Doi:10.12051/j.issn.1674-4942.2022.01.004

一种基于节点路径信息相似性的预测方法

黄寿孟[1,2]
(1.三亚学院 信息与智能工程学院,海南 三亚 572022;
2.三亚学院 陈国良院士团队创新中心,海南 三亚 572022)

摘 要:链路预测计算是在复杂网络分析任务中最重要和最具挑战性的任务之一,能根据网络中现有的链接预测缺失的链接并广泛应用于多种学科领域,包括社会网络分析、推荐系统和生物网络等。文中提出一种基于路径节点信息相似性的预测方法,该预测方法是利用节点共有的特征信息来推测下一个相关的路径节点信息,从而优化现有的基于路径预测方法。首先,由于网络中一对可达节点之间的最短距离始终是唯一的,所以通过合并本地路径节点信息计算出网络中所有节点对之间最短路径中的路径最大长度,从而提高节点路径相似性的准确性;接着,以类似的方式计算遍历所有未连接的节点对,计算它们预测分配的权重得分,得到将来可能会链接起来的节点相似性得分,即是这些节点中某两个节点的最短路径;最后,采用来自不同领域的真实世界数据集进行实验实证。结果表明,与现有的最先进的预测方法相比,该方法具有较高的预测精度。

关键词:复杂网络;链路预测;节点信息
中图分类号:TP393 **文献标志码**:A **文章编号**:1674-4942(2022)01-0025-06

A Link Prediction Method Based on Similarity of Node Path Information

HUANG Shoumeng[1,2]
(1. *School of Information & Intelligence Engineering*, *University of Sanya*, *Sanya* 572022, *China*;
2. *Academician Guoliang Chen Team Innovation Center*, *University of Sanya*, *Sanya* 572022, *China*)

Abstract: Link prediction computing is one of the most important and challenging tasks in complex network analysis. It predicts missing links based on existing links in networks, and is widely used in various scientific fields, including social network analysis, recommendation systems and biological networks. In this paper, a prediction method based on the similarity of path node information was proposed. The prediction method used the common feature information of nodes to predict the next related path node information, so as to optimize the existing path prediction method. As the shortest distance between a pair of nodes in network is always unique, by combining the local path node information, calculating the shortest path between all nodes in network to the maximum length of path and to improve the accuracy of node path similarity, and then calculating traverse all the unconnected nodes in a similar way and weighted score of distribution, the similarity score of nodes that may be linked in the future was obtained, that is, the shortest path of two nodes in these nodes. Finally, the real world data sets from different fields were used for experimental verification. The results show that the method has higher prediction accuracy, compared with the most advanced prediction methods.

Keywords: complex networks; link prediction; node information

许多现实世界的复杂系统可以通过复杂网络来表示,其节点表示实体,链路表示节点之间的交互[1]。因

收稿日期:2021-07-29
基金项目:海南省高等学校科学研究一般项目(Hnky2021-51)
作者简介:黄寿孟(1975—),广东湛江人,副教授,主要研究方向为信息技术与信息安全。E-mail:huang123888@126.com

此复杂网络被人们广泛研究与应用,其中研究复杂网络的链路预测是学者们热点问题之一,它的研究目标是估计两个尚未连接的节点之间存在链路的可能性[2]。目前,许多链路预测算法关注的都是基于节点相似性的思想[3],利用节点的基本属性定义节点相似性,也就是说如果两个节点具有许多共同的拓扑特征,那么它们就被认为是相似的[4]。比如共同邻居算法[5]、Adamic Adar[6]、Jaccard系数[7]等局部相似度量算法可以非常有效地计算,并在许多情况下都有良好的性能;还有全局度量算法[8-9]、基于本地路径信息算法[10]都有良好Katz指数、通勤时间、节点信息的预测性能。文献[11]针对加权网络综合考虑网络中边的聚类和扩散特性提出基于链接拓扑权重的含权预测指标的预测方法。文献[12]提出一种基于社团节点相关性的链路预测算法,把节点对扩展到二阶局部社团获取更多的网络结构信息。文献[13]根据复杂网络样本邻接矩阵化处理以及采用AdaBoost算法进行分类训练获取权重投票预测结果。文献[14]设计开发了结合网络拓扑特征、基本特征和附加特征的TNTlink模型,并结合物理学和计算机科学的领域知识,利用深度神经网络将这些特征集成到一个深度学习框架中从而解决链路预测问题。为了优化现有的基于路径预测方法,本文提出了两个节点u和v之间的新相似性索引,该索引是根据距离u和v固定距离内的节点的局部信息计算得出的。换句话说,索引信息不仅包含了直接连接到u与v的节点,还合并了位于u到v的所有可能长度较小的路径上的节点信息,最后比较两个使用最广泛的相似性指标,Adamic Adar和Katz Index的性能。

