引用格式：胡友鑫, 林茂彦, 罗剪秋, 等. 复杂网络高阶结构的关联规则挖掘及其应用 [J]. 电子科技大学学报, 2025, 54(1): 152-160.
HU Y X, LIN M Y, LUO J Q, et al. Association rule mining and applications based on higher-order structures in complex networks[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(1): 152-160.

复杂网络高阶结构的关联规则挖掘及其应用

胡友鑫，林茂彦，罗剪秋，陈 超，黄金煜*

（四川轻化工大学 计算机科学与工程学院，宜宾 644000）

摘要：网络高阶结构即满足特定条件的子网络，是网络科学领域重要的研究内容。近年来，关于高阶结构的研究不断增加，但是关于高阶结构之间内在联系的研究还相对较少。基于此，根据传统关联规则，定义了高阶结构之间的关联规则评判指标，并提出了一种有效挖掘高阶结构之间关联规则的通用算法框架。利用该算法，在 6 个真实世界网络中进行了 3 阶高阶结构（即高阶结构包含 3 个结点）间的关联规则挖掘。实验结果表明，真实世界网络中高阶结构之间存在强关联规则，且不同真实世界网络中高阶结构之间的关联规则存在差异。此外，将挖掘出的强关联规则应用于链路预测当中，进而实现了链路预测方法。相比于基线方法，所实现的链路预测方法在 4 个真实世界网络中取得了最好的性能表现。

关键词：关联规则；复杂网络；高阶结构；链路预测
中图分类号：TP311 **文献标志码：**A DOI: 10.12178/1001-0548.2023248

Association rule mining and applications based on higher-order structures in complex networks

HU Youxin, LIN Maoyan, LUO Jianqiu, CHEN Chao, and HUANG Jinyu*

(School of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin 644000, China)

Abstract: The study of higher-order structures, which refer to subnetworks within a network, is a crucial research topic in network science. In recent years, although the research on higher-order structures has been increasing, there has been relatively little research on the internal connections between higher-order structures. In light of conventional association rules, the evaluation criteria of association rules between higher-order structures are defined, and a general algorithm framework for effectively mining these association rules is proposed. The proposed approach has been applyed to mine association rules among three-order structures in six real-world networks. The results demonstrate strong association rules between higher-order structures in real-world networks, as well as variations in these rules across different networks. Additionally, we apply strong association rules to link prediction, resulting in a new link prediction method. This method outperforms the baseline methods in four real networks.

Key words: association rule; complex network; higher-order structure; link prediction

通过复杂网络研究，可以深入了解真实世界中复杂系统的结构特性和运行机制[1]。如分析谣言如何在社交网络中传播，解析各生态系统网络的异同点，通过蛋白质相互作用网络预测各蛋白质之间的关系等[2-4]。然而，目前的相关研究大多局限于描述成对的结点之间的二元关系。现实世界中很多复杂系统内部单元之间深层次的关系包含 3 个或者更多单元的高阶结构[5]。网络高阶结构可以揭示更多社交网络、计算机网络和生物网络等真实世界网络内部结点之间的内在联系[6]。

网络高阶结构指复杂网络中的子网络，如超图、网络模体等结构[7]。其中，网络模体是复杂网络中的一种重要的高阶结构，其被定义为网络中具有统计学意义的子图结构[8]。高阶结构是网络科学

收稿日期：2023−10−07
基金项目：四川省科技厅苗子工程重点项目（2019JDRC0085）；四川轻化工大学人才引进项目（2021RC13）
作者简介：胡友鑫，主要从事网络科学、链路预测等方面的研究。
*通信作者 E-mail: huangjy@suse.edu.cn

领域中的一个热门研究方向，近年来出现了许多有关高阶结构的研究。文献 [9] 在大肠杆菌的转录调控网络中发现不同的网络模体在该网络中具有不同的特定作用。文献 [10] 使用图论方法分析了大脑神经网络的连接模式，探讨了神经网络中的基本结构模式，揭示了在神经网络中存在的高阶结构。针对共词网络，文献 [11] 提出了一种基于网络模体的关键词提取算法，与其他传统算法相比拥有更好的性能。文献 [12] 系统性地研究了网络中的超图结构，提出了一种算法框架用来挖掘复杂网络中的超图模体结构，并分析了其在真实世界网络中的重要性。

然而，上述研究并没有考虑高阶结构之间可能存在的联系。复杂网络中的高阶结构之间可能会存在内在关联，相关研究有助于更加深入地了解复杂网络中的高阶关联关系、网络演化动力学，并可应用于链路预测、社交网络数据分析等。如聚类系数（clustering coefficient）可以看作无向网络中高阶结构关联的度量。文献 [13] 探讨了生物网络中由物种构成的高阶结构之间的交互作用对塑造生态系统多样性的影响。文献 [14] 通过相关实验验证了皮层神经网络中由神经元组织构成的高阶结构之间的固有联系。然而，目前针对高阶结构之间关联性的研究仍然较为有限，且多集中于特定领域（如皮层神经网络和生物多样性网络），这些研究尚未体现通用性。因此，亟须设计通用的算法框架，以深入挖掘高阶结构之间的内在联系。

关联规则作为数据挖掘的一个重要分支，其主要作用是在数据背后发现事物与事物之间可能存在的联系[15]。关联规则也被广泛应用于网络挖掘当中。文献 [16] 提出的 Gspan 算法用于挖掘频繁子图，该算法主要利用剪枝等手段快速挖掘出图集中的频繁子图。文献 [17] 提出了一种在社交媒体网络中挖掘用户之间关联规则的算法框架。该算法挖掘了用户之间的关联规则。文献 [18] 提出了一种通过关联规则检测社交网络中具有影响力的用户的算法，该算法可以作用在大规模网络中。

