文章编号:2096-7543(2023)03-0059-09 DOI:10.11863/j.suse.2023.03.08

基于三阶路径自适应度惩罚的链路预测方法

陈广福[1,2], 连雁平[1,2], 李晓飞[1,2]

(1.武夷学院数学与计算机学院,福建 武夷山 353400;
2.福建省茶产业大数据应用与智能化重点实验室,福建 武夷山 353400)

摘 要:链路预测目标是根据已知网络结构来预测缺失链接。现存大部分基于相似度方法仅考虑二阶路径而忽略三阶路径与网络拓扑特征关联程度,这将导致预测准确度下降而不适用于稀疏网络。针对以上不足,提出基于三阶路径自适应度惩罚(THADP)的链路预测算法改善稀疏网络预测精度。首先,统一泛化基于三阶路径相似度方法包括CN-L3、AA-L3和RA-L3,构建THADP框架保持节点所有三阶路径信息;其次,该框架与节点平均最短路径融合有利于捕获整个网络节点信息增强THADP鲁棒性;最后,在8个真实网络上,采用*AUC*和*F*1评价所提指标和基准方法性能,实验结果表明所提指标*AUC*和*F*1值最高分别提升了35.5%和21.5%。

关键词:复杂网络;链路预测;3阶路径;最短路径;自适应度惩罚

中图分类号:TP391 **文献标志码**:A

引 言

真实世界大量的复杂系统可由复杂网络来描述和表示,其中节点代表实体、连边表示实体间的关系。然而,由于真实网络数据收集总是部分的,因此如何寻找缺失链接间关系是复杂网络研究中最有挑战的问题。链路预测目标是根据已知网络结构及其节点属性等信息去推断节点对形成链接的可能性[1]。此外,链路预测还具有以下功能:1)预测缺失链接、识别虚假链接及消除网络噪音;2)根据当前网络结构信息探寻网络演化机制。因此,链路预测广泛应用于不同的领域。例如在社会网络,链路预测可用于检测用户异常信息,保护用户信息安全[2];社交网为不同用户推荐新朋友,可获得更好的用户体验[3];此外,在生物网络中,可用于预测蛋白质间先前未知的相互作用,从而降低经验方法的成本等[4]。

当前,大部分链路预测算法的研究都集中在基于相似度方法,该方法根据网络结构信息、节点聚类系数及节点属性等信息计算每个节点对相似度分数,分数越高产生链接可能性就越高。基于相似度算法又可以划分为局部、半局部和全局方法。局部相似度方法通过节点间的共同邻居数量来描述节点间产生链接的可能性,其中具有代表性的是共同邻居(Common Neighbors, CN)[5]、资源分配(Resource Allocation, RA)[6]和Adamic-Adar(AA)指

收稿日期:2022-03-27
基金项目:福建省自然科学基金项目(2021J011146);武夷学院引进人才科研启动基金项目(YJ202017)
作者简介:陈广福(1979-),男,讲师,博士,研究方向为链路预测和网络表示等,(E-mail)cgf21st@163.com

标[7]。王凯等[8]分析CN所有链接、节点首位端点和拓扑有效性提出了高效CN预测指标;王鑫等[9]提出了融合CN节点紧密度和贡献链路预测指标;Lee等[10]提出了协同过滤框架并与CN、AA和RA相融合提高节点间获得CN的能力;李星等[11]融合CN和邻域拓扑两类信息刻画了节点CN数量。然而上述研究讨论的都是二阶路径方法,存在着脆弱于稀疏网络而偏好于稠密网路的问题。最近有研究认为三阶路径相似度在大部分真实网络性能优于二阶路径方法。文献[12]针对蛋白质网结构特性提出三阶路径的相似度算法,该方法有效改善了蛋白质网中潜在相互作用关系的预测准确度。文献[13]在文献[12]基础上融合了CN、AA和RA构建三阶路径的局部相似度算法,结果表明在部分网络中三阶路径预测准确度胜过二阶路径方法。

