf{D} \mathbf{D}^{-1/2} \mathbf{F}, \mathbf{R}^T \mathbf{D} \mathbf{R} &= \\ (\mathbf{D}^{-1/2} \mathbf{F})^T \mathbf{D} \mathbf{D}^{-1/2} \mathbf{F} = \\ \mathbf{F}^T \mathbf{D}^{-1/2} \mathbf{D} \mathbf{D}^{-1/2} \mathbf{F} &= \mathbf{F}^T \mathbf{I} \mathbf{F} = \mathbf{I} \end{aligned} \quad (12)$$

优化目标最终确定为:

$$\begin{aligned} \arg \min_{\mathbf{F}} \quad & \text{tr}(\mathbf{F}^T \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} \mathbf{F}), \\ \text{s.t.} \quad & \mathbf{F}^T \mathbf{F} = \mathbf{I} \end{aligned} \quad (13)$$

2.2.2 求解

求解式(13)中的目标 $\mathbf{F}$ 的过程相当于求 $\mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$ 的前 k' 个特征值所对应的特征向量, 标准化后即可得到特征矩阵 $\mathbf{F}_{n \times k'}$ 。需要注意的

是, 在求解 F 的过程中进行了维度规约, 造成了信息损失, 因此得到的 $F_{n \times k}$ 。并不能利用式子 $R = D^{-1/2}F$ 直接转化 R , 需要进一步对 F 矩阵的行向量进行 k -means 聚类。聚类后即可得到降维后的 k 个簇划分 $S_1, S_2, \dots, S_k$, 分别对应子图序列 $H_1, H_2, \dots, H_k$ 。至此, 谱聚类通过求解规范化后的拉普拉斯矩阵的特征向量成功地将超图 H 分割成了不同类型的子图。

2.3 SCL-CMM模型

CMM (coordination matrix minimization) 模型^[14]是一种统计学的方法, 通过在顶点的邻接空间中进行非负矩阵分解和最小二乘匹配, 来筛选候选集中不完整网络所缺失的超链路。它具备优秀的重建超网络的能力, 并且在理论层面也具有可解释性。本文将超图谱聚类解析器与超链路预测器 CMM 相结合来应对异质超图预测结果覆盖不全面的挑战, 提出了基于 SCL-CMM 模型的超网络链路预测方法。

2.3.1 损失函数

在 2.1 节中, 本文利用聚类解析器对超图进行切割, 得到 k 个子图的关联矩阵序列 $S_1, S_2, \dots, S_k$ 。超链路预测器作为生成模型, 其目的是从 k 个子图的邻接空间中学习超链路的随机分布, 从中重建出缺失超链路集的超图拓扑。下面通过数学公式定义寻找不同子图缺失超链路的过程。

首先根据观察到的超图结构随机生成一些超边, 模拟不存在的超链路, 与 k 个子图 $S_1, S_2, \dots, S_k$ 共同组成候选集 $U_1, U_2, \dots, U_k$, 每个子图也删除一定比例的超链路作为缺失的超链路, 子图序列 $S_1, S_2, \dots, S_k$ 相应转化成不完整子图 $S'_1, S'_2, \dots, S'_k$, 显然每个不完整子图 S'_i 缺失的超链路都包含在候选集 U_i 中。此时 SCL-CMM 模型将分别从候选集 U_i 中筛选出最有可能是该不完整子图 S'_i 缺失的超链路, 成功将研究对象从整个超图细化为每个子图。

引入第 i 个不完整子图 S'_i 的标识矩阵 $A_i = \text{diag}([\lambda_1, \lambda_2, \dots, \lambda_n])$ 表示对应候选集 U_i 的筛选结果。 $\lambda_j=1$ 时表示候选集中第 j 条超链路 u_j 是 S'_i 缺失的链路, 否则 $\lambda_j \neq 1$ 时表示不是缺失的链路。本文用 ΔS_i , ΔA_i 分别表示第 i 个子图的缺失链路集对应的关联矩阵、邻接矩阵, 如果 A_i 已知, 那么缺失矩阵 ΔS_i , ΔA_i 可以通过以下公式计算。

$$\Delta S_i = A_i U_i,$$

$$\Delta A_i = U_i A_i U_i^T \quad (14)$$

因此, 每个子图完整的邻接矩阵可以表示为 $S'_i S_i'^T + U_i A_i U_i^T$, 为了求解 A_i , 本文使用非负矩阵分解方法求解。假设非负矩阵 $Z_i = [Z_1, Z_2, \dots, Z_n]^T$ 表示第 i 个子图的潜在因子矩阵, Z_i 中的每一行是包含顶点 V_i 的 l 个潜在特征的行向量 ($l \leq n$)。子图完整的邻接矩阵存在一定噪声, 表示为:

$$S'_i S_i'^T + U_i A_i U_i^T \approx Z_i Z_i^T \quad (15)$$

为了找到每个子图缺失的链路, 本文对每个子图提出以下优化问题。

$$\begin{aligned} \arg \min_{A_i, Z_i} & \|S'_i S_i'^T + U_i A_i U_i^T - Z_i Z_i^T\|_F^2, \\ \text{s.t. } & \lambda_j \in \{0, 1\} \quad j = 1, \dots, n^i, Z_i \geq 0 \end{aligned} \quad (16)$$

