基于局部社团和节点相关性的链路预测算法

杨旭华 俞 佳 张 端
(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 基于网络拓扑结构信息的链路预测算法是预测网络未知连边或未来连边的有效方法。在实际应用中,通过进一步提取网络结构信息可以提高网络链路预测结果的精度。文中提出了一种基于局部社团和节点相关性的链路预测算法(HCRP)。该算法把种子节点对的一阶局部社团扩展到二阶局部社团,获得了比一阶局部社团更多的网络结构信息;在用皮尔逊系数计算两个种子节点的相关系数时,该算法也考虑了二阶局部社团的最短路径、边聚类系数和连边密度对两个种子节点相似度的影响,获得了良好的预测网络连边的效果。实验采用了10个真实网络的数据,并对比了HCRP算法和11种知名算法,数值实验结果表明所提算法具有优良的链路预测性能。

关键词 链路预测,皮尔逊系数,二阶局部社团
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.024

Link Prediction Method Based on Local Community and Nodes' Relativity

YANG Xu-hua YU Jia ZHANG Duan
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Link prediction methods based on the network topology information are effect ways to predict unknown or future network edges. In real applications, further extraction and analyzation of network topology is helpful to improve the precision of network link prediction. This paper proposed a new link prediction method based on local community and nodes' relativity(HCRP). By expanding the first-level local communities to second-level ones, this method reveals more network topological information compared with first-level local communities. This method takes the shortest path of the second-level local community, coefficients of edge clustering and the impact of linking edge density on the similarity of two seed nodes into consideration when calculating relative coefficients between two seed nodes by using Pearson correlation coefficient, thus obtaining excellent prediction results of network linking edges. The algorithm was tested and compared with 11 well-known algorithms on 10 real network data sets. Results show that this algorithm has excellent performance of link prediction.

Keywords Link prediction, Pearson correlation coefficients, Second-level local community

1 引言

链路预测指通过网络中已知的节点信息来预测网络中没有直接连接的节点间产生连接的概率[1],其对人们理解复杂网络的演变过程和探索具有重要作用[2-3],例如生物网络中的食物链网络可以用来预测生物网络的演化[4];交通网络的交通数据可以用来预测交通流;电商网络中的数据可以用来预测用户的偏好,从而进行商品推荐;社会网络中的数据可以用来预测人们社交关系的演化等[5-7]。

链路预测算法的一般过程是基于已知的网络信息来预测未知的网络连边。网络信息丰富多样,包括网络节点属性、网络结构信息等。从信息论的角度分析,结构特征比节点特征更易获得,且更可靠[8]。如果想要通过节点信息来获得高精度的预测结果,则需要获取待测节点的最近邻节点的信息。例如,从网络资源分配的角度提出的RA(Resource Allocation)指标[9]与AA(Adamix-Adar)指标[10]均考虑了网络节点度信息,将两个节点之间的资源传递信息量作为相似性指标。再比如,基于共同邻居的CN(Common Neighbors)指标[11],将共同邻居个数作为衡量相似性的标准。另外,在共同邻居的基础上考虑两端节点度的影响,从而衍生出基于局部信息的相似性指标,如Salton指标[12]、Jaccard指标[13]、Sørensen指标[14]、HPI(Hub Promoted Index)指标[15]、HDI(Hub Depressed Iindex)指标[9]以及LHN-I指标[16]等。相比使用网络节点属性,这类算法更容易操作,且其结果具有更高的鲁棒性,因此被广大研究者所关注;但其缺点是此类算法缺乏对未来时刻网络链路的时序性预测[17]。除了结构信息的提取,考

到稿日期:2017-12-08 返修日期:2018-02-24 本文受国家自然科学基金(61374152,61773348),浙江省自然科学基金(LY17F030016,LY16F030014)资助。

杨旭华(1971-),男,博士,教授,主要研究方向为网络科学、智能交通、机器学习,E-mail:xhyang@zjut.edu.cn(通信作者);俞 佳(1993-),女,硕士生,主要研究方向为网络科学、链路预测,E-mail:522607620@qq.com;张 端(1972-),男,博士,副教授,主要研究方向为机器学习。

虑基于路径信息的相似性指标也是链路预测领域的一大热点。基于路径信息的相似性指标可以深度提取高阶局部信息,例如,Katz 指标考虑所有的路径[18],并赋予短路径大权重,同时赋予长路径小权重,将计算路径权重的贡献作为相似性指标。此外,LP(Local Path)指标是在共同邻居的基础上结合路径信息的一种综合性指标,LP 指标提取待测节点之间的三阶邻居的路径信息作为相似性指标[19]。

近年来,链路预测领域的研究者们在不同背景、不同研究的基础上提出了多种链路预测方法。基于共同邻居算法,Cannistraci 等于 2013 年提出了复杂网络中的局部社团概念与 CAR 算法[20],但只提取了一阶共同邻居。2015 年,Daminelli 等又在局部社团的基础上提出了四边形局部社团预测二分网络的概念与 LCP(Local-Community-Paradigm)算法[21],提取了二阶局部社团的信息,但该算法被证明只适用于二分网络。Yang 等于 2016 年在此基础上提出了局部社团的简谐平均路径与边聚类系数的概念,并提出了 LCAR(Local-CAR)算法[22],其在前人的基础上增加了局部路径对预测结果的影响。另外,Liao 等于 2015 年提出采用皮尔逊系数来提取高阶信息的方法,大大减小了提取高阶信息的计算量[23]。