全局相似度方法利用整个网络的拓扑结构信息去计算节点间相似度分数。如顾秋阳等[14]提出了高阶相似度预测算法,该算法以高阶路径信息为判别特征并惩罚节点有效长路径保持网络全局结构;文献[15]提出了线性优化(Linear Optimization,LO)指标捕获高阶路径信息。该类指标存在设计复杂且耗时的缺点,因而有研究者提出用半局部相似度方法融合局部和全局优点来调节性能与时间复杂度的平衡关系。如文献[16]提出自适应度惩罚(Adapative Degree Penalization,ADP)的链路预测方法,该方法通过泛化局部相似度算法,并融合节点聚类系数以提高预测准确度,同时验证了与网络结构的关联强度;文献[17]在文献[16]基础上提出共同邻居度惩罚算法(Common Neighbors Degree Penalization,CNDP),考虑了CN数量是否影响网络拓扑结构对泛化的局部相似度的影响;刘树新等[18]考虑任意节点双向资源匹配量,提出了资源传输匹配度算法。文献[16-17]提出自适应度惩罚的链路预测算法证实了共同邻居度与网络拓扑结构存在密切联系。然而,此类方法有以下两个明显的不足之处:1)泛化基于二阶路径局部方法无法捕获更多的节点路径信息导致预测准确度低;2)泛化基于二阶路径局部相似度框架性能依赖于网络拓扑结构信息,而平均节点聚类信息仅适用于稠密网络而脆弱于稀疏网络。针对以上不足,本文要解决以下两个问题:1)泛化基于三阶路径相似度构建一个泛化三阶路径自适应惩罚框架;2)通过融合该框架与平均最短距离改善预测准确度。围绕这两个待解决的问题展开研究,具体思路如下:首先在三阶路径(L3)指标基础上融合CN、AA和RA 3个局部指标,提出CN-L3、AA-L3与RA-L3;然后将以上3个方法泛化成统一的带参数三阶路径自适应度惩罚(Three-Hop Adapative Degree Penalization,THADP)链路预测框架。本文将平均最短路径融合三阶泛化框架中,提出THADP链路预测指标,通过惩罚三阶路径获得更多网络结构信息,再融合平均最短路径去挖掘更多网络结构信息。

1 基于THADP链路预测方法

1.1 问题描述

给定一个有向无权网络$\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$,其中$\boldsymbol{V}=\{v_i\}_{i=1}^{N}$是节点集,$\boldsymbol{E}$表示链接集,且每条边$e\in\boldsymbol{E}$是一个有序对$e=(v_i,v_j)$。这里不允许多个链接和自循环存在。用$\boldsymbol{A}=[a_{ij}]_{N\times N}$来表示$\boldsymbol{G}$的邻接矩阵。$\boldsymbol{G}$是无向无权网络,如果节点之间存在链接,则$a_{ij}=1$,否则$a_{ij}=0$。

用$\boldsymbol{U}$表示网络所有可能边的集合且$\boldsymbol{U}=\frac{N(N-1)}{2}$,显然,不存在边集合可表示为$\boldsymbol{U}-\boldsymbol{E}$。链路预测目标是从集合$\boldsymbol{U}-\boldsymbol{E}$中查找缺失链路。为了验证算法的性能,将观测到的链路集$\boldsymbol{E}$分成两部分:训练集$\boldsymbol{E}_{\mathrm{T}}$和测试集$\boldsymbol{E}_{\mathrm{P}}$,前者是已知信息,后者仅用于测试。显然,$\boldsymbol{E}_{\mathrm{T}}\cap\boldsymbol{E}_{\mathrm{P}}=\varnothing$和$\boldsymbol{E}_{\mathrm{T}}\cup\boldsymbol{E}_{\mathrm{P}}=\boldsymbol{E}$。

1.2 泛化三阶路径相似度

大部分真实网络是稀疏的,基于二阶路径相似度方法依赖于节点CN数量,对于稀疏网络来说,获得节点CN数量是有限的。因此,基于三阶路径相似度方法尽可能获得所有节点三阶路径信息弥补以上方法的不足。文献[16]考虑三阶路径长度提出3个基于三阶路径相似度方法分别是CN-L3、

AA-L3 和 RA-L3。任意节点 i 和 j 的基于三阶路径相似度的定义如式(1)~(3)所示。

(a)CN-L3 指标，该指标是尽可能寻找节点间所有三阶路径的数，其定义如下：

$$S_{ij}^{\text{CN-L3}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} a_{xy} \tag{1}$$

(b)AA-L3 指标，该指标类似 AA 指标惩罚较大的三阶节点度，其定义如下：

$$S_{ij}^{\text{AA-L3}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} \frac{a_{xy}}{\sqrt{\log k_x \log k_y}} \tag{2}$$

(c)RA-L3 指标，该指标类似 RA 指标三阶节点间资源传递，其定义如下：

$$S_{ij}^{\text{RA-L3}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} \frac{a_{xy}}{\sqrt{k_x k_y}} \tag{3}$$

其中，i 和 j 的邻居集合 Γ_i 和 Γ_j 中分别存在着节点 x 和 y。S 表示节点 i 和 j 相似度，a_{ij} 表示节点 x 和 y 的边，k 是表示节点度。

以上 3 个指标在不同真实网络中所获得节点间三阶路径信息是不一致的，表明链路预测准确度取决于不同网络结构特征。为提高以上方法适用于不同网络结构而获得最佳的预测结构，采自适应度惩罚方法去泛化基于三阶路径相似度方法构建一个统一的三阶泛化链路预测框架，其定义如下：

$$S_{ij}^{\text{ADP}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} a_{xy}\left(k_x k_y\right)^{-\frac{1}{2}\beta} \tag{4}$$