1 相关工作

网络$G=(V,E)$由节点的有限非空集合V和无序顶点对(称为链接)的有限集合E组成。网络可以是有向的(在每个边缘都分配有方向的地方),也可以是无向的(没有为边缘分配任何方向的地方)。本文只考虑无向简单网络,其中不允许多个链接和自环。顶点的度数$v\in V$是与v相连的邻居数。网络$G=(V,E)$中的行走w是一系列交替的顶点和边$v_0,e_1,v_1,e_2,v_2,\cdots,e_k,v_k$。其中$v_i\in V$和$e_i=(v_{i-1},v_i)$。该行走具有长度$k$,其定义为行走中的顶点数。无向简单网络$G=(V,E)$的邻接矩阵类型是$|V|\times|V|$。如果$u$和$v$有链接,则邻接矩阵$(u,v)$值为1,否则为0。邻接矩阵$(A^k)_{uv}$的第$k$次幂表示行走的次数,即从$u$到$v$的链接路径的长度为$k$。目前常用的本地和准本地/全局链接预测方法有:

1.1 公共邻居(CN)

如果两个节点存在许多公共邻居,则它们最有可能具有链接,用这种方法计算两个节点的公共邻居数[5],其计算公式为

$$\mathrm{CN}(u,v)=|\Gamma(u)\cap\Gamma(v)| \tag{1}$$

CN算法在许多情况下是评估其他方法性能的基准,CN也可以表示为$\mathrm{CN}(u,v)=(A^2)_{uv}$,其中$A$是网络的邻接矩阵。

1.2 Adamic Adar(AA)

该算法目标是在通过为较少联系的邻居分配更多权重来提高普通邻居的准确性[6],计算公式为

$$\mathrm{AA}(u,v)=\sum_{w\in \mathrm{CN}(u,v)}\frac{1}{\log|\Gamma(w)|} \tag{2}$$

其中w是u和v的共同邻居。在社交网络中,此算法可以计算不受欢迎的人(朋友数量较少的人)可能更容易将一对特定的朋友彼此介绍。

1.3 Jaccard(JC)

Jaccard算法是从两个节点的公共邻居计算出的另一种相似性度量[7],它定义为

$$\mathrm{JC}(u,v)=\frac{\mathrm{CN}(u,v)}{|\Gamma(u)\cup\Gamma(v)|} \tag{3}$$

JC的值可以解释为节点u和v的公共邻居的概率。

1.4 Katz

该算法基于整个网络的拓扑考虑节点对之间所有路径的集合[10]。路径的长度会以指数方式衰减,从而为较短的路径赋予更大的权重,它计算公式为

$$\mathrm{Katz}(u,v)=\sum_{l=1}^{\infty}\beta^{l}\left[path_{uv}^{l}\right] \tag{4}$$

其中$0<\beta<1$是控制不同长度路径的权重的参数。β的值非常小将导致惩罚较长长度的路径，并且在这种情况下将度量减小为CN。注意，相似度矩阵S也可以计算为$(I-\beta A)^{-1}-I$，其中，I表示大小为$|V|\times|V|$的单位矩阵。

1.5 局部路径（LP）

为了在准确性和复杂性之间取得良好的平衡，考虑较短长度的局部路径的LP算法[15]，计算公式为

$$\mathrm{LP}(u,v)=\sum_{i=2}^{l}\beta^{i-2}A^{i} \tag{5}$$

其中A是网络的邻接矩阵。与Katz算法一样，β设置为较小的值，以便较短的路径获得更大的权重。请注意，如果$\beta=0$，则该算法会退化为公共邻居。实际上，由于其计算较复杂，该度量通常在$L=3$时使用，从而将其减小为$\mathrm{LP}(u,v)=A^{2}+\beta A^{3}$。

