基于节点连接模式相关性的链接预测方法

单 娜 李龙杰 刘昱阳 陈晓云
(兰州大学信息科学与工程学院 兰州 730000)

摘 要 作为复杂网络分析中的一个研究热点,链接预测在许多领域中都有重要的应用价值,得到了广泛的关注。使用网络中的已知结构信息来计算未连接的节点对之间的相似性,进而评估其存在链接的可能性是目前最常用的方法。不同网络具有不同的结构特征,节点之间的特征对链接的形成具有重要影响。为了提高链接预测的性能,文中定义了节点的连接模式,并基于节点连接模式的相关性(Correlation of Nodes' Connecting Patterns,CNCP)设计了一个新的链接预测模型。该模型将 CNCP 与基本相似性指标相结合,通过综合节点的相似性与节点连接模式的相关性进行链接预测。文中将 CNCP 与 CN(Common Neighbors),RA(Resource Allocation),AA(Adamic-Adar)及 PA(Preferential Attachment)4 个相似性指标相结合,提出了 CNCP-CN,CNCP-RA,CNCP-AA 和 CNCP-PA 4 个新的链接预测指标。在 6 个真实数据集上的实验结果表明,所提方法在 AUC 和 Precision 2 个评价标准上的性能优于对比方法。

关键词 复杂网络,链接预测,节点连接模式相关性,相似性指标
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.190700057

Link Prediction Based on Correlation of Nodes' Connecting Patterns

SHAN Na LI Long-jie LIU Yu-yang CHEN Xiao-yun
(School of Information Science and Engineering, Lanzhou University, Lanzhou, 730000, China)

Abstract As a research hotspot in complex network analysis, link prediction has a wide range of applications in many fields, and hence has captured much attention of researchers. Similarity-based methods, which compute the similarity scores between unconnected node pairs based on the known network structures and estimate their connection likelihood according to the similarity scores, are commonly used. In general, different kinds of networks have diverse structural characteristics, and hence the correlation of characteristics between nodes has an important influence on the formation of links. To enhance the performance of link prediction, this paper defined the connecting pattern of a node, and proposed a new link prediction model based on the Correlation of Nodes' Connecting Patterns (CNCP). By combining CNCP with a similarity-based method, this model can take both similarity and correlation between nodes into account. In this paper, four CNCP-based methods, i. e., CNCP-CN, CNCP-RA, CNCP-AA and CNCP-PA, are derived from the model, in which similarity indexes are CN(Common Neighbors), RA(Resource Allocation), AA(Adamic-Adar) and PA(Preferential Attachment), respectively. The experimental results on six networks show that the proposed methods are superior to the compared ones under the criteria of AUC and Precision.

Keywords Complex networks, Link prediction, Correlation of nodes' connecting patterns, Similarity index

1 引言

链接预测作为复杂网络分析中的一个基本问题,在诸多领域中得到了广泛的应用。链接预测根据已知的网络结构信息来预测尚未连边的节点对之间产生连接的可能性[1]。在现实世界中,复杂网络可以用于描述包括社会、生物、信息和技术等在内的许多复杂系统,其中节点代表个体或代理,链接或边表示节点之间的关系或交互。在一些网络中,如蛋白质相互作用网络[2]、电网[3]和航空运输网络[4],如何发现哪些实体之间在不久的将来可能产生新的链接是一个非常重要又具有挑战的问题。此类问题可以定义为链接预测问题。

近年来,学者们提出了大量的链接预测算法,主要包括:基于节点相似度的算法[5]、基于似然分析的算法[6]以及基于概率论模型的算法[7]。其中,基于节点相似度的算法最直观、最流行。此类算法认为节点间的相似度越高,其存在链接的可能性就越大。基于节点相似度的算法又可分为基于邻居的算法和基于路径的算法。基于邻居的相似度算法包括 Salton 指标[8]、Sørensen 指标[9]和 CN 指标[10]等。其优点在于计算

到稿日期:2019-04-06 返修日期:2019-07-30 本文受国家自然科学基金(61602225),兰州大学中央高校基本科研业务费专项资金(lzujbky-2019-90)资助。

单 娜(1996—),女,硕士生,主要研究方向为网络信息挖掘;**李龙杰**(1981—),男,博士,工程师,硕士生导师,CCF 会员,主要研究方向为数据挖掘、机器学习、网络信息挖掘等,E-mail:ljli@lzu.edu.cn(通信作者);**刘昱阳**(1996—),男,硕士生,主要研究方向为网络信息挖掘;**陈晓云**(1954—),女,硕士,教授,博士生导师,CCF 高级会员,主要研究方向为数据挖掘、大数据分析、网络信息挖掘等。

复杂度低,在聚类系数高的网络中能取得令人满意的结果。然而,由于它们不能计算没有共同邻居的节点间的相似性,因此在聚类系数低的稀疏网络中很难得到高的准确率。基于路径的相似度算法有 LP(Local Paths)指标[11]、Katz 指标[12]和 LHN-II 指标[13]等。其中一些算法可以解决基于邻居的相似度算法在低聚类系数网络中低准确率的问题。但是,一些基于距离的相似性指标对观察到的边的比例很敏感[14],这意味着在算法的训练集中,如果观察到的边的比例下降,其预测精度会明显降低。

