基于节点度异质性惩罚的链路预测方法

陈广福[1,2],江 玲[1],韩辉珍[1]
(1. 武夷学院 数学与计算机学院,福建 武夷山 353400;
2. 认知计算与智能信息处理福建省高校重点实验室,福建 武夷山 353400)

摘 要:针对大部分现存的链路预测方法仅关注规则网络以及偏好连接现象而导致在稀疏网络获得低质量性能,提出一种节点度异质性惩罚的链路预测框架(NDHP),该框架最优预测准确度与网络拓扑特征有密切关联。首先,计算整个网络节点度获得所有节点对的度异质性相似度;其次,采用惩罚节点度较大机制去惩罚度异质性权重较大的节点抑制节点间差异;最后,通过可调参数将平均节点聚类系数和平均最短路径分别和基于度异质性惩罚框架相关联,获取网络结构信息来弥补网络稀疏信息不足,并提出基于节点度异质性惩罚的平均聚类系数指标(NDHP_AC)和基于节点度异质性惩罚的平均距离指标(NDHP_AD)。此外,在 8 个真实无向无权网络上与最近代表性的方法相比较,所提两个指标在预测缺失链接和鲁棒性两方面性能优于基准指标。尤其在高度稀疏网络中,所提指标的 AUC 和 AUPR 分别最大提高了 15.3% 和 8.6%。

关键词:复杂网络;链路预测;度异质性;平均节点聚类系数;平均最短路径
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)12-0081-07
doi:10.3969/j.issn.1673-629X.2022.12.013

Link Prediction Method Based on Node Degree Heterogeneity Penalization

CHEN Guang-fu[1,2],JIANG Ling[1],HAN Hui-zhen[1]
(1. School of Mathematics and Computer,Wuyi University,Wuyishan 353400,China;
2. Key Laboratory of Cognitive Computing and Intelligent Information Processing in Fujian Education Institutions,
Wuyishan 353400,China)

Abstract:Most of existing link prediction methods only focus on regular networks and preferential attachment phenomenon,which results in low quality performance in highly sparse networks. We propose a link prediction framework based on node degree heterogeneity penalty. The optimal prediction accuracy of the framework is closely related to network topology characteristics. Firstly,the node degree of the whole network is calculated to obtain the degree heterogeneity similarity of the predicted node pairs. Secondly,the mechanism of punishing nodes with higher degree of heterogeneity is used to suppress the differences between nodes. Finally,the average node clustering coefficient and the average shortest path are associated with the degree heterogeneity based punishment framework respectively by the adjustable parameters to obtain the network structure information to make up for the lack of sparse network information,and the Node Degree Heterogeneity Penalization via Average Clustering coefficient (NDHP_AC) and the Node Degree Heterogeneity Penalization via Average Distance (NDHP_AD) are proposed. In addition,compared with the most recent representative method,the performance of the proposed two indicators is better than the benchmark indicators in predicting missing links and robustness on eight real undirected and unweighted networks. Especially in highly sparse networks,the proposed method improves the maximum AUC and AUPR by 15.3% and 8.6%,respectively.

Key words:complex network;link prediction;degree heterogeneity;average node clustering coefficient;average shortest path

0 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,其中节点代表实体和链接表示实体间的关系。由于真实网络数据收集总是不完整及受噪声影响,如何寻找缺失链接间的关系是复杂网络研究中最有挑战的问题。链路预测的目标是根据已知网络结构及其节

收稿日期:2021-12-10 **修回日期**:2022-04-13
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(YJ202017)
作者简介:陈广福(1979-),男,博士,讲师,CCF 会员(J7630M),通讯作者,研究方向为链路预测、网络表示。

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)预测缺失的无向、加权和有向链接,识别虚假链接及消除网络噪声;(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。例如在电子邮件系统,链路预测可阻止和过滤不相关的和广告的邮件[2];在社交网络,链路预测启用信任度量保护用户的隐私信息[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,从而显著降低经验方法的成本等[4]。