1.6 优先联系PA

该算法表示在社交网络中有很多朋友的用户将来倾向于获得更多的联系[16-19]，其计算公式为

$$\mathrm{PA}(u,v)=|\Gamma(u)|\cdot|\Gamma(v)| \tag{6}$$

2 预测模型

2.1 设计背景

基于相似度的链接预测方法通常取决于网络的拓扑结构，这种方法的目的是基于节点的本地连接结构预测一对节点之间的丢失链接。本地链路预测方法使用两个节点的直接邻居的节点信息来预测它们之间的链接。例如，AA算法将较高的权重分配给程度较小的公共邻居。直觉是，度数较小的公共邻居在预测两个节点之间的链接时更为重要。虽然局部相似性算法仅考虑直接连接到查询节点的那些节点，但是全局方法（例如Katz算法和LP算法）考虑整个网络的拓扑。但是全局方法不考虑连接节点的程度。因此需要一种可以同时利用局部和全局相似性指标来测量相似性的链路预测方法。

2.2 PSI预测方法

若未加权网络$G=(V,E)$存在$(u,v)\in E'$，如何预测两个节点u和v之间的链接。为了估计节点u和v之间潜在连接的可能性，考虑到网络中所有与节点u和v处于固定距离内的节点，笔者提出一种相似性指数（简称为PSI）计算方法。

定义1　若$\Gamma^{k}(u)$代表与节点u的最短距离为k的所有节点的集合，即节点u和v之间的相似性指数$\mathrm{PSI}(u,v)$定义为

$$\mathrm{PSI}(u,v)=\sum_{\substack{i+j=2\\ i,j>0}}^{l}\beta^{i+j-2}\left[\sum_{w\in\{\Gamma i(u)\cap\Gamma j(v)\}}\frac{1}{\log|\Gamma(w)|}\right] \tag{7}$$

其中β是权重系数，若距离大于1的节点的权重较小，则L是在计算$\mathrm{PSI}(u,v)$时考虑的最长路径的长度。此方法结合了Adamic Adar和LP算法的优点，考虑水平范围比AA宽的本地路径（AA仅考虑两个节点的直接邻居），同时也考虑了每个节点的程度（LP算法仅计算本地路径，而忽略程度）。

由于网络中一对可达节点之间的最短距离始终是唯一的，因此建议在相似性指数PSI计算中每个节点仅被考虑一次，从而得出PSI的等效定义。

定义2　两个节点u和v之间的相似性指数$\mathrm{PSI}(u,v)$定义为

$$\mathrm{PSI}(u,v)=\sum_{\substack{w\in V\{u,v\}\\ 1<d=\delta(u,w)+\delta(v,w)}}\frac{\beta^{d-2}}{\log|\Gamma(w)|} \tag{8}$$

其中$\delta(u,v)$表示节点u和v之间的最短距离。

2.3 PSI算法

对于$L=2$，PSI简化为AA，LP指数简化为CN。因此假设$L>2$，$\mathrm{PSI}(u,v)$的值计算如下：

若节点u和v未连接,首先确定节点u和v的公共邻居,并通过取其度的对数的倒数来计算其分数。该分数被添加到相似性指数PSI(u,v)。接着计算连接到两个节点u和v中的一个但距另一个节点的最短路径恰好是两个的节点。它们的分数是以类似的方式计算的,即取其度的对数的倒数。但是它们在预测得分的计算中分配的权重较小。这是通过将他们的分数乘以$\beta(0<\beta<1)$来完成的,然后将其值添加到PSI(u,v)。该过程一直持续到计算出与节点u和v的距离小于或等于L的所有节点的分数。

注意这时L的最大值可以是$2d$,其中d是网络的直径,其定义为网络中所有节点对之间最短路径中的路径最大长度。

下面是计算相似性指数PSI的算法过程,其中G^T表示用于训练学习的网络集合,E^T表示用于训练学习的邻接节点对集合:

1:输入:网络$G^T=(V,E^T)$,具有邻接矩阵A,距离长度L值和权重参数β。

2:输出:相似度得分矩阵S。

3:$s \leftarrow 0$

(将相似矩阵初始化为0)

4:$\delta \leftarrow$ All Pair Shortest Paths(G^T)

(计算最短路径)