针对以上算法的不足,文中提出一种基于局部社团和节点相关性的链路预测算法 HCRP(High-level Community and Robust Pearson's),基于 CAR[20]算法给出了二阶局部社团信息,用皮尔逊系数计算两个种子节点的相关系数作为对节点间高阶信息的提取方式,同时考虑了二阶局部社团的最短路径、边聚类系数和连边密度对两个种子节点相似度的影响。此外,本文还给出了一些知名的网络模型用于测试算法的精确性与适用性,具体包括 Jazz[24],Football[25],Dolphins[26],Taro[27-28],ContiguousUSA[29],Physicians[30],Karate[31],Zebra[32],Euroroad[33],Hamsterster Full[34] 共 10 个网络,使用 AUC[35] 与 Precision[36] 指标判断本文算法对这 10 个网络的预测效果。

本文第 2 节介绍算法的过程;第 3 节给出实验数据;第 4 节通过 11 个经典指标来对比本文提出的 HCRP 算法以及 10 个网络的实验数据说明,并给出算法的实验结果对比;最后总结全文。

2 基于局部社团和节点相关性的链路预测算法(HCRP)

在实际应用中,传统算法(如 CN[11],RA[10])获取的网络结构共同邻居、度数等特性有助于预测节点对之间是否存在连边。然而单一的特征结构往往不能排除一些隐藏的影响因素,如预测交通拥堵状况时,会出现雨天、雾霾、交通事故等。因此结合多方面特征的算法(如 LCP[21],CAR[19])往往比单特征方法更为有效和鲁棒。

文中提出的 HCRP 算法首先将种子节点对之间的一阶局部社团扩展到二阶局部社团,获得了比一阶局部社团更多的网络信息,如图 1 所示;然后通过使用连边密度的概念描述局部社团集聚性,提高了检测局部社团聚集度的敏感性;通过计算皮尔逊系数来提取高阶局部社团的信息。由于高阶信息提取的复杂性高,且提取时常带有噪声,因此皮尔逊系数适用于对高阶路径信息进行归纳。本文以 x,y 种子节点为例计算 x,y 种子节点之间的相似性分数:

$$HCRP_{xy}=LCT_{xy}*CN_{xy}*SCS_{xy}+\lambda*S_{xy} \tag{1}$$

其中,LCT_{xy} 是二阶社团系数,结合共同邻居以及共同邻居间的连边,能检测种子节点对之间局部社团中连边的聚集度。λ 为可调参数,此处取值为 0.01,S_{xy} 为两个节点的皮尔逊系数。

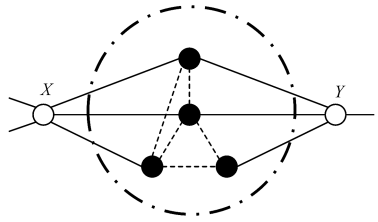


图 1 种子节点对之间的二阶局部社团

Fig. 1 Second-level local community between seed nodes

$HCRP_{xy}$ 指标由两部分组成,$LCT_{xy}*CN_{xy}*SCS_{xy}$ 部分描述了局部社团的二阶局部社团的信息,$\lambda*S_{xy}$ 部分提取了两个种子节点之间的高阶信息。其中,LCT_{xy} 作为二社团系数包含了边聚类系数(提取二阶局部社团边密集度信息)、平均最短路径距离(提取二阶局部社团节点密集度信息)和连边密度(提取同时考虑二阶局部社团节点和边的密集度信息);CN_{xy} 提取种子节点对的二阶共同邻居和一阶共同邻居的信息;SCS_{xy} 提取了共同邻居之间的连边信息;$LCT_{xy}*CN_{xy}*SCS_{xy}$ 部分以乘积的形式表示同时考虑了二阶局部社团的 LCT_{xy},CN_{xy},SCS_{xy} 3 种因素对种子节点对相似性的贡献,有利于提高预测的精确度。

二阶社团系数 LCT_{xy} 的计算涉及到局部二阶社团的信息提取,首先记录局部二阶社团的二阶共同邻居为 CN_{xy},然后记录共同邻居之间的内部边数为 SCS_{xy}。为了描述共同邻居之间连边的密集性,本文使用边聚类系数,计算方法为:

$$K_{xy}=\frac{SCS_{xy}}{\frac{CN_{xy}*(CN_{xy}-1)}{2}} \tag{2}$$

此外,本文考虑局部社团的一阶路径与二阶路径的平均距离,计算方法为:

$$\overline{L}=\frac{1}{AVG} \tag{3}$$

$$AVG_{xy}=\frac{1}{CN_{xy}*(CN_{xy}-1)/2}\sum_{g\leqslant h}\frac{1}{d_{gh}} \tag{4}$$