其中 β 为可调参数。

由式(4)可知，当 $\beta = 0$ 时，式(4)就退化为：$S_{ij}^{\text{CN-L3}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} a_{xy}$；当 $\beta = 1$ 时，式(4)就退化为：$S_{ij}^{\text{RA-L3}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} \frac{a_{xy}}{\sqrt{k_x k_y}}$；

当 $\beta \in (0,1)$ 时，式(4)就退化为：$S_{ij}^{\text{AA-L3}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} \frac{a_{xy}}{\sqrt{\log k_x \log k_y}}$。

1.3　THADP 指标

三阶泛化链路预测框架(式(4))预测的准确度依赖于 β，而与 β 有直接关联的是平均节点最短路径以及平均聚类系数，本文采用节点平均最短路径方法。复杂网络求解最短路径方法较多，由于是无向无权网络，若节点间存在链接，那么值均大于 0，因此使用经典的 Dijkstra 算法计算所有节点对间的最短路径，设节点对间最短路径的矩阵为 D，其节点平均最短距离定义如下：

$$\boldsymbol{D}_{\text{avg}} = \frac{\boldsymbol{D}}{N(N-1)} \tag{5}$$

其中 N 为网络节点数。

由三阶泛化链路预测框架，将平均最短距离融合泛化框架构建统一链路预测指标 THADP，定义如下：

$$S_{ij}^{\text{THADP}} = \sum_{x\in\Gamma_i, y\in\Gamma_j} a_{xy}\left(k_x k_y\right)^{-\frac{1}{2}\beta D_{avg}} \tag{6}$$

泛化框架构建统一链路预测指标 THADP 有以下两优点：1)克服二阶路径相似度方法仅考虑节点 CN，THADP 可从特别稀疏网络中获得更多节点的信息；2)利用平均最短路径获得网络全局结构信息提高预测准确度。

为更深入理解泛化框架构建统一链路预测指标 THADP，设 5 个节点的部分网络来举例说明，如图 1 所示。使用式(6)方法计算节点 i 和 j 相似度具体过程如下：节点 i 和 j 间存在以下 3 条三阶路径：i–a–x–j，i–x–a–j 和 i–x–b–j。节点 a、b 与 x 的度分别为：$k_a = 6$，$k_b = 3$，$k_x = 4$，节点 i 与 j 最短路径为 2，因此节点 i 和 j 相似度分数为：

$$\begin{aligned} S_{ij}^{\text{THADP}} &= (k_a k_x)^{-\frac{1}{2}\times 2\beta} + (k_a k_x)^{-\frac{1}{2}\times 2\beta} + (k_b k_x)^{-\frac{1}{2}\times 2\beta} \\ &= 2\times(0.0417)^{\beta} + (0.0833)^{\beta}。 \end{aligned}$$

最后节点 i 和 j 相似度分数依赖于可调参数 β。

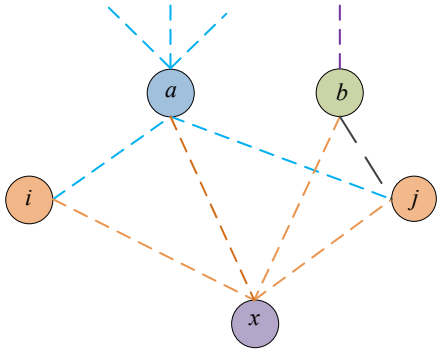


图 1　5 节点网络示意图

2 实验结果与分析

2.1 评价度量

本文用*AUC*和*F*1两个度量来衡量所有方法的性能[19],以*F*1值作为综合性指标。两个度量的值越高表示该方法预测准确度越高。

1)*AUC*为在测试集$\boldsymbol{E}_{\mathrm{p}}$中的链接分数大于随机选择的一个不存在集$\boldsymbol{U}-\boldsymbol{E}$中的链接分数的概率。独立地比较$m$次,若有$m_1$次测试集中的链接的分数值大于不存在集中的链接的分数,有m_2次两者分数值相等,*AUC*定义为:

$$AUC=\frac{m_1+0.5\times m_2}{m} \tag{7}$$

2)*F*1为召回率(*Recall*)和准确率(*Precision*)综合性度量,可更全面而有效地评价算法性能,其定义为:

$$F1=\frac{2\times Recall\times Precision}{Recall+Precision} \tag{8}$$

2.2 数据集

2.2.1 无向无权网络的数据集

1)美国航空运输网络(USAir,USA)[20]:该网络由332个链接和2126个节点组成链接。节点表示机场,链接代表两个机场之间联系。

2)犯罪网络(CRIme,CRI)[20]是二部图犯罪网,左节点表示一个人,右节点表示一个罪行。两个节点之间的"边"代表左节点参与了由右节点所代表的犯罪案件。

3)空中交通管制网(Air Traffic Control,ATC)[21]:该网络由美国联邦航空管理局国家飞行数据中心(NFDC)选航线数据库构建。该网络中的节点表示机场或服务中心,链接由NFDC推荐的首选路线字符串创建。

