

基于模体的朴素贝叶斯链路预测方法

曾 茜, 韩 华, 马媛媛

(武汉理工大学 理学院, 武汉 430070)

摘 要: 在具有模体特征的食物链网络、社交网络中,局部朴素贝叶斯(LNB)的链路预测方法通过准确区分每个共邻节点的贡献以提高链路预测的精确度,但忽略了每个共邻节点对所在路径的贡献不同以及网络模体结构对链接形成的作用。针对LNB链路预测方法存在的局限性问题,结合路径模体特征与朴素贝叶斯理论,提出基于模体的朴素贝叶斯链路预测方法。定义模体密度以量化路径结构上模体的聚集程度。考虑路径结构上模体密度对链接形成的影响,构建每条路径的角色贡献函数,以量化每条路径结构的模体特征对节点相似性的影响。在此基础上,根据朴素贝叶斯理论与角色贡献函数推导节点相似性指标。在Football、USAir、C.elegans、FWMW、FWEW和FWFW 6个真实网络上进行实验,结果表明,该方法能有效提高预测性能且具有较优的鲁棒性,其中在具有显著模体特征的FWMW、FWEW、FWFW网络上,相比现有相似性指标中较优的Katz指标,所提相似性指标的AUC值提升了2%~7%。

关键词: 复杂网络;链路预测;朴素贝叶斯;相似性指标;模体密度

开放科学(资源服务)标志码(OSID):



中文引用格式:曾茜,韩华,马媛媛.基于模体的朴素贝叶斯链路预测方法[J].计算机工程,2022,48(10):95-102.

英文引用格式:ZENG X, HAN H, MA Y Y. Naive Bayes link prediction method based on motif[J]. Computer Engineering, 2022, 48(10): 95-102.

Naive Bayes Link Prediction Method Based on Motif

ZENG Xi, HAN Hua, MA Yuanyuan

(School of Science, Wuhan University of Technology, Wuhan 430070, China)

[Abstract] In food chain networks and social networks with motif features, Local Naive Bayes (LNB) link prediction method improves the accuracy of link prediction by accurately distinguishing the contribution of each common neighbor node, but neglects the different contributions of each common neighbor node to the path and the role of the model structure in the network on the link formation. To address the limitations of LNB link prediction methods, this study proposes a motif-based naive Bayes link prediction method by combining path-motif features and naive Bayes theory. The motif density is defined to quantify the degree of aggregation of motifs on the path structure. Considering the influence of the motif density on the path structure on the link formation, the role contribution function of each path is constructed to quantify the impact of the motif features of each path structure on the similarity of nodes. Then, the similarity index of nodes is derived according to the naive Bayes theory and the role contribution function. Experiments on Football, USAir, C.elegans, FWMW, FWEW and FWFW networks show that the proposed method can effectively improve the prediction performance and has better robustness. On the FWMW, FWEW, and FWFW networks with obvious motif features, the AUC value of the proposed similarity index increased by 2%~7% compared with the suboptimal Katz index among the existing similarity indexes.

[Key words] complex network; link prediction; naive Bayes; similarity index; motif density

DOI: 10.19678/j.issn.1000-3428.0062847

0 概述

随着网络科学的不断发展,在社会科学、自然科学、信息科学等领域中的复杂关系问题都可以通过复杂网络来描述^[1]。在复杂网络中通常出现网络未知或部分未知的情况,因此,对缺失信息的还原和预测是网络研究过程的关键。链路预测是根据已知的

网络信息,预测尚未形成连边的两个节点之间产生链接的可能性,以预测未知链接^[2]和未来链接^[3]。链路预测被广泛应用在不同领域中,例如,在蛋白质网络中探究蛋白质之间的相互作用^[4-5],在电商网络中推送客户感兴趣的产品^[6],在社交网络中推荐可能认识的人^[7],在航空网络中推测影响网络演化的重要因素^[8]等。

基金项目:国家自然科学基金(12071364);国家自然科学基金青年科学基金项目(11701435)。

作者简介:曾茜(1997—),女,硕士研究生,主研方向为复杂网络分析;韩华,教授、博士;马媛媛,硕士研究生。

收稿日期:2021-09-29 **修回日期:**2021-11-18 **E-mail:** zengxi0201@163.com

现有链路预测方法从结构相似性^[9]角度可以分为基于局部结构的方法(如共邻节点指标(CN)^[10]、Adamic-Adar(AA)^[11]、资源分配指标(RA)^[12]等)、基于半局部结构的方法(如局部路径指标(LP)^[13])和基于全部结构的方法(如考虑全局路径的Katz指标^[14]、节点偏好性随机游走指标(DRW)^[15])。基于局部结构的方法因计算复杂度低、适用性广等优点备受研究人员的关注^[16],但这类方法大多简单地假设每个共邻节点对链路形成的贡献一致。为此,文献[17]提出局部朴素贝叶斯(LNB)模型,引入朴素贝叶斯理论来区分不同共邻节点的贡献,取得了较优的预测效果。文献[18]提出树增广朴素贝叶斯预测(TAN)方法,引入树增强朴素贝叶斯概率模型,缓解在共邻节点之间强独立性假设的问题。文献[19]提出扩展局部朴素贝叶斯(ELNB)方法,通过对共邻节点的角色贡献函数进行扩展,揭示了微观尺度下节点聚类系数对链路形成的作用。但这些方法都是基于共邻节点的角色贡献展开研究,忽略了路径结构特征的贡献。近年来,已有研究表明,考虑路径结构周围的拓扑信息能有效提升预测性能^[20-21]。

此外,LNB模型及其相关改进方法并未考虑网络中的模体结构特征,而真实网络(如食物链网络、社交网络)存在大量模体结构。针对含有模体特征的网络,LNB模型仍存在局限性。网络模体的概念最早由文献[22]提出,即在真实网络中富集出现的由少量节点形成的小规模同构子图。文献[23]提出模体顶点度和边度来衡量网络中点和边的重要性。文献[24]提出三角模体度和四边模体度的代数算法,设计基于模体特征的攻击策略。文献[25]依据朴素贝叶斯理论解释了使用模体边度进行链路预测的可行性,提出基于单模体边度和双模体边度的链路预测方法,揭示了模体对链路形成的重要作用。随着模体在复杂网络中深入研究,模体顶点度和边度已经不足以描述更复杂的拓扑特征,基于路径结构的模体测量有待提出。

本文结合路径模体特征与朴素贝叶斯理论,提出一种基于模体的局部朴素贝叶斯链路预测方法MLNB,以分析路径结构的模体特征对节点相似性的影响。定义路径结构上三角模体的聚集程度——模体密度,考虑模体密度对待测连边的影响,通过构建路径的角色贡献函数分析路径的相似性贡献,同时推导出基于共邻节点的扩展指标。