网络中节点间相隔的距离越远,彼此的关系就越弱。基于此思想,Wang 等[15]通过定义节点的坐标矩阵将网络中的节点转化为向量,然后通过计算向量间的余弦值得到 CD(Cosine Distance)指标,并在此基础上提出了一系列基于 CD 的链接预测方法。受其启发,本文提出了基于节点连接模式的链接预测方法,该方法根据节点到其他节点的距离定义其连接模式,然后利用皮尔逊相关系数定义节点连接模式之间的相关性,从而得到 CNCP 指标,在此基础上将 CNCP 与已有的相似性指标相结合得到一系列基于 CNCP 的方法。

2　相关工作

2.1　问题描述

令 $G=(V,E)$表示无权无向连通网络,其中不存在自链接和重复链接,V 是网络 G 中节点的集合,节点数为 $|V|$,E 是 G 中的边(或链接)的集合,边数为 $|E|$。给定不相连的节点对(v_i,v_j),节点 v_i 和 v_j 之间的相似性定义为 $s_{i,j}$,其值越高,节点 v_i 和 v_j 之间的链接就越有可能存在,因此 $s_{i,j}$ 也可以看作节点 v_i 和 v_j 之间存在链接的得分。网络 G 的链接可以表示为邻接矩阵 $\boldsymbol{A}$,当节点 v_i 和 v_j 之间存在链接时,邻接矩阵 $\boldsymbol{A}$ 中的元素 $a_{i,j}$ 等于 1,否则等于 0。如果节点 v_i 和 v_j 之间存在链接,则两个节点互为邻居节点。在矩阵 $\boldsymbol{A}$ 中,第 i 行元素的和是节点 v_i 的度,表示为 k_i。

2.2　常用链接预测方法

(1)CN 指标[10]利用共同邻居的数量来度量节点间的相似性。两个节点的共同邻居数越多,它们就越相似。假设节点 v_i 和 v_j 是网络 G 中的两个节点,则 v_i 和 v_j 的相似性定义为:

$$s_{i,j}^{\mathrm{CN}}=|\Gamma(i)\cap\Gamma(j)| \tag{1}$$

其中,$\Gamma(i)$表示 v_i 的邻居集合,$\Gamma(j)$表示 v_j 的邻居集合。

(2)RA 指标[5]是基于资源分配的思想的。设每个节点拥有一定数量的资源,在进行资源分配时,节点之间的资源通过它们的共同邻居进行传输,每个邻居节点在传输资源时将其接收的资源平均分配给其邻居。给定节点 v_i 和 v_j,v_i 通过它们的每个共同邻居向 v_j 传递资源,假设每个邻居传递一个单位的资源。RA 定义了 v_i 与 v_j 之间的相似度为 v_j 接收到的资源的数量,如式(2)所示:

$$s_{i,j}^{\mathrm{RA}}=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{k_z} \tag{2}$$

其中,k_z 是共同邻居 z 的度。

CN 指标和 RA 指标的不同之处在于:前者不区分共同邻居,即 CN 认为每个共同邻居对相似性计算的贡献相同;后者根据邻居的度来区分其贡献,邻居的度越高,其贡献就越小。

(3)AA 指标[16]与 RA 指标相似,它们都依据共同邻居的度来区分其对相似性的贡献。不同之处在于 AA 使用度的对数来区分共同邻居对相似度的贡献,而 RA 则直接使用度。在某些网络中,节点的度往往很高,若直接使用度的倒数计算节点的相似度,则得到的相似度将很小,而使用度的对数则有效避免了此问题。AA 定义节点 v_i 和 v_j 的相似性为:

$$s_{i,j}^{\mathrm{AA}}=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{\log k_z} \tag{3}$$

(4)PA 指标[17]认为一条边连接到一个节点的概率正比于该节点的度。PA 定义节点 v_i 和 v_j 的相似性为:

$$s_{i,j}^{\mathrm{PA}}=k_i\times k_j \tag{4}$$

3　提出的方法

本文提出利用节点间连接模式的相关性进行链接预测。首先定义每个节点的连接模式,这一概念参考了文献[15]中的 k-距离矩阵和坐标矩阵;然后利用皮尔逊相关系数定义节点连接模式的相关性(CNCP);最后将 CNCP 与基本的相似度指标相结合,得到新的链接预测方法。

3.1　节点连接模式的相关性

本节首先定义节点的连接模式,然后利用皮尔逊相关系数评估节点间连接模式的相关性。

定义 1(k-距离矩阵 $\boldsymbol{L}$[15])　给定网络 $G=(V,E)$,$n=|V|$是网络中节点的个数。给定距离参数 k,网络 G 的 k-距离矩阵 $\boldsymbol{L}$ 是一个 $n\times n$ 的矩阵,$l_{ij}\in\boldsymbol{L}$ 表示节点 v_i 与 v_j 之间的关系。设 d_{ij} 为 v_i 与 v_j 之间的最短距离,如果 $d_{ij}\leqslant k$,定义 $l_{ij}=d_{ij}$,否则 $l_{ij}=\infty$。