4)论文引用网络(KOHonen,KOH)[21]:该网络是"自组织映射"引用网。节点表示论文,链接代表论文间引用关系。

5)EPA[21],该网络将200页的响应集扩展到一个搜索引擎查询来构建。

6)电力网(POwerGrid,POG)[21]:该网络是美国西部各州电网的信息。一条边代表一条电源线路。一个节点可以是发电机、变压器或变电站。

7)蛋白质交互网络(VIDal,VID)[21]是人类蛋白质之间相互作用的网络,节点表示蛋白质,方向链接代表蛋白质间交互关系。

8)论文引用网络(SmaGr,SG)[21]:该网络是关于网络理论与实验的引用网络,它由1024个节点和4916条链接组成。链接方向表示引用关系。

2.2.2 真实世界无向网络拓扑特征

本文采用8个真实世界无向无权网络评价所有方法的性能,其拓扑结构特征统计见表1。

表1 8个真实世界无向网络拓扑特征统计

网络	*N*	$\lvert\boldsymbol{E}\rvert$	$<k>$	$<d>$	*AC*	*Density*
USA	332	2126	12.8072	2.7381	0.6252	0.0387
ATC	1226	1307	2.1313	5.2182	0.0576	0.0017
KOH	4469	12 719	5.6921	2.5212	0.2105	0.0013
CRI	829	1474	3.5561	5.0400	0.0058	0.0043
EPA	4471	8890	3.7267	3.5568	0.0637	0.0007
POG	4941	6594	2.6691	18.9892	0.0801	0.0005
VID	3133	6438	4.1095	3.8188	0.0635	0.0013
SG	1024	4916	9.6016	2.9814	0.3071	0.0094

表1中,*N*为节点数,$\lvert\boldsymbol{E}\rvert$表示$\boldsymbol{E}$链接集合的链接数,$<k>$表示平均度,$<d>$表示平均最短距离,*AC*表示节点平均聚类系数,*Density*表示网络稀疏程度。

2.3 基准方法

为验证所提算法性能,本文用11个最近几年的代表性方法与之比较。11个链路预测方法介绍如下。

1)3个基于三阶路径相似度(CN-L3,AA-L3,RA-L3)已在本文1.2节作了详细说明。

2)3个基于协同过滤框架融合CN、AA与RA构成SCF-CN、SCF-AA和SCF-RA 3个预测指标[10],其协同过滤框架定义为:

$$\boldsymbol{S}^{\mathrm{SCF}}=(\boldsymbol{A}+\boldsymbol{I})S+[(\boldsymbol{A}+\boldsymbol{I})S]^{\mathrm{T}} \tag{9}$$

其中,$\boldsymbol{A}$为任意网络邻接矩阵,$\boldsymbol{I}$是单位矩阵,S为任意相似度。

3）线性最优化（Linear Optimization，LO）[15]：该指标假设两个节点之间存在链接的可能性可以通过相邻节点贡献的线性求和来展开，其定义为：

$$S^{LO} = \alpha A^3 - \alpha^2 A^5 + \alpha^3 A^7 - \alpha^4 A^9 + \cdots \tag{10}$$

其中$0 < \alpha < 1$。

4）自适应度惩罚指标（Adapative Degree Penalization，ADP）[16]：该方法通过泛化基于局部相似度（CN、AA和RA）构建统一框架并融合节点聚类系数，任意节点相似度定义如下：

$$S^{ADP}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} |\Gamma_z|^{-\beta C} \tag{11}$$

其中C为节点平均聚类系数。

5）共同邻居度惩罚指标（Common Neighbors Degree Penalization，CNDP）[17]：该方法是在ADP方法基础考虑CN数，任意节点相似度定义为：

$$S^{CNDP}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} |C_z| \left(|\Gamma_z|\right)^{-\beta C} \tag{12}$$

其中$|C_z|$是共同邻居数。

6）局部路径指标（Local Path，LP）[6]：该指标扩展CN指标，考虑三阶路径因素，其定义为：

$$S_{xy}^{LP} = A^2 + \varepsilon A^3 \tag{13}$$

其中ε为可调参数。

7）Katz指标[22]：该方法考虑整个网络所有节点的路径，其定义为：

$$S = (I - \alpha \times A)^{-1} - I \tag{14}$$

其中，I为单位矩阵，α为可调参数。

2.4　结果分析

本实验硬件平台为Intel Core i5-7200U CPU笔记本，主频2.71 GHz，内存4 G，操作系统为Windows 10，所有方法使用Matlab R2016b实现。此外，所提方法含可调参数β，为公平比较所有的方法，所有数据集设$\beta = 0.5$。对于LO方法的参数设为0.1，LP的参数设为0.001，ADP和CNDP的参数设为1.5及Katz参数设为0.01。

