ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 5,Sep. 2018,pp. 900－910
DOI:10.16337/j.1004-9037.2018.05.015

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

结合共同邻居贡献度的节点相似性链路预测算法

王　鑫[1,2]　陈　喜[1,2]　钱付兰[1,2]　张燕平[1,2]

(1.安徽大学计算机科学与技术学院,合肥,230601;2.安徽大学计算机智能与信号处理教育部重点实验室,合肥,230601)

摘　要:链路预测是复杂网络的一个重要研究方向,基于节点相似性的链路预测方法是最为常用的一种方法。目前大部分使用节点链接紧密度的节点相似性链路预测方法,未考虑每个共同邻居节点的差异性,即不同的节点对连边的贡献度是不同的。本文提出一种结合共同邻居节点之间的节点贡献度和链接紧密度的链路预测算法。该算法首先计算共同邻居节点之间的链接信息作为节点的链接紧密度,再定义耦合度聚簇系数表示共同邻居节点贡献度,最终将二者结合。在实际数据集上的实验结果表明,该算法比4种经典的链路预测算法(CN,AA,RA和Jaccard)和基于节点链接密度的算法CNBIDE具有更好的预测精度。

关键词:复杂网络;链路预测;贡献度;紧密度;节点相似性

中图分类号:TP391　　**文献标志码**:A

Node-Similarity Link Prediction Algorithm Combined Common Neighbor Contribution

Wang Xin[1,2], Chen Xi[1,2], Qian Fulan[1,2], Zhang Yanping[1,2]

(1. School of Computer Science and Technology, Anhui University, Hefei, 230601, China; 2. Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei, 230601, China)

Abstract: Link prediction is an important research direction of complex networks, and the method based on the node similarity is one of the most popular methods. So far, most of the node similarity prediction methods using link density have not considered the difference of each common neighbor node, that is, the contribution of different nodes to the link is different. Therefore, this paper proposes a link prediction algorithm based on the node contribution and link density of the common neighbor nodes(LDNC). The algorithm first calculates the link information between the common neighbor nodes as the link density of the nodes, and then defines node-coupling clustering coefficient to describe the contribution of the common neighbor nodes, and finally combines the two parameters. Experiments based on the real-world datasets show that the LDNC is more accurate compared with four baseline link prediction algorithms (CN, AA, RA and Jaccard) and the CNBIDE algorithm based on the node link density.

Key words: complex networks; link prediction; contribution; density; node similarity

---

基金项目:国家自然科学基金(61673020,61602003,61402006)资助项目;教育部人文社科青年基金(14YJC860020)资助项目。
收稿日期:2017-05-19;**修订日期**:2017-07-09

引　言

随着互联网的发展,各种复杂网络如雨后春笋般出现,链路预测作为数据挖掘领域的研究方向之一,得到越来越多的关注。链路预测是指如何通过已知的网络节点以及网络结构等信息,预测网络中尚未产生连边的两个节点之间产生链接的可能性[1-3]。这种预测既包含了对未知链接的预测,也包含了对未来链接的预测。链路预测对一些实际问题具有重要的理论研究意义,如网络的动态演化等。刘宏鲲等利用链路预测的方法推断出影响航空网络演化的重要因素[4]。链路预测也具有广泛的实际应用价值,如电子商务网站向用户推荐感兴趣的商品,在线社交网络中的好友推荐[5];在蛋白质相互作用网络中,利用链路预测算法预测尚未发现的蛋白质分子之间的相互作用关系,有助于加快揭开蛋白质网络的真实面目[6,7];在科学家合作网络中识别科学家之间潜在的合作可能[8];此外链路预测方法在识别犯罪网络、检测和发现恐怖袭击等社会安全领域也能发挥重要作用。

1　链路预测研究现状

链路预测早期的研究方法主要是基于马尔科夫链和机器学习。文献[9]应用马尔科夫链进行网络的链路预测与路径分析。之后文献[10]将基于马尔科夫链的预测方法扩展到自适应网站的预测中。通过提取网络的特征,机器学习方法通过训练模型来进行链路预测。O′Madadhain 等应用了几种分类器来预测网络中的潜在链接[11]。以上的两种方法比较复杂,从复杂网络的角度进行链路预测研究是一种新的方式,这种方法较简单,且具有普适性。基于拓扑相似性的链路预测算法是近几年的主流方法,这种方法的一个重要假设就是两个节点越相似,则节点间存在链接的可能性越大,因此该方法的的一个关键问题就是如何来定义节点之间的相似性[12]。

通常可以将拓扑网络中的各种信息融合在一起来定义节点之间的相似性,其中最重要的信息是节点属性和网络结构,节点属性信息有较高的预测精度,但是收集这些信息不易。即使可以获得节点属性信息,从这些复杂的信息中鉴别出哪些信息对预测有用也是一件困难的事。与节点属性信息相比较,网络结构信息更容易获取。并且,基于网络结构信息的链路预测方法对结构相似的网络具有普遍适用性。因此,近些年来基于网络结构相似性链路预测方法受到越来越多的关注。