定义 2(节点连接模式)　根据 k-距离矩阵,定义一个 $n\times n$ 的矩阵 $\boldsymbol{P}$。当 $i\neq j$ 时,$\boldsymbol{P}$ 中元素 $p_{ij}=\frac{1}{l_{ij}}$,否则 $p_{ij}=1$。矩阵 $\boldsymbol{P}$ 的第 i 个行向量 $\boldsymbol{P}_i$ 定义为节点 v_i 的连接模式。

皮尔逊相关系数可以用来计算两个向量的相关程度,其取值在[−1,1]之间。该值大于零时,表示两个向量是正相关的;该值小于零时,表示两个向量是负相关的;该值等于零时,表示两个向量间不是线性相关的。本文使用皮尔逊相关系数来衡量两个节点的相关性。

定义 3(节点连接模式的相关性)　两个节点的连接模式的皮尔逊相关系数被定义为节点连接模式的相关性,表示为:

$$r_{i,j}=\frac{\mathrm{cov}(\boldsymbol{P}_i,\boldsymbol{P}_j)}{\sigma_i\sigma_j} \tag{5}$$

其中,$\boldsymbol{P}_i$ 表示节点 v_i 的连接模式向量,$\mathrm{cov}(\boldsymbol{P}_i,\boldsymbol{P}_j)$为 $\boldsymbol{P}_i$ 和 $\boldsymbol{P}_j$ 的协方差,σ_i 为 $\boldsymbol{P}_i$ 的标准差。

3.2　基于节点连接模式相关性的相似性指标

节点连接模式的相关性可以作为链接预测的重要补充,但节点本身的相似性才是最基本的。因此本文提出的框架通过结合两个方面的信息来解决链接预测问题。给定两个不相连的节点 v_i 和 v_j,$r_{i,j}$表示这两个节点的连接模式的相关性,$s_{i,j}^{*}$代表通过相似性算法 s^* 计算得到的 v_i 和 v_j 之间的相似性。本文的链接预测框架将 v_i 和 v_j 之间存在链接的可能性定义为:

$$s_{i,j}^{\mathrm{CNCP-}*}=(1+r_{i,j})^{\alpha}\cdot s_{i,j}^{*} \tag{6}$$

其中,α 是一个自由参数,用于调节 CNCP 的影响。显然,当 $\alpha=0$ 时,$r_{i,j}$ 的影响忽略不计,$s^{\mathrm{CNCP-}*}$ 与 s^* 等价。

本文考虑将 CNCP 与 CN,RA,AA 和 PA 4 种现有的相

似性指标相结合，分别得到 CNCP-CN，CNCP-RA，CNCP-AA 和 CNCP-PA 4 种基于 CNCP 的指标，表示为：

$$s_{i,j}^{\text{CNCP-CN}}=(1+r_{i,j})^{\alpha}\cdot s_{i,j}^{\text{CN}} \tag{7}$$

$$s_{i,j}^{\text{CNCP-RA}}=(1+r_{i,j})^{\alpha}\cdot s_{i,j}^{\text{RA}} \tag{8}$$

$$s_{i,j}^{\text{CNCP-AA}}=(1+r_{i,j})^{\alpha}\cdot s_{i,j}^{\text{AA}} \tag{9}$$

$$s_{i,j}^{\text{CNCP-PA}}=(1+r_{i,j})^{\alpha}\cdot s_{i,j}^{\text{PA}} \tag{10}$$

3.3 与基于 CD 的相似性指标的对比

网络中节点间的最短距离越远，彼此的联系就越弱。CD (Cosine Distance)指标[15]根据节点间的最短距离计算每个节点的坐标，并通过节点间坐标的余弦距离定义节点之间的相似性。本节将对 CD 指标以及 CD 与 PA 相结合产生的 CDI 指标进行介绍，并描述上述指标与本文方法的区别。

节点坐标矩阵 $\boldsymbol{C}$ 的定义与定义 2 中矩阵 $\boldsymbol{P}$ 的定义相同，当 $i\neq j$ 时，元素 $c_{ij}=\frac{1}{l_{ij}}$，否则 $c_{ij}=1$。矩阵 $\boldsymbol{C}$ 的第 i 个行向量 $\boldsymbol{C}_i$ 定义为节点 v_i 的坐标。CD 指标[15]通过计算节点坐标间的余弦值得到两个节点间的相似度，定义为：

$$s_{i,j}^{\text{CD}}=\frac{(\boldsymbol{C}_i,\boldsymbol{C}_j)}{\|\boldsymbol{C}_i\|\times\|\boldsymbol{C}_j\|} \tag{11}$$