1 基本概念

1.1 问题描述

本文给定一个无权无向网络 $G=(V,E)$,不考虑网络中的重边和自环,其中 V 和 E 分别表示网络中所有节点的集合和所有边的集合,全集 U 表示所有可能边的集合。 $e_{xy} \in E$ 表示节点 x 和节点 y 存在连边, $\bar{e}_{xy} \notin E$ 表示节点 x 和节点 y 不相连。网络中节点

总数为 $|V|=n$,边的总数为 $|E|=m$,最大可能的边的数量为 $|U|=\frac{n(n-1)}{2}$ 。节点 x 的邻居集合用 $\Gamma(x)$ 来表示,节点 x 和节点 y 的共邻节点集合表示为 $\Gamma(x,y)=\Gamma(x) \cap \Gamma(y)$ 。链路预测的目标是寻找集合 $U-E$ 中缺失或暂未连接的边。

1.2 相关指标

在现有链路预测方法中常用的相似性指标主要有8个:

1)共邻节点指标。对于待预测节点 x 和节点 y ,如果共邻节点个数越多,则 x 和 y 连边的可能性越大。通过节点对 x 和 y 的相似性得分计算它们共邻节点的个数,如式(1)所示:

$$S_{xy}^{CN} = |\Gamma(x,y)| \quad (1)$$

2)Adamic-Adar指标。AA指标考虑共邻节点的度越小对链路形成的贡献越大,在CN指标的基础上为每个共邻节点分配一个权重,权值定义为该共邻节点度的对数的倒数。AA指标定义如式(2)所示:

$$S_{xy}^{AA} = \sum_{\omega \in \Gamma(x,y)} \frac{1}{\log_a k_{\omega}} \quad (2)$$

其中: k_{ω} 为节点 ω 的度。

3)资源分配指标。对于未连接的节点 x 和节点 y ,在资源从节点 x 传递到节点 y 的过程中,每个共邻节点的资源传输量与其度值成反比。RA指标定义如式(3)所示:

$$S_{xy}^{RA} = \sum_{\omega \in \Gamma(x,y)} \frac{1}{k_{\omega}} \quad (3)$$

4)基于CN指标的局部朴素贝叶斯相似性指标(LNBCN)。在CN指标的基础上,LNBCN^[17]考虑不同共邻节点对链路形成的贡献不同,定义了角色函数 R_{ω} 来度量每个共邻节点的贡献,如式(4)所示:

$$S_{xy}^{LNBCN} = \sum_{\omega \in \Gamma(x,y)} (\log_a s + \log_a R_{\omega}) \quad (4)$$

其中: s 为节点 x 和 y 相连的概率与不相连概率的比值。

5)基于AA指标的局部朴素贝叶斯相似性指标(LNBAA)。将局部朴素贝叶斯原理与AA指标相结合,LNBAA定义如式(5)所示:

$$S_{xy}^{LNBAA} = \sum_{\omega \in \Gamma(x,y)} \frac{1}{\log_a k_{\omega}} (\log_a s + \log_a R_{\omega}) \quad (5)$$

6)基于RA指标的局部朴素贝叶斯相似性指标(LNBRA)。将局部朴素贝叶斯原理与RA指标相结合,LNBRA定义如式(6)所示:

$$S_{xy}^{LNBRA} = \sum_{\omega \in \Gamma(x,y)} \frac{1}{k_{\omega}} (\log_a s + \log_a R_{\omega}) \quad (6)$$

7)局部路径指标。不同于上述6种局部指标,LP指标是一种半局部指标,不仅考虑了二阶路径(即共邻节点)的贡献,也考虑了三阶路径对节点相似性的贡献。LP指标计算如式(7)所示:

$$S^{LP} = A^2 + \alpha A^3 \quad (7)$$

其中: A 为网络的邻接矩阵; A^2 和 A^3 分别为待测节点对的二阶路径数和三阶路径数; α 为控制三阶路径权

重的可调参数,一般取值为0.01。

8) Katz 指标。全局相似性指标 Katz 计算节点 x 和 y 之间所有路径的贡献,其定义如式(8)所示:

$$S_{xy}^{\text{Katz}} = \sum_{l=1}^{\infty} \alpha^l \cdot |\text{path}_{x,y}^l| = \alpha^1 A_{xy}^1 + \alpha^2 A_{xy}^2 + \cdots + \alpha^n A_{xy}^n \quad (8)$$

其中: $|\text{path}_{x,y}^l|$ 为节点 x 和 y 之间 l 阶路径的数目; α 为控制各个路径权重的参数。

2 本文方法

2.1 模体理论

模体是一种介于节点和社团之间的网络子图,是在真实网络中频繁出现的结构单元,可以很好地反映网络的结构和功能。模体的基本特性是在真实网络中出现的频率远高于其在规模相同的随机网络中出现的频率。模体的基本统计特征如下:

1) 模体的频率。对于给定的含有 n 个节点的子图 M , 如果子图 M 是网络的模体, 那么模体 M 的频率^[22]定义如式(9)所示:

$$f(M) = \frac{n(M)}{N} \quad (9)$$

其中: $n(M)$ 表示该子图在真实网络中出现的次数; N 表示含有 n 个节点的子图出现的总次数。

2) 模体的 P 值。模体 M 在随机网络中出现次数的频率大于节点数量相同的真实网络中出现次数的频率。 P 值越小,说明真实网络的模体特征越明显^[22]。

3) 模体的 Z 得分。对于模体 M_i , $N_{\text{real } i}$ 表示该模体在真实网络中出现的次数, $N_{\text{rand } i}$ 表示该模体在随机网络中出现的次数。 $N_{\text{rand } i}$ 的平均值为 $\langle N_{\text{rand } i} \rangle$, 标准差为 $\sigma_{\text{rand } i}$, 则模体 M_i 在真实网络中的 Z 得分^[22]如式(10)所示:

$$Z_i = \frac{N_{\text{real } i} - \langle N_{\text{rand } i} \rangle}{\sigma_{\text{rand } i}} \quad (10)$$

网络中模体的存在性可以根据上述模体概念和基本统计特征来检验。在大多数无向网络中,三节点模体是最常见的模体结构。三节点构成的两种模体结构如图1所示。 Δ 型三角模体作为网络中最小的完全图,构成网络中信息传播的局部单元,能反映网络局部结构中的聚集特性。本文基于网络的三角模体特征,提出针对三角模体结构的链路预测方法。