其中,g 和 h 表示二阶局部群落中的任意两个节点,d_{gh} 为 g 和 h 两个节点之间一阶与二阶的路径长度。

另外,本文探究局部社团节点数与节点之间的连边数之间的关系,提出了局部社团连边密度的概念,节点数越小,连边数越多,局部社团的密度就越大。局部社团的连边密度定义为:

$$D_{xy}=SCS_{xy}/CN_{xy} \tag{5}$$

由以上公式计算二阶社团系数的计算方法为:

$$LCT_{xy}=K_{xy}*D_{xy}/\overline{L} \tag{6}$$

二阶社团系数获取了二阶局部社团的信息,提高了共同邻居算法的敏感性。然而二阶以上的有效信息依旧没有被提

取利用。因此，本文将计算皮尔逊系数，作为高阶信息的提取手段，皮尔逊系数在统计学中被称为皮尔逊积矩相关系数(Pearson Product-moment Correlation Coefficient)，常用于度量两个变量是否存在线性相关性，其值介于－1 与 1 之间。若接近 1 则正相关，若为－1 则负相关，若为 0 则意味着两个变量之间没有线性关系。皮尔逊积矩相关系数的定义如下：

$$S_{xy}=\frac{\sum_k (A_{xk}-\langle A_x\rangle)*(A_{yk}-\langle A_y\rangle)}{\sqrt{\sum_k (A_{xk}-\langle A_x\rangle)^2}*\sqrt{\sum_k (A_{yk}-\langle A_y\rangle)^2}} \tag{7}$$

其中，$\langle A_x\rangle$表示邻接矩阵中第 x 行元素的均值，其中 k 表示邻接矩阵 A 含 k 个节点。

图 1 为二阶局部社团说明，标为 x，y 的种子节点为网络中任意一对无直接连边的节点，x 和 y 的一阶共同邻居节点和二阶共同邻居节点以及这些节点之间的连边构成二阶局部社团。其中 x 和 y 之间长度为 2 的路径的中间的一个节点为一阶共同邻居；长度为 3 的路径的中间的 2 个节点为二阶共同邻居；图 1 中一阶共同邻居和二阶共同邻居用黑色圆点表示，共有 4 个。同时获取一阶共同邻居和二阶共同邻居之间存在的连边数，图 1 中用虚线表示，共有 5 条。

3　实验数据

表 1 列出了 10 个无权无向真实网络数据的网络特性，NN 为网络节点个数，EN 为网络连边数，GC 为度分布基尼系数，CC 为集聚系数。实验中，本文将每个网络数据用随机算法分为训练集数据和测试集数据，其中训练集数据占全网络数据的 90%，测试集数据占全网络数据的 10%。然后使用 AUC 指标和 Precision 指标衡量 12 种算法在不同网络中的预测效果。

表 1　10 个真实网络的网络特性

Table 1　Features of ten real networks

Network	NN	EN	GC	CC
JA	198	5484	0.346	0.52
FB	115	615	0.325	0.309
DO	62	159	0.325	0.309
TA	22	78	0.118	0.275
CU	49	107	0.201	0.406
PH	241	1098	0.244	0.304
KA	34	78	0.385	0.256
ZE	27	111	0.308	0.845
ER	1174	1417	0.241	0.0339
HF	2426	16631	0.589	0.231

表 1 中 Jazz[24] 网络(JA)是一个爵士音乐社团网络，其中每个节点代表一位音乐家，连边代表两位音乐家在乐队中共同演奏。Football[25] 网络(FB)是美国大学足球队网络，节点代表足球队队员，边表示球员之间的互动。Dolphins[26] 网络(DO)是宽吻海豚的社会网络，节点为宽吻海豚，边表示宽吻海豚间频繁的关联。Taro[27-28] 网络(TA)是巴布亚村户之间的礼物赠送网络。一个节点代表一个家庭，两个家庭之间存在边表示赠送了礼物。ContiguousUSA[29] 网络(CU)是美国哥伦比亚的 48 个相邻州和特区，边表示两个州共享一个边界。Physicians[30] 网络(PH)是在城镇伊利诺斯、皮奥里亚、布卢明顿、昆西、盖尔斯堡的 246 名医生之间创新知识传播的网络，一个节点表示一个医生，边表示医生之间的互动。Karate[31] 网络(KA)是空手道俱乐部成员网络，每个节点代表俱乐部的一个成员，每个边代表俱乐部两个成员之间的联系。Zebra[32] 网络(ZA)是 28 头斑马的关系网络，一个节点代表斑马，两个斑马之间的边表明它们在研究过程中有相互作用。Euroroad[33] 网络(ER)是欧洲城市间的路网，每个节点表示一个城市，边表示两个城市之间存在路。Hamsterster Full[34] 网络(HF)是 hamsterster.com 网站中用户的亲友关系网络，每个节点表示一个用户，连边表示用户之间存在亲友关系。

4　数值仿真

4.1　11 种经典链路预测指标

本文选取了几种典型链路预测算法(CN 算法[11]、RA 算法[9]、AA 算法[10]、LCP 算法[21]、CAR 算法[20]、Salton 算法[12]、Jaccard 算法[13]、Sørensen 算法[14]、HPI 算法[15]、HDI 算法[9]以及 LHN-I 算法[16])与本文提出的 HCRP 算法进行比较。