5:for $(u,v) \in E \setminus E^{\mathrm{T}}$ do

6: for $w \in V \setminus \{u,v\}$ do

7: $d=\delta(u,w)+\delta(v,w)$

8: if $d \leqslant 1$ then

9: PSI$(u,v) \leftarrow$ PSI$(u,v)+\beta^{d-2}(\log|w|)^{-1}$ (更新相似性得分)

10: end if

11: end for

12:end for

在PSI算法中,接受网络G^T的邻接矩阵作为输入,该矩阵是通过去除E^P中的链接而获得的。它还接受权重参数β和L(L的最大值为最大路径的长度)。首先计算网络中所有节点对之间的最短路径,for循环(第5至12行)遍历所有未连接的节点对,将来可能会链接起来。通过考虑网络中所有节点的相似性得分,这些节点从两个节点到最短路径的总和小于L,而相似性得分根据等式在第11行中更新。

3 实验结果

3.1 数据集

本文使用KONECT网络上12个公开可用的数据集[1],各数据集网络的拓扑特性如表1所示,其中$|V|$和$|E|$分别是节点和链接的数量,CC是聚类系数,$|k|$和$|d|$是平均度和平均路径长度,ρ表示网络的密度,H是网络的异质性。

3.2 评估指标

为了估计预测算法的准确性,本文使用标准度量AUC精度指标[1]。如果在n个独立预测比较中,n'是缺失链接具有较高分数的次数,n''是缺失链接和不存在具有相同分数的链接的次数,则将精度AUC$=(n'+0.5n'')/n$如果所有链接分数均根据独立的概率相同分布随机生成,则预测算法的AUC精度指标应约为0.5。因此,若AUC>0.5,即表示该算法的预测性能效果好,且AUC值越接近1表示预测效果越好。

3.3 结果

本文将数据集网络的链路集合E随机分为两组,即训练集E^T和测试集E^P。训练集包含90%的链接,而其余10%的链接用于测试目的,使用相同的集合E^T和E^P来评估各种预测方法的性能。同时在PSI,Katz和LP算法中参数值$\beta=0.005$,$L=3$,每个数据集进行100次实验,从而得到每个算法平均精度AUC值见表2。

表1　数据集网络的拓扑特性

Table 1　Topological characteristics of the data set network

数据集	$\|V\|$	$\|E\|$	CC	$\|k\|$	$\|d\|$	ρ	H
Karate	34	78	0.588	4.588	1.204	0.139	7.769
US-Roads	49	107	0.507	4.367	2.082	0.091	4.935
Dolphin	62	159	0.303	5.129	1.678	0.084	6.805
Train-Bombing	64	243	0.711	7.594	1.345	0.121	12.597
Neurons	279	2 287	0.337	16.394	1.218	0.059	25.916
E.Coli	329	456	0.222	2.772	2.421	0.008	12.314
Netscience	379	914	0.798	4.823	3.021	0.013	8.021
Infectious	410	17 298	0.467	84.380	1.815	0.206	2.992
Metabolic	453	4 596	0.782	20.291	1.332	0.045	17.903
USAir	500	2 980	0.726	11.920	1.496	0.024	53.785
Email	1 133	5 451	0.254	9.622	1.803	0.009	18.688
Yeast	2 375	11 693	0.388	9.847	2.548	0.004	34.223

表2　各算法在不同数据集下的预测精度AUC

Table 2　AUC for each algorithm under different data sets

数据集	PSI	Katz	LP	CN	AA	PA	JC
Karate	0.784 3	0.762 8	0.766 2	0.702 8	0.736 6	0.731 8	0.613 8
US-Roads	0.925 0	0.894 4	0.895 5	0.897 5	0.904 2	0.440 6	0.918 7
Dolphins	0.842 5	0.838 4	0.826 9	0.786 3	0.790 2	0.667 4	0.785 0
Train-Bombing	0.948 5	0.930 9	0.931 2	0.932 2	0.943 8	0.798 5	0.929 8
Neurons	0.882 8	0.866 0	0.867 0	0.859 3	0.874 7	0.723 8	0.830 5
E.Coli	0.894 9	0.884 6	0.866 3	0.621 3	0.628 1	0.878 8	0.612 0
Netscience	0.992 9	0.986 1	0.986 0	0.981 1	0.984 9	0.661 3	0.978 2
Infectious	0.955 4	0.948 3	0.947 7	0.912 3	0.915 1	0.696 9	0.914 9
Metabolic	0.927 7	0.899 5	0.900 3	0.867 1	0.905 5	0.848 0	0.751 7
US-Air	0.966 5	0.951 3	0.952 2	0.952 2	0.962 1	0.919 6	0.911 1
Email	0.936 6	0.933 6	0.925 4	0.865 4	0.867 8	0.819 4	0.862 4
Yeast	0.973 1	0.970 6	0.969 5	0.914 2	0.914 9	0.864 8	0.913 2