当前,不同类型链路预测算法被提出应用在不同场景。基于相似度算法是最简单和有效的方法,该方法是根据已知网络结构和节点属性计算节点对的相似度分数,然后按升序对节点分数进行排序,分数越排在前面节点对越有可能形成链接。一般而言,基于相似度方法可分为基于局部相似度、全局方法和半局部方法。基于局部相似度算法利用局部结构(如共同邻居数量、节点聚类及度相关聚类等)信息去计算未链接节点对分数,其中共同邻居(Common Neighbors, CN)[5]、资源分配(Resource Allocation, RA)[6]和 Adamic-Adar(AA)[7]是典型代表。AA 与 RA 指标核心思想类似,都是启用惩罚度较大节点,两者区别是 AA 指标是惩罚度权重大的节点而 RA 指标是惩罚获得资源多的节点。此外,文献[8]协同过滤和自包含协同过滤框架融合局部相似度算法 CN、AA 和 RA 极大改善了预测精度。文献[9]提出一种基于共同邻居邻域拓扑稠密性加权的链路预测方法,该方法利用共同邻居的节点度和邻域拓扑相对稠密指数刻画共同邻居及其邻域拓扑的相似性贡献;基于节点聚类系数方法利用共同邻居的节点聚类能有效提取网络局部结构信息。例如文献[10]在文献[9]的基础上考虑节点和链接聚类系数可获得可观察链接和节点聚类系数信息以及文献[11]提出度相关聚类系数和度相关聚类能力路径指标衡量节点聚类能力。上述方法的优点是设计简单及良好的可扩展性,可获得较好的预测精度,缺点是仅考虑局部信息,无法获得更多网络结构或节点聚类信息导致敏感于稀疏网络[12]。

全局方法启用整个网络结构信息去计算未链接节点对分数。例如文献[13]考虑整个网络路径信息;文献[14]提出基于高阶路径相似度算法,该算法利用高阶路径作为判别特征,对种子节点对间的可用长路径实施惩罚;文献[15]提出线性最优化(Linear Optimization,LO),该方法利用节点邻居贡献捕获高阶路径信息。上述方法的优点是显著改善了预测准确度,但缺点是耗时。为平衡局部与全局方法的不足,提出半局部方法,该方法可同时保持局部及三阶路径信息,其中局部路径方法(Local Path,LP)[6]是典型的代表。文献[16]提出一种基于资源传输路径有效性的链路预测方法,该方法分析节点间潜在的资源传输路径对资源传输量的影响。

现存大部分方法仅在宏观上利用网络局部和全局结构信息而忽略了微观上节点度与网络拓扑结构的关联程度。例如 Shang 等人[17]提出节点差异性指标衡量整个网络节点对异质性处理稀疏和树状网络链路预测问题,该方法仅考虑节点对的异质性,预测精度不够理想。此外,文献[18]提出自适应度惩罚的链路预测方法,该方法泛化 CN、AA 和 RA 构建一个统一泛化框架,该框架取得最优预测准确度依赖于网络拓扑特征。文献[19]在文献[18]的基础上考虑共同邻居数量改善预测结果。以上方法通过惩罚度大的节点与网络结构有着密切关联,尤其文献[18-19]考虑局部结构信息而敏感于稀疏网络。

大部分真实世界的网络呈现稀疏性,然而当前大部分网络在稀疏网络获得低质量性能。因此,该文要解决以下两个问题:(1)如何衡量网络被预测节点对的异质性;(2)节点对异质性如何与网络拓扑特征相关联。以下围绕这两个问题,提出度异质性惩罚的链路预测指标。具体地,节点度表示节点与其他节点的联系和关联程度。节点度越高,节点就越重要。然而,节点度高的未必贡献大,相反节点度小的反而贡献大。网络节点对异质性反映节点间邻居信息的差异,更全面衡量整个网络节点度分布情况。首先计算各节点度,再计算节点对异质性,然后惩罚节点度异质性大的节点对,构建节点度异质性惩罚框架(Node Degree Heterogeneity Penalization,NDHP)。节点平均聚类系数和平均最短路径衡量网络节点聚集能力,通过可调参数与上述框架相融合提出两个新颖指标,分别是基于节点度异质性惩罚的平均聚类系数指标(Node Degree Heterogeneity Penalization via Average Clustering coefficient,NDHP_AC)和基于节点度异质性惩罚的平均距离(Node Degree Heterogeneity Penalization via Average Distance,NDHP_AD)。

总之,该文贡献如下:

(1)针对稀疏网络,提出节点度异质性惩罚的链路预测指标,该方法启用节点异质性衡量节点对差异并与网络拓扑特征相融合改善稀疏网络预测准确度;