2.3.2 优化

该优化问题的目标为最适合不完整网络的缺失链路集 (由 k 个 A_i 给出) 和最能解释完整网络的 l 个潜在因子矩阵。通过 EM 框架迭代地执行两步优化算法, 直到算法收敛或达到最大迭代次数终止运行, EM 框架的 E 步骤为固定 Z_i 的前提下最小化式 (17), 相当于整数的最小二乘匹配问题。

$$\begin{aligned} \arg \min_{A_i} & \|S'_i S_i'^T + U_i A_i U_i^T - Z_i Z_i^T\|_F^2, \\ \text{s.t. } & \lambda_j \in \{0, 1\} \quad j = 1, \dots, n^i \end{aligned} \quad (17)$$

EM 框架的 M 步骤为固定 A_i 的前提下最小化式 (18), 相当于对称的非负矩阵分解问题。

$$\begin{aligned} \arg \min_{Z_i} & \|S'_i S_i'^T + U_i A_i U_i^T - Z_i Z_i^T\|_F^2, \\ \text{s.t. } & Z_i \geq 0 \end{aligned} \quad (18)$$

这样每个子图都可通过 A_i 在各自的候选集中筛选缺失超链路集 ΔS_i , $\Delta S_1 + \Delta S_2 + \dots + \Delta S_k$ 的和就是 SCL-CMM 模型预测到的所有超链路。算法 1 总结了 SCL-CMM 模型的完整过程。

首先利用超图的聚类解析器对超图进行子图划分, 需要进行 n_1^2 次矩阵乘法操作, 对应的时间复杂度为 $O(n_1^2)$ 。聚类后得到 k 个子图, 每个子图运行 CMM 模型时, 时间复杂度为 $O(n_2^3)$ 。因此最终 SCL-CMM 模型的时间复杂度为 $O(n_1^2 + kn_2^3)$, n_1 代表整个超图的超链路数量, n_2 代表聚类后的子图的超链路数量。

算法1 SCL-CMM模型

输入 不完整超图的关联矩阵 $S_{m \times n}$, 聚类数

目 k , 训练集比例 p , 高斯核函数带宽 σ , 降维后的维度 k'

输出 预测的缺失链路集的关联矩阵 ΔS

1) ΔS 初始化为空

2) 通过式(2)计算超边的相似度矩阵 $\text{Sim}_{|E| \times |E|}$

3) 通过式(3)和式(4)计算超边的权重矩阵 W 和度矩阵

4) 通过式(5)计算拉普拉斯矩阵 L

5) 计算标准化后的拉普拉斯矩阵 $L_{\text{sym}} = D^{-1/2} L D^{-1/2}$

6) 计算 L_{sym} 最小的 k' 个特征向值所对应的特征向量矩阵 $T_{n \times k'} = [t_1, t_2, \dots, t_{k'}]$

7) for $i=1$ to k' do

8) $f_i = \frac{t_i}{|t_i|}$

9) end for

10) 拼接 $f_1, f_2, \dots, f_{k'}$ 得到特征矩阵 $F_{n \times k'}$

11) 根据 $S_1, S_2, \dots, S_k$ 随机生成不存在的超链路, 与 k 个子图构建候选超链路集 $U_1, U_2, \dots, U_k$, 并且对每个子图随机删除 $(1-p)$ 比例的超链路构造不完整子图序列 $S'_1, S'_2, \dots, S'_k$

12) for $i=1$ to k do

13) $\Delta S_i = \text{CMM}(S'_i, U_i)$

14) end for

15) $\Delta S = \Delta S_1 \cup \Delta S_2 \cup \dots \cup \Delta S_k$

3 实验

重建代谢网络有利于理解人类疾病的代谢基础、提高生物工程系统的产量和发现新的药物靶点的重要工具。但是真实的代谢网络往往缺失一些重要的反应, 严重损害了它们的效用。因此, 充分利用已知超链路信息全面准确地还原代谢网络具有十分重要的现实意义。本文通过预测不同类型超网络的缺失反应的实验来验证基于 SCL-CMM 模型的超链路预测方法的有效性。

3.1 实验数据集与预分析

本文选用了代谢网络(多种物质参与代谢反应构成超边)、GPS 网络(用户在某位置从事某活动构成超边)、引文网络(作者在会议上发表论文构成超边)、药物网络(患者服用某药物出现某反应构成超边)进行实验, 所含的节点数和

超链路数存在明显差异, 数据集规模如表 1 所示, 以此检测所提模型对不同规模 and 不同类型网络的灵敏度。实验开始之前, 本文对代谢网络的超链路数据进行了一系列规约降维操作, 最终根据数据的分布进行可视化, 6 个代谢超网络数据集的超链路分布图如图 2 所示。

表 1 数据集规模对比

Table 1 Comparison of datasets' sizes

数据集	节点数	超链路数
iABRBC283	342	469
iIT341	485	554
iAF692	628	690
iHN637	698	698
iAF1260b	1 668	2 388
iJO1366	1 805	2 583
GPS	221	1 436
DBLP	8 329	14 677
Drug	7 456	171 756

从图 2 中可以看出, 每个数据集都形成了规格不同的小团体, 群体内部间连接紧密, 趋向群体中心, 不同群体间数据分布稀疏, 说明代谢网络数据集呈现明显的集聚特征。可以用来测试 SCL-CMM 模型的准确性。且在图 2(b)、图 2(c)、图 2(d) 数据集中, 不同类别之间的界限比较清晰, 小集团内部数据排布稠密, 具有更明显的集聚特性, 笔者猜测挖掘集聚特征的超链路预测算法在性能上更能显著区别于基准算法。而其他 3 个数据集都存在一个比较大的集团, 数据分布不够紧密, 且与其他集团的界限比较模糊, 可能会影响聚类的效果, 导致整个算法的提升效果不够明显。