本文根据传统的关联规则算法，在复杂网络中定义了高阶结构之间的关联规则评判指标，并提出了可以有效挖掘高阶结构间关联规则的通用算法框架。此外，还将所设计的关联规则挖掘算法应用于链路预测问题当中，进而实现了相应的链路预测算法。实验结果表明，相较于基于相似性指标的链路预测基线方法，本文提出的链路预测算法可以提升链路预测准确性。

1　相关理论

1.1　图论相关概念

1）子图

设 $G=(V,E)$，$G'=(V',E')$ 为两个图。其中，V 为 G 的结点集，E 为 G 的边集；V' 为 G' 的结点集，E' 为 G' 的边集。若 $V'\subseteq V$ 且 $E'\subseteq E$，则称 G' 是 G 的子图[19]。

2）导出子图

设 $G=(V,E)$，V' 是 V 的一个非空子集。以 V' 为结点集，以两端点均在 V' 中的边的全体为边集所组成的子图，称为 G 的由 V' 导出的子图，记为 $G[V']$，简称为 G 的导出子图[19]。

3）图的同构

设 $G_1=(V_1,E_1)$，$G_2=(V_2,E_2)$ 为两个图。若存在从 V_1 到 V_2 的双射，使得边之间有如下的关系：设 $u_1\leftrightarrow u_2$，$v_1\leftrightarrow v_2$，$u_1\in V_1$，$v_1\in V_1$，$u_2\in V_2$，$v_2\in V_2$，则 $(u_1,v_1)\in E_1$，当且仅当 $(u_2,v_2)\in E_2$，且 (u_1,v_1) 的重数与 (u_2,v_2) 的重数相同，则称两图同构，记为 $G_1\cong G_2$[19]。

1.2　高阶结构相关概念

1）高阶结构

网络高阶结构指复杂网络中的子网络，可以根据其包含的结点数量对网络高阶结构进行划分。网络中包含 $n(n\geqslant 3)$ 个结点的子网络称为 n 阶高阶结构。网络高阶结构种类随着阶数的增加而增加，有向网络和无向网络中 n 阶高阶结构的种类数量见表 1[20]。

表 1　高阶结构的种类数量

阶数	无向网络	有向网络
3	2	13
4	6	199
5	21	9 364
6	112	1 530 834

如表 1 所示，有向网络中的 3 阶高阶结构种类共有 13 种。有向网络中具体的 13 种 3 阶高阶结构种类信息如图 1 所示。

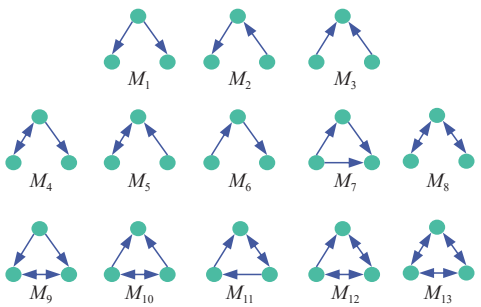


图 1　有向网络中的 3 阶高阶结构种类

将所有有向网络中的 3 阶高阶结构种类的集合用M_s表示，即：

$$M_s = \{M_1, M_2, \cdots, M_{13}\} \tag{1}$$

2）高阶结构之间的关联规则

在真实世界网络中，网络高阶结构之间存在着的一些具有重要意义的内在联系。如图 2 所示，在社交网络中，若已知用户A和用户B是朋友，并且用户B和用户C也是朋友，那么用户A和用户C很有可能也是朋友（社交网络中未被标记的朋友关系），或以后可能成为朋友（社交网络中以后可能存在的朋友关系）。在论文引用网络中，若已知论文A引用了论文B，并且论文B引用了论文C，那么论文A很有可能还引用了论文C，或者说论文A引用了论文C的概率将会提高。

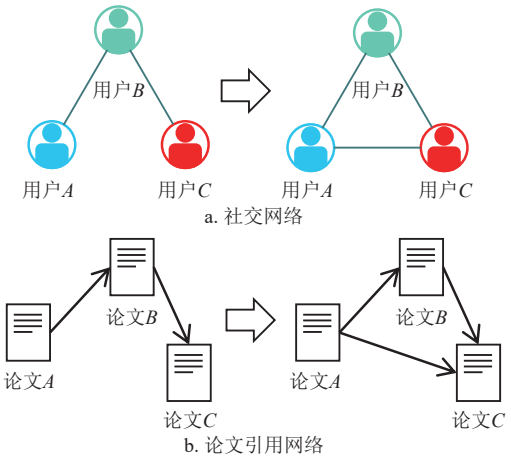


图 2 社交网络和论文引用网络的高阶交互示例

网络中高阶结构之间的内在联系往往错综复杂，研究所有的高阶结构之间的联系十分困难。为此，本文仅对上述这些有着重要意义的高阶结构之间的交互关系进行系统性的研究，并将其定义为高阶结构之间的关联规则，表示为：

$$M_i \Rightarrow M_j \tag{2}$$

式中，高阶结构M_i与高阶结构M_j的阶数相同，并且M_i是M_j的子图。

2 基于高阶结构的关联规则

根据传统的关联规则[21]，本研究定义了高阶结构之间的关联规则评判指标，进而提出了关联规则挖掘算法框架。

2.1 高阶结构之间的关联规则定义

高阶结构之间的关联规则算法评判指标包含支持度、置信度和提升度，具体定义如下。

1）支持度（Support）

支持度用于衡量关联规则的重要性，反映了关联规则中的高阶结构在网络中出现的频繁程度。支持度的计算公式为：

$$\text{Support}(M_i \Rightarrow M_j) = \frac{N(M_i \Rightarrow M_j)}{N_{\text{all}}} \tag{3}$$