(2)在 8 个真实世界稀疏网络上,启用 AUC 和 AUPR 度量评价 NDHP_AD 和 NDHP_AC,结果表明这两种方法性能优于当前现存代表性方法。

1 方 法

1.1 问题描述

给定一个无向无权网络 $G(V,E)$,其中 V 表示节

点集,E 表示链接集,该文不允许多个链接和自循环存在。用 $\boldsymbol{X}=[x_{ij}]_{n\times n}$ 表示 G 的邻接矩阵。G 是无向无权网络,如果节点之间存在链接,则 $x_{ij}=1$,否则 $x_{ij}=0$。此外,网络任意节点 x 的度表示为 k_x,$\Gamma(x)$ 表示节点 x 的邻居集合,$z\in\Gamma(x)\cap\Gamma(y)$ 表示节点 z 是节点 x 和 y 的共同邻居。

接下来进一步将所有可能的 $\frac{|V|(|V|-1)}{2}$ 链接表示为 U,则 $U-E$ 表示不存在的链接集。链路预测的目标是从集合 $U-E$ 中查找缺失链接。为了测试算法性能,将观测到的链接集 E 随机分成两部分:训练集 E^{T} 和测试集 E^{P}。前者是已知信息而后者仅用于测试。显然,$E^{\mathrm{T}}\cap E^{\mathrm{P}}=\varnothing$ 和 $E^{\mathrm{T}}\cup E^{\mathrm{P}}=E$。

1.2　节点度异质性框架

不同真实世界网络中的节点扮演不同角色。例如在社交网络中,有大 V 节点也有普通粉丝。在交通运输网,有交通枢纽重要的节点也有普通站点。节点的度是反映节点与其他节点的关联程度,度越大,该节点越重要。真实世界网络中度的分布是不均地存在着偏好链接的现象。节点度较高的偏好与度较高的相链接,也存在着节点度较高与较低的相链接。因此,这种现象就造成节点间异质性。然而真实世界网络节点度较大的贡献不如节点度小的贡献,并采用惩罚度较高节点方法获得较好的预测准确度。然而,这种方法无法全面评估网络所有节点度差异仅关注节点共同邻居的度差异。因此,该文关注每个节点度的异质性去捕获网络全局结构信息。设任意网络中两个节点 x 和 y,那么它们的度分别 k_x 和 k_y,则度异质性(Degree Heterogeneity,DH)定义如下:

$$\mathrm{DH}(x,y)=|k_x-k_y| \tag{1}$$

然而在异配网中,式(1)无法准确衡量节点间的异质性。因此,为更全面体现度贡献,重写式(1)如下:

$$S_{xy}^{\mathrm{DH}}=\left(\frac{1}{\mathrm{DH}(x,y)}\right)^{a}=\left(\frac{1}{|k_x-k_y|}\right)^{a} \tag{2}$$

式(2)的作用是抑制节点度的差异。

1.3　所提指标(NDHP_AC 和 NDHP_AD)

式(2)利用度异质性衡量被预测节点产生链接的可能性,该框架的预测准确度与可调参数 α 有密切关联。为弥补惩罚节点间差异导致节点度不均衡,通过可调参数 α 与网络拓扑特征相融合获得更多网络结构信息。当前研究已证实网络结构信息是链路预测重要的组成部分,而聚类系数和最短路径是网络拓扑特征最重要的指标。节点聚类系数反映了节点间聚集程度,表明节点聚类系数越高,节点与其他节点关联程度就越高;而平均最短路径具有“小世界”特性,节点间产生链接最多不会超过 6 节点。在社交网络中,聚类系数可以衡量给定用户的朋友之间成为朋友的倾向,平均最短路径意味着要成为朋友最多不会超过 6 个人。因此,该文将平均节点聚类系数和平均最短路径融合到式(2)中弥补稀疏网络节点邻居信息的不足。由于任意网络平均节点聚类系数和平均最短路径都是一个常数,先计算每个节点聚类系数和最短路径。设任意网络节点 z,聚类系数定义如下:

$$C_z=\frac{2t_z}{k_z(k_z-1)} \tag{3}$$

其中,t_z 表示经过节点 z 闭合三角形个数,k_z 表示节点 z 的度。由式(3)可得局部平均节点聚类系数,有:

$$\bar{C}=\frac{C_z}{|V|} \tag{4}$$

设网络任意节点间最短距离矩阵为 $\boldsymbol{D}$,那么平均最短距离为:

$$\bar{\boldsymbol{D}}=\frac{\boldsymbol{D}}{|\boldsymbol{V}|(|\boldsymbol{V}|-1)} \tag{5}$$