3.2 评价指标

AUC 是一种衡量链路预测精确度的常用指标, 计算存在的未知边的分数值大于不存在的未知边的分数值的概率, 表示为:

$$\text{AUC} = \frac{n' + n''}{n} \quad (19)$$

其中, n 代表总实验次数, n' 为测试集中的分数值大于不存在的边的次数, n'' 为测试集中的分数值等于不存在的边的次数。AUC 的值越大, 说明算法提升的效果越明显。

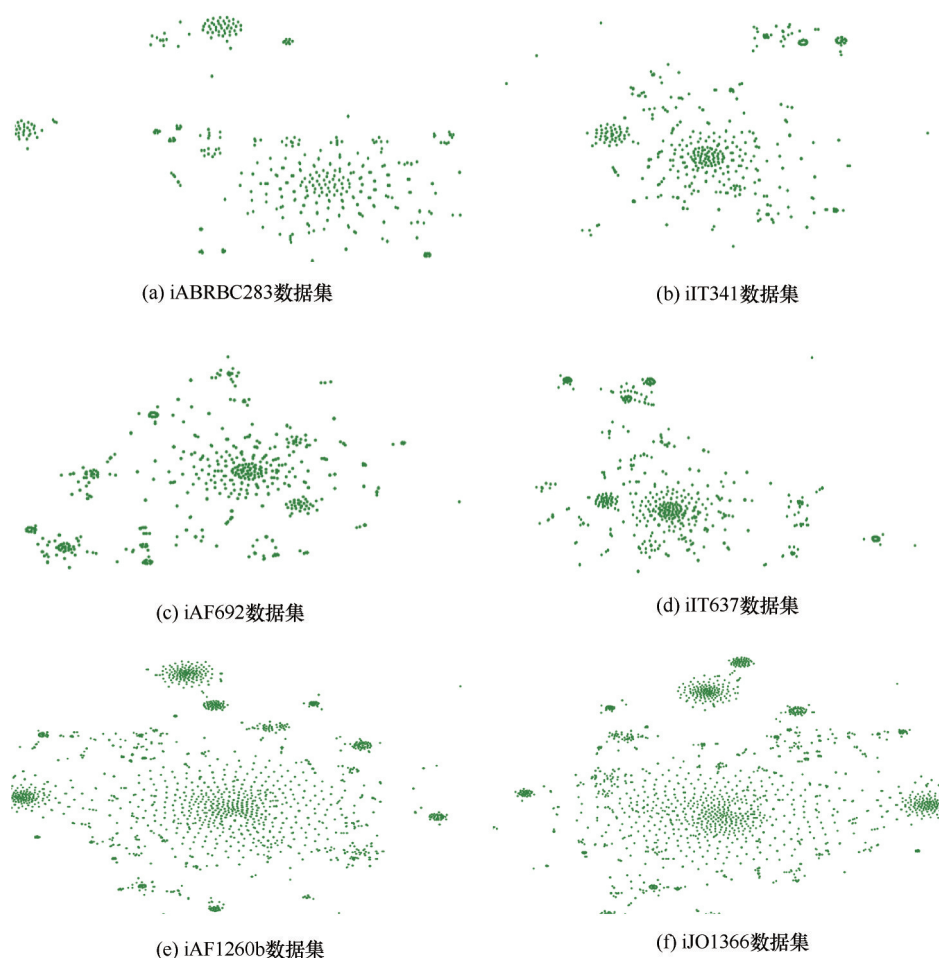


图2 6个代谢超网络数据集的超链路分布图

Figure 2 Hyperlink distribution map of six metabolic hypernetwork datasets

不同于准确性评估指标AUC，召回率是评估算法预测是否全面的指标，即网络中缺失的超链路有多少条被预测出来，表示为：

$$\text{Recall} = \frac{n}{L} \quad (20)$$

其中， L 代表网络中真实存在但却观察不到的超链路数目， n 为算法预测出的真实缺失的超链路数目，即召回条数。召回条数越多，召回率越高，预测的结果覆盖更全面。综合考虑AUC、召回条数和精确度能更为全面地评价算法的性能。

3.3 基准算法

为了充分量化所提模型的有效性和准确性，本文选择了基于相似性的链路预测模型、基于信息熵的逻辑回归模型、基于深度学习的超链路预测模型和基于统计学的超链路预测模型作为基准

模型进行对比实验。

1) BS (bayesian sets) [22]针对的是从概念或集群中检索项目的问题。给定一个查询 $S = \{s_1, s_2, \dots, s_m\}$ ，BS 计算所有超链路 $u \in U$ 的得分，并检索得分最高的超链路。

2) Katz [23]是全局属性的链路预测指标。对于一个有 m 个节点的超链路会产生 $m(m-1)/2$ 个成对的Katz索引，计算它们的平均值作为超链路的Katz指标得分。

3) CN (common neighbor) [24]是局部属性的链路预测指标，推广到超链路的过程与Katz相同。

4) HPLSF [25]是使用监督学习的超链路预测方法。它基于学习到潜在特征的熵值，训练逻辑回归模型来输出超链路的预测分数。

5) NHP-U [12]是神经超链路预测器，利用图

卷积神经网络学习超链路的表示来预测无向超图的超链路。

6) CMM (coordinated matrix minimization) 方法^[14]使用 EM 框架来寻找最适合超网络的候选超链路。

7) C3MM (clique-closure based hyperlink prediction)^[26]方法提出了集团封闭假说, 基于超链路更有可能由近团而不是非团形成的理论, 改进了 CMM 方法。