式中，N_{all}表示网络中所有n阶子图的数量；$N(M_i \Rightarrow M_j)$表示网络G中满足下述条件的子图数量：该子图H和高阶结构M_i同构并且由H的结点集V_1生成的导出子图$G[V_1]$和高阶结构M_j同构。最小支持度用min_sup表示，为按照实际意义设置的支持度阈值。

2）置信度（Confidence）

置信度用于衡量关联规则的可靠性，反映关联规则中后项对于前项的依赖程度。置信度的计算公式为：

$$\text{Confidence}(M_i \Rightarrow M_j) = \frac{N(M_i \Rightarrow M_j)}{N(M_i)} \tag{4}$$

式中，$N(M_i)$表示网络中和高阶结构M_i同构的子图数量。最小置信度用min_conf表示，为按照实际意义设置的置信度阈值。

3）提升度（Lift）

提升度用于衡量关联规则的相关性，反映关联规则前项对于后项的影响程度。提升度的计算公式为：

$$\text{Lift}(M_i \Rightarrow M_j) = \frac{N(M_i \Rightarrow M_j) \times N_{\text{all}}}{N(M_i) \times N(M_j)} \tag{5}$$

最小提升度用min_lift表示，为按照实际意义设置的提升度阈值。

挖掘的n阶子图看作一条数据，所有n阶子图的集合即为总数据集。并且，每条数据包含的信息除了n阶子图本身外，还包含该n阶子图在网络中对应的导出子图信息。

2.2 高阶结构关联规则挖掘算法框架

根据高阶结构之间的关联规则评判指标，通过子图枚举可以挖掘出给定网络中任意高阶结构之间的关联规则，但在挖掘较大的网络时，时间开销会过大。基于此，在关联规则挖掘算法中使用如下方法降低时间复杂度。该算法在挖掘n阶高阶结构之间的关联规则时，仅通过n阶导出子图枚举实现（可视为n层结点遍历）。该方法可以在一定程度上降低时间复杂度。

以挖掘 3 阶高阶结构之间的关联规则为例，具体的算法框架如图 3 所示。图 3 最左边展示了有向网络中 3 阶高阶结构之间的子图包含关系矩阵，该矩阵记录了高阶结构之间的子图关系。如坐标(3,5)对应的值为 1，表示高阶结构M_5的所有子图中包含 1 个和高阶结构M_3同构的子图。若坐标(i,j)对应的值为 0，表示对应的两个高阶结构之间不存在子图关系。将该矩阵定义为：

$$\boldsymbol{F} \in \mathbb{R}^{13\times 13} \tag{6}$$

矩阵$\boldsymbol{F}$的大小是13×13，矩阵$\boldsymbol{F}$中的坐标(i,j)对应的值表示M_j的所有子图中和M_i同构的子图数量，且$i=1,2,\cdots,13$；$j=1,2,\cdots,13$。矩阵$\boldsymbol{F}$中的数值是固定的，在算法框架中用于结合导出子图的枚举结果，简化关联规则各评判指标的计算。图 3 还展示了关联规则挖掘算法框架。在算法框架中，计算$M_i \Rightarrow M_j$关联规则的步骤如下。

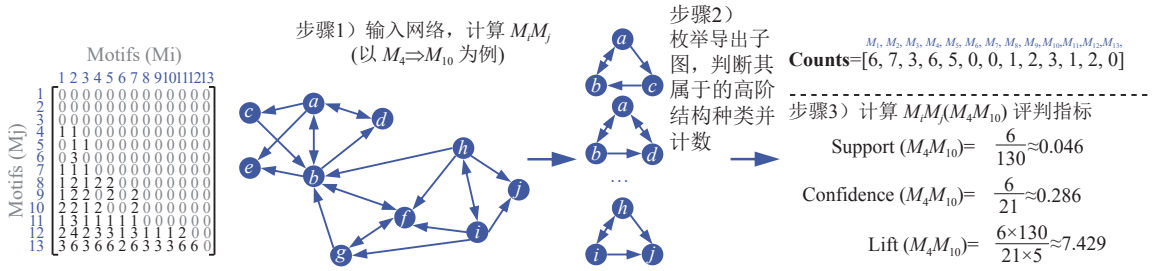


图 3　高阶结构关联规则挖掘算法框架

1）将网络做为输入数据，设定需要计算的关联规则。

2）枚举导出子图。对于每个 3 阶导出子图H，判断H是否与M_s中某个高阶结构$M_x(x=1,2,\cdots,13;\ M_x \in M_s)$同构，并计数。实验中使用计数矩阵 **Counts** 记录和特定高阶结构同构的导出子图数量。该矩阵定义为：