节点 x 和节点 y 为网络中的任意一对无直接连边的节点对，算法计算二者之间的相似性分数，相似性分数越高，连接的可能性就越大。

1)CN[11] 指标将两个节点之间的共同邻居数作为该节点对的相似性分数。将 $\Gamma(x)$ 记作 x 节点的一阶邻居节点集，$\Gamma(y)$ 记作 y 节点的一阶邻居节点集，则该指标定义为：

$$CN(x,y)=|\Gamma(x)\bigcap\Gamma(y)| \tag{8}$$

2)RA[9] 指标出于网络资源分配的角度考虑。资源通过种子节点的共同邻居节点传递，将从 x 节点传递到 y 节点的资源量作为相似性指标，k_z 即表示种子节点 z 的度，该指标定义为：

$$RA(x,y)=\sum_{z\in\Gamma(x)\bigcap\Gamma(y)}\frac{1}{k_z} \tag{9}$$

3)AA[10] 指标认为度小的节点的贡献大于度大的节点，因此种子节点对之间的共同节点的贡献取决于共同节点的度的大小，k_z 即表示种子节点 z 的度，该指标定义为：

$$AA(x,y)=\sum_{z\in\Gamma(x)\bigcap\Gamma(y)}\frac{1}{\lg k_z} \tag{10}$$

4)LCP[21] 指标结合了二阶共同邻居的网络结构信息，形成了一个局部社团。其中 $TCN(x,y)$ 为二阶共同邻居个数，$TSCS(x,y)$ 为二阶共同邻居之间的连边数，该指标定义为：

$$LCP(x,y)=TCN(x,y)*TSCS(x,y) \tag{11}$$

5)CAR[20] 指标结合了一阶共同邻居的网络结构信息，形成了一个一阶局部社团，其中 $CN(x,y)$ 为一阶共同邻居个数，$SCS(x,y)$ 为一阶共同邻居之间的连边数，该指标定义为：

$$CAR(x,y)=CN(x,y)*SCS(x,y) \tag{12}$$

6)Salton[12] 指标(Sal)又名余弦相似性指标，余弦相似性分数越接近 1，说明两个节点的结构性质越吻合，越相似，k_x 即表示种子节点 x 的度，该指标定义为：

$$SAL_{xy}=\frac{|\Gamma(x)\bigcap\Gamma(y)|}{\sqrt{k_x k_y}} \tag{13}$$

7)Jaccard[13] 指标(Jac)是由 Jaccard 提出的指标，其考虑了种子节点之间共同邻居数占种子节点各自一阶邻居总数和

的比例，$|\Gamma(x)\cup\Gamma(y)|$ 表示种子节点 x，y 度的总数，该指标定义为：

$$JAC_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \tag{14}$$

8)Sørensen[14] 指标(Sør)常用于生态学数据研究，考虑共同邻居与种子节点度总和的比值，k_x 即表示种子节点 x 的度，该指标定义为：

$$S\phi r_{xy}=\frac{2*|\Gamma(x)\cap\Gamma(y)|}{k_x+k_y} \tag{15}$$

9)HPI[15] 指标被称为大度节点有利指标，k_x 即表示种子节点 x 的度，该指标定义为：

$$HPI_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{\min\{k_x,k_y\}} \tag{16}$$

10)HDI[9] 指标被称为大度节点不利指标，k_x 即表示种子节点 x 的度，该指标定义为：

$$HDI_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{\max\{k_x,k_y\}} \tag{17}$$

11)LHN-I[16] 指标由 Leicht 等提出，该指标定义为：

$$LHN-I_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{k_x*k_y} \tag{18}$$

其中，分子 $|\Gamma(x)\cap\Gamma(y)|$ 为共同邻居指标，分母 k_x*k_y 为两个种子节点度的成绩，它们之间成正比关系。

为了更全面地比较 12 种链路预测算法，本文使用邻接矩阵的编码方式，针对一对未直接相连的节点连接可能性的计算过程，评估 12 种链路预测算法的时间复杂度。其中 LCP 算法和本文的 HCRP 算法的复杂度为 $O(n^2)$，比其他 10 种算法的复杂度高。这是因为这两种算法都获取了二阶局部社团的信息，导致计算量稍大，而其余的 10 种算法只获取了一阶局部社团的信息，复杂度均为 $O(n)$。为了提高算法的稳定性和精确度，在很多情况下，一定程度的时间代价是值得的。

4.2 指标度量

为了测试算法的准确性，实验中将已知连边数据分成训练集 P 和测试集 P' 两部分，其中训练集占总数据的 90%，测试集占总数据的 10%。训练集 P 中的网络信息为已知信息，测试集 P' 中的信息为待测信息，用于验证测试的准确性。取数据训练集 P 构建内部连通的无向无权网络 $G(V,E)$，E 为连边，V 为节点，$(x,y)\in V$。本文给出 AUC 指标和 Precision 指标来衡量链路预测算法的精确度。AUC 指标从整体上来衡量链路预测算法的准确性，Precision 指标仅仅考虑将相似性分数降序排列后，前 N 个节点对预测的准确性。