其中，$\|\boldsymbol{C}_i\|$ 表示 $\boldsymbol{C}_i$ 的模，$(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 表示 $\boldsymbol{C}_i$ 和 $\boldsymbol{C}_j$ 的内积。可以发现，$s_{i,j}^{\text{CD}}$ 值越大，向量 $\boldsymbol{C}_i$ 和 $\boldsymbol{C}_j$ 之间的夹角越小，节点 v_i 和 v_j 就越相似。

CD 指标不仅可以单独使用，还可以与其他方法相结合产生新的链接预测方法。文献[15]定义了一个改进的 CD 指标 CDI，该方法结合了 CD 与 PA 两个指标的优势，定义如下：

$$s_{i,j}^{\text{CDI}}=\frac{(\boldsymbol{C}_i,\boldsymbol{C}_j)}{\|\boldsymbol{C}_i\|\times\|\boldsymbol{C}_j\|}\times(k_i\times k_j) \tag{12}$$

从上述定义中可以看出，CD 指标通过节点坐标的余弦距离来度量其相似性，本文方法则通过计算节点连接模式的皮尔逊相关系数衡量节点间的相关性，将相关性与基本相似性算法相结合来度量节点间的相似性。与 CD 指标相比，本文提出的指标有以下优点：1)结合了节点连接模式的相关性和节点间相似性来度量节点间存在链接的可能性，考虑得更加全面；2)本文的方法能够通过调节自由参数 α 的值来控制相关性的影响，使得本文方法能够在不同的网络中得到较优的预测结果，有更广泛的适用性。

4 实验数据集与评价方法

4.1 数据集

本文使用来自不同领域的 6 个真实网络对链接预测算法的性能进行评估。这些网络的简要描述如下：1)Football 网络[18]，2000 年秋季常规赛期间，美国爱荷华州院校之间的橄榄球比赛网络；2)Jazz 网络[19]，爵士乐音乐家之间的协作网络；3)Mangwet 网络[20]，红树林河口湿季的食物链网络；4)Foodweb 网络[21]，美国佛罗里达海湾雨季的食物链网络；5)Dolphins 网络[22]，62 只海豚之间频繁联系的社交网络；6)Email 网络[23]，大学成员之间的电子邮件交换网络。

本文将所有的网络视为无向无权网络，表 1 列出了这 6 个网络的基本拓扑结构。其中，e 表示网络效率[24]，c 表示网络聚类系数，r 表示网络同配系数[5]，$\langle k\rangle$ 表示网络平均度，p 表示所有最短路径的平均长度，$\langle H\rangle$ 表示度异质性[1]，d 表示网络直径。

表 1 6 个网络的基本拓扑特征

Table 1 Basic topological features of six networks

	Football	Jazz	Dolphins	Email	Mangwet	Foodweb
$\lvert V\rvert$	115	198	62	1133	97	128
$\lvert E\rvert$	613	2742	159	5451	1446	2075
e	0.450	0.513	0.379	0.300	0.655	0.622
c	0.403	0.633	0.259	0.220	0.468	0.3346
r	0.162	0.020	−0.044	0.078	−0.151	−0.112
$\langle k\rangle$	10.66	27.697	5.129	9.622	29.814	32.422
p	2.508	2.235	3.357	3.606	1.693	1.776
$\langle H\rangle$	1.007	1.395	1.327	1.942	1.266	1.237
d	4	6	8	8	3	3

4.2 评价方法

为了验证链接预测算法的准确性，我们将网络的链接 E 随机划分为两部分：训练集 E_T 和测试集 E_P，其满足：$E_T\cap E_P=\varnothing$，$E_T\cup E_P=E$。评价链接预测算法性能的指标有很多，本文采用常用的 Precision 指标[25]以及 AUC 指标[1]来评价链接预测算法的性能。

为了计算 Precision 值，将网络中所有不存在的链接根据它们的相似度按降序排列，然后计算得分最高的前 L 条链接中预测正确的链接数 l 所占的比例，如式(15)所示：

$$Precision=\frac{l}{L} \tag{13}$$

AUC 在链接预测中的定义如下：随机从测试集和不存在的链接中各取一条链接，比较这两条链接的分数，设在 n 次独立比较中，测试集中的链接比不存在的链接拥有更高分数的次数为 n_1，两者拥有相同分数的次数为 n_2，则 AUC 的计算公式为：

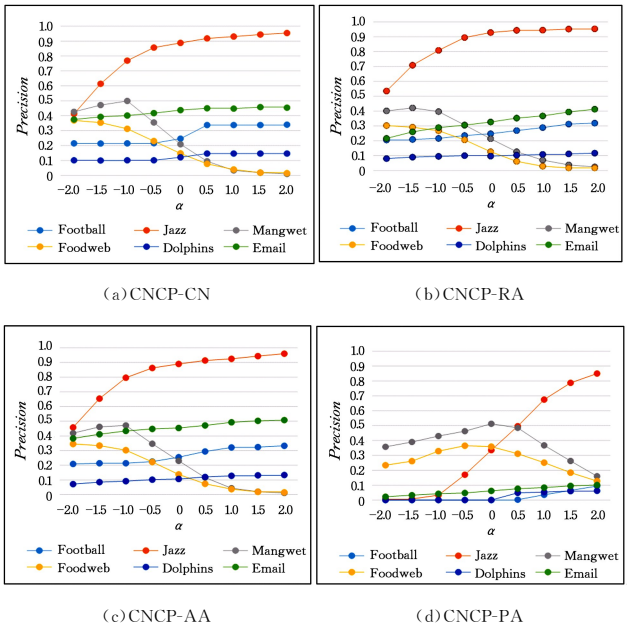
$$AUC=\frac{n_1+0.5\cdot n_2}{n} \tag{14}$$