$$\textbf{Counts} \in \mathbb{R}^{1\times 13} \tag{7}$$

式中，矩阵 **Counts** 中所有元素初始值为 0。在枚举过程中，矩阵 **Counts** 的具体计数操作表示为：

$$\text{Counts}[x] = \text{Counts}[x] + 1 \tag{8}$$

式中，$x=1,2,\cdots,13$。枚举完所有导出子图后，得到矩阵 **Counts** 的值。

3）计算关联规则评判指标。根据计数矩阵 **Counts** 和矩阵$\boldsymbol{F}$，首先计算出式 (9)、式 (10) 和 (11) 中的变量N_{all}、$N(M_i \Rightarrow M_j)$和$N(M_x)$的值，再根据式 (3)、式 (4) 和式 (5) 最终计算出支持度、置信度、提升度。变量N_{all}、$N(M_i \Rightarrow M_j)$和$N(M_x)$三者的计算公式为：

$$N_{\text{all}} = \sum_{q=1}^{13}\left(\left(\sum_{p=1}^{13} F[q][p]\right)\times \text{Counts}[q]\right) \tag{9}$$

$$N(M_i \Rightarrow M_j) = F[j][i]\times \text{Counts}[j] \tag{10}$$

$$N(M_x) = \sum_{p=1}^{13}(F[p][x]\times \text{Counts}[p]) + \text{Counts}[x] \tag{11}$$

式中，$i=1,2,\cdots,13$；$j=1,2,\cdots,13$；$x=1,2,\cdots,13$。如式 (9) 所示，如果已知网络中每种模体的数量，又由于每个模体中包含的子图数量是固定的，就可以直接计算出网络中所有子图的数量。式 (10) 和式 (11) 可以用类似方式推导出，根据 2.1 节关于$N(M_i \Rightarrow M_j)$和$N(M_x)$的定义，在已知矩阵 **Counts** 和矩阵$\boldsymbol{F}$的情况下可以直接计算出数值。

2.3　算法伪代码

算法 1 给出了本文提出的有向网络中 3 阶高阶结构关联规则挖掘算法的伪代码。通过算法 1 可以挖掘出 3 阶高阶结构之间的关联规则。

算法 1 高阶结构关联规则挖掘

输入：有向网络$G=(V,E)$，模体信息M_s，子图包含关系矩阵$\boldsymbol{F}$，所需挖掘的关联规则$M_i \Rightarrow M_j$

输出：$M_i \Rightarrow M_j$的支持度、置信度和提升度

创建并初始化计数矩阵 **Counts**

for v_i, v_j, v_k in V do

　$g \leftarrow$ get Subgraph From Nodes(G, v_i, v_j, v_k)

　for M_x in M_s do　　//判断同构并计数

　　if　　g与M_x同构

　　then　Counts$[x] \leftarrow$ Counts$[x]+1$

Support, Confidence, Lift $\leftarrow$ calculate Indicators (Counts, $\boldsymbol{F}$)//计算关联规则

return Support, Confidence, Lift

算法 1 中的函数get Subgraph From Nodes()获取并返回由网络G的结点子集$V'=\{v_i,v_j,v_k\}$构成的导

出子图$G[V']$。在实验中，该函数以及算法 1 中判断g与M_x同构的操作，均依靠 Igraph 库进行实现。算法 1 中的函数calculate Indicators()根据式 (3)、式 (4)、式 (5)、式 (9)、式 (10) 和式 (11) 计算出所需关联规则的支持度、置信度和提升度。

时间复杂度计算如下所示。算法 1 主要分为枚举导出子图并计数模块和计算关联规则指标模块。在枚举导出子图并计数模块，算法的主要时间复杂度是$O(|V|^3|M_s|T)$。其中$|V|$表示结点集的大小，$|M_s|$表示 3 阶模体集合的大小（该数值为 13），T为同构判断操作所需的时间。同构判断的时间复杂取决于图的大小和结构，在算法 1 中，进行同构判断的对象均为 3 阶高阶结构（仅包含 3 个结点的图），故可以将其视为常数时间复杂度。所以，该模块的时间复杂度为$O(|V|^3)$。此外，在计算关联规则指标模块，直接根据公式进行计算各指标，时间复杂度为$O(1)$。综上所述，算法 1 的整体时间复杂度为$O(|V|^3)$。

3 关联规则挖掘实验及分析

3.1 数据集

本文选择了 6 个不同的真实有向网络进行实验，网络数据集信息如下。

Cat[22]：猫上表皮细胞的神经网络。结点表示猫的神经元细胞，边表示这些细胞之间的相互映射关系。

Celegans[23]：一种蠕虫的神经网络。结点表示蛋白质等物质，边代表各物质之间的相互作用。

Yeast[8]：酵母菌代谢网络。结点表示不同的酵母菌，边代表酵母菌之间的代谢关系。

Email[24]：欧洲一家大型研究机构生成的电子邮件通信网络。结点表示不同的成员，边表示成员之间的电子邮件通信关系。

Small[25]：基于论文《The small world problem》[26]的引文网络。每个结点表示不同的论文，边代表论文之间的引用关系。

Wikivote[27]：维基百科中用户选举管理员的投票网络。结点表示不同的用户，边代表投票关系。

表 2 展示了这 6 个真实世界网络的拓扑结构信息。其中，$|V|$表示网络中结点的数量，$|E|$表示网络中有向边的数量，$<k>$表示平均度数，$<d>$表示平均最短路径长度，$<C>$表示平均聚集系数。

表 2 网络数据集信息

网络	$\|V\|$	$\|E\|$	$<k>$	$<d>$	$<C>$
Cat	65	1 139	17.523	1.862	0.661
Celegans	297	2 345	7.896	3.992	0.292
Yeast	688	1 079	1.568	1.440	0.047
Email	1 005	25 571	25.444	2.653	0.399
Small	1 059	4 921	4.647	3.242	0.297
WikiVote	7 115	103 689	14.573	3.341	0.141

