doi: 10.12068/j.issn.1005-3026.2020.01.006

时态网络节点相似性度量及链路预测算法

陈东明, 袁泽枝, 黄新宇, 王冬琦
(东北大学 软件学院, 辽宁 沈阳 110169)

摘 要: 详细分析和阐述了时态网络中的链路预测问题,将时态网络按时间顺序划分为具有相同时间间隔的多层网络快照序列.针对基于共同邻居的相似性指标对网络链路刻画粒度较粗糙的问题,提出了基于邻居节点聚类系数的相似性度量指标NCC和NCCP,并基于此提出时态网络链路预测算法.通过在真实数据集上的对比实验验证了利用邻居节点的聚类信息可以提高预测精度.利用真实邮件数据集验证了所提出的链路预测算法预测效果的优越性,并且实验结果证明越接近预测时间的网络结构对预测结果影响越大.

关 键 词: 时态网络;链路预测;多层网络;聚类;相似性
中图分类号: TP 391 **文献标志码**: A **文章编号**: 1005-3026(2020)01-0029-07

Node Similarity Measurement and Link Prediction Algorithm in Temporal Networks

CHEN Dong-ming, *YUAN Ze-zhi*, *HUANG Xin-yu*, *WANG Dong-qi*
(School of Software, Northeastern University, Shenyang 110169, China. Corresponding author: WANG Dong-qi, E-mail: wangdq@swc.neu.edu.cn)

Abstract: Link prediction in temporal networks was analyzed and discussed in detail. The temporal network was divided into multilayer network snapshot sequences with the same time interval in chronological order. Aiming at solving the problem of rough granularity obtained by the common-neighbor-based similarity index, similarity indexes NCC and NCCP based on neighbor node clustering coefficient were proposed. Then a link prediction algorithm for temporal networks was designed for networks based on these two indicators. The comparison experiments on real datasets showed that the cluster information of neighbor nodes can improve the prediction accuracy. The superiority of the proposed link prediction algorithm was verified by a real mail dataset, and the experimental results showed that the closer the network structure is to the prediction time, the greater the impact on the prediction results.

Key words: temporal networks; link prediction; multilayer network; clustering; similarity

网络科学的研究不仅是在宏观上挖掘不同复杂网络之间的共性以及它们所遵循的普适性规律,从中观层面对网络群组结构和层次结构进行研究,而且在微观层面也提出节点的度及其度分布、最短距离等来表示网络的测度.然而根据已有的信息构建网络模型时,所得到的观测数据并不一定真实有效,或部分缺失、或掺杂错误数据等,有时还会因时间因素导致不能够获得潜在的网络信息,在仿真实验时得不到准确的数据和理想的研究结论.因此,链路预测成为网络信息挖掘的一个研究热点[1].

链路预测(link prediction)是通过对已知的网络拓扑结构进行分析,构建预测算法以发现网络中尚未存在连边的节点对之间产生连边的概率[2],本质上是从网络链路的微观层面解释网络结构形成的原因.链路预测解决的是网络中缺失信息的还原与预测问题.所谓还原,指的是对网络中实际存在的但尚未被探测到的链路的发现,这

收稿日期: 2019-05-29
基金项目: 辽宁省自然科学基金资助项目(20170540320); 辽宁省博士启动基金资助项目(20170520358); 中央高校基本科研业务费专项资金资助项目(N172415005-2).
作者简介: 陈东明(1968-),男,安徽怀宁人,东北大学教授.

种链路也被称为未知链接(unknown links);所谓预测,指的是对网络中目前不存在但是未来很可能存在的链路的预测,这种链路也被称为未来链接(future links).

复杂网络中的链路预测算法主要是基于网络静态图,即网络规模以及节点间相互作用不变,然后分析其拓扑结构,推断网络的真实情况. 然而现实网络是动态变化的,网络结构随时间推移不断变化,时态网络(temporal networks)[3]中节点间产生连边的时间信息对预测将来产生新链接的概率有重要的意义. Tang[4]将时态网络切片,提出了时态距离、可达性等概念研究时态网络. Paranjape 等[5]定义 δ – motif 小子图作为分析时态网络的工具,大大提高了算法的时间效率. Lei 等[6]提出了一种非线性模型(GCN – GAN)来解决加权的时态网络链路预测问题. 该模型利用 GCN 计算各个时间切片的局部特征,然后将计算结果输入到 LSTM 模型,刻画网络的动态变化情况,再利用 GAN 生成预测结果. 由于该方法的训练过程较为复杂,因此还需要进一步优化以实现大规模网络上的链路预测. Yasami 等[7]利用随机多层网络模型解决网络中的链路缺失和未来链路的预测问题,并在仿真数据集和真实的 DBLP 数据集中表现出众. 然而,算法仅限于有向网络,其他类型的网络中缺少通用性. 此外,一些统计学方法,如整合移动自回归模型,即 ARIMA[8]也可以用于时态网络链路预测研究,但受限于网络数据的平稳性检验,因此对于网络结构随时间变化较大的预测效果并不是十分理想. 因此,时态网络链路预测问题还有很大的研究空间.