8) DHNE (deep hypernetwork embedding)^[27]是学习非线性簇相似函数的深度模型, 利用局部和全局近似形成的嵌入空间来保持超边的不可分解性。

9) HPNG (hyper-path-based representation learning for hyper-networks)^[28]针对异质超网络的多阶语义关系进行嵌入学习, 利用超路径的随机游走捕获超图中的成对关系和丛状关系。

3.4 实验结果与分析

本文针对不同类型的网络进行仿真实验。在执行 SCL-CMM 模型对超链路进行聚类时, 将控制高斯核函数作用范围的带宽 σ 设置为 1, 描述超边特征的拉普拉斯矩阵的特征向量矩阵的维度 k' 设置为 4, 分别在代谢网络数据集、GBS 数据集、DBLP 数据集、Drug 数据集运行时, 将聚类数目 k 设定为 9、14、11、18、24、28、16、109、1 210 (此时算法的预测性能最好)。为了尽可能保证实验的准确性, 消除随机误差的影响,

对每个网络进行 100 次独立实验取平均结果。下面对 SCL-CMM 模型和 9 个基准模型的 AUC、召回条数和召回率进行分析。

从表 2~表 4 可以看出, 在不同数据集下, SCL-CMM 模型相比其他基准模型能够预测出更多缺失的超链路, 获得了最高的召回率和 AUC 值, 这说明本文所提方法在所有数据集上都实现了对基线的显著改进, 对超链路进行的聚类操作能够提升模型的预测能力。这是由于 SCL-CMM 模型充分利用了网络的集聚特性, 所有类型的超链路都可以参与到预测过程中, 扩大了预测结果的覆盖范围, 能够还原出更多类型的子网络中缺失的超链路。同时, 数据聚类的方法有很多, 这里之所以选用谱聚类, 是由于谱聚类方法对于数据集的样本空间和概率分布没有限制, 更具普适性; 同时谱聚类方法也是一个为图量身定做的方法, 对于一些样本数量更小的类别有更好的识别能力, 不会出现合并小类的情况。因此, 选用谱聚类方法对超图做聚类更能显著提升预测效果。除此之外, 进一步观察代谢网络的实验结果可以发现, SCL-CMM 模型的召回相对 CMM 模型在第 2~4 个数据集上提升最为明显, 约为 20%, 而在 iABRBC283、iAF1260b、iJO1366 数据集上提升了 12% 左右, 这与本文在对数据作预分析时的猜想一致, 网络集聚特征的显著与否将直接影响模型的预测能力, 集聚特征越明显, 模型预测越准确全面。

表 2 AUC 对比

Table 2 Comparison of AUC

模型	iABRBC283 数据集	iIT341 数据集	iAF692 数据集	iHN637 数据集	iAF1260b 数据集	iJO1366 数据集	GPS 数据集	DBLP 数据集	Drug 数据集
BS	0.676 9	0.503 5	0.480 6	0.507 2	0.654 4	0.686 6	0.625 4	0.698 0	0.687 2
Katz	0.443 4	0.443 4	0.521 7	0.496 1	0.608 0	0.546 3	0.578 9	0.612 3	0.591 7
CN	0.401 0	0.439 3	0.449 4	0.422 6	0.455 2	0.447 8	0.476 7	0.498 7	0.467 3
HPLSF	0.651 9	0.592 3	0.575 3	0.504 1	0.539 7	0.528 4	0.601 2	0.676 7	0.602 1
NHP-U	0.622 1	0.672 4	0.638 1	0.582 2	0.642 6	0.566 7	0.631 2	0.654 2	0.602 3
CMM	0.644 3	0.714 3	0.646 3	0.629 8	0.693 0	0.645 1	0.701 2	0.724 3	0.731 2
C3MM	0.667 2	0.721 1	0.701 2	0.689 0	0.701 2	0.702 1	0.723 1	0.786 5	0.765 2
DHNE	0.698 9	0.752 0	0.698 8	0.692 3	0.753 2	0.735 2	0.777 8	0.843 2	0.821 2
HPNG	0.703 3	0.741 3	0.712 1	0.698 7	0.767 8	0.783 2	0.791 2	0.864 2	0.840 7
SCL-CMM	0.727 7	0.757 8	0.725 4	0.741 7	0.789 3	0.791 9	0.8023 2	0.883 1	0.846 5

表3 召回条数对比

Table 3 Comparison of Recall number

模型	iABRBC283 数据集	iIT341 数据集	iAF692 数据集	iHN637 数据集	iAF1260b 数据集	iJO1366 数据集	GPS 数据集	DBLP 数据集	Drug 数据集
BS	44	19	37	43	219	253	144	1 494	16 217
Katz	14	5	32	26	133	145	65	838	13 297
CN	11	9	21	16	122	126	64	681	6 053
HPLSF	38	20	38	27	148	151	142	1 254	0.387 1
NHP-U	48	46	79	76	237	287	157	1 920	20 236
CMM	46	45	77	72	249	280	170	1 686	20 683
C3MM	50	61	84	99	274	306	179	1 853	24 121
DHNE	54	66	97	110	307	328	195	2 319	25 217
HPNG	55	71	108	114	297	348	194	2 278	26 289
SCL-CMM	57	73	111	115	309	354	201	2 357	27 017