3.2 结果与分析

通过实验挖掘了 6 个真实世界网络中所有 3 阶高阶结构之间的关联规则。6 个真实世界网络中挖掘出的关联规则的支持度、置信度和提升度散点分布图如图 4 所示。可以看出，不同网络的 3 阶高阶结构间的关联规则分布不同，存在不同程度的差异。6 个真实网络的散点大多集中于图中的左下角位置，这表明真实网络的高阶结构之间虽然存在着许多关联规则，但大部分关联规则的支持度、置信度和提升度均较低，意味着其为较弱的关联规则。6 个真实网络中均存在少量拥有较高支持度、置信度和提升度的关联规则，对应的散点一般分布在图中的左上角或右下角，这些关联规则一般被视为其中的强关联规则。此实验结果表明有向网络中部分高阶结构之间存在内在关联，进而揭示了有向网络内部深层次的结构特性。实验结果还表明不同类型网络中存在关联的高阶结构是不同的。因此，与聚类系数类似，高阶结构之间的关联规则是分析有向网络结构特性的重要度量。具体而言，6 个真实网络中的散点数量是不同的，部分网络中散点数量明显较少，这与网络的稀疏性和网络类型有关。

根据挖掘出的关联规则的支持度、置信度和提升度数值分布及相关经验，将最小支持度min_sup设置为 0.01，最小置信度min_conf设置为 0.1，最小提升度min_lift设置为 0.6，筛选出 6 个真实网络中的强关联规则。表 3 列出了 6 个真实网络中强关联规则按置信度降序的 TOP3 规则（当强关联规则不足 3 条时，展示筛选出的所有强关联规则）。可以看出，不同真实网络中的强关联规则的支持度、置信度和提升度数值是不相同的。在 Cat 网络中，$M_{12} \Rightarrow M_{13}$规则为置信度最高的强关联规则，该关联规则的支持度为 0.050，置信度为 0.821，提升度为 98.32。在 Celegans 网络中，$M_7 \Rightarrow M_{10}$规则为置信度最高的强关联规则，该关联规则的支持度为 0.032，置信度为 0.246，提升度为 11.034。在 Yeast 网络中，$M_2 \Rightarrow M_7$规则为置信

度最高的强关联规则，该关联规则的支持度为 0.265，置信度为 0.181，提升度为 0.655。在 Email 网络中，$M_{11} \Rightarrow M_{12}$规则为置信度最高的强关联规则，该关联规则的支持度为 0.011，置信度为 0.502，提升度为 34.557。在 Small 网络中，$M_3 \Rightarrow M_7$规则为置信度最高的强关联规则，该关联规则的支持度为 0.325，置信度为 0.214，提升度为 0.652。在 WikiVote 网络中，$M_5 \Rightarrow M_9$规则为置信度最高的强关联规则，该关联规则的支持度为 0.181，置信度为 0.156，提升度为 11.718。表 3 中网络的强关联规则可以应用于链路预测等多个领域。如 Small 网络中存在强关联规则$M_3 \Rightarrow M_7$，表明网络中两篇论文A和B同时引用论文C的话，论文A和B之间也有可能存在引用关系（A引用B或B引用A）。而关联规则的支持度、置信度和提升度可以量化这种可能性。