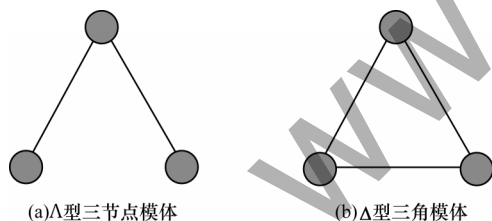


图1 在无向网络中三节点模体结构示例

Fig.1 Examples of three-nodes motif structure in undirected network

2.2 模体密度的分析与量化

LNBCN 指标在 CN 指标的基础上,区分每个共邻节点对待测连边的不同贡献,进一步提高预测精度,但忽略了经过不同共邻节点的路径对待测连边

贡献的差异性。在路径上三角模体的聚集程度即模体密度,对待测节点间连边的产生具有重要影响。本文引入不同的局部结构图进行对比分析,3种不同的节点连接方式如图2所示。

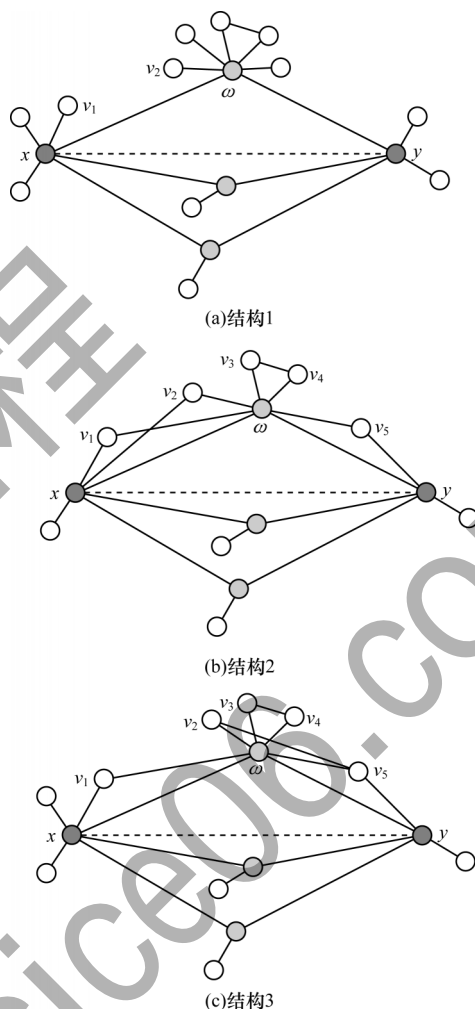


图2 3种不同的节点连接方式

Fig.2 Three different node connection methods

从图2可以看出,在3种结构中均有一对待测节点 x 和 y ,以及待分析的路径结构 w_{xoy} 。从图2(a)和图2(b)可以看出,节点 x 和节点 y 的度值都相同,3个共邻节点的度值也都相同。但结构1中路径 w_{xoy} 的 Δ 型模体都不构成 Δ 型模体,例如 $\Delta_{x\omega v_1}$ 和 $\Delta_{x\omega v_2}$ 。在结构2的路径 w_{xoy} 上 Δ 型模体的聚集程度更高,这说明结构1的路径 w_{xoy} 对 Δ 型模体形成 Δ 型模体起到抑制作用,结构2的路径 w_{xoy} 对形成 Δ 型模体有促进作用。对于三节点模体 Δ_{xoy} , 结构2比结构1形成三角模体 Δ_{xoy} 的可能性更大,即节点 x 和节点 y 产生连边的可能性更大。从图2(b)和图2(c)可以看出,3个共邻节点既具有相同的度值,邻居之间的连边数也相同。根据 LNBCN 原理,即用节点聚类系数来区分和量化不同共邻节点的贡献,由于在结构2和结构3中共邻节点 ω 的聚类系数相同,因此对待测边的贡献相同。从图2(c)可以看出,以节点 ω 为共邻节点的任意待测节点对,例如 (v_1, v_2) 和 (v_3, v_5) , 节点 ω 对它们连边的贡献度都是相同的。LNBCN 方法仅考虑共邻节点本身的局部特征属性的差异化,却

忽略了共邻节点与其待测节点之间连边的局部结构差异。在结构2中路径 w_{xoy} 上的 Δ 型模体个数明显更多,模体聚集程度明显高于 Δ 型模体,说明结构2中路径 w_{xoy} 对形成 Δ 型模体 $\Delta_{x,\omega,y}$ 的促进作用更大,节点 x 和节点 y 更有可能形成连边。在结构3中3对待测节点 (x,y) 、 (v_1,v_2) 和 (v_3,v_5) ,路径 w_{xoy} 、 $w_{v_1\omega v_2}$ 和 $w_{v_3\omega v_5}$ 上的 Δ 型模体聚集程度各不相同,分别对待测节点产生连边的影响程度也不相同,不能简单地用同一节点聚类系数来度量各自的贡献。此外,从信息传播路径的角度,共邻节点提供二阶传播路径,模体聚集程度越高的二阶路径结构也能提供更多的三阶传播路径,以图2(b)为例,三阶传播路径有 $w_{xv_1\omega y}$ 、 $w_{xv_2\omega y}$ 和 $w_{xv_3\omega y}$,为节点 x 和 y 之间信息传播提供更大的可能性。

上述分析表明,节点 x 和 y 产生连边的可能性会因路径模体特征的不同而不同。因此,本文考虑到每条路径的模体特征对相似性有一定的影响,通过定义新的模体测度来描述路径结构上模体的聚集程度。

在复杂网络中边聚类系数是刻画局部三角环聚集程度的重要参数。根据边聚类系数的定义,即一条边的两个端点与其共邻节点之间所构成的三角形数与所有可能包含该边的三角形数的比值^[26],图2(b)中节点 x 和节点 ω 形成的连边 $e_{x\omega}$ 的边聚类系数计算如式(11)所示:

$$C_{x\omega} = \frac{|\Gamma(x,\omega)| + 1}{\min\{k_x - 1, k_\omega - 1\}} \quad (11)$$

边聚类系数准确地描述了一条边上三角模体的聚集程度。

定义1 (模体密度 (Motif Density, MD)) 对于网络中任意的待测节点 x 和 y , ω 是节点对的共邻节点,将路径 w_{xoy} 的模体密度定义为包含路径 w_{xoy} 的所有三角模体数目与所有可能包含该路径的三角模体数目的比值,如式(12)所示:

$$M_{MD}(w_{xoy}) = \frac{|\Gamma(x,\omega)| + |\Gamma(\omega,y)| + 1}{\min\{k_x - 1, k_\omega - 1\} + \min\{k_\omega - 1, k_y - 1\} + 2} \quad (12)$$