基于局部信息、基于路径和基于随机游走的相似性算法是最主要的 3 种基于结构相似性的链路预测方法。共同邻居指标(Common neighbors,CN)[13]是最简单的基于局部信息的相似性指标,该指标认为两节点的相似度正比于共同邻居节点数量。与 CN 相似的 Jaccard 算法[14]就是 CN 算法中计算节点相似性的归一化形式。如果考虑共同邻居节点的度,著名的有 Admic-Adar 指标(AA)[15]。受到网络中资源分配过程的启发,周涛等提出了资源分配指标(Resource allocation,RA)[16]。基于路径的相似性指标有局部路径指标(Local path,LP)[16,17],它只考虑了二阶路径数目。Katz 指标[18]考虑了网络的所有路径。有相当数量的相似性指标基于随机游走过程定义,包括平均通勤时间(Average commute time,ACT)[19],余弦相似性指标 $\cos^{+}$[20],有重启的随机游走指标(Random walk with restart,RWR)[21]以及基于局部游走的 LRW(Local random walk)和 SRW(Superposed random walk)指标[22]。以上的链路预测方法除了 Katz 指标[18],其他的都是基于局部的相似性指标。这种指标由于其复杂度较低,得到了广泛的应用。

随着链路预测得到越来越多的关注,近年来各种新颖的链路预测方法相继被提出。文献[23]根据网络中路径上潜在的资源传播的思想,提出一种扩展的 RA 算法。文献[24]的目的是利用社团结构信息来提高基本的链路预测方法的性能。该信息来自于同一个社团内的两个节点所处的社团层次数目(利用了分层的社团发现算法)。结果显示有价值的社团信息可以提高链路预测的效果。但是无论是经典方法,还是最新方法,很多都忽略了每个节点的贡献度不同以及共同邻居节点之间的链接紧密度信

息。因此,本文提出一种新的链路预测算法,综合考虑每个共同邻居节点的差异性以及链接的紧密度。

2 问题描述及预测方法

2.1 问题描述

定义一个无向无权网络 $G(V,E)$,其中 V 代表节点的集合,E 为边的集合。网络的节点总数为 N,边的数量为 M,所有的节点对组成全集 U。这样,链路预测问题可描述为:定义某种相似性指标,对没有连边的节点对(x,y),计算其相似度 S_{xy},S_{xy} 越大,则节点对(x,y)出现连边的可能性越大。

为了测试算法精度,将集合 E 分为训练集 E^{T} 和测试集 E^{P}。将属于全集 U 但不属于集合 E 的边称为不存在的边,属于 U 但不属于训练集 E^{T} 的边称为未知边。在本文中,训练集所占比例为 90%,测试集为 10%。

2.2 链路预测方法

基于局部信息的相似性算法由于其计算复杂度低,适合规模较大的复杂网络,故得到广泛使用。在本文中,将会和 CN,AA,RA 和 Jaccard 这 4 种经典的相似性算法以及基于节点连接密度的算法 CNBIDE 进行比较。本文使用的符号说明如表 1 所示。

表 1 符号说明

Tab. 1 Symbolic representation

符号	含义
S_{xy}	节点 x 与 y 的相似性评分
K_x	节点 x 的度值,等于与节点 x 直接相连的节点个数
$\Gamma(x)$	与节点 x 直接相连的节点构成的集合,即节点 x 的邻居集合
$C(x)$	节点 x 的聚簇系数
$\lvert Z\rvert$	Z 集合中元素的数量

CN 算法[13]。该算法中,两个节点 x 和 y 的相似性就定义为它们共同邻居的数量,即

$$\boldsymbol{S}_{xy}^{\mathrm{CN}}=\lvert\Gamma(x)\cap\Gamma(y)\rvert \tag{1}$$

Jaccard 算法[13]。Jaccard 指标代表了一个随机选择的节点 x 或 y 的邻居是 x 和 y 的共同邻居的可能性,即

$$\boldsymbol{S}_{xy}^{\mathrm{Jaccard}}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert} \tag{2}$$

AA 算法[15]。该算法认为度小的共同邻居节点对连边影响更大,即

$$\boldsymbol{S}_{xy}^{\mathrm{AA}}=\sum_{z\subset\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k_z} \tag{3}$$

RA 算法[16]。受网络资源分配过程的启发,周涛等提出了 AA 算法。该算法惩罚了度大的共同邻居节点,即

$$\boldsymbol{S}_{xy}^{\mathrm{RA}}=\sum_{z\subset\Gamma(x)\cap\Gamma(y)}\frac{1}{k_z} \tag{4}$$