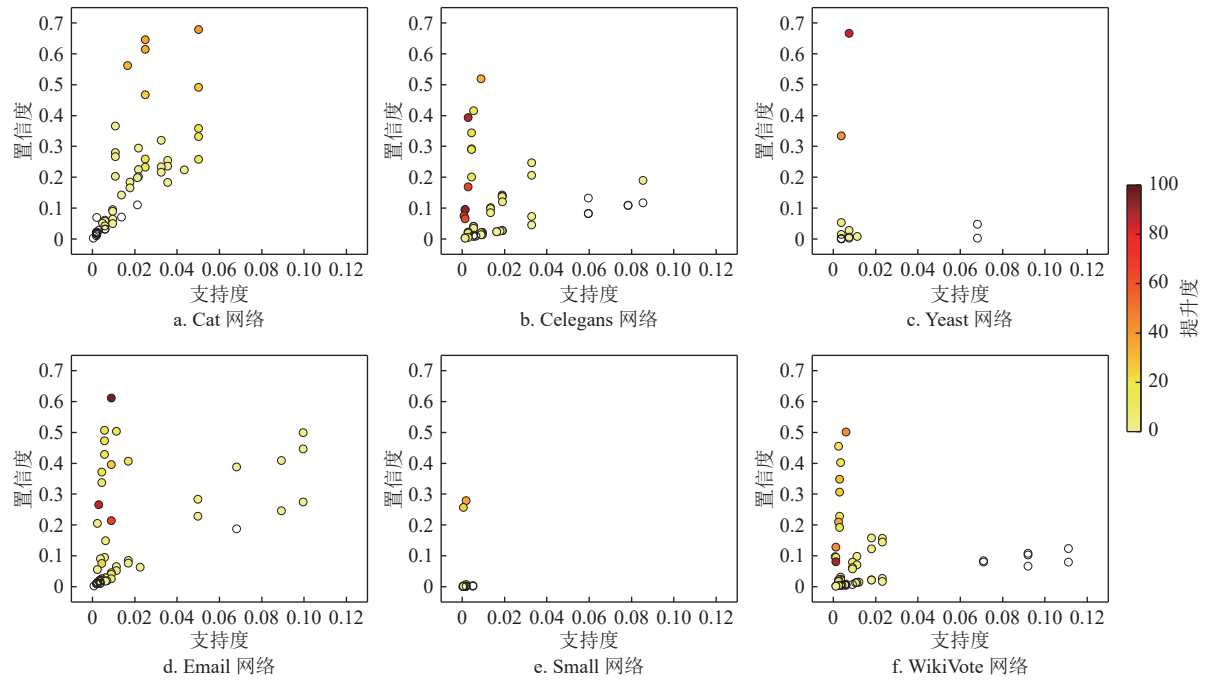


图 4　支持度、置信度和提升度散点分布图

表 3　强关联规则信息

网络	强关联规则TOP3	关联规则指标		
		支持度	置信度	提升度
Cat	$M_{12} \Rightarrow M_{13}$	0.050	0. 821	98.32
	$M_{11} \Rightarrow M_{13}$	0.050	0.678	81.148
	$M_9 \Rightarrow M_{13}$	0.025	0.644	77.168
Celegans	$M_7 \Rightarrow M_{10}$	0.032	0.246	11.034
	$M_4 \Rightarrow M_{10}$	0.032	0.206	9.226
Yeast	$M_1 \Rightarrow M_4$	0.085	0.188	1.186
	$M_2 \Rightarrow M_7$	0.265	0.181	0.655
Email	$M_{11} \Rightarrow M_{12}$	0.011	0.502	34.557
	$M_5 \Rightarrow M_8$	0.099	0.498	8.320
	$M_4 \Rightarrow M_8$	0.099	0.446	7.442
Small	$M_3 \Rightarrow M_7$	0.325	0.214	0.652
	$M_5 \Rightarrow M_9$	0.181	0.156	11.718
WikiVote	$M_7 \Rightarrow M_{10}$	0.023	0.156	9.814
	$M_4 \Rightarrow M_{10}$	0.023	0.144	9.070

此外，各网络中的强关联规则数量也不同。实验中，Cat 网络有 31 条强关联规则，Celegans 网络有 12 条强关联规则，Yeast 网络有 1 条强关联规则，Email 网络有 11 条强关联规则，Small 网络有 1 条强关联规则，WikiVote 网络有 7 条强关联规则。该实验结果与网络中设置的最小支持度、最小置信度和最小提升度阈值有关。这表明，针对具体的网络，设置更加适当的阈值，会得到更好的挖掘结果。

4　基于高阶结构关联规则的应用

本节将高阶结构之间的关联规则应用于链路预测，进而提出了基于高阶结构关联规则的链路预测方法，并与有向网络中的链路预测经典方法进行性能对比实验。

4.1　基于高阶结构关联规则的链路预测方法

本文提出的基于高阶结构关联规则的链路预测方法（link prediction based on association rules of higher-order structures, ARHS），将高阶结构之间的强关联规则作为预测器，置信度作为对应预测器

的预测分数，进行链路预测。

在链路预测方法 ARHS 中，预测(x,y)为链接可能性分数S_{xy}定义为：

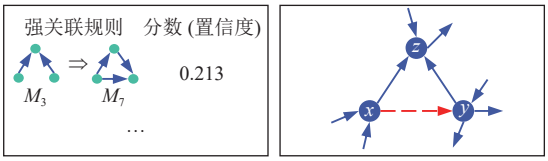
$$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}S(x,y,z) \tag{12}$$

式中，$\Gamma(x)$表示结点x在该有向网络的基图中的邻居结点集合；$S(x,y,z)$表示基于结点x、y和z对有序对(x,y)产生链接可能性的预测分数，表示为：

$$S(x,y,z)=\begin{cases}\text{confidence}, G[x,y,z]\Rightarrow G[x,y,z]+(x,y)\text{为强关联规则}\\ 0,\text{其他情况}\end{cases} \tag{13}$$

式中，$G[x,y,z]$表示在预测网络G中由结点集$\{x,y,z\}$构成的导出子图。(x,y)表示需要预测的有序对。$S(x,y,z)$表示将预测网络中的图$G[x,y,z]$视为M_i，假设(x,y)为边，将由图$G[x,y,z]$加上边(x,y)构成的图视为M_j。若关联规则$M_i\Rightarrow M_j$为强关联规则，则赋予该强关联规则对应的置信度分数，否则为 0。

提出的链路预测方法的预测流程如图 5 所示。



a. 预测器 b. 链路预测示例

图 5 链路预测流程

如图 5a 所示，将预测网络中的所有强关联规则作为预测器。如图 5b 所示，对于预测(x,y)为链接可能性时，由结点x和y的所有共同邻居（此处仅有结点z）决定。具体地，结点x、y和z在预测网络中对应的导出子图为$H_1=(\{x,y,z\},\{(x,z),(y,z)\})$，其与模体$M_3$同构。且该导出子图加上边$(x,y)$组成的图为$H_2=(\{x,y,z\},\{(x,z),(y,z),(x,y)\})$，其与模体$M_7$同构。由于$M_3\Rightarrow M_7$为预测网络中的强关联规则，故$(x,y)$为链接可能性分数为 0.213，即强关联规则$M_3\Rightarrow M_7$对应的置信度数值。

4.2 对比实验

链路预测方法已经被广泛地研究[28-30]，但大部分方法是基于无向网络的。基于相似性指标的方法由于其简洁性和具有较强的可解释性等特点，在多个相关领域中被广泛应用[31]。并且，此类算法可以改进后应用于有向网络链路预测中。本文选择其中的 7 种基于相似性指标的链路预测方法作为基准方法。每种方法公式如下所示。