从表2中明显得出,PSI算法在不同的数据集中的预测精度AUC值均高于其他算法在同一数据集上所得到的AUC值,预测效果较好,其通过合并有关可达节点的信息使预测算法的准确性大大提高。

4 结论

链路预测是复杂网络分析中最重要和最具挑战性的领域之一。链路预测算法的目标是基于网络中的现有链路来估计丢失链路的可能性。本文设计了一种基于本地路径索引的链接预测方法,结合有关路径上的节点信息预测网络中两个节点之间存在链路的可能性。与基于本地路径的频率来预测丢失链接的本地路径索引不同,本文提出的方法合并了有关路径上的节点信息,最后利用真实数据集进行实验分析,证明该方法比原有的方法具有更高的准确性。

参考文献:

[1] AZIZ F, GUL H, MUHAMMAD I, et al. Link prediction using node information on local paths[J]. Physica A: Statistical Mechanics and Its Applications, 2020, 557: 124980.

[2] HU W, LI J, CHENG J, et al. Security monitoring of heterogeneous networks for big data based on distributed association algorithm

[J]. Computer Communications,2020,152:206-214.
[3] KOVÁCS I A,LUCK K,SPIROHN K,et al. Network-based prediction of protein interactions[J]. Nature Communications,2019,10:1240.
[4] DAUD A,AHMAD M,MALIK M S I,et al. Using machine learning techniques for rising star prediction in co-author network[J]. Scientometrics,2015,102(2):1687-1711.
[5] SHI C,LI Y T,ZHANG J W,et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering,2017,29(1):17-37.
[6] SUN Y Z,HAN J W,YAN X F,et al. PathSim: meta path-based top-*K* similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment,2011,4(11):992-1003.
[7] JIANG L,YANG C C. User recommendation in healthcare social media by assessing user similarity in heterogeneous network[J]. Artificial Intelligence in Medicine,2017,81:63-77.
[8] LIANG W X,LI X,HE X S,et al. Supervised ranking framework for relationship prediction in heterogeneous information networks[J]. Applied Intelligence,2018,48(5):1111-1127.
[9] LI J C,ZHAO D L,GE B F,et al. A link prediction method for heterogeneous networks based on BP neural network[J]. Physica A: Statistical Mechanics and Its Applications,2018,495:1-17.
[10] 王慧,乐孜纯,龚轩,等. 基于特征分类的链路预测方法综述[J]. 计算机科学,2020,47(8):302-312.
[11] 袁榕,宋玉蓉,孟繁荣. 一种基于加权网络拓扑权重的链路预测方法[J]. 计算机科学,2020,47(5):265-270.
[12] 杨旭华,俞佳,张端. 基于局部社团和节点相关性的链路预测算法[J]. 计算机科学,2019,46(1):155-161.
[13] 龚追飞,魏传佳. 基于改进AdaBoost算法的复杂网络链路预测[J]. 计算机科学,2021,48(3):158-162.
[14] 王慧,乐孜纯,龚轩,等. 基于特征学习的链路预测模型TNTlink[J]. 计算机科学,2020,47(12):245-251.
[15] LI Z P,FANG X,BAI X,et al. Utility-based link recommendation for online social networks[J]. Economics of Networks eJournal,2017:6668325.
[16] 胡建涛. 社交网络结构分析与预测建模[D]. 成都:电子科技大学,2019.
[17] 黄寿孟,夏王霞. 基于LBSN中锚链接方法的链路预测模型[J]. 海南热带海洋学院学报,2021,28(5):72-77.
[18] 黄寿孟,夏王霞. 一种基于LSH技术的链路预测方法[J]. 信息记录材料,2021,22(7):139-142.
[19] 黄寿孟. 一种基于监督学习的异构网链路预测模型[J]. 计算机科学,2021,48(S2):111-116.

责任编辑:刘 红