以图2(b)为例,路径 w_{xoy} 的模体密度计算为:

$$M_{MD}(w_{xoy}) = \frac{2 + 1 + 1}{\min\{5, 6\} + \min\{6, 4\} + 2} = \frac{4}{11}$$

2.3 基于模体的朴素贝叶斯相似性指标

本文在对路径结构上模体密度进行分析和量化后,基于模体密度来分析路径的相似性贡献,进而在朴素贝叶斯指标的基础上提出改进的链路预测方法。

根据文献[17],LNB方法将共邻节点 ω 的相似性贡献函数计算为待测节点 x 和 y 之间连接与不连接的概率之比,如式(13)所示:

$$R_\omega = \frac{P(e_{xy}|\omega)}{P(\bar{e}_{xy}|\omega)} \quad (13)$$

其中: $P(e_{xy}|\omega)$ 为 ω 的节点聚类系数,且满足 $P(e_{xy}|\omega) + P(\bar{e}_{xy}|\omega) = 1$ 。 R_ω 表示共邻节点 ω 对待测节点产生连边和不产生连边的贡献比。由于这种贡献函数的定义方式无法区分因路径模体特征不同而产生的贡献,为量化每条路径对节点相似性的影响,采用模体密度来定义路径的角色贡献函数。

定义2 (基于路径模体密度的角色贡献函数) 将路径 w_{xoy} 的角色贡献函数 $R(w_{xoy})$ 定义为在路径条件下,待测节点产生连边和不产生连边概率的比值,连边概率用模体密度来表示,如式(14)所示:

$$R(w_{xoy}) = \frac{P(e_{xy}|w_{xoy})}{P(\bar{e}_{xy}|w_{xoy})} = \frac{M_{MD}(w_{xoy})}{1 - M_{MD}(w_{xoy})} \quad (14)$$

其中: $P(e_{xy}|w_{xoy})$ 表示路径 w_{xoy} 对链接形成有促进作用,用该路径的模体密度来计算; $P(\bar{e}_{xy}|w_{xoy})$ 表示路径 w_{xoy} 对链接形成有抑制作用。由式(14)可知, $M_{MD}(w_{xoy})$ 越大,路径 w_{xoy} 的促进作用越大,抑制作用越小,路径相似性贡献 $R(w_{xoy})$ 就越大,与2.2节的分析相符。因此,本文利用 $R(w_{xoy})$ 量化路径 w_{xoy} 对连边相似性的贡献是准确合理的。

定义3 (基于模体的朴素贝叶斯链路预测指标) 根据贝叶斯理论^[17-19],在所有经过共邻节点路径的条件下,节点 x 和 y 连边与不连边概率的计算如式(15)和式(16)所示:

$$P(e_{xy}|W(x,y)) = \frac{P(e_{xy}) \cdot P(W(x,y)|e_{xy})}{P(W(x,y))} \quad (15)$$

$$P(\bar{e}_{xy}|W(x,y)) = \frac{P(\bar{e}_{xy}) \cdot P(W(x,y)|\bar{e}_{xy})}{P(W(x,y))} \quad (16)$$

其中: $W(x,y)$ 表示经过共邻节点且连接节点 x 和 y 的所有路径的集合。

假设每条路径对待测连边的贡献是相互独立的,则:

$$P(W(x,y)|e_{xy}) = \prod_{\omega \in I(x,y)} P(w_{xoy}|e_{xy}) \quad (17)$$

$$P(W(x,y)|\bar{e}_{xy}) = \prod_{\omega \in I(x,y)} P(w_{xoy}|\bar{e}_{xy}) \quad (18)$$

通过式(15)和式(16)相除的方式构建相似性指标,如式(19)所示:

$$r_{xy}^{MLNBCN} = \frac{P(e_{xy}|W(x,y))}{P(\bar{e}_{xy}|W(x,y))} = \frac{P(e_{xy}) \prod_{\omega \in I(x,y)} P(w_{xoy}|e_{xy})}{P(\bar{e}_{xy}) \prod_{\omega \in I(x,y)} P(w_{xoy}|\bar{e}_{xy})} = \underbrace{\frac{P(e_{xy})}{P(\bar{e}_{xy})}}_{\text{constant value}} \underbrace{\prod_{\omega \in I(x,y)} \frac{P(w_{xoy}|e_{xy})}{P(w_{xoy}|\bar{e}_{xy})}}_{\text{role of } w_{xoy}} \quad (19)$$

其中: $P(e_{xy})$ 和 $P(\bar{e}_{xy})$ 分别表示整体网络中连边存在和不存在的概率,均为常数。 $P(e_{xy})$ 和 $P(\bar{e}_{xy})$ 的计算如式(20)和式(21)所示:

$$P(e_{xy}) = \frac{2|E|}{|V| \cdot (|V| - 1)} \quad (20)$$

$$P(\bar{e}_{xy}) = 1 - P(e_{xy}) \quad (21)$$

显然, $s^{-1} = \frac{P(e_{xy})}{P(\bar{e}_{xy})}$ 也为常数,表示网络中连边存在和不存在的比值,可以忽略。

将式(14)、式(20)和式(21)代入式(19)中,等式两边取对数,得到其简化形式,如式(22)所示:

$$s_{xy}^{MLNBCN} = \sum_{\omega \in I(x,y)} (\log_a s + \log_a R(w_{xoy})) \quad (22)$$

定义4 (扩展指标) 基于共邻节点的相似性预测模型有很多经典的预测指标。受LNB模型启发,为进一步验证MLNB方法的有效性,本文把MLNB思想应用到AA指标和RA指标上,得到MLNB扩展指标,如式(23)和式(24)所示:

$$S_{xy}^{MLNBAA} = \sum_{\omega \in I(x,y)} \frac{1}{\log_a k_{\omega}} (\log_a s + \log_a R(w_{xoy})) \quad (23)$$

$$S_{xy}^{MLNBRA} = \sum_{\omega \in I(x,y)} \frac{1}{k_{\omega}} (\log_a s + \log_a R(w_{xoy})) \quad (24)$$

3 实验与结果分析

3.1 实验数据与预分析

为了评价 MLNB 指标的预测准确性, 本文在 Football 网络^[27]、USAir 网络^[28]、C.elegans 网络^[29]、FWMW 网络^[30]、FWEW 网络^[31]和 FFWW 网络^[32]这 6 个真实网络上进行实验。不同的网络具有不同的模体特征。当网络具有显著的模体特征时, 挖掘模体特征的链路预测方法在性能上显著区别于传统的链路预测方法。因此, 本文在进行仿真实验之前, 首先要检验网络模体的存在性。本文基于 2.1 节的模体基本理论以及 Rand-ESU^[33]算法, 通过模体发现软件 FANMOD 对 6 个网络进行模体存在性检验。6 个网络的特征参数及模体存在性检验如表 1 所示。从表 1 可以看出, Z 得分为正数, P 值为 0, 说明以上 6 种网络都有三角模体特征, 可以用来测试 MLNB 方法的准确度。