5 实验

5.1 参数分析

本文提出的方法包含两个参数：k 和 α，k 用于控制节点之间的距离，α 用于调节 CNCP 对链接预测结果的影响。根据小世界效应，我们将距离参数 k 设为 5。在不同的网络上，CNCP 的影响也可能不同，因此需要通过实验确定不同网络上最优的 α 取值。

图 1 展示了本文的 4 种基于 CNCP 的链接预测指标的 Precision 值随 α 的变化情况。图 1 中的结果是 40 次独立实验结果的平均值。每次独立实验中，从每个网络中随机选择 90% 的链接作为训练集 E_T，剩余 10% 的链接作为测试集 E_P。为了计算 Precision 值，实验中将 L 的值设为 20。从图中可以看出：1)CNCP-CN，CNCP-RA 和 CNCP-AA 3 个指标的 Precision 值随 α 的变化趋势非常接近，因为 3 个指标都可以归为基于共同邻居的方法；2)在 Football，Jazz，Dolphins 和 Email 4 个网络中，本文的 4 种方法均在 $\alpha=2.0$ 时取得最高的 Precision 值；3)在 Foodweb 和 Mangwet 2 个网络中，CNCP-CN，CNCP-RA 和 CNCP-AA 3 个指标的 Precision 值随 α 的增加而逐渐降低，CNCP-PA 在 $\alpha=0$ 时取得最高的 Precision 值。



(a)CNCP-CN　(b)CNCP-RA

(c)CNCP-AA　(d)CNCP-PA

图 1　不同 α 值对应的 *Precision* 值

Fig. 1　*Precision* values versus different values of α

观察表 1 可知,Foodweb 和 Mangwet 网络具有以下共同特点:1)两个网络均为异配网络;2)两个网络都具有较大的平均度和较短的平均最短距离,均为稠密网络。根据文献[26]可知:异配网络具有大度节点倾向于与小度节点相连的特点。由此可知,在异配网络中相邻节点的拓扑结构往往差异较大,其连接模式的相关性也较低。因此应考虑在异配稠密网络中减弱节点连接模式相关性的影响。从图 1 中可以看出,当 α 为负值时,CNCP-CN,CNCP-RA 和 CNCP-AA 3 个指标取得了较高的 Precision 值。在异配网络中,PA 指标本身已经具有很好的性能[15],因此 CNCP 对其性能的影响有限。

通过以上分析,在接下来的实验中,对于 Foodweb 和 Mangwet 网络,将 α 值设为 −2.0;在其余 4 个网络中,将 α 值设为 2.0。

5.2　结果分析

首先对比本文的 4 种方法与相应的 4 个基本相似性方法(CN,RA,AA 和 PA)的性能。表 2 列出了相应的 Precision 与 AUC 结果。其结果也是 40 次独立实验结果的平均值,每次独立实验中,训练集与测试集的比例为 9∶1。计算 Precision 值时,L 的值为 20。表 2 中,每个网络的最优值用斜体表示,基本指标与 CNCP 结合前后的较优值用加粗表示。