1）有向共同邻居指标（DCN）[32]：

$$S_{xy}^{\text{DCN}}=|\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)| \tag{14}$$

式中，$\Gamma_{\text{out}}(x)$表示以结点x为起始点的邻居结点集；$\Gamma_{\text{in}}(x)$表示以结点y为终点的邻居结点集。

2）有向 Adamic-Adar 指标（DAA）[32]：

$$S_{xy}^{\text{DAA}}=\sum_{z\in\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)}\frac{1}{\log(k_{\text{out}}(z))} \tag{15}$$

式中，$k_{\text{out}}(z)$表示以结点z为始点的邻居结点集的大小，又称出度。

3）有向资源分配指标（DRA）[32]：

$$S_{xy}^{\text{DRA}}=\sum_{z\in\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)}\frac{1}{k_{\text{out}}(z)} \tag{16}$$

4）有向 Jaccard 指标（DJA）[33]：

$$S_{xy}^{\text{DJA}}=\frac{|\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)|}{|\Gamma_{\text{out}}(x)\cup\Gamma_{\text{in}}(y)|} \tag{17}$$

5）有向 Sorensen 指标（DSO）[33]：

$$S_{xy}^{\text{DSO}}=\frac{|\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)|}{|\Gamma_{\text{out}}(x)+\Gamma_{\text{in}}(y)|} \tag{18}$$

6）有向 HPI 指标（DHPI）[34]：

$$S_{xy}^{\text{DHPI}}=\frac{|\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)|}{\min\{\Gamma_{\text{out}}(x),\Gamma_{\text{in}}(y)\}} \tag{19}$$

7）有向 HDI 指标（DHDI）[35]：

$$S_{xy}^{\text{DHDI}}=\frac{|\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)|}{\max\{\Gamma_{\text{out}}(x),\Gamma_{\text{in}}(y)\}} \tag{20}$$

在对比实验中，性能度量采用 AUC（area under curve）指标[36]。对比实验采用 6 个真实世界的有向网络。将网络已有边集按照 9:1 的比例划分为训练集和测试集，分别抽取 20 次。通过 20 次实验得到的 AUC 平均值来衡量方法的性能，实验结果如表 4 所示（表中加粗数字表示对应网络中 AUC 最大值）。从表 4 的预测结果可以看出，提出的 ARHS 方法在其中 4 个网络（Cat 网络、Celegans 网络、Email 网络和 WikiVote 网络）中 AUC 值最大，表示其预测性能最优，且在 Celegans 网络和 Cat 网络中预测精度有显著提高，分别比最优基准方法的 AUC 提升了 3.5% 和 6.8%。上述结果表明，将高阶结构应用于链路预测中能够得到有效的链路预测方法。

表 4　AUC 结果对比

网络	方法							
	DCN	DAA	DRA	DJA	DSO	DHPI	DHDI	ARHS
Cat	0.871	0.886	0.887	0.887	0.887	0.862	0.875	**0.892**
Celegans	0.796	0.783	0.789	0.770	0.770	0.779	0.767	**0.819**
Yeast	0.518	0.513	0.516	**0.519**	0.514	0.513	0.518	0.518
Email	0.942	0.921	0.948	0.935	0.918	0.923	0.929	**0.952**
Small	**0.687**	0.683	0.686	0.685	0.685	0.686	0.685	0.683
WikiVote	0.924	0.923	0.922	0.921	0.920	0.921	0.916	**0.928**

5　结束语

复杂网络中的高阶结构之间可能存在着复杂的内在联系，通过分析高阶结构间的内在联系可以更加深入地了解复杂网络。本文根据传统的关联规则算法，在复杂网络中定义了高阶结构之间的关联规则评判指标，并提出了可以有效挖掘高阶结构间关联规则的算法框架。为了验证算法，在 6 个真实网络中进行了相关实验，挖掘出了高阶结构间的关联规则。实验结果表明，不同网络中高阶结构之间存在关联规则，并且不同网络中的关联规则存在差异性。在此基础上，本研究将高阶结构之间的关联规则应用于链路预测当中，进而实现了链路预测算法。对比实验结果验证了所实现的链路预测算法的有效性。本研究挖掘出的高阶结构之间的关联规则可揭示复杂网络中高阶结构之间存在的一些重要的内在联系，且可用于推断网络中潜在的关系和分析网络演化动力学。

本文提出的挖掘高阶结构之间关联规则的算法框架存在局限性。该算法不太适用于具有较小平均度数等性质的复杂网络。且在挖掘更高阶的高阶结构之间的关联规则时，算法的时间复杂度会随高阶结构的阶数提高而提高。

参 考 文 献

[1] SBARABASI A L. The network takeover[J]. Nature Physics, 2012, 8(1): 14-16.

[2] VESPIGNANI A. Modelling dynamical processes in complex socio-technical systems[J]. Nature Physics, 2012, 8(1): 32-39.

[3] YU Z, LU S, WANG D, et al. Modeling and analysis of rumor propagation in social networks[J]. Information Sciences, 2021, 580(1): 857-873.

[4] YOUNIS H, ANWAR M W, KHAN M U G, et al. A new sequential forward feature selection algorithm for mining best topological and biological features to predict protein complexes from protein–protein interaction networks[J]. Interdisciplinary Sciences: Computational Life Sciences, 2021, 13(3): 371-388.