CNBIDE 算法[25]。该算法考虑了共同邻居的链接紧密度,然后和 CN 算法相结合,即

$$\boldsymbol{S}_{xy}^{\mathrm{CNBIDE}}=\lvert\Gamma(x)\cap\Gamma(y)\rvert*\left(1+\frac{\lvert Z\rvert}{\frac{1}{2}\lvert\Gamma(x)\cap\Gamma(y)\rvert(\lvert\Gamma(x)\cap\Gamma(y)\rvert-1)}\right) \tag{5}$$

式中 $Z=\{(a,b)\mid(a,b)\in E,a\in\Gamma(x)\cap\Gamma(y),b\in\Gamma(x)\cap\Gamma(y)\}$。

这几种链路预测算法都是从共同邻居节点的角度提出。CN 算法只考虑了共同邻居的数量,未对

它们的重要性做任何区分。Jaccard 考虑了除共同邻居之外的其他邻居节点,但也只是考虑到了邻居节点的数量对链接产生的影响。AA 和 RA 算法比较相似,区别在于定义共同邻居节点权重的方式,前者以 $1/\log k$ 的形式递减,后者以 $1/k$ 的形式递减。CNBIDE 算法考虑了共同邻居数量,也考虑了共同邻居之间的链接密度。然而,这几种算法并未考虑到每个共同邻居节点对连边的贡献度是不同的,即不能简单地将邻居节点同等对待。本文针对以上问题提出一个新的节点相似性算法,既考虑了共同邻居的链接紧密度,也考虑了共同邻居节点的贡献度。

3　结合共同邻居贡献度和连接紧密度的链路预测方法

3.1　共同邻居节点的链接紧密度

在基于节点相似性的链路预测算法中,计算节点间的连边概率通常是基于它们的共同邻居节点信息。但是,在以往的算法中,利用的大多是共同邻居的数量信息或者度信息,忽略了共同邻居之间的相互关系,而共同邻居之间的链接关系往往对连边有极大的影响。

图 1 表示 2 个简单的网络拓扑结构图及抽取的 x 和 y 节点共同邻居的拓扑结构。目的是通过图 1,分析节点 x 和 y 之间存在链接的概率大小。图 1 中(a),(b)两个网络拓扑图中,节点 x 和 y 的共同邻居节点都是 4 个,都是 a,b,c,d,且两图中的 a,b,c,d 节点的度值均相等。如果利用基于共同邻居的链路预测算法来进行预测(如 CN 等),则两图中 x 和 y 节点的相似性概率相同。如果利用基于共同邻居的度信息的链路预测算法(如 AA,RA 等)进行预测,两图中 x 和 y 节点之间存在链接的概率也相等。图 1 中(c),(d)分别是由(a),(b)中节点 x 和 y 的共同邻居节点 a,b,c,d 及它们之间的连边组成的子网络。通过比较(c)和(d),可以看出图(c)中节点之间两两互连,而图(d)的各个节点之间并无关联。假设节点 x 和 y 代表现实中的两个人,他们都有 a,b,c,d 4 个相同的朋友,但是一种情况是这 4 个共同朋友都认识,彼此是朋友关系,可以理解为都在同一个单位或者有共同的爱好,在这种情况下,x 和 y 通过共同朋友结识或者产生联系的概率大大增大。而如果共同朋友没有任何交集,彼此孤立存在,则 x 和 y 认识的机会则大为降低。

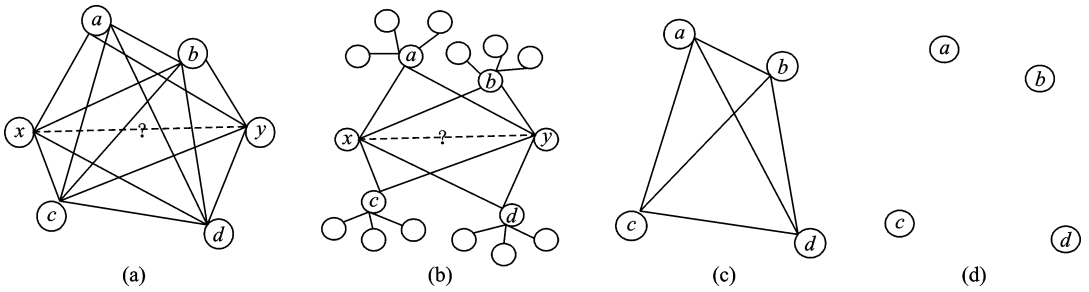

图 1　x,y 节点及其共同邻居组成的小网络

Fig. 1　Small network composed of x, y nodes and their neighbors

考虑到共同邻居的链接紧密程度对最终的链接产生的影响,本文引入了链接紧密度的概念。如果共同邻居节点之间有更多的连边,也就是链接更紧密,则认为被预测节点之间存在链接的概率更高。本文定义一个表示邻居节点集合中各节点链接紧密程度的指标—链接紧密度,具体定义如下

$$LD(x,y)=\begin{cases}0 & \Gamma(x)\cap\Gamma(y)\in\varnothing\\ \dfrac{|Z|}{|\Gamma(x)\cap\Gamma(y)|} & \Gamma(x)\cap\Gamma(y)\notin\varnothing\end{cases}\tag{6}$$