式(4)和式(5)与式(2)相融合,提出节点度异质性惩罚的平均聚类系数指标和节点度异质性惩罚的平均最短距离指标:

$$S_{xy}^{\mathrm{NDHP\_AC}}=(|k_x-k_y|)\times\left(\frac{1}{|k_x-k_y|}\right)^{\alpha\bar{C}} \tag{6}$$

$$S_{xy}^{\mathrm{NDHP\_AD}}=(|k_x-k_y|)\times\left(\frac{1}{|k_x-k_y|}\right)^{\alpha\bar{D}} \tag{7}$$

其中,α 是可调参数。

为便于理解等式,取任意网络的一部分,如图 1 所示,该网络由 7 个节点和 7 条链接构成。根据等式,计算节点 x 和 y 的相似度,$k_x=4$,$k_y=2$,节点 x 和 y 间的异质性 DH=4-2=2,又根据等式(4)可计算整个网络的平均节点聚类系数 $\bar{C}=0.3095$,因此 x 和 y 的相似度 $S_{xy}^{\mathrm{NDHP\_AC}}=2\times\left(\frac{1}{2}\right)^{\alpha\bar{C}}=2\times\left(\frac{1}{2}\right)^{0.3095\times\alpha}$。同理 NDHP_AD 指标,由等式(5)可计算出整个部分网络的平均最短距离 $\bar{D}=2.1429$,因此 x 和 y 的相似度 $S_{xy}^{\mathrm{NDHP\_AD}}=2\times\left(\frac{1}{2}\right)^{\alpha\bar{D}}=2\times\left(\frac{1}{2}\right)^{2.1429\times\alpha}$

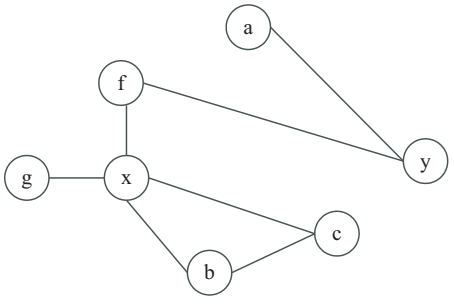

图 1　例证 NDHP_AC 和 NDHP_AD 相似度计算过程

两个指标最终预测结果通过调整参数来获得。

为实现以上过程,所提指标执行过程如下:

算法 1:NDHP_AC 和 NDHP_AD。

输入:任意无向无权网络的邻接矩阵 $\boldsymbol{A}$ 及可调参数 α;

输出: AUC 和 AUPR 平均值。

(1)将任意无向无权数据集转为邻接矩阵;

(2)将邻接矩阵划分为训练集和测试集;

(3)根据式(2)惩罚节点对差异大的度;

(4)根据式(4)和式(5)计算节点平均节点聚类系数和平均最短距离;

(5)根据式(6)和式(7)融合网络结构特征计算被预测节点对 x 和 y 的相似度分数 S_{xy} 并生成预测分数矩阵;

(6)根据 S_{xy} 计算 AUC 和 AUPR。

2 实验结果

本节主要介绍评价度量、基准方法、数据集和实验结果分析。

2.1 评价度量

启用 AUC[20] 和 AUPR[21] 二个度量去衡量所有方法的性能,其中 AUPR 是综合性指标。AUC 和 AUPR 值越高表示该方法预测准确度越高,两个度量具体定义如下:

(1) AUC(Area Under ROC Curve)是 ROC 曲线下的面积,可以理解为在测试集 E^{P} 中的链接分数大于随机选择的一个不存在集 $U - E$ 中的链接分数的概率。独立地比较 n 次,若有 n_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 n_2 次两分数值相等,AUC 定义如下:

$$\mathrm{AUC} = \frac{n_1 + 0.5 * n_2}{n}$$

(2) AUPR(Area Under Precision-Recall curve)是精度-召回率曲线下的面积。精度-召回率曲线即是阈值曲线,该曲线每一个点都对应着不同的分数阈值,具有不同的精度和召回率。

2.2 数据集

采用 8 真实世界无向无权网络评价所有指标性能,其拓扑结构特征统计在表 1 中。

其中, $|V|$ 是节点数, $|E|$ 是链接数,$< k >$表示平均度, r 表示网络同配系数,AC 表示节点平均聚类系数,Density 表示网络稀疏程度。