表 2　6 个网络中的 *Precision* 值和 *AUC* 值

Table 2　*Precision* values and *AUC* values over six networks

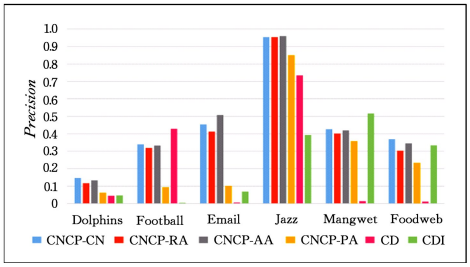
	Precision						*AUC*					
	Football	Jazz	Dolphins	Email	Mangwet	Foodweb	Football	Jazz	Dolphins	Email	Mangwet	Foodweb
CN	0.2462	0.8875	0.1217	0.4375	0.2088	0.1475	0.8396	0.9540	0.7712	0.8444	0.7099	0.6091
CNCP-CN	*0.3400*	**0.9537**	*0.1467*	**0.4537**	**0.4263**	*0.3688*	*0.8462*	**0.9614**	**0.7719**	**0.8453**	**0.7716**	**0.6817**
	38.10%↑	7.46%↑	20.54%↑	3.7%↑	104.17%↑	150.03%↑	0.79%↑	0.78%↑	0.09%↑	0.11%↑	8.69%↑	11.92%↑
RA	0.2475	0.9287	0.0967	0.3263	0.2138	0.1263	0.8396	0.9702	0.7717	0.8460	0.7153	0.6135
CNCP-RA	**0.3187**	**0.9537**	**0.1167**	**0.4125**	**0.4012**	**0.3025**	**0.8449**	*0.9724*	*0.7733*	0.8462	0.7703	0.6742
	28.77%↑	2.69%↑	20.68%↑	26.42%↑	87.65%↑	139.51%↑	0.63%↑	0.23%↑	0.21%↑	0.02%↑	7.69%↑	9.89%↑
AA	0.2550	0.8900	0.1067	0.4537	0.2300	0.1375	0.8379	0.9613	0.7727	0.8464	0.7126	0.6110
CNCP-AA	**0.3325**	*0.9600*	**0.1333**	*0.5075*	**0.4188**	**0.3450**	**0.8460**	**0.9663**	**0.7731**	*0.8465*	**0.7726**	**0.6803**
	30.39%↑	7.87%↑	24.93%↑	11.86%↑	82.9%↑	150.9%↑	0.97%↑	0.52%↑	0.05%↑	0.01%↑	8.42%↑	11.34%↑
PA	0.0000	0.3350	0.0250	0.0613	*0.5125*	**0.3588**	0.2676	0.7673	0.6210	0.7784	0.7782	0.7317
CNCP-PA	**0.0938**	**0.8500**	**0.0617**	**0.1012**	0.3575	0.2337	**0.7831**	**0.8878**	**0.7438**	**0.8147**	*0.7922*	*0.7508*
	—	153.73%↑	146.8%↑	65.09%↑	30.24%↓	34.87%↓	192.64%↑	15.70%↑	19.77%↑	4.66%↑	1.80%↑	2.61%↑

从表 2 的 *Precision* 结果可以发现:1)CNCP-CN,CNCP-RA 和 CNCP-AA 3 个链接预测方法在 6 个实验网络上的 *Precision* 值明显高于 CN,RA 和 AA 3 个基本的相似性指标。在 Mangwet 与 Foodweb 两个异配稠密网络上,与 CN,RA 和 AA 相比,CNCP-CN,CNCP-RA 和 CNCP-AA 的精度提升十分显著,提升均在 80%以上,最高达到 150%,而且其在其他 4 个网络上,也都有不同程度的提升。产生这一现象的原因在于这 3 个基于 CNCP 的预测方法不但考虑了节点之间的相似性,还融入了节点连接模式的相关性,通过控制参数 α 的值可以调节相关性对链接预测的影响。2)在 Football,Jazz,Dolphins 与 Email 4 个网络上,CNCP-PA 的精度优于 PA,而在 Mangwet 与 Foodweb 这两个网络上,CNCP-PA 的精度不如 PA。这是因为 PA 指标的优先依附思想符合网络 Mangwet 与 Foodweb 的演化机制,因此其在这两个网络上能够取得较好的性能。3)除了 Mangwet 网络,其他 5 个网络上的最高 *Precision* 值都由本文提出的方法取得。由表 1 可以发现,Mangwet 网络的异配性最高,PA 在该网络中的表现最好。

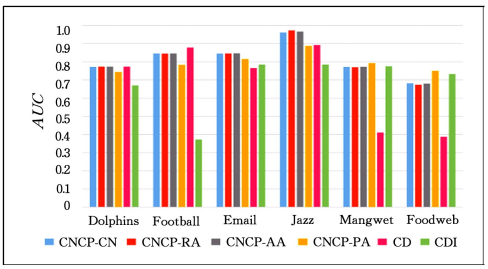
从表 2 中相同设置下的 *AUC* 结果可以发现,在所有网络中,现有相似性指标与 CNCP 结合后得到的 *AUC* 值都有不同程度的提升。并且在 Mangwet 与 Foodweb 两个网络上,CNCP-PA 的 *AUC* 值不但高于 PA 的结果,也优于其他所有方法。这一现象与表 2 中的 *Precision* 结果正好相反。CNCP-PA 在 Mangwet 与 Foodweb 上的 *Precision* 值明显低于 PA 的结果。这是因为 Precision 只关注前 L 条预测的链接,而 AUC 从全局的角度评价预测方法的性能。总体而言,基于 CNCP 的相似性指标可以有效提升现有链接预测指标的预测精度。

由于本文方法受到了 CD 指标的启发,因此我们将本文方法与 CD 及 CDI 指标进行对比,结果如图 2 所示。其结果也是 40 次独立实验结果的平均值,训练集与测试集的比例为 9∶1。在计算 *Precision* 时,L 的值为 20。图 2(a)中,除 Football 和 Mangwet 两个网络外,其余所有网络中的最优 *Precision* 值均由基于 CNCP 的算法取得。在 Email,Mangwet 和 Foodweb 3 个网络上,CD 指标的 *Precision* 值非常低,而本文方法的表现都还不错。图 2(b)中,Email,Jazz,Mangwet 和 Foodweb 4 个网络上的最高 *AUC* 值均由本文提出的方法获得。在 Dolphins 上,CNCP-CN,CNCP-RA,CNCP-AA 和 CD 得到了几乎相同的 *AUC* 值。在 Football 上,本文方法的性能