表4 召回率对比

Table 4 Comparison of Recall

模型	iABRBC283 数据集	iIT341 数据集	iAF692 数据集	iHN637 数据集	iAF1260b 数据集	iJO1366 数据集	GPS 数据集	DBLP 数据集	Drug 数据集
BS	0.468 1	0.171 2	0.268 1	0.273 9	0.458 2	0.489 4	0.501 8	0.509 0	0.472 1
Katz	0.148 9	0.108 1	0.231 9	0.165 6	0.278 2	0.280 5	0.231 7	0.285 4	0.303 5
CN	0.117 0	0.081 1	0.152 2	0.101 9	0.255 2	0.243 7	0.223 1	0.232 1	0.176 2
HPLSF	0.404 2	0.180 2	0.275 4	0.172 0	0.309 6	0.292 1	0.416 2	0.427 1	0.387 1
NHP-U	0.510 6	0.414 4	0.572 5	0.484 1	0.495 8	0.555 1	0.547 9	0.654 2	0.589 1
CMM	0.489 4	0.405 4	0.558 0	0.458 6	0.520 9	0.541 6	0.592 1	0.574 3	0.602 1
C3MM	0.531 2	0.547 7	0.612 3	0.632 1	0.574 1	0.591 7	0.623 8	0.631 2	0.702 2
DHNE	0.576 5	0.591 9	0.701 2	0.698 2	0.642 1	0.634 1	0.677 7	0.790 2	0.734 1
HPNG	0.591 2	0.643 1	0.782 1	0.728 9	0.621 3	0.672 9	0.678 1	0.776 2	0.775 27
SCL-CMM	0.606 4	0.657 7	0.804 3	0.732 5	0.646 4	0.684 7	0.701 4	0.803 1	0.786 5

3.5 参数影响

聚类数目作为超参数,对预测结果有显著影响,需要不断实验寻找最佳值。图3展示了设定不同聚类数目时 SCL-CMM 模型的召回条数的变化。从总体来看,这6幅图的走向是一致的,算法在达到最佳结果之前,随着聚类数目的不断增加,越来越多小类别的超链路被发现,网络的聚集特性在超链路预测中可以得到更好的应用,因此预测结果覆盖的种类更加全面,召回条数随之增加;但当达到最佳效果后,继续增加聚类数目会使得隶属同一个类别的超链路被错误分开,不

仅会增加算法计算负担,同时这种过度拆分超链路类别的做法会明显降低算法的预测效果,召回条数也就随之减少。观察每张图的预测结果可以发现, SCL-CMM 模型在 iABRBC283、iIT341、iAF692、iHN637、iAF1260b、iJO1366、GPS、DBLP、Drug 这9个数据集上运行时,分别在聚类数目 $k=9$ 、14、11、18、24、28、16、109、1 210 时召回条数最多,取得了最佳的预测效果。

3.6 鲁棒性分析

为了进一步分析基于聚类的超链路预测模型的鲁棒性,本文在设定不同的训练集比例下对比

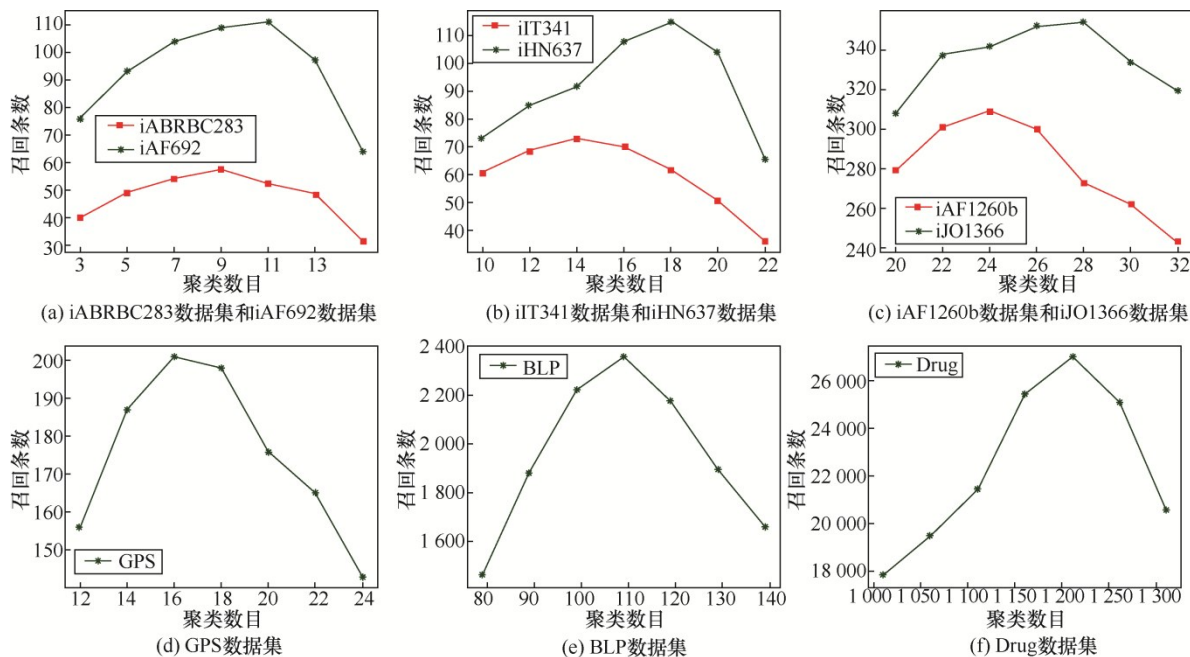


图3 设定不同聚类数目时SCL-CMM模型的召回条数的变化

Figure 3 Changes in the number of recalls of the SCL-CMM model when different numbers of clusters are set