表 1 8 个真实世界无向无权网络拓扑特征统计

Network	$\lvert V\rvert$	$\lvert E\rvert$	$< k >$	r	AC	Density
CEL	297	2 148	14.464 6	-0.163 2	0.292 4	0.048 9
HAM	1 858	12 534	13.491 9	-0.084 7	0.141 4	0.007 3
BNF	1 781	8 911	10.006 7	-0.094 2	0.262 8	0.005 6
ROAD	1 174	1 417	2.414	0.126 7	0.016 7	0.002 1
YEA	2 361	6 646	5.629 8	-0.099 1	0.130 1	0.002 4
EMA	1 133	5 451	9.622 2	0.078 2	0.220 2	0.008 5
ADO	2 539	10 455	8.235 5	0.251 3	0.146 7	0.003 2
SM	3 084	10 400	6.744 2	-0.031 6	0.150 7	0.002 2

8 个无向无权网络的数据集介绍如下:

(1)线虫的神经网络(CELegans, CEL)[22] 是由 297 个节点和 2 148 条链接构成。节点表示线虫神经元,节点间链接表示神经元突触。

(2)蛋白质相互作用网络(YEAst, YEA)[22] 是由 2 361个节点和 6 646 条链接组成。节点表示蛋白质,节点间链接表示相互作用关系。

(3)青少年健康网络(ADOlescent, ADO)[22] 是根据 1994/1995 年的一项调查创建的。一个节点代表一个学生,两个学生之间的边表明左学生选择了右学生作为朋友。

(4)论文引用网络(SciMet, SM)[23] 是来自 Garfield 使用 HistCite 软件生成的引用网络数据集。节点表示论文,链接表示论文间引用关系。

(5)脑神经网络(NBFly, NBF)[23] 是来源脑网络,由 1 781 个节点和 8 911 条链接组成。节点表示纤维束,链接表示纤维束之间关系。

(6)国际电子公路网络(ROAD)[22],该网络节点代表城市,两个节点之间的边表示它们由一条电子公路连接。

(7)朋友网络(HAMster, HAM)[20] 包含了 hamster. com 网站用户之间的友谊和家庭联系。

(8)电子邮件网络(EMAil, EMA)[23] 是 Rovirai Virgili 大学成员之间电子邮件交流网络。节点表示成员,链接表示成员间电子邮件间次数。

2.3 基准方法

为验证所提算法性能,启用 10 个最近几年的代表性方法与之比较。10 个链路预测方法介绍如下:

(1)3 个基于自包含协同过滤框架(Self-included Collaborative Filtering,SCF)融合(CN、AA 和 RA)的 SCF-CN、SCF-AA 和 SCF-RA 的指标[8],其框架定义如下:

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

(2)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,其定义如下:

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

(3)偏好连接指标(Preferential Attachment,PA)[24],该指标核心思想是节点度大的更容易产生链接,其定义如下:

$$S_{xy}^{\mathrm{PA}} = k_x \times k_y$$

(4)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,其定义如下:

$$S_{xy}^{\mathrm{LP}} = A^2 + \alpha A^3$$

(5)节点聚类系数链路预测指标(Clustering Coefficient for Link Prediction,CCLP)[9],该方法利用节点共同邻居聚类系数捕获局部结构信息,其定义如下:

$$S_{xy}^{\mathrm{CCLP}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{t_z}{k_z(k_z - 1)/2}$$

其中,t_z 表示经过 z 的三角形数。

(6)共同邻居度惩罚指标(Common Neighbors Degree Penalization,CNDP)[19],该方法是在自适应惩罚方法基础上考虑共同邻居数,任意节点相似度定义如下:

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

其中,$|C_z|$ 是共同邻居数。

(7)Katz 指标[13],该方法考虑整个网络所有节点的路径,其定义如下:

$$S = (\boldsymbol{I} - \alpha \cdot A)^{-1} - \boldsymbol{I}$$

其中,$\boldsymbol{I}$ 是单位矩阵,α 是可调参数。

(8)矩阵森林指标(Matrix-Forest Index,MFI)[25],该指标基于矩阵森林提出的全局指标,其定义如下:

$$S = (\boldsymbol{I} + \alpha \boldsymbol{L})^{-1}$$

其中,$\boldsymbol{L}$ 是拉普拉斯矩阵,α 是可调参数。

2.4　实验结果分析

实验硬件平台为 Intel Core i5-7200U CPU 笔记本,主频 2.71 GHz,内存 4 GB,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,为公平比较所有的方法,在所有数据集中设 $\alpha = 0.5$。而其余指标 LO 可调参数为 0.1,LP 可调参数为 0.001,CNDP 可调参数为 1.5,Katz 可调参数为 0.01。

通过两个实验评估所提方法的性能。首先,启用 AUC 和 AUPR 度量全面评估所有 12 个指标性能;然后,评估可调参数 α 对所提方法性能的影响。

对于第一实验,启用 AUC 和 AUPR 度量评估所有 12 指标性能,实验结果见表 2,并观察到以下四个现象:

(1)所提两个指标(NDHP_AD 和 NDHP_AC)在 8 个数据集中 AUC 和 AUPR 获得最优,表明充分利用网络结构信息可以有效改善稀疏网络的性能。NDHP_AD 性能略优于 NDHP_AC,其主要原因是聚类系数在 8 个稀疏网络无法获得更多节点邻居信息而最短路径获得更多节点邻域信息。通过表 1 可以观察到数据集 CEL、HAM、BNF、YEA 和 SM 的同配系数为负数,表明以上网络中存在大量度大的节点与度小的节点相链接,由此造成节点间度异质性,所提方法采用惩罚机制,结果表明该机制有效可行。此外,除了 CEL 数据集外,其余数据集均为高度稀疏网络,所提指标在高度稀疏网络中均获得高性能。

(2)所提指标与五个局部相似度方法(SCF-CN、SCF-AA、SCF-RA、PA 和 CCLP)相对比,性能获得显著的改善。在 AUC 度量,NDHP_AD 和 NDHP_AC 指标与五个指标中最好第二指标在数据集上 CEL、HAM、BNF、ROAD、EMA、ADO、YEA 和 SM 相比较,分别提升了 10%、3.1%、9.2%、22%、13%、6.3%、7.2%和 6.2%;同理,在 AUPR 度量,在数据集上 CEL、HAM、BNF、ROAD、EMA、ADO、YEA 和 SM 分别提高了 7.9%、7.7%、10.3%、20%、26%、13%、27%和 20%。上述五个指标获得低质量性能的主要原因是稀疏网络中无法获得足够的结构信息,尽管基于协同过滤框架利用对称性增强了获取局部结构信息能力,然而高度稀疏网络中局部结构信息是有限的。

(3)NDHP_AD 和 NDHP_AC 与三个全局指标(MFI、Katz 和 LO)相比较,所提指标获得最优性能。三个全局指标性能最接近所提指标,主要原因是全局指标获得整个网络拓扑结构信息,如 Katz 可获得整个网络节点路径信息,LO 可获得整个网络节点的邻居贡献值,因此全局指标可以用全局结构信息弥补稀疏性不足。MFI 指标优于 LO 和 Katz 的原因是该方法核心任意一个为根节点生成森林再统计与此根节点相似的节点的比例。

(4)半局部相似度算法(LP 和 CNDP)的核心平衡时间复杂度和性能关系。该方法的缺点是无法完全捕获整个网络结构信息而导致敏感于稀疏网络。通过表 2 可观察到,以上两种方法仅在数据集 CEL 中获得相对较好的性能,而在其余数据集中与全局方法相比

均获得低质量性能,主要原因是 LP 方法仅考虑三阶路径信息,而 CNDP 仅利用整个网络平均节点聚类系数。此外,文中方法与 LP 和 CNDP 在高度稀疏网络中相比性能显著提升。在数据集 YEA 中,文中方法 AUC 和 AUPR 分别提高了 14% 和 46% 。

表 2 8 个无向无权网络上 12 个不同指标对应 AUC 和 AUPR 值

指标	度量	CEL	HAM	BRA	ROAD	YEA	EMA	ADO	SM
SCF-CN	AUPR	0.392	0.314	0.303	0.256	0.163	0.244	0.135	0.167
	AUC	0.846	0.939	0.881	0.558	0.835	0.894	0.845	0.893
SCF-AA	AUPR	0.593	0.592	0.568	0.116	0.404	0.524	0.357	0.456
	AUC	0.866	0.944	0.885	0.548	0.843	0.890	0.845	0.896
SCF-RA	AUPR	0.594	0.603	0.575	0.120	0.405	0.540	0.352	0.463
	AUC	0.867	0.954	0.894	0.551	0.836	0.905	0.842	0.899
PA	AUPR	0.458	0.407	0.428	0.175	0.313	0.378	0.304	0.302
	AUC	0.739	0.864	0.867	0.342	0.784	0.781	0.616	0.821
CCLP	AUPR	0.461	0.311	0.361	0.290	0.195	0.350	0.204	0.255
	AUC	0.870	0.808	0.868	0.505	0.695	0.841	0.754	0.779
MFI	AUPR	0.603	0.631	0.598	0.357	0.542	0.619	0.630	0.607
	AUC	0.876	0.923	0.874	0.634	0.806	0.895	0.904	0.911
Katz	AUPR	0.591	0.622	0.609	0.341	0.551	0.620	0.620	0.603
	AUC	0.852	0.912	0.880	0.626	0.817	0.892	0.886	0.905
LO	AUPR	0.580	0.644	0.572	0.410	0.541	0.574	0.525	0.549
	AUC	0.827	0.936	0.817	0.516	0.774	0.813	0.721	0.789
LP	AUPR	0.490	0.386	0.385	0.220	0.203	0.313	0.170	0.228
	AUC	0.875	0.916	0.887	0.555	0.833	0.899	0.847	0.899
CNDP	AUPR	0.447	0.295	0.354	0.175	0.189	0.320	0.173	0.244
	AUC	0.864	0.806	0.878	0.525	0.708	0.853	0.765	0.795
DHPLP_AC	AUPR	0.673	0.680	0.678	0.496	0.665	0.669	0.634	0.663
	AUC	0.971	0.985	0.986	0.779	0.971	0.968	0.917	0.961
DHPLP_AD	AUPR	0.670	0.682	0.677	0.498	0.666	0.673	0.636	0.664
	AUC	0.966	0.989	0.985	0.782	0.973	0.973	0.919	0.962