式中 $Z=\{(a,b)\,|\,(a,b)\in E,a\in\Gamma(x)\cap\Gamma(y),b\in\Gamma(x)\cap\Gamma(y)\}$,$\Gamma(x)\cap\Gamma(y)$ 表示节点 x 和 y 的共同邻

居节点集合。

3.2 节点贡献度

基于相似性的的链路预测方法大都仅考虑共同邻居节点的数量或度值信息来评估节点间连边的概率。这些方法将每个共同邻居节点都看成是没有差别的,但是在实际的网络中,不同节点的影响力是不同的。例如在社交网络中,外向的人往往要比内向的人引导两个人相识的作用更强。因此,如果对每个共同邻居节点的差异性不加以区分,将它们对节点连边的贡献度看成是相同的,可能对链路预测的准确性产生影响。

在社交网络中,两个节点如果都和同一个节点相链接,把两节点间的节点称为两节点的共同邻居节点,不同的共同邻居节点对节点连边的贡献度是不同的,本文把节点耦合聚簇系数[26]看成每个共同邻居节点对连边产生的贡献度。该指标综合了每个节点的度信息和聚簇系数信息,体现了每个节点的差异性。

节点贡献度公式为

$$NC(n)=\frac{\sum_{i\in C_n}\left(\frac{1}{k_i}+C(i)\right)}{\sum_{j\in\Gamma_n}\left(\frac{1}{k_j}+C(j)\right)} \tag{7}$$

式中 $\Gamma(n)$是节点 n 的邻居节点集合。如果(M,N)代表待预测的节点对,则 $n\in\Gamma(M)\cap\Gamma(N)$。$C_n$表示节点 n 的邻居节点和待预测节点 M,N 的邻居节点的交集再加上 M 和 N 节点所组成的集合,也即是 $C_n=\Gamma(M)\cap\Gamma(N)\cap\Gamma(n)\cup\{M,N\}$。

$C(n)$表示节点 n 的聚簇系数,其计算公式为

$$C(n)=\frac{2\times E_n}{k_n\times(k_n-1)} \tag{8}$$

式中 E_n 表示节点 n 的邻居节点之间的链接数目。k_n 表示节点 n 的度值。

在式(2)中,因为 $C_n\subseteq\Gamma(n)$,则有 $\sum_{i\in C_n}\left(\frac{1}{k_i}+C(i)\right)\leqslant\sum_{j\in\Gamma_n}\left(\frac{1}{k_j}+C(j)\right)$。因此,$NC(n)\in(0,1]$。特殊情况下,当 $C_n=\Gamma(n)$时,$NC(n)=1$。

3.3 结合共同邻居节点贡献度和链接紧密度算法

传统的链路预测方法大多考虑的网络结构较少,有的只考虑共同邻居个数和度信息,有的只考虑了路径的数量。如果考虑更多的网络结构信息,对预测的准确率可能会有一个提升。基于此,本文从待预测节点的共同邻居节点出发,既考虑了每个节点的差异性,用节点耦合聚簇系数来表示每个共同邻居节点对待预测节点产生连边的贡献度。同时,共同邻居节点之间的紧密程度也可以用来衡量待预测节点之间的相似程度。如果共同邻居节点之间链接越紧密,说明待预测节点之间有更多的相似性性质,相似性程度越高。

将节点贡献度和链接紧密度相结合,本文提出了结合共同邻居贡献度和连接紧密度的链路预测(Node contribution and link density of the common neighbor nodes, LDNC)算法。LDNC 算法中节点 x 和 y 的相似度矩阵 $\boldsymbol{S}_{xy}$ 定义为

$$\boldsymbol{S}_{xy}=\frac{1}{1+\mathrm{e}^{-LD(x,y)}}+\sum_{z\in\Gamma(x)\cap\Gamma(y)}NC(z) \tag{9}$$

式中:$LD(x;y)$是 x,y 节点的共同邻居之间的链接紧密度;$NC(z)$表示 $x;y$ 节点的共同邻居节点 z 对 x,y 节点连边的贡献度。由于网络中连边密度的差异,节点之间的链接紧密度差异较大,经反复实验,将节点的链接紧密度进行归一化后,预测精确度最高。

LDNC 算法的具体过程如下:

输入:无向无权网络 Network

输出:评价指标结果(AUC,Precision)

开始:

(1) 统计网络中被预测节点对(x,y)的共同邻居个数以及它们之间的连边个数,使用式(6)得出节点对(x,y)的共同邻居节点连边紧密度。

(2) 使用式(7)得出被预测节点对(x,y)的每个共同邻居节点对其产生连边的贡献度。

(3) 根据式(9)计算被预测节点对(x,y)之间的相似度 S_{xy}。 最后求得相似度矩阵 SIM。

(4) 根据相似度矩阵 SIM 计算评价指标(AUC,Precision)的结果。

结束

3.4 算法复杂度分析

链路预测算法一般分为基于局部信息和基于全局信息。前者的优势在于时间复杂度较低,适合大规模的网络应用,而后者的时间复杂度则较高。基于局部信息的最简单的链路预测算法是 CN 算法[13],它的时间复杂度为 $O(n^2)$。 Katz 算法[18]是一种代表性的基于全局信息的链路预测算法,它的时间复杂度为 $O(n^3)$。 而本文提出的 LDNC 算法的时间复杂度为 $O(n^2)$,和 CN 算法相同,可适用于大规模网络。

4 实验结果与分析

4.1 评价指标

链路预测的评价指标主要有 AUC(Area under the receiver operating characteristic curve)、精确度(Precision)。AUC 是比较常用的一种评价指标,它是从整体上衡量算法的精确度[27]。Precision 只考虑排在前 L 位的边是否预测准确[28]。

AUC 指标[27]。AUC 可以理解为在测试集中边的分数值比随机选择的一个不存在边的分数值高的概率,也就是说,每次随机从测试集中选取一条边与随机选择的不存在的边进行比较,如果测试集中的边的分数值大于不存在边的分数值,就加 1 分;如果两个分数值相等,就加 0.5 分。这样独立地比较 n 次,如果 n'次测试集中的边的分数值大于不存在的边的分数,n''次两个分数值相等,则 AUC 定义为

$$\mathrm{AUC}=\frac{n'+0.5n''}{n} \tag{10}$$

若所有分数都随机产生,则 AUC=0.5。因此,AUC 大于 0.5 的程度衡量了算法比随机算法准确的程度。

Precision 指标[28]。Precision 指标定义为在前 L 个预测连边中被预测准确的比例。如果有 m 个预测准确,即排在前 L 的连边中有 m 个在测试集中,则 Precision 定义为

$$\mathrm{Precision}=m/L \tag{11}$$

Precision 值的大小与参数 L 有关,对于给定的 L,Precision 值越大,则预测结果越准确。

4.2 实验数据集

实验时,在保证网络连通的情况下,将网络中的连边随机的划分为训练集和测试集,其中训练集占 90%,测试集占 10%。为了验证 LDNC 算法是否有效,实验时选取了 9 个涵盖不同领域的真实网络数据集,分别是:

(1) 食物链网络 1(FWMW)[29]

这是红树林河口湿季的食物链网络,包含 97 个节点和 1 492 条边,其中节点表示生物,边表示生物

之间的捕食关系。本文将该网络进行无向处理。

(2) 食物链网络 2(FWFW)[30]

这是佛罗里达海湾雨季的食物链网络。包含 128 种节点和 2 075 条边,其中节点表示生物,边表示生物之间的捕食关系。本文将该网络进行无向处理。

(3) 爵士音乐家合作网络(Jazz)[31]

这是一个爵士音乐家合作网络,包含 198 个节点和 2 742 条边,其中音乐家用节点表示,节点间的连边表示两个音乐家是朋友关系。

(4) 线虫的神经网络(C. elegans)[32]

这是秀丽隐杆菌(C. elegans)的神经网络,包含了 297 个节点和 2 345 条边,其中神经元用节点来表示,边代表神经元突触或者间隙连接。本文将该网络进行无向处理。

(5) 美国航空网络(USAir)[33]

这是美国航空网络,有 332 个节点和 2 126 条边,其中机场用边表示,连边表示航线。如果两个机场有直飞航线,则机场所对应的两个节点之间有一条连边。

(6) 美国政治博客网络(PB)[34]

这是一个由博客网页之间的链接关系组成的网络,包含 1 490 个节点和 19 090 条有向边,其中节点为博客网页,边表示网页之间的超链接。本文将该网络进行无向处理。

(7) 科学家合作网络(NS)[35]

这个一个科学家合著网络,包含了 1 589 个节点和 2 742 条边,其中节点代表科学家,边代表科学家之间的合作关系。

(8) 蛋白质相互作用网络(Yeast)[36]

节点表示蛋白质,边表示其相互作用关系。包含 2 375 个节点和 11 693 条边。

(9) 电力网络(Power)[32]

这是美国西部电力网络,包含 4 941 个节点和 6 594 条边。其中节点代表变电站或换流站,连边表示它们之间的高压线。

表 2 列出刻画网络特征常用的统计量符号及其含义,表 3 列出 9 个真实网络数据集的统计量信息。

表 2 统计量符号及其含义

Tab. 2 Statistical symbols and meaning

符号	含义
N	网络的节点数
M	网络的边数
$\langle k\rangle$	网络的平均度
C	网络的聚类系数
P	平均最短路径
D	网络直径

表 3 9 个真实网络的统计量

Tab. 3 Statistical measure of nine real networks

网络	N	M	$\langle k\rangle$	C	P	D
FWMW	97	1 492	15.4	0.468	1.692 9	3
FWFW	128	2 075	32.42	0.334 6	1.776	3
Jazz	198	2 742	27.70	0.618	2.235	6
C. elegans	297	2 345	7.9	0.292	2.46	5
USAir	332	2 126	12.81	0.749	2.738	6
PB	1 490	19 090	27.36	0.36	2.74	8
NS	1 589	2 742	3.451 2	0.637 8	—	—
Yeast	2 375	11 693	9.85	0.388	5.10	15
Power	4 941	6 594	2.67	0.107	18.99	46