本文做3个实验评估所提方法的性能。首先，用*AUC*和*F*1两个度量全面评估所有12个指标性能；其次，对12个算法健壮性进行对比；最后，对可调参数敏感性分析。

对于第1个实验，将原始网络按9∶1的比例划分为训练集和测试集，再用*AUC*和*F*1度量评估所有方法性能，其实验结果见表2。通过分析基准方法与THADP在8个网络上对应*AUC*和*F*1值可以得到以下结果。

1）本文所提方法在8个数据集上与11个指标相比较，*AUC*和*F*1度量值均获得最优。从表2可知，8个真实网络均是十分稀疏，表明本文方法通过泛化三阶路径相似度能够有效捕获更多节点三阶路径信息，并可以融合平均节点最短路径获得全局结构信息。

2）分析表2中列出所有数据集节点平均最短距离路径，可观察到节点平均最短距离相对较小是USA、KOH和SG，较大的是POG。从表2可知，本文所提方法在POG数据集*AUC*和*F*1性能最差，这表明平均最短相对较大时THADP无法捕获更多三阶路径信息；而在USA、KOH和SG上，THADP获得最佳预测准确度，这表明节点平均最短距离直接影响THADP是否能有效捕获三阶路径信息。

3）THADP与最优基准算法相比较，在USA、ATC、KOH、CRI、EPA、POG、VID和SG上，*AUC*分别提升1.8%、12.3%、8.3%、35.5%、18.9%、15.3%、20.2%、9.8%，*F*1分别提升0.9%、7.6%、6.6%、21.5%、13.6%、9.5%、13.7%、6.0%。

4）基于三阶路径相似度（CN-L3、AA-L3和RA-L3）几乎在所有的数据集中获得最差的性能，其主要原因是由于网络稀疏性导致无法获得足够节点三阶路径的信息。相比之下，本文所提方法的性能获得显著提升。例如在ATC数据集中，THADP指标与基于三阶路径相似度的最优指标相比，*AUC*和*F*1度量预测准确度分别提升了27%和35%；在CRI数据集中，THADP指标与基于三阶路径相似度的最优指标相比，*AUC*和*F*1度量预测准确度分别提升了35%和40%。在泛化基础引入网络拓扑特征-节点平均最短路径后*AUC*和*F*1值均获得显著的提升，这

表明泛化后预测准确度与网络拓扑特征有着强关联。

5)基于自包含协同过滤系列链路预测方法性能较基于三阶路径相似度系列方法有所提高,主要原因是该系列方法采用自包含协同过滤的对称性方法计算节点相似度,可获得更多局部节点信息。然而,与THADP相比较,THADP获得的质量性能更高,其主要原因是THADP能够捕获更多三阶路径信息。例如在EPA数据集中,THADP方法与其他性能最优异方法相比*AUC*和*F*1度量预测准确度分别提升了21%和34%,在其余数据集均有显著提高。