表 1 6 个网络的特征参数与模体存在性检验

Table 1 Feature parameters and motif existence test of six networks

网络	顶点数	边数	频率/%	平均频率/%	标准差	Z 得分	P 值
Football	115	613	18.634	0.042 7	0.000 973 2	191.030	0
USAir	332	2 126	17.959	1.033 6	0.008 865 1	19.092	0
C.elegans	297	2 148	6.849	0.187 2	0.001 650 2	40.369	0
FWMW	97	1 446	19.654	8.059 5	0.009 962 4	12.048	0
FWEW	69	880	22.842	16.360 0	0.005 978 6	10.842	0
FFWW	128	2 075	13.127	5.060 7	0.004 915 7	16.409	0

3.2 评价指标

为了量化链路预测方法的准确性, 一般将边集 E 随机划分为训练集 E^T 和测试集 E^P, 满足 E = E^T ∪ E^P, 并且 E^T ∩ E^P = ∅。训练集 E^T 作为可观察到的已知网络信息用于计算待测节点对的相似性分数。测试集 E^P 作为待预测的网络信息用于验证预测的准确性。本文使用 AUC (Area Under the Curve) 值^[34]、精确度 (P)^[35]来评价链路预测方法。AUC 从整体上衡量方法的准确性, P 衡量局部预测的准确性。

AUC 值可解释为随机选择一条缺失边 (即 E^P 中的边) 的分数值大于随机选择一条不存在边 (即 U - E 中的边) 的分数值的概率。本文进行 n 次独立抽取, 如果有 n' 次缺失边的分数值更高, n'' 次抽取两条边的分数值相等。AUC 值如式 (25) 所示:

$$A_{AUC} = \frac{n' + 0.5n''}{n} \quad (25)$$

精确度 (P) 计算前 L 条边的预测准确率。将预测边的相似性得分按照降序进行排序, 如果在测试集中排名前 L 的有 m 条边, 那么精确度计算如式 (26) 所示:

$$P = \frac{m}{L} \quad (26)$$

由于本文选取 6 个真实网络的规模不同, 因此统一将各数据集边数的 10% 作为 L 的值。

3.3 仿真实验结果分析

本文实验针对 6 个具有模体特征的网络进行仿真实验, 采用随机抽样方法按 9:1 划分训练集和测试集。为了消除随机误差的影响, 对每个网络进行 100 次独立实验并取平均值。本文将提出的 MLNBs 指标与局部属性的 CN 指标、AA 指标、RA 指标、LNBs 指标, 半局部属性的 LP 指标和全局属性的 Katz 指标进行对比。在多个不同类型网络中 MLNBs 指标与现有指标的 AUC 值对比如表 2 所示。

表 2 不同指标的 AUC 值对比

Table 2 AUC values comparison among different indexes

指标	Football	USAir	C.ele-gans	FWMW	FWEW	FFWW
CN	0.844 5	0.953 3	0.850 7	0.713 4	0.683 9	0.607 1
AA	0.843 2	0.965 8	0.866 8	0.715 5	0.693 3	0.609 8
RA	0.845 2	0.971 6	0.869 8	0.718 1	0.700 8	0.612 7
LNBCN	0.841 5	0.960 3	0.861 1	0.708 1	0.553 6	0.628 5
LNBA A	0.841 5	0.967 6	0.864 7	0.719 8	0.557 8	0.638 1
LNBRA	0.842 6	0.971 8	0.865 9	0.731 8	0.564 8	0.651 6
MLNBCN	0.851 9	0.961 0	0.861 2	0.791 8	0.752 4	0.727 5
MLNBAA	0.853 4	0.967 7	0.866 8	0.799 4	0.762 9	0.739 2
MLNBRA	0.853 5	0.972 3	0.875 7	0.811 8	0.782 3	0.765 5
LP	0.858 1	0.952 2	0.869 4	0.724 7	0.702 0	0.622 5
Katz	0.859 2	0.950 5	0.867 0	0.751 3	0.731 5	0.675 9

从表 2 可以看出, 在每个网络上 MLNBs 系列指标 (MLNBCN、MLNBAA、MLNBRA) 的 AUC 值均优于对应的原始指标 (CN、AA、RA) 和 LNBs 指标 (LNBCN、LNBA A、LNBRA), 表明路径模体密度对链路形成的可能性是有一定的影响。在 MLNBs 系列指标中, MLNBRA 指标的 AUC 值最高, MLNBAA 指标次之, MLNBCN 指标最低, 这说明惩罚度大的共邻节点能够有效提高预测精度。MLNBRA 指标的 AUC 值在 Football 网络中仅次于 LP 和 Katz 指标, 相差不超过 1%, 在剩余的网络中 MLNBRA 指标均具有较优的 AUC 值。LP 指标在共同邻居的基础上考虑了三阶路径信息。Katz 指标考虑全局信息, 时间复杂度相对较高。因此, LP 指标和 Katz 指标的预测精度有一定优势。而 MLNBs 系列指标计算二阶路径的局部信息, 时间复杂度低于 LP 和 Katz 指标。本文提出的 MLNB 方法能解释路径模体聚集程度与节点对链接的关系, 且复杂度低于半局部和全局方法, 在含有模体的网络上具有较优的适用性。

在 FWMW、FWEW、FFWW 这 3 个食物链网络中, MLNBs 系列指标的 AUC 值均优于所有的基准指标。若以次优的全局 Katz 指标为基准, MLNBs 系列指标的 AUC 值在 FWMW 网络中至少提升了 5%, 在 FWEW 网络中至少提升了 2%, 在 FFWW 网络中至少提升了 7%。从表 1 可以看出, FWMW、FWEW、FFWW 网络的三角模体平均频率较大, 说明食物链网络中存在大量的三角模体, 对于这类模体特征较为明显的网络, 基于模体特征的 MLNB 方法预测的效果更好, 进一步验证了 MLNB 指标针对此类网络