注:—表示该网络不是连通网络。

4.3 实验结果分析

本文采用 AUC 和 Precision 作为衡量链路预测算法准确性的指标。LDNC 算法和 CN, AA, RA, Jaccard 和 CNBIDE 算法的比较结果如表 4 和表 5 所示。表 4 是 AUC 指标的结果,表 5 是 Precision 指

标($L=100$)的结果,加黑字体表示最优结果。图 2 给出了 6 种算法的 Precision 值随 L 值变化的结果。

观察表 4 AUC 指标的结果,在 FWMW,FWFW,USAir,NS,Yeast 和 Power 这 6 个网络上,LDNC 算法均取得了最优值,而在 C. elegans 网络上,LDNC 和 RA 的 AUC 值相同,也是最优。从表 4 也可以看出,除了 LDNC 算法,RA 算法在各个数据集上表现得最好,这说明网络中节点度对预测结果的影响是巨大的。而当网络的平均度较大时,RA 指标表现最好。Jazz 和 PB 这两个网络的平均度较大,在 AUC 评价指标上也好于 LDNC,但是用 Precision 指标进行比较,LDNC 预测准确性好于 RA。

观察表 5 Precision 指标的结果,在 FWMW,FWFW,Jazz,C. elegans,USAir,PB,NS 和 Yeast 这 8 个网络中,LDNC 算法的预测准确率最高。其中在 NS 和 Yeast 网络中,预测精度提升超过了 10%。在 Power 网络中 CNBIDE 取得了最优的预测精度,但 LDNC 算法的预测精度也超过了 4 种经典算法。Precison 定义为前 L 个预测边中预测准确的比例,是一种局部评价指标。而 LDNC 算法在 Precision 上表现较好,说明 LDNC 算法对排名靠前的边预测准确率高。

实验结果表明,和 CN,AA,RA 和 Jaccard 4 种传统链路预测算法相比,LDNC 算法预测准确性更高,具有较好的预测效果。同时和基于链接紧密度的链路预测算法 CNBIDE 算法相比,也取得了更优的预测效果,说明不同的邻居节点对节点连边的贡献度是不同的,LDNC 算法考虑了这种差异性,取得了更好的效果。另外,由于 Precision 评价指标的最终结果和 L 的取值有关,在图 2 中列出了 LDNC 算法和 CN,AA,RA,Jaccard 和 CNBIDE 算法的 Precision 值随 L 值变化的情况。其中 L 的取值范围是 [50,300],间隔是 50。结果表明,随着 L 值的增加,各个算法的 Precision 值总体呈下降趋势,但是 LDNC 算法总体上是最优的。

表 4　评价指标 AUC 的结果

Tab. 4　The results of AUC on 9 real-networks

网络	CN	AA	RA	Jaccard	CNBIDE	LDNC
FWMW	0.710 1	0.714 2	0.719 4	0.627 3	0.719 7	0.732 5
FWFW	0.608 3	0.610 9	0.616 1	0.529 7	0.624 1	0.642 4
Jazz	0.954 3	0.961 6	0.970 8	0.960 8	0.954 9	0.967 6
C. elegans	0.846 6	0.863 6	0.868 6	0.795 1	0.851 8	0.868 6
USAir	0.9371	0.9478	0.952 7	0.914 0	0.951 2	0.967 1
PB	0.942 8	0.944 8	0.945 7	0.911 6	0.9421	0.9436
NS	0.9913	0.9916	0.9917	0.9914	0.991 3	0.992 6
Yeast	0.914 0	0.914 8	0.913 0	0.913 2	0.914 2	0.914 9
Power	0.625 0	0.624 9	0.624 8	0.625 0	0.627 2	0.627 4

表 5　评价指标 Precision 的结果($L=100$)

Tab. 5　The result of precision on 9 real-networks ($L=100$)

网络	CN	AA	RA	Jaccard	CNBIDE	LDNC
FWMW	0.150 0	0.157 0	0.158 0	0.024 0	0.146 0	0.162 0
FWFW	0.084 0	0.087 0	0.076 0	0.010 0	0.086 0	0.096 0
Jazz	0.822 0	0.831 0	0.822 0	0.728 0	0.856 0	0.899 0
C. elegans	0.138 0	0.138 0	0.130 0	0.014 0	0.132 0	0.149 0
USAir	0.601 0	0.611 0	0.646 0	0.012 0	0.591 0	0.650 0
PB	0.421 0	0.378 0	0.257 0	0.002 0	0.458 0	0.466 0
NS	0.866 0	0.976 0	0.978 0	0.566 0	0.833 0	0.984 0
Yeast	0.644 0	0.670 0	0.493 0	0.029 0	0.630 0	0.743 0
Power	0.098 0	0.079 0	0.001 0	0.164 0	0.188 0	0.129 0