[5] BATTISTON F, CENCETTI G, IACOPINI I, et al. Networks beyond pairwise interactions: Structure and dynamics[J]. Physics Reports, 2020, 874(1): 1-92.

[6] BATTISTON F, AMICO E, BARRAT A, et al. The physics of higher-order interactions in complex systems[J]. Nature Physics, 2021, 17(10): 1093-1098.

[7] BENSON A R, GLEICH D F, LESKOVEC J. Higher-order organization of complex networks[J]. Science, 2016, 353(6295): 163-166.

[8] MILO R, SHEN ORR S, ITZKOVITZ S, et al. Network motifs: Simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.

[9] BALAZSI G, BARABASI A L, OLTVAI Z N. Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli[J]. Proceedings of the National Academy of Sciences, 2005, 102(22): 7841-7846.

[10] BATTISTON F, NICOSIA V, CHAVEZ M, et al. Multilayer motif analysis of brain networks[J]. Chaos (Woodbury, NY), 2017, 27(4): 047404.

[11] CHEN Y, WANG J, LI P, et al. Single document keyword extraction via quantifying higher-order structural features of word co-occurrence graph[J]. Computer Speech and Language, 2019, 57(1): 98-107.

[12] LOTITO Q F, MUSCIOTTO F, MONTRESOR A, et al. Higher-order motif analysis in hypergraphs[J]. Communications Physics, 2022, 5(1): 79.

[13] BAIREY E, KELSIC E D, KISHONY R. High-order species interactions shape ecosystem diversity[J]. Nature Communications, 2016, 7(1): 12285.

[14] YU S, YANG H, NAKAHARA H, et al. Higher-order interactions characterized in cortical activity[J]. Journal of Neuroscience, 2011, 31(48): 17514-17526.

[15] BERLIN. Association rule mining: models and algorithms[M]. Berlin: Springer Berlin Heidelberg, 2002.

[16] YAN X, HAN J. gSpan: Graph-based substructure pattern mining[C]//the 2nd IEEE International Conference on Data Mining. Maebashi: IEEE, 2002: 721-724.

[17] KRUSE R, LOKUKATAGODA T, ALKHUSHAYNI S. A framework for association rule learning with social media networks[J]. IOP SciNotes, 2022, 3(1): 15001.

[18] AGOUTI T. Graph-based modeling using association rule mining to detect influential users in social networks[J]. Expert Systems with Applications, 2022: 117436.

[19] GROSS J L, YELLEN J. Graph theory and its applications[M]. Boca Raton: CRC Press, 2005.

[20] HARARY F, PALMER E M. Graphical enumeration[M]. [S.l.]: Academic Press, 1973.

[21] LAROSE D T, LAROSE C D. Discovering knowledge in data: An introduction to data mining[M]. New Jersey: Wiley, 2014.

[22] DE REUS M A, VAN DEN HEUVEL M P. Rich club organization and intermodule communication in the cat connectome[J]. Journal of Neuroscience, 2013, 33(32): 12929-12939.

[23] WATTS D J, STROGATZ S H. Collective dynamics of ‘small-world’ networks[J]. Nature, 1998, 393(6684): 440-

442.
[24] YIN H, BENSON A R, LESKOVEC J, et al. Local higher-order graph clustering[C]//the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM, 2017: 555-564.
[25] GARFIELD E. From computational linguistics to algorithmic historiography[J]. Knowledge and Language: Building Large-Scale Knowledge Bases for Intelligent Applications, 2001, 18(1): 1-3.
[26] MILGRAM S. The small world problem[J]. Psychology Today, 1967, 2(1): 60-67.
[27] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Predicting positive and negative links in online social networks[C]//the 19th International Conference on World Wide Web. North Carolina: ACM, 2010: 641-650.
[28] 李治成, 吉立新, 刘树新, 等. 基于拓扑有效连通路径的有向网络链路预测方法[J]. 电子科技大学学报, 2021, 50(1): 127-137.
LI Z C, JI L X, LIU S X, et al. A method of link prediction in directed network based on effective connectivity path[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(1): 127-137.
[29] 方祺娜, 许小可. 基于异质模体特征的社交网络链路预测[J]. 电子科技大学学报, 2022, 51(2): 274-281.
FANG Q N, XU X K. Link prediction by hetergeneous motifs in social networks[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(2): 274-281.
[30] 佘美富, 王逸伟, 张建章等. 超图环境下链路预测问题的探究[J]. 西南大学学报 (自然科学版), 2023, 45(8): 61-75.
SHE M F, WANG Y W, ZHANG J Z, et al. Exploration of link prediction via hypergraph[J]. Journal of Southwest University(Natural Science Edition), 2023, 45(8): 61-75.
[31] 吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社, 2013.
LYU L Y, ZHOU T. Link prediction[M]. Beijing: Higher Education Press, 2013.
[32] ZHANG X, ZHAO C, WANG X, et al. Identifying missing and spurious interactions in directed networks[J]. International Journal of Distributed Sensor Networks, 2015, 11(9): 507386.
[33] LIBEN-NOWELL D, KLEINBERG J. The link prediction problem for social networks[C]//Proceedings of the 12th International Conference on Information and Knowledge Management. Halifax: ACM, 2003: 556-559.
[34] RAVASZ E, SOMERA A L, MONGRU D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1551-1555.
[35] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661.
LYU L Y. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.
[36] HANLEY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic curve[J]. Radiology, 1982, 143(1): 29-36.

编 辑 叶 芳