具有一定的有效性和可行性。

在不同类型网络中MLNBs系列指标与现有指标的精确度对比如表3所示,所有指标的预测结果在0~0.28,大多数指标的精确度小于0.2。

表3 不同指标的精确度对比

Table 3 Precision comparison among different indexes

指标	Football	USAir	C.ele-gans	FWMW	FWEW	FFWW
CN	0.082 9	0.204 4	0.050 8	0.077 6	0.086 4	0.046 8
AA	0.083 0	0.213 3	0.054 5	0.078 0	0.078 0	0.047 6
RA	0.083 0	0.228 2	0.054 1	0.081 6	0.089 7	0.050 3
LNBCN	0.078 4	0.208 5	0.057 3	0.119 7	0.004 6	0.110 1
LNBA A	0.078 6	0.217 2	0.057 4	0.123 8	0.009 4	0.111 9
LNBR A	0.079 5	0.228 2	0.052 4	0.121 2	0.023 7	0.111 0
MLNBCN	0.085 6	0.213 6	0.076 2	0.146 7	0.118 2	0.120 4
MLNB A A	0.085 6	0.223 5	0.089 3	0.149 0	0.119 9	0.124 1
MLNB R A	0.085 8	0.271 8	0.088 4	0.154 3	0.124 7	0.131 1
LP	0.082 6	0.204 7	0.052 6	0.082 1	0.092 9	0.049 5
Katz	0.082 4	0.201 2	0.057 0	0.096 6	0.102 0	0.067 5

在USAir网络中,MLNBs系列指标(MLNBCN、MLNBAA、MLNBRA)的精确度不仅优于对应的原始指标(CN、AA、RA)和LNBs系列指标(LNBCN、LNBA A、LNBR A),而且相对半局部LP指标和全局Katz指标也有明显提升。其中MLNBRA指标的精确度最优,RA和LNBR A指标的精确度次优。精确度分析结果说明在USAir网络中,共邻节点的度对链路的形式具有较大作用。在剩余的网络中,MLNBs系列指标的精确度均优于所有的基准指标。在所有网络中,MLNBs系列指标的精确度由大到小排序:MLNBRA>MLNBAA>MLNBCN。从表3可以看出,基于模体特征的MLNB指标的精确度相比局部和全局指标有较大幅度提升,证明了计算模体结构信息有助于提升预测性能。

3.4 鲁棒性分析

为了进一步分析MLNBs指标的鲁棒性,本节在不同的训练集比例下研究MLNBs指标与基准指标预测结果的变化情况。在不同网络中各指标的预测值仍然是取100次独立实验的平均值。当训练集比例从0.6开始每次增加0.1直到0.9时,各指标的AUC值对比如图3所示。

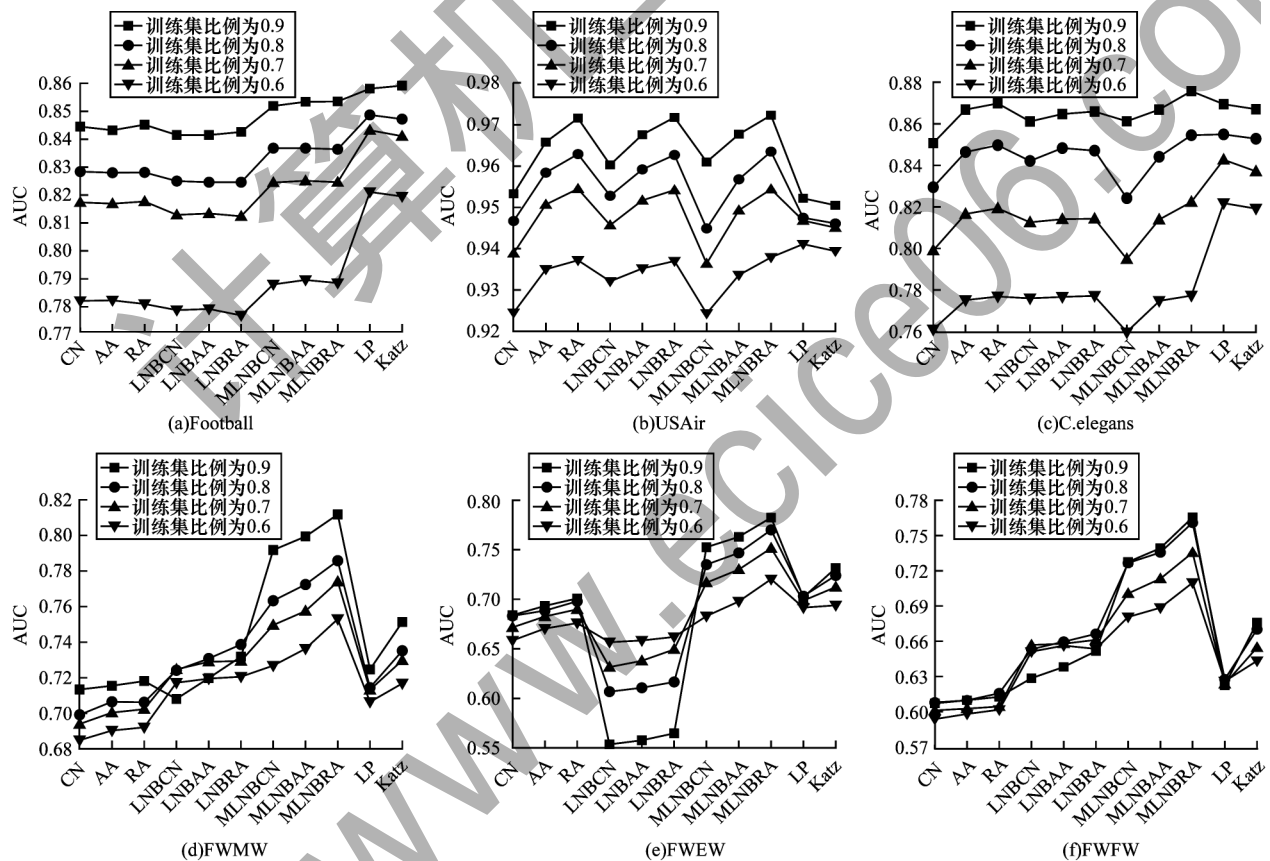


图3 在不同的训练集比例下各指标的AUC值对比

Fig.3 AUC values comparison among various indexes under different training set proportions

从图3可以看出,当训练集比例增加时,多数指标的AUC值随之增大,这是由于训练集比例增加使网络中共邻节点数目和已知拓扑信息增加,预测性能也随之提升。MLNBs指标的AUC值随训练集比例增大而增大。当可观测数据仅有60%时,除了USAir和C.elegans网络中MLNBs指标相对局部相似性指标变化范围不大,在其余网络中仍取得相对

较优的预测结果,表明MLNBs指标在不同网络中具有较优的鲁棒性。LNBs指标在FWMW、FWEW、FFWW网络中并不遵从预测值随训练集比例增大而增大的规律,说明这3个网络中LNBs指标对训练集比例变化的敏感程度不同。