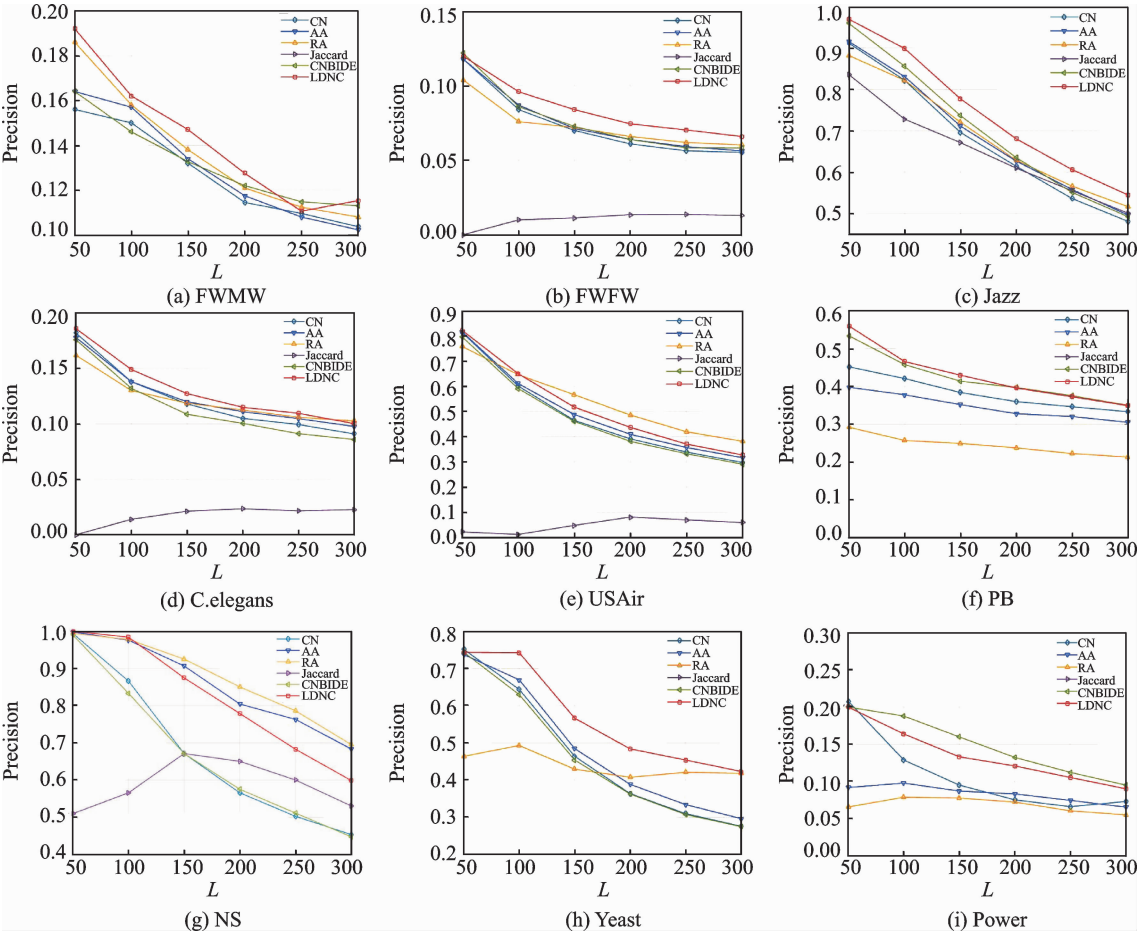

图 2 精确度 Precision 与 L 的关系

Fig. 2 Relationship between Precision and L

5 结束语

本文提出了一种新的链路预测算法，为一种结合共同邻居贡献度的节点相似性链路预测算法。该算法既考虑了共同邻居节点的差异性，用节点耦合聚簇系数来定义每个节点对连边产生的贡献度，又考虑了共同邻居之间的连边情况，用连边的紧密程度来衡量待预测节点之间的相似程度。与 4 种传统的预测算法和基于节点紧密度的算法 CNBIDE 在 9 组真实网络数据集上进行实验，结果表明，LDNC 算法具有更高的预测准确度。今后的研究主要从两方面进行。首先，可以考虑更多的节点边信息，不仅要考虑连边数量，也要考虑路径信息。其次，社团也是网络的一个重要结构，如果两节点在同一社团内，它们之间的相似性程度则相对较高，则它们之间的连边概率则相对较大，因此如何将社团信息与其他的网络信息相结合，也是未来的一个研究方向。

参考文献：

[1] Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A Statistical Mechanics & Its Applications, 2011, 390(6): 1150-1170.