4.2.1 AUC

简单来说，AUC 指标是指通过预测算法得到的测试集 P' 中存在的节点对的相似性分数比随机选择的不存在的节点对的相似性分数高的概率。也就是说，在利用训练集 P 中的网络信息时，使用链路预测算法得到测试集 P' 中的节点对之间连接的相似性分数。每次随机从测试集中选取一条边的相似性分数与随机选择不存在的边的相似性分数进行比较。记录测试集中的链路相似性分数大于不存在的边的相似性分数为 N_1，测试集中的链路相似性分数等于不存在的边的相似性分数为 N_2，比较总次数为 N，则 AUC 指标定义为：

$$AUC=\frac{N_1+0.5*N_2}{N} \tag{19}$$

AUC 值越大，证明预测节点之间连接的可能性越大，链路预测算法越好。

4.2.2 Precision

Precision 指标是指通过链路预测算法计算出的测试集中相似性分数最高的前 m 个数据被准确预测的比例，如果有 n 个预测是准确的，则 Precision 指标为：

$$Precision=\frac{n}{m} \tag{20}$$

$Precision$ 值越大，证明预测越准确。在链路预测的应用中，有时不需要衡量预测的整体水平，仅仅需要知道前几个最大概率的预测结果是否正确即可。例如，在电商推荐喜好商品时，只需要给顾客提供其最可能喜爱的商品就能满足需求。

4.3 仿真结果比较

表 2 列出了 12 种算法在 10 个网络中的 AUC 指标数据。在链路预测中，AUC 作为模型评价指标，可以很好地描述算法的整体预测效果。表 3 列出了 12 种算法在 10 个网络中的 Precision 指标数据。Precision 指标将算法的相似性分数从大到小排列，取前 m 个数据，若其中有 n 个数据被准确预测，则准确率为 n/m。在实验中，取 m 为测试集数据的 10%。表 2 和表 3 中的数据均为计算 50 次结果的平均值。为了测评算法的稳定性，图 2 给出 12 种链路预测算法在 10 个真实网络中的 AUC 指标结果，图 3 给出 12 种链路预测算法在 10 个真实网络中的 AUC 指标堆积柱形图，图 4 给出 12 种链路预测算法在 10 个真实网络中的 Precision 指标柱形图，为了测评算法的稳定性，图 5 给出表 2 中 AUC 指标计算 50 次数据的标准差，图 6 给出表 3 中 Precision 指标计算 50 次数据的标准差。

表 2 12 种链路预测算法在 10 个真实网络中的 AUC 指标对比

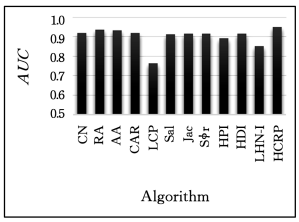
Table 2 Comparision of AUC index of twelve algorithms on ten real networks

	CN	RA	AA	CAR	LCP	Sal	Jac	Sør	HPI	HDI	LHN-I	HCRP
JA	0.920	0.937	0.932	0.920	0.764	0.914	0.915	0.915	0.893	0.916	0.850	**0.950**
FB	0.803	0.800	0.808	0.801	0.759	0.815	0.819	0.816	0.817	0.815	0.820	**0.877**
DO	0.741	0.715	0.743	0.711	0.761	0.734	0.738	0.738	0.728	0.741	0.728	**0.776**
TA	0.555	0.564	0.558	0.555	0.277	0.588	0.661	0.582	0.585	0.640	0.675	**0.697**
CU	0.822	0.830	0.818	0.820	0.790	0.833	0.833	0.833	0.776	0.832	0.835	**0.907**
PH	0.769	0.771	0.776	0.769	**0.894**	0.776	0.775	0.775	0.776	0.774	0.773	0.782
KA	0.673	0.682	0.677	0.674	**0.722**	0.658	0.672	0.672	0.660	0.660	0.668	0.697
ZE	0.937	0.927	0.748	0.939	0.883	0.725	0.774	0.774	0.774	0.774	0.771	**0.969**
ER	0.529	0.541	0.527	0.531	0.528	0.529	0.528	0.531	0.531	0.530	0.531	**0.560**
HF	0.926	0.926	0.923	0.926	0.928	0.926	0.925	0.924	0.924	0.923	0.922	**0.930**

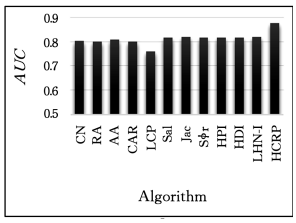
表 3　12 种链路预测算法在 10 个真实网络中的 Precision 指标对比

Table 3　Comparision of Precision index of twelve algorithms on ten real networks

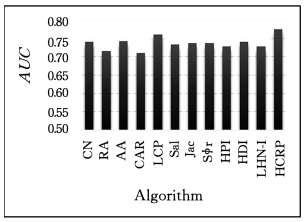
	CN	RA	AA	CAR	LCP	Sal	Jac	Sør	HPI	HDI	LHN-I	HCRP
JA	0.261	0.305	0.296	0.257	0.069	0.245	0.242	0.242	0.187	0.231	0.09	**0.305**
FB	0.266	0.274	0.267	0.305	0.010	0.307	0.301	0.308	0.293	0.309	0.334	**0.351**
DO	0.100	0.057	0.080	0.080	0.047	0.040	0.053	0.053	0.010	0.063	0.010	**0.113**
TA	0.007	0.029	0.036	0.007	0.001	**0.114**	0.079	0.100	0.079	0.071	0.093	0.100
CU	0.105	0.130	0.100	0.010	0.020	0.150	0.120	0.120	0.088	0.12	0.160	**0.165**
PH	0.089	0.096	0.104	0.092	0.070	**0.128**	0.090	0.090	0.088	0.079	0.047	0.100
KA	0.167	0.133	0.142	0.167	0.025	0.010	0.001	0.002	**0.183**	0.001	0.002	0.167
ZE	0.655	0.686	0.182	0.659	0.359	0.200	0.195	0.195	0.195	0.195	0.200	**0.686**
ER	**0.008**	0.006	0.007	0.007	0.007	0.001	0.001	0.001	0.002	0.001	0.002	0.007
HF	0.149	0.226	0.213	0.160	0.132	0.190	0.082	0.083	0.039	0.037	0.077	**0.227**