当训练集比例从0.6开始每次增加0.1直到0.9时,各指标的精确度对比如图4所示。

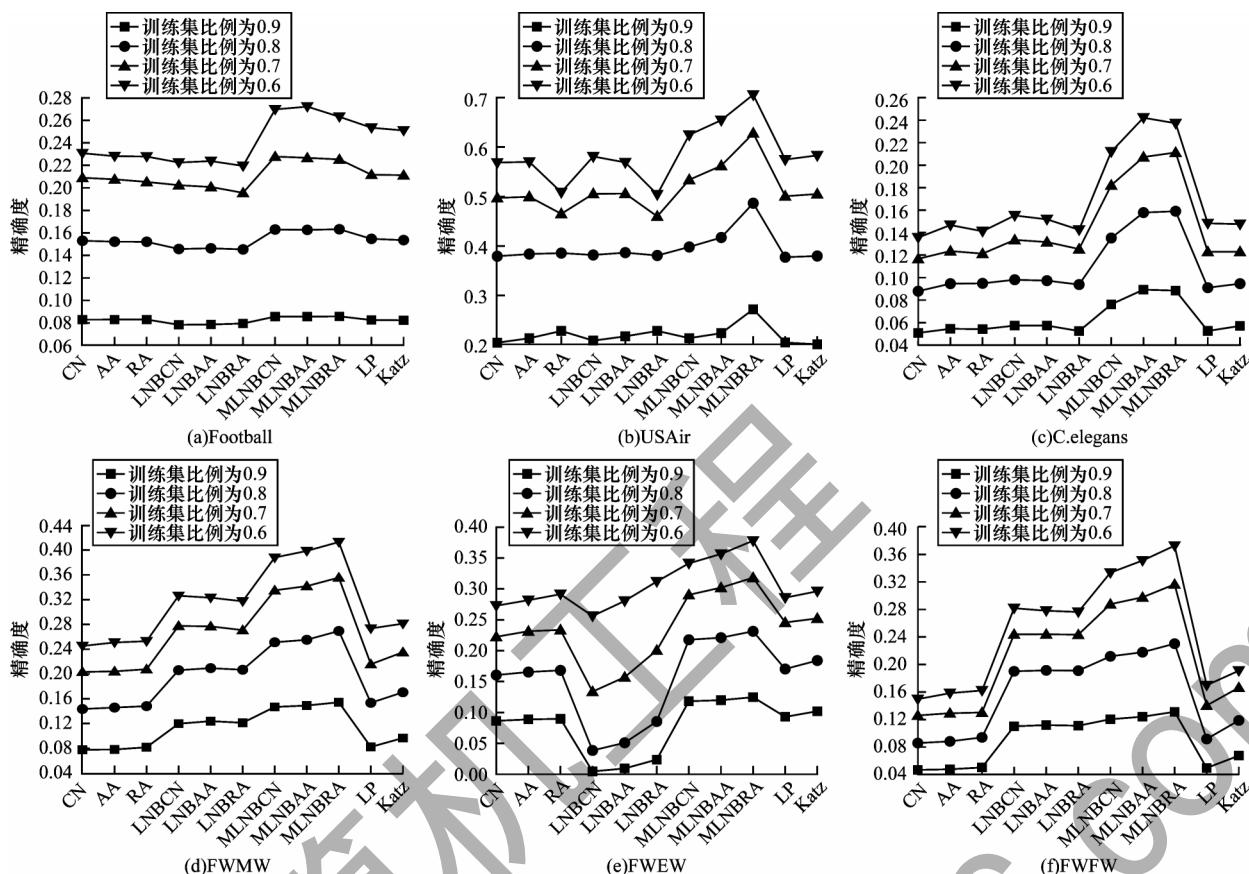


图4 在不同的训练集比例下各指标的精确度对比

Fig.4 Precision comparison among various indexes under different training set proportions

与AUC值的变化规律相反,图4中各指标的精确度随训练集比例的增加而下降,这是由于精确度计算前 L 条边预测的准确率,训练集比例越大,前 L 条预测边在测试集中的可能性越小,精确度就越小。当可观测数据仅有60%时,在各网络中MLNBs指标(MLNBCN、MLNBAA、MLNBRA)的精确度均优于对应的原始指标(CN、AA、RA)和LNBs指标(LNBCN、LNBAA、LNBRA),相对LP和Katz指标也有明显提升,这表明MLNBs指标具有较优的鲁棒性。

无论训练集如何划分,在图3中Football和C.elegans网络上的LP、Katz指标相较于MLNBs指标的AUC值有一定的优势,而图4中LP、Katz指标的精确度都低于MLNBs指标,说明LP和Katz指标并不是在所有评价指标下都表现良好。因此,MLNBs指标在AUC值和精确度两种评价指标测试下具有较优的性能,与不同基准指标相比,在各网络中具有较优的鲁棒性。

4 结束语

本文针对具有模体特征的网络提出一种基于模体的朴素贝叶斯链路预测方法。从模体角度定义模体密度来描述路径结构上模体的聚集程度。考虑到路径结构上模体密度对链接形成的作用,构建基于路径的角色贡献函数,以量化路径的相似性贡献。在此基础上,结合朴素贝叶斯理论,推导MLNBCN及其扩展指标。实验结果表明,本文方法具有较优

的鲁棒性,所提相似性指标的AUC值和精确度均优于LNBs指标和CN、AA、RA等基准指标。本文所提的链路预测方法仅针对无权无向、含有模体结构的网络,后续将本文方法MLNB应用到加权有向网络中,研究加权有向网络的模体特征对链路预测准确度的影响。此外,设计适用于更多不同领域网络的链路预测方法也是下一步的重点研究方向。

参考文献

- [1] BATOOL K, NIAZI M A. Modeling the Internet of Things: a hybrid modeling approach using complex networks and agent-based models[EB/OL]. [2021-07-25]. <http://link.springer.com/content/pdf/10.1186%2Fs40294-017-0043-1.pdf>.
- [2] YANG Y, LICHTENWALTER R N, CHAWLA N V. Evaluating link prediction methods[J]. Knowledge and Information Systems, 2015, 45(3): 751-782.
- [3] LI S B, HUANG J W, ZHANG Z G, et al. Similarity-based future common neighbors model for link prediction in complex networks[J]. Scientific Reports, 2018, 8(1): 1-14.
- [4] JEONG H, MASON S P, BARABÁSI A L, et al. Lethality and centrality in protein networks[J]. Nature, 2001, 411(6833): 41-42.
- [5] HUANG Z A, HUANG Y, YOU Z H, et al. Novel link prediction for large-scale miRNA-lncRNA interaction network in a bipartite graph[J]. BMC Medical Genomics, 2018, 11(6): 113.