[2] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661.

Lü Linyuan. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5):651-661.

[3] Getoor L, Diehl C P. Link mining: A survey[J]. ACM Sigkdd Explorations Newsletter, 2005, 7(2):3-12.

[4] 刘宏鲲, 吕琳媛, 周涛. 利用链路预测推断网络演化机制[J]. 中国科学:物理学 力学 天文学, 2011,41:816-823.
Liu Hongkun, Lü Linyuan, Zhou Tao. Uncovering the network evolution mechanism by link prediction[J]. Sci Sin Phys Mech Astron, 2011, 41: 816-823.

[5] Aiello L M, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media[J]. ACM Transactions on the Web, 2012, 6(2):1-33.

[6] 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.

[7] Stumpf M P H, Thorne T, Silva E D, et al. Estimating the size of human interactome[J]. Proceedings of the National Academy of Sciences, 2008, 105(19):6959-6964.

[8] Gu Q, Zhou J, Ding C H Q. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs[C]// SIAM International Conference on Data Mining. Columbus, Ohio, USA:[s. n.],2010:199-210.

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

[10] Zhu J, Hong J, Hughes J G. Using markov chains for link prediction in adaptive web sites[J]. Lecture Notes in Computer Science, 2004, 2311:60-73.

[11] Lü Linyuan. Research status and prospect of link prediction[EB/OL]. http://blog.sciencenet.cn/home.php?mod=space&uid=329471&do=blog&id=318268,2010-04-30.

[12] Lorrain F, White H C. Structural equivalence of individuals in social networks[J]. Journal of Mathematical Sociology, 1971, 1(1):49-80.

[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] O'Madadhain J, Hutchins J, Smyth P. Prediction and ranking algorithms for event-based network data[J]. ACM Sigkdd Explorations Newsletter, 2005, 7 (2):23-30.

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

[16] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4):623-630.

[17] Lü L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80(2):593-598.

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

[19] Klein D J, Randic M. Resistance distance[J]. Journal of Mathematical Chemistry, 1993, 12(1):81-95.

[20] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Transactions on Knowledge & Data Engineering, 2007, 19(3):355-369.

[21] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks, 2012, 56(18): 3825-3833.

[22] Liu W, Lu L. Link prediction based on local random walk[J]. EPL, 2010, 89(5):58007-58012.

[23] Liu S, Ji X, Liu C, et al. Extended resource allocation index for link prediction of complex network[J]. Physica A Statistical Mechanics & Its Applications, 2017, 479:174-183.

[24] Deylami H A, Asadpour M. Link prediction in social networks using hierarchical community detection[C]//Information and Knowledge Technology (IKT), 2015 7th Conference on.[S.l.]:IEEE, 2015: 1-5.

[25] 李淑玲. 基于相似性的链接预测方法研究[D]. 哈尔滨:哈尔滨工程大学,2012.
Li Shuling. Research on link prediction methods based on the similarity[D]. Harbin:Harbin Engineering University,2012.

[26] Li F, He J, Huang G, et al. Node-coupling clustering approaches for link prediction[J]. Knowledge-Based Systems, 2015, 89(C):669-680.

[27] 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.

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

[29] 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 & Shelf Science, 1998, 47(3):329-349.

[30] Ulanowicz R E, Bondavalli C, Egnotovich M S. Network analysis of trophic dynamics in south Florida ecosystem, fy 97: The florida bay ecosystem[EB/OL]. http://www.cbl.umces.edu/~atlss/FBay701.html, 1998.

[31] Gleiser P M, Danon L. Community structure in jazz[J]. Advances in Complex Systems, 2003, 6(4): 565-573.

[32] Watts D J, Strogatz S H. Collectivedynamics of 'small-world' networks[J]. Nature, 1998:440-442.

[33] Batagelj V, Mrvar A. Pajek-program for large network analysis[J]. Connections, 1998, 21(2):47-57.

[34] Adamic L A, Glance N. The political blogosphere and the 2004 US election: Divided they blog[C]//Proceedings of the 3rd International Workshop on Link Discovery. New York: ACM, 2005: 36-43.

[35] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006, 74(3): 036104.

[36] Von M C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions[J]. Nature, 2002, 417(6887):399-403.

作者简介:

王鑫(1992-),男,硕士研究生,研究方向:机器学习、链路预测等,E-mail:wangxintodd@163.com。

陈喜(1978-),男,讲师,研究方向:智能计算、链路预测等。

钱付兰(1978-),女,讲师,研究方向:商空间、社交网络和推荐系统等。

张燕平(1962-),女,教授,博士生导师,研究方向:计算智能与商空间理论、机器学习及应用、人工神经网络与智能信息处理等。

(编辑:刘彦东)