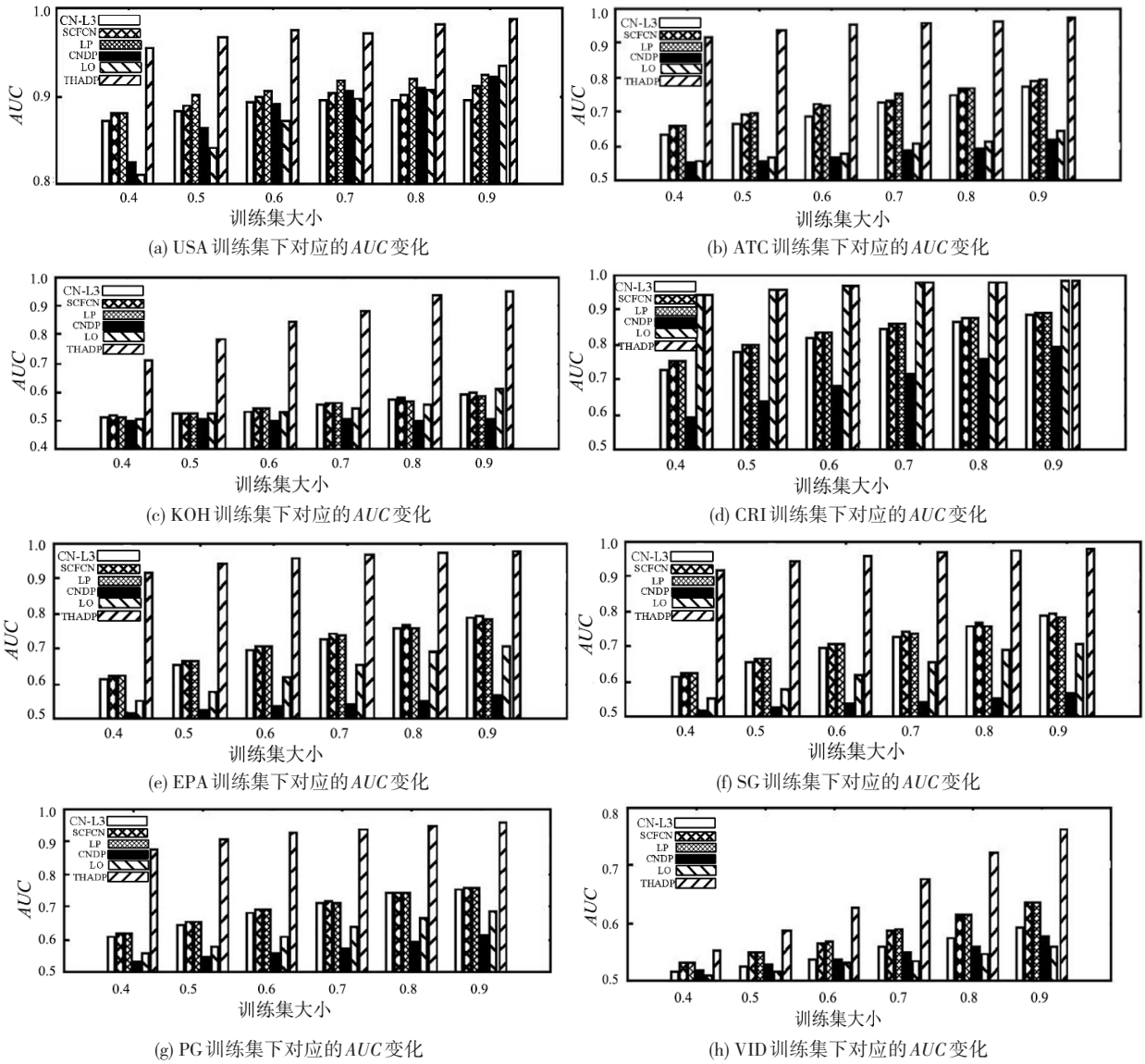
6)LO和Katz是全局相似度方法,其中Katz最接近本文方法,该方法利用整个网络节点对的信息获得相似度分数矩阵来弥补网络稀疏性的不足。而LO性能略低于Katz,主要原因是该方法仅考虑节点所有邻居贡献值线性求和。LO和Katz与THADP比较,后者预测准确度高于前者。从网络结构角度分析,THADP指标同时保持高阶路径长度和全局结构信息。

7)THADP、CNDP和ADP3个指标均采用类似技术路线,THADP性能显著优于CNDP和ADP方法。CNDP和ADP性能较差的主要原因是泛化基于二阶路径局部相似度仅考虑局部结构信息而平均节点聚类系数也只能反映网络局部节点关系。尤其在特别稀疏网络CRI中表现尤为明显,THADP指标的*AUC*和*F*1最多提高44%和57%。

表2 基准方法与THADP在8个网络上对应*AUC*和*F*1值

指标	USA		ATC		KOH		CRI		EPA		POG		VID		SG	
	AUC	*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*	*F*1	*AUC*	*F*1
CN-L3	0.896	0.662	0.663	0.421	0.881	0.297	0.607	0.348	0.789	0.227	0.596	0.183	0.759	0.223	0.865	0.434
AA-L3	0.892	0.707	0.678	0.405	0.852	0.579	0.598	0.269	0.788	0.495	0.568	0.143	0.734	0.403	0.848	0.670
RA-L3	0.889	0.719	0.698	0.416	0.860	0.580	0.591	0.261	0.796	0.506	0.599	0.192	0.752	0.425	0.831	0.655
SCF-CN	0.907	0.669	0.774	0.474	0.890	0.304	0.599	0.336	0.777	0.221	0.633	0.190	0.758	0.230	0.872	0.440
SCF-AA	0.909	0.741	0.779	0.497	0.898	0.657	0.586	0.242	0.738	0.405	0.641	0.213	0.762	0.465	0.895	0.726
SCF-RA	0.934	0.755	0.789	0.502	0.903	0.663	0.593	0.258	0.768	0.449	0.636	0.210	0.762	0.467	0.899	0.729
LP	0.969	0.781	0.711	0.441	0.892	0.408	0.582	0.280	0.786	0.279	0.636	0.164	0.760	0.282	0.874	0.531
LO	0.925	0.709	0.597	0.605	0.800	0.690	0.565	0.563	0.715	0.651	0.553	0.578	0.700	0.644	0.816	0.702
CNDP	0.939	0.707	0.783	0.499	0.870	0.683	0.524	0.206	0.742	0.514	0.653	0.219	0.766	0.528	0.839	0.584
ADP	0.942	0.716	0.626	0.372	0.819	0.428	0.509	0.246	0.588	0.192	0.594	0.185	0.620	0.230	0.842	0.506
Katz	0.925	0.757	0.848	0.697	0.887	0.722	0.558	0.532	0.735	0.635	0.668	0.613	0.747	0.641	0.880	0.734
本文方法	0.987	0.789	0.971	0.773	0.986	0.788	0.962	0.778	0.985	0.787	0.821	0.708	0.968	0.781	0.997	0.794