聚类后的超链路模型与传统超链路模型预测结果的变化情况。图4分别展示了设定不同训练集比例(10%、20%、30%、40%、50%)时基于谱聚类的超链路预测模型(SCL-HPLSF、SCL-NHP-U、SCL-CMM)和传统超链路预测模型(HPLSF、NHP-U、CMM)的召回率对比,每个数据集的预测结果仍然是取100次独立实验的平均值。从总体分析,所有图都呈现了缓慢上升的趋势,当训练集比例降低时,超网络缺失的比例随之增加,从候选集中筛选的超链路在测试集的可能性就会提高,因此算法的召回率随训练集比例的降低而增加;同时纵向观察,在所有数据集的不同训练集比例下,基于谱聚类的超链路预测模型(SCL-HPLSF、SCL-NHP-U、SCL-CMM)的召回率都显著优于未做聚类处理的传统超链路预测模型(HPLSF、NHP-U、CMM),能够预测出更多缺失的超链路,这充分说明不同的超链路预测技术在聚类后的子网络上运行能够显著提升预测效果,且基于谱聚类的超链路预测模型具有良好的鲁棒性。

4 结束语

本文提出了基于SCL-CMM模型的超网络链路预测方法,同时融合了网络的结构信息以及集

聚特征,打破了单纯利用超网络的拓扑特性进行链路预测的格局。真实世界中超边的异质性使超网络内部呈现不同程度的集聚属性,针对这一物理现象,本文开发了超图的聚类解析器来挖掘超链路间的特征差异,从而重构原始的异质超图为多个高阶的子图,然后基于子图的非线性行为利用超链路预测器生成潜在链路集的拓扑结构。该方法捕获了所有类别的超链路的高阶语义信息,弥补了传统算法的不足,全面准确地还原了网络的全貌。实验结果表明,这种结合网络特征的算法能够更好地预测网络的状态,在预测精确性和全面性方面,SCL-CMM在多个具有挑战性的数据集上始终优于现有的方法,并且具有良好的鲁棒性。

挖掘真实的有效信息对于全方位多元化地重构超网络来说至关重要,尤其是由超网络本身具有的高阶交互性和异质性引起的内部结构和机制的改变,能够显著影响对超网络动态演化和链路预测的研究进展。未来,设计更多融合网络拓扑信息和功能特征的超链路预测方法值得进一步探究,同时考虑如果将该方法运用到时序的超网络中,预测未来的超链路也是亟待解决的问题。