弱于 CD 指标，但显著优于 CDI 指标。由此可知，本文算法与基于 CD 的算法相比具有更好的链接预测效果。



(a) Precision

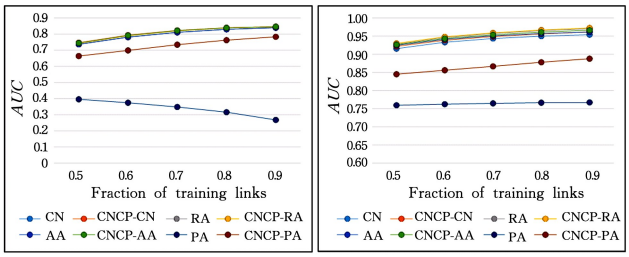


(b) AUC

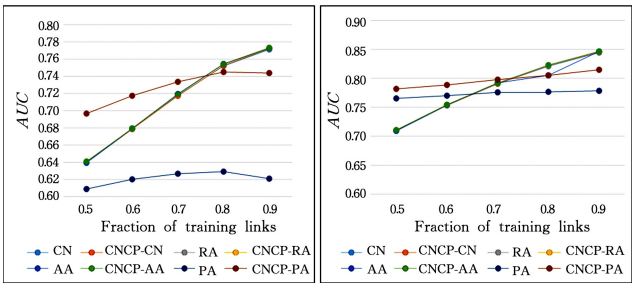
图 2 基于 CNCP 的指标和基于 CD 的指标的性能对比

Fig. 2 Comparison of performance between CNCP-based index and CD-based index

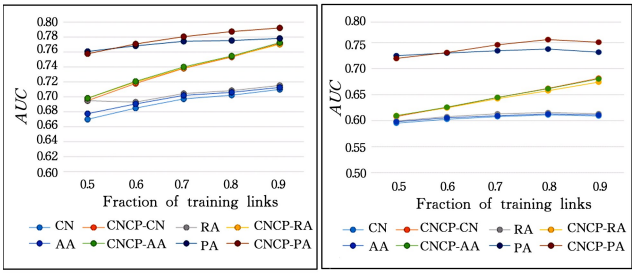
图 3 和图 4 分别给出了在不同训练集比例(从 0.5 增加到 0.9)下，*AUC* 和 *Precision* 的变化。



(a) Football (b) Jazz



(c) Dolphins (d) Email

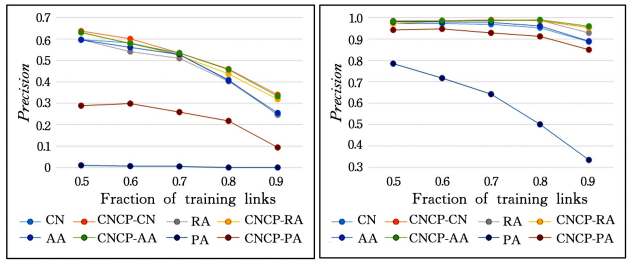


(e) Mangwet (f) Foodweb

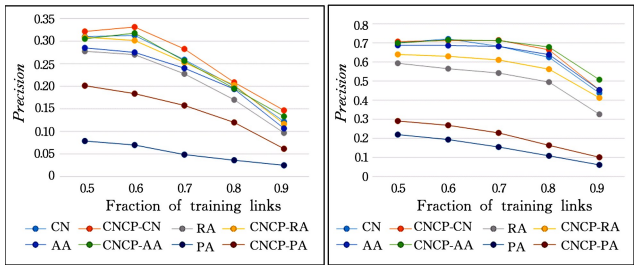
图 3 6 个网络中训练集比例由 50%增长到 90%时 *AUC* 的变化

Fig. 3 Changes of *AUC* when ratio of training set increases from 50% to 90% in six networks

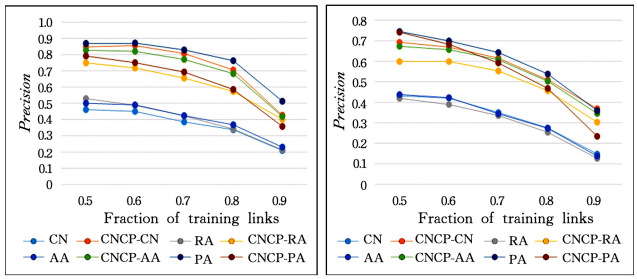
从图 3 中可以发现，当训练集比例从 0.5 增加到 0.9 时，*AUC* 值呈整体上升趋势。这是因为，训练集比例的增加能够提供更多的训练信息，从而提高了 *AUC* 值，相反，训练集比例的降低将增加链接预测的难度。从图 4 中可以看出，与 *AUC* 相比，*Precision* 表现出相反的变化趋势，当训练集比例从 0.5 增加到 0.9 时，*Precision* 分数呈现整体下降的趋势。文献[27]也观察到了这一现象。根据 AUC 的定义，训练集的减少将导致 n_1 的弱化和 n_2 的强化(见式(16))，从而降低 *AUC* 值。而随着测试集的增加(训练集减少)，获取测试集中链接的概率会增加，从而更容易找到缺失链路。因此，在实际应用中需要结合两个度量指标来评价链接预测方法的效果。