对于第二个实验,测试6种代表性指标鲁棒性。通过改变训练集大小来改变原始网络的稀疏性。本实验设训练集大小范围为0.4、0.5、0.6、0.7、0.8、0.9,若训练集比例太小会破坏网络结构导致预测不准确,其实验结果如图2所示。从图2中可以看出:1)本文所提指标在不同训练集下*AUC*值的波动不明显,而有些指标在训练集变化时产生较大的波动,例如在ATC和POG数据集中CN-L3指标。2)当训练集仅占40%时、测试集占60%的情况下,本文所提指标*AUC*值胜过其余5个指标,这表明在极端稀疏情况下,THADP指标依然获得最佳的预测准度,同时也表明该方法适用于处理稀疏网络的链路预测问题。3)本文所提方法在训练集仅占40%情况下,在所有稀疏网络中*AUC*值均优于其余5个指标,表明该方法具备良好的鲁棒性。



(a) USA训练集下对应的*AUC*变化　(b) ATC训练集下对应的*AUC*变化

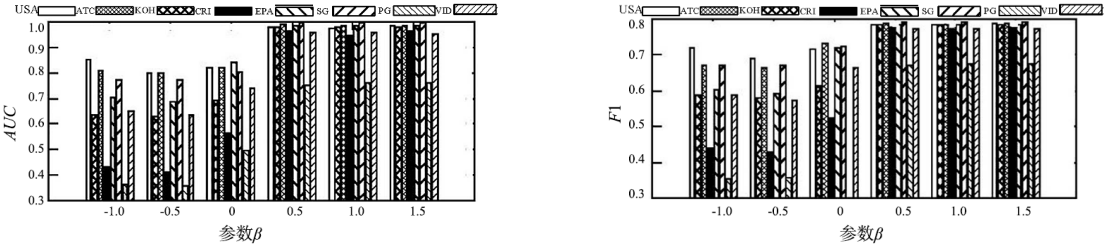
(c) KOH训练集下对应的*AUC*变化　(d) CRI训练集下对应的*AUC*变化

(e) EPA训练集下对应的*AUC*变化　(f) SG训练集下对应的*AUC*变化

(g) PG训练集下对应的*AUC*变化　(h) VID训练集下对应的*AUC*变化

图2　不同训练集下对应*AUC*变化

最后,分析可调参数β对性能的影响。β可调参数与网络结构特征有着强关联,通过调整β使THADP获得最佳性能。设置β变化为-1.0、-0.5、0、0.5、1.0和1.5,其实验结果如图3所示。由图3可看出:当β为负数时,*AUC*和*F*1值均处于最差,主要是因为当β为负数时无有效惩罚度,导致获得的性能最差;当β=0时式(6)退化为式(1),性能有所提高;当β大于0时,性能开始快速提高并获得恒定水平,主要原因是可产生有效惩罚度并获得全局结构信息。上述表明泛化后框架的最佳性能依赖于网络结构信息。因此,当β=0.5时所有网络均获得最优性能。



(a) 不同参数β在8个网络上对应的*AUC*　(b) 不同参数β在8个网络上对应的*F*1值

图3　不同参数β在8个网络上对应的*AUC*和*F*1值

3 结束语

如何融合不同网络的不同类型拓扑结构信息来提高链路预测的准确度是当前研究的热点之一。本文提出了基于三阶路径自适应度惩罚的链路预测指标,该方法核心是泛化基于三阶路径相似度算法(CN-L3,AA-L3,RA-L3)统一构建三阶路径自适应度惩罚的框架。将平均最短路径与自适应度可调参数相结合组成三阶路径自适应度惩罚的链路预测指标。在8个真实网络上与最近的代表性方法比较,结果表明本文方法更具优势。此外,实验结果表明所提指标预测准确度与网络拓扑的拓扑特征有密切关联。

相对于无向网络的链路预测,如何泛化三阶路径相似方法以适用于有向网络链路预测,以及验证有向网络结构是否同样与泛化后框架有关联程度,是值得进一步研究的课题。

参考文献:

[1] 李艳丽,周涛.链路预测中的局部相似性指标[J].电子科技大学学报,2021,50(3):422-427.