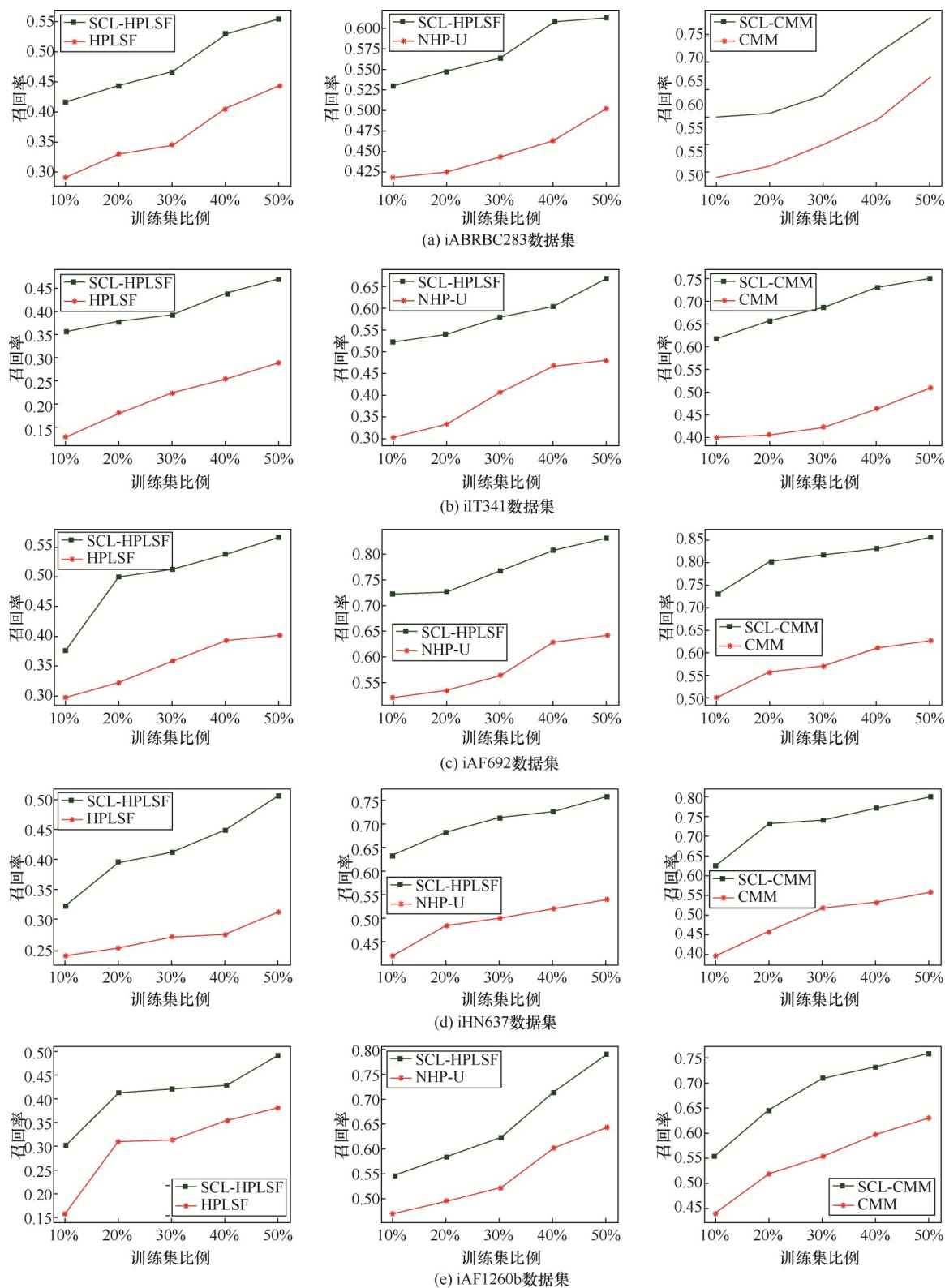


图4 设定不同训练集比例时基于谱聚类的超链路预测模型和传统的超链路预测模型的召回率对比
 Figure 4 Comparison of recall rates between the hyperlink prediction model based on spectral clustering and the hyperlink prediction model when different training set ratios are set

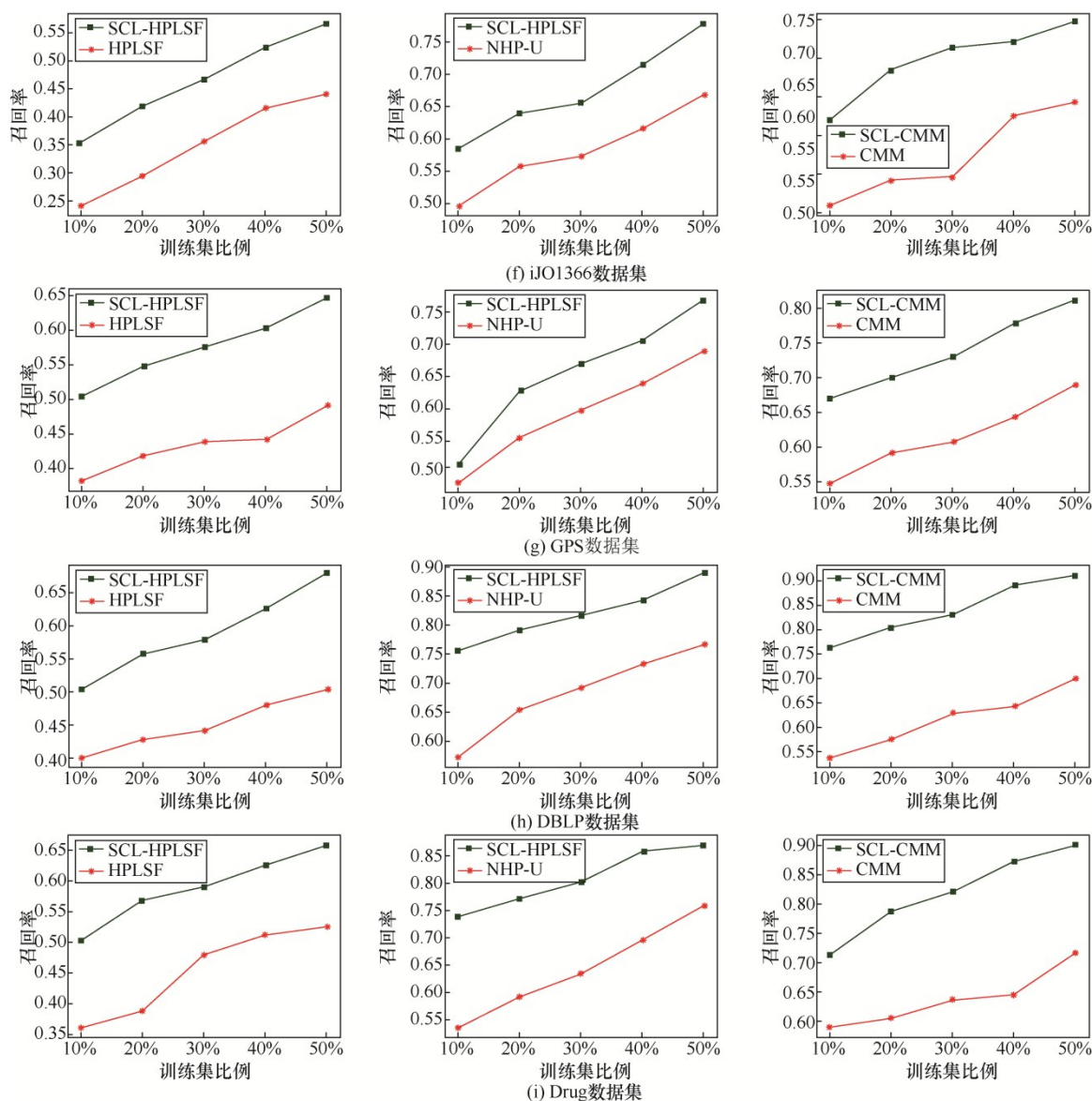


图4 设定不同训练集比例时基于谱聚类的超链路预测模型和传统的超链路预测模型的召回率对比(续)