(a) Football (b) Jazz



(c) Dolphins (d) Email



(e) Mangwet (f) Foodweb

图 4 6 个网络中训练集比例由 50%增长到 90%时 *Precision* 的变化

Fig. 4 Changes of *Precision* when ratio of training set increases from 50% to 90% in six networks

结束语 本文提出了基于节点连接模式相关性(CNCP)的链接预测模型，该模型结合 CNCP 与基本相似性指标进行链接预测，因此能够通过集成节点的相似性和相关性提高预测的性能。本文将 CNCP 与 CN，RA，AA 及 PA 分别进行结合，设计了 4 种基于 CNCP 的链接预测方法，即 CNCP-CN，CNCP-RA，CNCP-AA 及 CNCP-PA。为了验证提出方法的性能，本文在 6 个真实数据集上进行了实验。实验结果表明，基于 CNCP 的预测方法在评价指标 AUC 和 Precision 上均有较好的表现。在本文的模型中，参数用于调节节点连接模式相关性对链接预测的影响，进一步的研究希望能够根据网络的结构特征自动调整取值，或者至少给出建议的取值范围。

参 考 文 献

[1] LV L,ZHOU T. Link prediction in complex networks:A survey [J]. Physica A:statistical mechanics and its applications,2011,390(6):1150-1170.

[2] CANNISTRACI C V,ALANIS-LOBATO G,RAVASI T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks[J]. Scientific Reports,2013,3:1613.

[3] SHERKAT E,RAHGOZAR M,ASADPOUR M. Structural link prediction based on ant colony approach in social networks [J]. Physica A:Statistical Mechanics and its Applications,2015,419:80-94.

[4] LI F,HE J,HUANG G,et al. Retracted:A clustering-based link prediction method in social networks[J]. Procedia Computer Science,2014,29:432-442.

[5] ZHOU T,LV L,ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B,2009,71(4):623-630.

[6] CLAUSET A,MOORE C,NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature,2008,453(7191):98.

[7] SARUKKAI R R. Link prediction and path analysis using Markov chains[J]. Computer Networks,2000,33(1-6):377-386.

[8] SALTON G,MCGILL M J. Introduction to Modern Information Retrieval [M]. Auckland:McGraw-Hill,1983.

[9] SØRENSEN T A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons[J]. Biologiske Skrifter,1948,5(4):1-34.

[10] NEWMAN M E J. Clustering and preferential attachment in growing networks[J]. Physical Review E,2001,64(2):025102.

[11] LV L,JIN C H,ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E,2009,80(4):046122.

[12] KATZ L. A new status index derived from sociometric analysis [J]. Psychometrika,1953,18(1):39-43.

[13] LEICHT E A,HOLME P,NEWMAN M E J. Vertex similarity in networks[J]. Physical Review E,2006,73(2):026120.

[14] LIU W,LV L. Link prediction based on local random walk[J]. EPL (Europhysics Letters),2010,89(5):58007.

[15] WANG T,WANG H,WANG X. CD-Based indices for link prediction in complex network [J]. PloS One,2016,11(1):e0146727.

[16] ADAMIC L A,ADAR E. Friends and neighbors on the web[J]. Social Networks,2003,25(3):211-230.

[17] BARABÁSI A L,ALBERT R. Emergence of scaling in random networks[J]. Science,1999,286(5439):509-512.

[18] GIRVAN M,NEWMAN M M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences,2002,99(12):7821-7826.

[19] KUNEGIS J. Jazz musicians network dataset [OL]. http://konect. uni-koblenz. de/networks/arenas-jazz.

[20] 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.

[21] ULANOWICZ R E,BONDAVALLI C,EGNOTOVICH M S. Technical Report:CBL 98-123 (1998)[R/OL]. http://www.cbl. umces. edu/atlss/FBay701. html.

[22] LUSSEAU D,SCHNEIDER K,BOISSEAU O J,et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology and Sociobiology,2003,54(4):396-405.

[23] GUIMERA R,DANON L,DIAZ-GUILERA A,et al. Self-similar community structure in a network of human interactions[J]. Physical Review E,2003,68(6):065103.

[24] LATORA V,MARCHIORI M. Efficient behavior of small-world networks[J]. Physical Review Letters,2001,87(19):198701.

[25] LICHTENWALTER R N,LUSSIER J T,CHAWLA N V. New perspectives and methods in link prediction[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2010:243-252.

[26] NEWMAN M E J. Assortative mixing in networks[J]. Physical Review Letters,2002,89(20):208701.

[27] YANG J,ZHANG X D. Predicting missing links in complex networks based on common neighbors and distance[J]. Scientific Reports,2016,6:38208.