[2] KAGAN D,ELOVICHI Y,FIRE M.Generic anomalous vertices detection utilizing a link prediction algorithm[J].Social Network Analysis and Mining,2018,8(1):1-13.

[3] WU S,SUN J,TANG J.Patent partner recommendation in enterprise social networks[C]//Proceedings of the Sixth ACM International Conference on Web Search and Data Mining,Italy,Rome,February 4-8,2013:43-52.

[4] YU S,ZHAO M,FU C,et al.Target defense against link-prediction-based attacks via evolutionary perturbations[J].IEEE Transactions on Knowledge and Data Engineering,2019,33(2):754-767.

[5] 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.

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

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

[8] 王凯,刘树新,于洪涛,等.基于共同邻居有效性的复杂网络链路预测算法[J].电子科技大学学报,2019,48(3):432-439.

[9] 王鑫,陈喜,钱付兰,等.结合共同邻居贡献度的节点相似性链路预测算法[J].数据采集与处理,2018,33(5):900-910.

[10] LEE Y L,ZHOU T.Collaborative filtering approach to link prediction[J].Physica A:Statistical Mechanics and its Applications,2021:126107.

[11] 李星,朱宇航,柏溢,等.基于共同邻居邻域拓扑稠密性加权的链路预测方法[J].计算机应用研究,2021,38(5):1503-1507.

[12] KOVÁCS I A,LUCK K,SPIROHN K,et al.Network- based prediction of protein interactions[J].Nature Communications,2019,10(1):1240.

[13] ZHOU T,LEE Y L,WANG G.Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms[J].Physica A:Statistical Mechanics and its Applications,2021,564:125532.

[14] 顾秋阳,吴宝,池仁勇.基于高阶路径相似度的复杂网络链路预测方法[J].通信学报,2021,42(7):61-69.

[15] PECH R,HAO D,LEE Y L,et al.Link prediction via linear optimization[J].Physica A:Statistical Mechanics and its Applications,2019,528:121319.

[16] MARTÍNEZ V,BERZAL F,CUBERO J C.Adaptive degree penalization for link prediction[J].Journal of Computational Science,2016,13:1-9.

[17] RAFIEE S,SALAVATI C,ABDOLLAHPOURI A.CNDP:link prediction based on common neighbors degree penalization[J].Physica A:Statistical Mechanics and its Applications,2020,539:122950.

[18] 刘树新,李星,陈鸿昶,等.基于资源传输匹配度的复杂网络链路预测方法[J].通信学报,2020,41(6):70-79.

[19] YANG Y,LICHTENWALTER R N,CHAWLA N V.Evaluating link prediction methods[J].Knowledge and Information Systems,2015,45(3):

751-782.

[20] ROSSI R A,AHMED N K.The network data repository with interactive graph analytics and visualization[C]//Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,Texas,USA,January 25-30,2015:4292-4293.

[21] KUNEGIS J.Konect:the koblenz network collection[C]//Proceedings of the 22nd International Conference on World Wide Web Companion, Rio de Janeiro,Brazil,May 13-17,2013:1343-1350.

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

引用格式:

中 文:陈广福,连雁平,李晓飞.基于三阶路径自适应度惩罚的链路预测方法[J].四川轻化工大学学报(自然科学版),2023,36(3):59-67.

英 文:CHEN G F,LIAN Y P,LI X F.Link prediction method based on three-hop adaptive degree penalization[J].Journal of Sichuan University of Science & Engineering (Natural Science Edition),2023,36(3):59-67.

Link Prediction Method Based on Three-Hop Path Adaptive Degree Penalization

CHEN Guangfu[1,2], LIAN Yanping[1,2], LI Xiaofei[1,2]

(1.College of Mathematics and Computer, Wuyi University, Wuyishan 354300, China; 2. Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyishan 354300, China)

Abstract: The goal of link prediction is to predict missing links according to the known network structure. Most of the existing methods based similarity only focus on the 2-hop path while ignore the association between the three-hop path and the topology, which leads to the reduction of prediction accuracy and is not suitable for sparse networks. In view of this shortcomings, a link prediction algorithm based on three-hop path adaptive degree penalty (THADP) has been proposed to improve the prediction accuracy of sparse networks. Firstly, the methods based on three-hop path similarity including CN-L3, AA-L3 and RA-L3 are generally unified, which are used to construct THADP framework to maintain all three-hop path information of nodes. Secondly, the framework and the average shortest path of nodes are fused, which is conducive to capture the node information of the whole network and enhance the robustness of THADP. Finally, the *AUC* and *F*1 are used to evaluate the performances of the proposed metrics and the baseline method on eight real networks, and the experimental results show that the *AUC* and *F*1 of the proposed metrics are improved by 35.5% and 21.5%, respectively.

Key words: complex network; link prediction; three-hop path; shortest path; adaptive degree penalization