NDHP_AC 和 NDHP_AD 带有可调参数 α ,该参数的作用是平衡节点度异质性和网络结构信息关联程度。设可调参数 α 的范围为[-1, -0.5,0,0.5,1,1.5],实验结果如图 2 和图 3 所示。

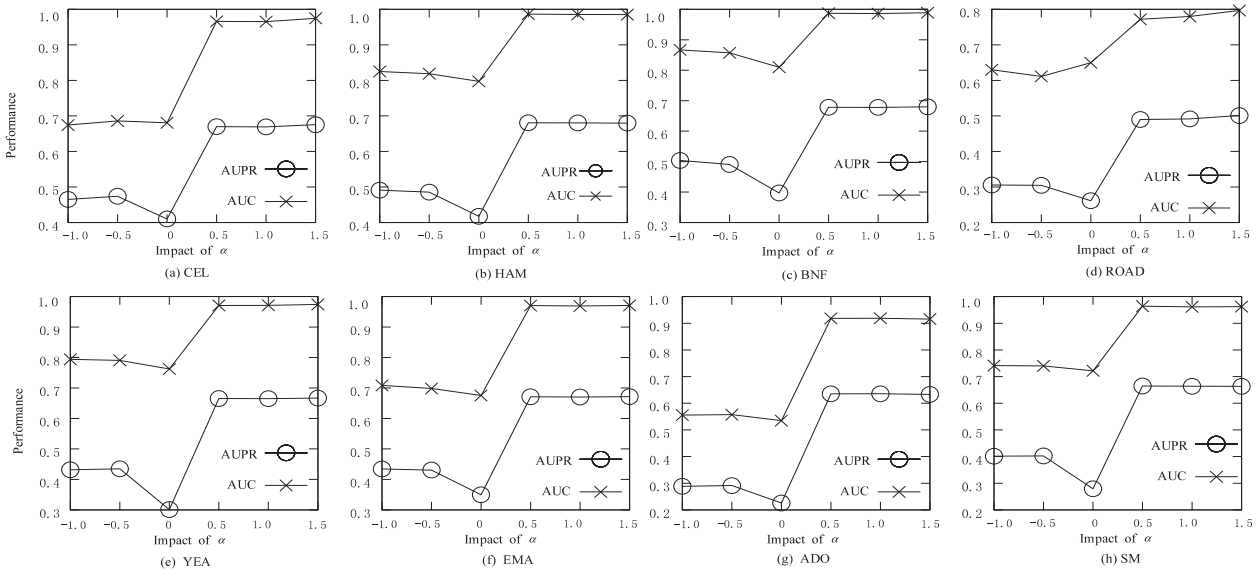

图 2 参数 α 变化对 NDHP_AC 性能影响

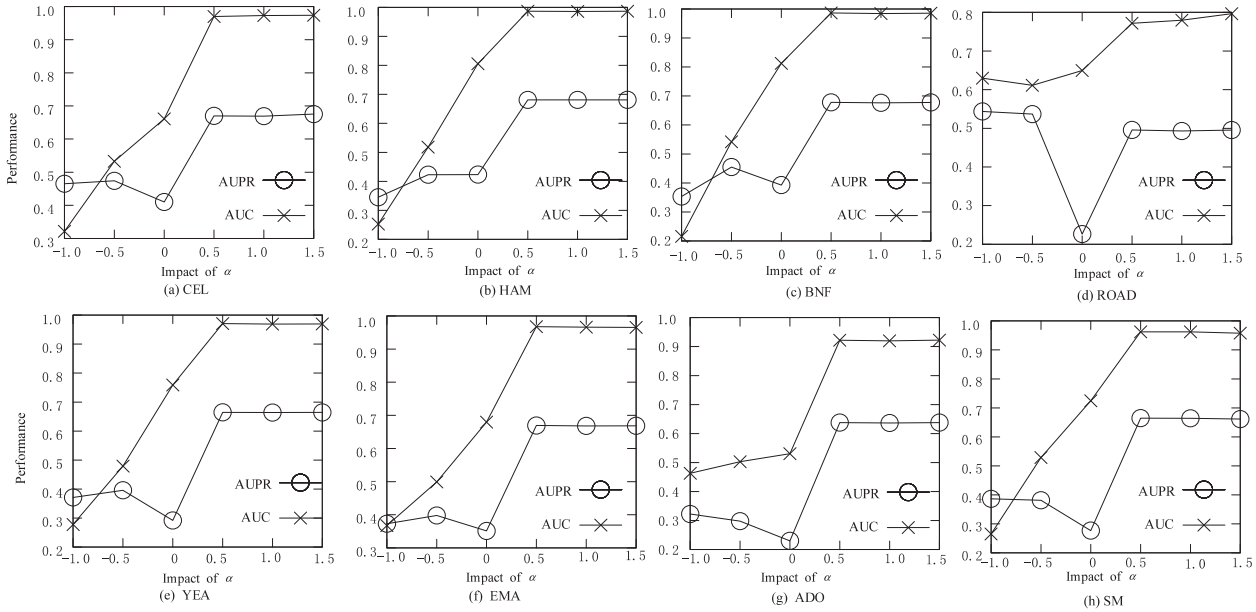

图 3　参数 α 变化对 NDHP_AD 性能影响

可观察到当 α 为负数时,AUC 和 AUPR 值最低,主要原因是未调用惩罚机制去抑制度异质性;当 $\alpha=0$ 时,文中方法退化为节点度差异代表相似度分数,此时 AUC 和 AUPR 值最低,主要原因是在无法启用惩罚机制及利用网络结构信息;当 α 为正数时,文中方法可以调用惩罚机制及充分利用网络结构信息获得最优的性能。因此,当 $\alpha=0.5$ 时,NDHP_AC 性能最优。

3　结束语

探索和利用网络结构是链路预测的重要组成部分,如何将网络结构和度差异相结合适用于稀疏网络是一个挑战性问题。该文提出节点度异质性惩罚的链路预测指标,该指标利用节点度异质性捕获网路全局结构,再利用节点聚类系数衡量节点间聚合能力,然后结合平均节点聚类系数保持网络局部结构。采用 8 个稀疏网络与 10 个最近代表性方法测试所提方法的性能,结果表明该方法性能上远超其他基准方法,同时健壮于高度稀疏网络。

未来尝试将网络社区结构与节点属性信息融入该方法,此外可以扩展该方法到加权和有向网络。

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] HUANG Z,ZENG D D. A link prediction approach to anomalous email detection[C]//Proceedings of the 2006 IEEE international conference on systems,man and cybernetics. Taipei,China:IEEE,2006:1131-1136.

[3] HUO Z, HUANG X, HU X. Link prediction with personalized social influence[C]//Proceedings of the thirty-second AAAI conference on artificial intelligence. Louisiana:AAAI, 2018:2289-2296.

[4] YU S,ZHAO M,FUC,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] LEE Y L,ZHOU T. Collaborative filtering approach to link prediction[J]. Physica A:Statistical Mechanics and Its Applications,2021,578:126107.

[9] 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.

[10] WU Z,LIN Y,WAN H,et al. Predicting top-L missing links with node and link clustering information in large-scale networks[J]. Journal of Statistical Mechanics:Theory and Experiment,2016,2016(8):083202.

[11] LIU Y,ZHAO C,WANG X,et al. The degree-related clustering coefficient and its application to link prediction[J]. Physica A:Statistical Mechanics and Its Applications,2016, 454:24-33.

[12] LÜ L,ZHOU T. Link prediction in complex networks:a survey[J]. Physica A:Statistical Mechanics and Its Applications,2011,390(6):1150-1170.

[13] KATZ L. A new status index derived from sociometric analy-

(下转第 158 页)