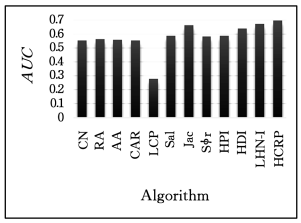
(a)Jazz 网络(JA)仿真下 AUC 指标对比



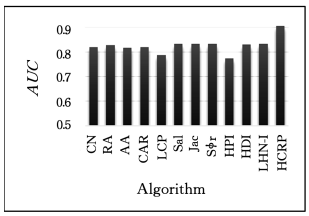
(b)Football 网络(FB)仿真下 AUC 指标对比



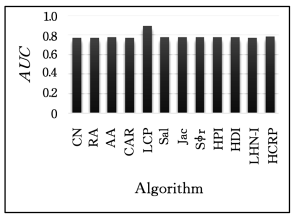
(c)Dolphins 网络(Do)仿真下 AUC 指标对比



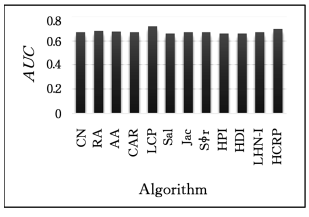
(d)Taro 网络(TA)仿真下 AUC 指标对比



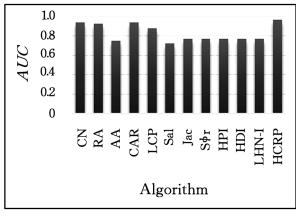
(e)ContiguousUSA 网络(CU)仿真下 AUC 指标对比



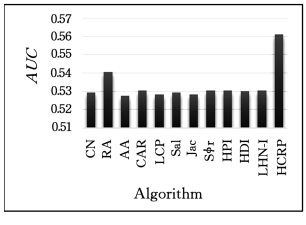
(f)Physicians 网络(PH)仿真下 AUC 指标对比



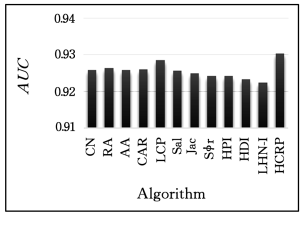
(g)Karate 网络(KA)仿真下 AUC 指标对比



(h)Zebra 网络(ZE)仿真下 AUC 指标对比



(i)Euroroad 网络仿真下 AUC 指标对比



(j)Hamsterster Full 网络(HF)仿真下 AUC 指标对比

图 2　12 种链路预测算法在 10 个真实网络中的 AUC 指标结果图

Fig. 2　AUC index results of twelve algorithms on ten real networks

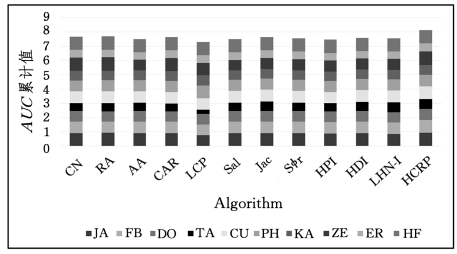


图 3　12 种链路预测算法在 10 个真实网络中的 AUC 指标堆积柱形图

Fig. 3　Stacked clumn of AUC index of twelve algorithms on ten real networks

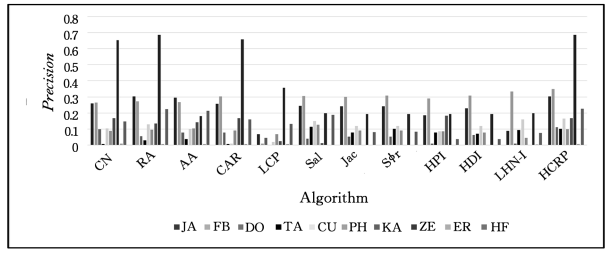


图 4　12 种链路预测算法在 10 个真实网络中的 Precision 指标柱形图

Fig. 4　Histogram of Precision index of twelve algorithms on ten real networks

从表 2 和表 3 可以发现，HCRP 算法在大部分网络中的表现优于其他指标，而 Physicians 网络(PH)与 Karate(KA)网络的 AUC 指标表现较差，说明了 HCRP 算法指标的优越性。

在图 2 中，HCRP 算法对大部分网络的预测效果都优于其他传统算法，其中 ContiguousUSA 网络的精确度提高得最明显。但 Physicians 网络和 Karate 网络的 *LCP* 指标略大于 HCRP 指标。另外，从图中可以看出，HCRP 算法对于小网络表现较好，解决了因为网络小、信息提取少而导致的精确度低的问题。