- [6] ZHANG L L, LI J, ZHANG Q L, et al. Domain knowledge-based link prediction in customer-product bipartite graph for product recommendation[J]. *International Journal of Information Technology & Decision Making*, 2019, 18(1): 311-338.
- [7] MA C, ZHOU T, ZHANG H F. Playing the role of weak clique property in link prediction: a friend recommendation model[J]. *Scientific Reports*, 2016, 6: 30098.
- [8] 刘宏鲲, 吕琳媛, 周涛. 利用链路预测推断网络演化机制[J]. *中国科学: 物理学 力学 天文学*, 2011, 41(7): 816-823.
LIU H K, LÜ L Y, ZHOU T. Uncovering the network evolution mechanism by link prediction[J]. *Scientia Sinica (Physica, Mechanica & Astronomica)*, 2011, 41(7): 816-823. (in Chinese)
- [9] YU C M, ZHAO X L, AN L, et al. Similarity-based link prediction in social networks: a path and node combined approach[J]. *Journal of Information Science*, 2017, 43(5): 683-695.
- [10] 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.
- [11] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. *Social Networks*, 2003, 25(3): 211-230.
- [12] ZHOU T, LÜ L, ZHANG Y C. Predicting missing links via local information[J]. *The European Physical Journal B*, 2009, 71(4): 623-630.
- [13] LYV L, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, 2009, 80(4): 1-22.
- [14] KATZ L. A new status index derived from sociometric analysis[J]. *Psychometrika*, 1953, 18(1): 39-43.
- [15] ZHANG Y Y, SHI Z, FENG D, et al. Degree-biased random walk for large-scale network embedding[J]. *Future Generation Computer Systems*, 2019, 100: 198-209.
- [16] LU Y D, GUO Y F, KORHONEN A. Link prediction in drug-target interactions network using similarity indices[J]. *BMC Bioinformatics*, 2017, 18(1): 39.
- [17] LIU Z, ZHANG Q M, LÜ L, et al. Link prediction in complex networks: a local naive Bayes model[J]. *Europhysics Letters*, 2011, 96(4): 48007.
- [18] WU J H. A generalized tree augmented naive Bayes link prediction model[J]. *Journal of Computational Science*, 2018, 27: 206-217.
- [19] ZHANG H F, JIA M M, XIANG B B, et al. Predicting missing links in complex networks via an extended local naive Bayes model[J]. *Europhysics Letters*, 2020, 130(3): 38002.
- [20] 刘英杰, 刘士虎, 徐伟华. 基于有效路径拓扑稳定性的链路预测方法[J]. *计算机应用研究*, 2022, 39(1): 90-95.
LIU Y J, LIU S H, XU W H. Link prediction method based on topology stability of effective path[J]. *Application Research of Computers*, 2022, 39(1): 90-95. (in Chinese)
- [21] 李英乐, 何赞园, 王凯, 等. 基于资源传输节点拓扑紧密性的链路预测方法[J]. *计算机工程*, 2021, 47(1): 50-57.
LI Y L, HE Z Y, WANG K, et al. Link prediction method based on topological tightness of resource transmission nodes[J]. *Computer Engineering*, 2021, 47(1): 50-57. (in Chinese)
- [22] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: simple building blocks of complex networks[J]. *Science*, 2002, 298(5594): 824-827.
- [23] 韩华, 刘婉璐, 吴翎燕. 基于模体的复杂网络测度研究[J]. *物理学报*, 2013, 62(16): 1-8.
HAN H, LIU W L, WU L Y. The measurement of complex network based on motif[J]. *Acta Physica Sinica*, 2013, 62(16): 1-8. (in Chinese)
- [24] 贾承丰, 韩华, 完颜娟, 等. 基于网络模体特征攻击的网络抗毁性研究[J]. *复杂系统与复杂性科学*, 2017, 14(4): 43-50.
JIA C F, HAN H, WAN Y J, et al. Network destruction resistance based on network motif feature[J]. *Complex Systems and Complexity Science*, 2017, 14(4): 43-50. (in Chinese)
- [25] 柳娟, 刘亚芳, 许爽, 等. 基于多模体边度的科学家合作关系预测[J]. *计算机学报*, 2020, 43(12): 2372-2384.
LIU J, LIU Y F, XU S, et al. Predicting scientific collaboration by edge degree of multiple motifs[J]. *Chinese Journal of Computers*, 2020, 43(12): 2372-2384. (in Chinese)
- [26] 胡健, 杨炳儒. 基于边聚集系数的社区结构发现算法[J]. *计算机应用研究*, 2009, 26(3): 858-859.
HU J, YANG B R. Community structure discovery algorithm based on edge clustering coefficient[J]. *Application Research of Computers*, 2009, 26(3): 858-859. (in Chinese)
- [27] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. *PNAS*, 2001, 99(12): 7821-7826.
- [28] BATAGELJ V, MRVAR A. Pajek-program for large network analysis[J]. *Connections*, 1998, 21(2): 47-57.
- [29] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks[J]. *Nature*, 1998, 393(6684): 440-442.
- [30] BAIRD D, LUCZKOVICH J, CHRISTIAN R R. Assessment of spatial and temporal variability in ecosystem attributes of the St. Marks national wildlife refuge, apalachee bay, Florida[J]. *Estuarine, Coastal and Shelf Science*, 1998, 47(3): 329-349.
- [31] ULANOWIC R E, DEANGELIS D L. Network analysis of trophic dynamics in south Florida ecosystems [M]// GEROULD S, HIGER A L U S. Geological survey program on the south Florida ecosystem. Washington D. C., USA: Government Printing Office, 1999: 114-115.
- [32] MICHALSKI R, PALUS S, KAZIENKO P. Matching organizational structure and social network extracted from email communication [M]// ABRAMOWICZ W. Business information systems. Berlin, Germany: Springer, 2011: 197-206.
- [33] WERNICKE S, RASCHE F. FANMOD: a tool for fast network motif detection[J]. *Bioinformatics*, 2006, 22(9): 1152-1153.
- [34] ZENG G P, ZENG E. On the three-way equivalence of AUC in credit scoring with tied scores[J]. *Communications in Statistics-Theory and Methods*, 2019, 48(7): 1635-1650.
- [35] WU Z H, LIN Y F, ZHAO Y J, et al. Improving local clustering based top-L link prediction methods via asymmetric link clustering information [J]. *Physica A: Statistical Mechanics and Its Applications*, 2018, 492: 1859-1874.