Figure 4 Comparison of recall rates between the hyperlink prediction model based on spectral clustering and the hyperlink prediction model when different training set ratios are set

参考文献:

- [1] 马扬. 基于谱聚类的复杂网络时序链路预测技术研究[D]. 长沙: 国防科学技术大学, 2016.
MA Y. Research on complex networks time-series link prediction based on spectral clustering[D]. Changsha: National University of Defense Science and Technology, 2016.
- [2] LI C, JI X, LIU S, et al. A new method for predicting future links in temporal networks based on node influence[J]. International Journal of Modern Physics C, 2021, 32(12): 2150160.
- [3] LIAO L, HE X, ZHANG H, et al. Attributed social network embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2257-2270.
- [4] WANG P, XU B W, WU Y R, et al. Link prediction in social networks: the state-of-the-art[J]. Science China Information Sciences, 2015, 58(1): 1-38.
- [5] HAGHANI S, KEYVANPOUR M R. A systemic analysis of link prediction in social network[J]. Artificial Intelligence Review, 2019, 52(3): 1961-1995.
- [6] SHI C, HU B, ZHAO W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357-370.
- [7] ZHU X, YANG X, YING C, et al. A new classification algorithm recommendation method based on link prediction[J]. Knowledge-Based Systems, 2018, 159: 171-185.
- [8] KAYA B. A hotel recommendation system based on customer location: a link prediction approach[J]. Multimedia Tools and Applications, 2019, 78(12): 3215-3230.

- tions, 2020, 79(3): 1745-1758.
- [9] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2d knowledge graph embeddings[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 32.
- [10] SULAIMANY S, KHANSARI M, NEJAD A M. Link prediction potentials for biological networks[J]. International Journal of Data Mining and Bioinformatics, 2018, 20(2): 161-184.
- [11] BREIT A, OTT S, AGIBETOV A, et al. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction[J]. Bioinformatics, 2020, 36(13): 4097-4098.
- [12] YADATI N, NITIN V, NIMISHAKAVI M, et al. Link prediction in hypergraphs using graph convolutional networks[R]. 2018.
- [13] 刘怡君, 唐先一, 李倩倩, 等. 超链路预测[J]. 管理评论, 2012, 24(12):9.
- LIU Y J, TANG X Y, LI Q Q, et al. Superlink prediction [J] Management Review, 2012, 24 (12): 9
- [14] ZHANG M, CUI Z, JIANG S, et al. Beyond link prediction: predicting hyperlinks in adjacency space[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 32.
- [15] LI D, XU Z, LI S, et al. Link prediction in social networks based on hypergraph[C]//Proceedings of the 22nd International Conference on World Wide Web. 2013: 41-42.
- [16] ARYA D, WORRING M. Exploiting relational information in social networks using geometric deep learning on hypergraphs[C]//Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. 2018: 117-125.
- [17] LUGO-MARTINEZ J, ZEIBERG D, GAUDELET T, et al. Classification in biological networks with hypergraphlet kernels[J]. Bioinformatics, 2021, 37(7): 1000-1007.
- [18] FAN H, ZHANG F, WEI Y, et al. Heterogeneous hypergraph variational autoencoder for link prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
- [19] PATIL P, SHARMA G, MURTY M N. Negative sampling for hyperlink prediction in networks[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2020: 607-619.
- [20] ZHANG M, CUI Z, OYETUNDE T, et al. Recovering metabolic networks using a novel hyperlink prediction method[J]. arXiv Preprint arXiv: 1610.06941, 2016.
- [21] FATEMI B, TASLAKIAN P, VAZQUEZ D, et al. Knowledge hypergraphs: prediction beyond binary relations[EB]. arXiv Preprint arXiv: 1906.00137, 2019.
- [22] GHAHRAMANI Z, HELLER K A. Bayesian sets[C]//Proceedings of Advances in Neural Information Processing Systems. 2005: 18.
- [23] KATZ L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.
- [24] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031.
- [25] XU Y, ROCKMORE D, KLEINBAUM A M. Hyperlink prediction in hypernetworks using latent social features[C]//Proceedings of International Conference on Discovery Science. Springer, Berlin, Heidelberg, 2013: 324-339.
- [26] SHARMA G, PATIL P, MURTY M N. C3MM: clique-closure based hyperlink prediction[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021: 3364-3370.
- [27] TU K, CUI P, WANG X, et al. Structural deep embedding for hypernetworks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
- [28] HUANG J, LIU X, SONG Y. Hyper-path-based representation learning for hyper-networks[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 449-458.

[作者简介]



任玉媛 (1997-), 女, 河北张家口人, 信息工程大学硕士生, 主要研究方向为链路预测。



马宏 (1968-), 男, 江苏东台人, 国家数字交换系统工程技术研究中心研究员, 主要研究方向为社会网络分析、电信网关防护。



刘树新 (1987-), 男, 山东潍坊人, 博士, 国家数字交换系统工程技术研究中心助理研究员, 主要研究方向为复杂网络、网络信息挖掘。



王凯 (1980-), 男, 博士, 国家数字交换系统工程技术研究中心副研究员, 主要研究方向链路预测、社会网络分析。