在图 3 中，本文使用 AUC 指标的堆积柱形图来衡量 12 种算法的预测效果。从图 3 中可看出，HCRP 算法的每种颜色面积均匀，其次是 CN 算法、RA 算法，说明这 3 种算法能够稳定预测各网络。从纵坐标上看，HCRP 算法的总量第一，其次是 RA 算法。

在图 4 中，本文给出了 Precision 指标的柱形图。其中 Zebar 网络的预测效果普遍较高，Taro 网络的预测效果普遍较低，相比于其他算法，HCRP 算法的预测结果最稳定。

为了测评算法的稳定性，本文在图 5 中给出了 12 种算法的 AUC 指标 50 次计算结果的标准差。在图 6 中，给出了 12 种算法的 Precision 指标 50 次计算结果的标准差。其中

Hamsterster Full,Euroroad,Jazz,Physicians 这 4 种网络的标准差较小,其余 6 种网络的标准差相对稍大。从表 1 中可知,10 种真实网络的节点大小排列为:Hamsterster Full>Euroroad>Physicians>Jazz>其他 6 种网络;连边数大小排列为:Hamsterster Full>Jazz>Euroroad>Physicians>其他 6 种网络。因此,我们可以认为网络越大,算法的预测结果就越稳定。

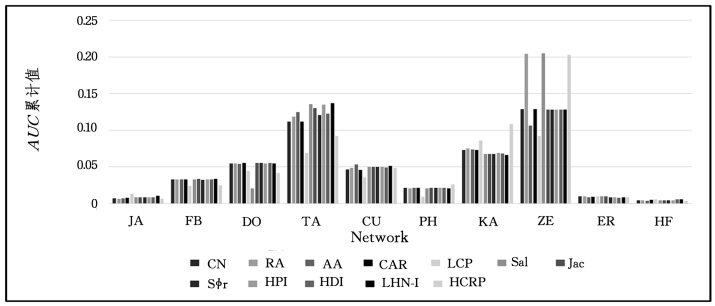


图 5 12 种链路预测算法在 10 个真实网络中的 AUC 指标标准差

Fig. 5 Standard deviation of AUC index of twelve algorithms on ten real networks

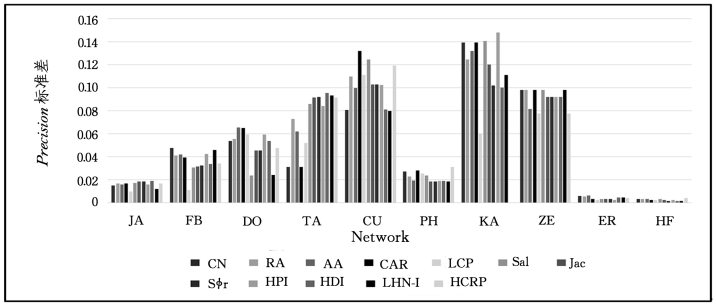


图 6 12 种链路预测算法在 10 个真实网络中的 Precision 指标标准差

Fig. 6 Standard deviations of Precision index of twelve algorithms on ten real networks

4.4 实验分析

本实验中选取 11 种经典算法与本文算法进行对比。这 12 种算法都是基于共同邻居算法(CN 算法)提出的,其中 RA 算法和 AA 算法着重关注共同邻居节点的度为待测节点的资源分配做出的贡献。LCP 算法和 CAR 算法更关注共同邻居之间的结构紧密性,为待测节点提供了预测信息。而 Salt 算法、Jac 算法、Sør 算法、HPI 算法、HDI 算法以及 LHN-I 算法将 CN 算法作为分子,通过获取待测节点周边的结构信息作为分母,调节了共同邻居对预测结果影响的比重,使得在预测两对拥有相同共同邻居个数的待测节点时有更精确的区分。本文的 HCRP 算法中二阶社团系数从平均路径距离、连边密度、边聚类系数等方向尽力提取结构紧密性的信息,又通过皮尔逊系数提取了节点之间的高阶信息,从而提高了链路预测算法的精确度。由图 2 分析,针对不同网络,LCP 算法的预测结果不是非常稳定,说明 LCP 算法的信息提取受网络结构的影响大,提取的信息随网络结构的不同掺杂着不同程度的噪声;而 CAR 算法的预测结果比较中庸,相比于 HCRP 算法,其信息提取不足。由图 3 和图 4 可知,HCRP 算法在不同网络条件下的预测效果比较稳定。在 Precision 指标测试下,相似性分数高的节点对的预测结果也相对稳定。总体上看,本文算法同时考虑了局部社团信息以及高阶信息,较全面地提取了两个待测节点的结构信息,因此总体上的效果优于其他算法。

结束语 链路预测具有广泛的应用价值,虽然链路预测领域日趋发展成熟,但现在仍没有百分百预测准确的算法存在,在预测过程中往往包含不能去除的噪声。因此,不断提升链路预测算法的精度是一项长远的工程。本文算法为了增强二阶局部社团信息的提取,提出了二阶社团系数的概念。另外,本文算法使用皮尔逊系数提取高阶局部社团的信息,大大降低了高阶信息提取的复杂度,且算法的稳定性较好,对于小网络仍然具有良好的预测效果。

参 考 文 献

[1] ZHAO J,MIAO L,YANG J,et al. Prediction of Links and Weights in Networks by Reliable Routes[J]. Scientific Reports,2015,5:12261.

[2] WANG W Q,ZHANG Q M,ZHOU T. Evaluating Network Models:A Likelihood Analysis[J]. Epl,2011,98(2):28004.

[3] ZHANG Q M,XU X K,ZHU Y X,et al. Measuring multiple evolution mechanisms of complex networks[J]. Scientific Reports,2015,5(1):10350.

[4] YU H,BRAUN P,YILDIRIM M A,et al. High-quality binary protein interaction map of the yeast interactome network[J]. Science,2008,322(5898):104-110.

[5] WANG P,XU B W,WU Y R,et al. Link prediction in social networks:the state-of-the-art[J]. Science China,2015,58(1):011101.

[6] HU W B,PENG C,LIANG H L,et al. Event Detection Method Based on Link Prediction for Social Network Evolution[J]. Journal of Software,2015,26(9):2339-2355. (in Chinese)
胡文斌,彭超,梁欢乐,等. 基于链路预测的社会网络事件检测方

法[J]. 软件学报,2015,26(9):2339-2355.

[7] ZENG A,XU X Q. Hybrid Collaborative Filtering Recommendation Algorithm Based on Friendships and Tag[J]. Computer Science,2017,44(8):246-251. (in Chinese)
曾安,徐小强. 基于好友关系和标签的混合协同过滤算法[J]. 计算机科学,2017,44(8):246-251.

[8] ZHU B,XIA Y. An information-theoretic model for link prediction in complex networks[J]. Scientific Reports,2015,5:13707.

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

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

[11] FRANÇOIS L,HARRISON C. White. Structural equivalence of individuals in social networks[J]. Social Networks,1977,1(1):67-98.

[12] DILLON M. Introduction to modern information retrieval[J]. Information Processing & Management,1983,19(6):402-403.

[13] JACCARD P. Etude de la distribution florale dans une portion des Alpes et du Jura[J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles,1901,37(142):547-579.

[14] SØRENSEN T J. 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.

[15] RAVASZ E. Hierarchical organization of modularity in metabolic networks[J]. Science,2002,297(5586):1551-1555.

[16] LEICHT E A,HOLME P,NEWMAN M E. Vertex similarity in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2006,73(2 Pt 2):026120.

[17] ZHANG Y Q,LU Y L,YANG G Z. Link Prediction of AS Level Internet Based on Association Rule of Frequent Closed Graphs [J]. Computer Science,2016,43(s1):314-318. (in Chinese)
张岩庆,陆余良,杨国正. 基于频繁闭图关联规则的 AS 级 Internet 链路预测方法[J]. 计算机科学,2016,43(s1):314-318.

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

[19] LÜL,JIN C H,ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E,2009,80(2):046122.

[20] CANNISTRACI C V,ALANISLOBATO 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(4):1613.

[21] DAMINELLI S,THOMAS J M,DURÁN C,et al . Common neighbours and the local-community-paradigm for link prediction in bipartite networks [J]. New Journal of Physics,2015,17(11):1-8.

[22] YANG X H,ZHANG H F,LING F,et al. Link prediction based on local community properties[J]. International Journal of Modern Physics B,2016,30(31):12.

[23] LIAO H,ZENG A,ZHANG Y C. Predicting missing links via correlation between nodes[J]. Physica A Statistical Mechanics & Its Applications,2015,436(1):216-223.

[24] PABLO M. GLEISER,LEON D. Community structure in jazzZ [J]. Advances in Complex Systems,2003,6(4):565-573.

[25] BULDYREV S V. Robustness of interdependent networks under targeted attack[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2011,83(2):065101.

[26] 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 & Sociobiology,2003,54(4):396-405.

[27] HAGE P,HARARY F. Structural models in anthropology[M]. Cambridge:Cambridge University Press,1983:705-714.

[28] SCHWIMMER E G. Exchange in the social structure of the Orokaiva[D]. Columbia:University of British Columbia,1961.

[29] KNUTH D E. The Stanford GraphBase:a platform for combinatorial computing[M]. New York: Association for Computing Machinery,1993.

[30] BURT R S. Social Contagion and Innovation:Cohesion Versus Structural Equivalence [J]. American Journal of Sociology,1987,92(6):1287-1335.

[31] ZACHARY W W. An Information Flow Model for Conflict and Fission in Small Groups [J]. Journal of Anthropological Research,1977,33(4):452-473.

[32] SUNDARESAN S R,FISCHHOFF I R,DUSHOFF J,et al. Network metrics reveal differences in social organization between two fission-fusion species,Grevy's zebra and onager[J]. Oecologia,2007,151(1):140-149.

[33] ŠUBELJ L,BAJEC M. Robust network community detection using balanced propagation[J]. European Physical Journal B,2011,81(3):353-362.

[34] JÉRÔME K. KONECT:the Koblenz network collection[C]// International Conference on World Wide Web Companion. New York:ACM,2013:1343-1350.

[35] HANLEY J A,MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology,1982,143(1):29-36.

[36] HERLOCKER J L. Evaluating collaborative filtering recommender systems[J]. Acm Transactions on Information Systems,2004,22(1):5-53.