1 问题描述

定义一个无权网络,$G=\{V,E,T\}$,$V=\{v_1,v_2,\cdots,v_n\}$是网络中所有节点的集合,T 是网络的时间跨度. 假设节点产生连边没有时间的延迟,即事件是在一瞬间发生的,那么可以用 $e_{xy}^t=(v_x,v_y,t)$ 表示在第 t 时刻节点 v_x 和节点 v_y 之间产生了连边,则 $E=\{e^1,e^2,\cdots,e^t\mid t=1,2,3,\cdots\}$ 是所有时态网络中的连边的集合. 将网络按照时间间隔切分成 m 个时间切片,设每个时间窗口长度为 L,则这一系列的网络状态可以描述为

$$G=\{G_t\mid t=1,2,\cdots,m\}. \tag{1}$$

其中,G_t 表示网络在 t 时刻的拓扑结构.

时态网络中的链路预测问题可以描述为:已知 0 ~ T 时刻的网络拓扑结构变化情况,预测第 $T+1$ 时刻的网络中节点的连边情况,简单来说,就是根据网络历史信息预测下一时刻网络中的连边情况. 以数学形式表示为

$$G=\{G_t\mid t=0,1,2,\cdots,T\}\Rightarrow G_{T+1}. \tag{2}$$

2 算法设计与分析

2.1 算法提出

链路的存在与当前网络的拓扑结构有着密不可分的关系,如果在每一层的网络时间快照中计算节点对的相似性值,则会得到对应的一系列按时间顺序排列的节点对之间的相似性值,记为

$$S=\{S_t(v_x,v_y)\mid t=1,2,\cdots,m\}. \tag{3}$$

其中,$S_t(v_x,v_y)$ 表示在第 t 时刻的网络结构中节点对 (v_x,v_y) 的相似性值. 在第 t 时刻,网络可以视为静态网络,则可以利用静态网络中的相似性指标计算相似性.

基于共同邻居的 JC 指标[9]、AA 指标[10]等在实验中都有很好的表现,而且算法复杂度低且适用于大型网络. 尽管表现良好,但是由于所使用的网络信息有限,因此预测准确度不够理想. 该方法的另一个劣势在于,链路预测指标对于连边的刻画粒度比较粗糙. 当网络中子图结构相似时,只利用共同邻居这一信息会忽略邻居节点间的连边关系这一重要信息,具有不同结构的节点对之间的相似性指标值区分度不大. 图 1 所示为具有相同共同邻居信息的网络 A 和 B,有两对未连接的种子节点 (A,a) 和 (B,b) 都只有 3 个共同邻居节点.

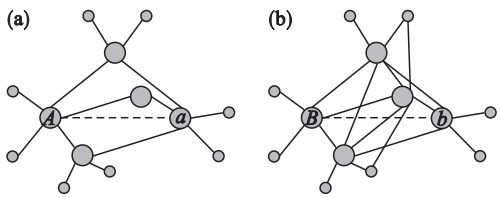

图 1 具有相同共同邻居信息的网络示意图

Fig. 1 Two networks with the same common neighbor information

(a)—网络 A;(b)—网络 B.

分别计算图 1 所示网络中两个种子节点的连接概率,如果以 Jaccard 指标值表示 Score 分数值,那么在网络 A 中 $S(v_A,v_a)=3/7$,在网络 B 中 $S(v_B,v_b)=3/7$. 显然,JC 指标赋予了它们相同的值,然而网络 B 中的两个种子节点显然比网络 A 中的两个种子节点有更紧密的关系,局部相似性更高. 因此可知,所利用的局部信息不能轻易地区分这两对节点,也不足以完整地表示节点之间的

相似性.

为了更加完备地利用网络的结构信息,本文利用邻居节点的局部聚类信息来表达链路的结构信息,这样的优势在于可以表达与目标链路具有相同结构的一些其他重要链路的结构信息. 如何利用节点的聚集特性来描述种子节点之间产生连边的可能性,提出以下两种假设思路:

1)假设种子节点间产生链接的概率(或分值)等于节点间的相似性值,而相似性可以用种子节点的共同邻居节点的聚类性表示.

2)如果种子节点间产生链接的概率(或分值)等于节点间的相似性值,假设邻居节点的聚类性可以增强种子节点间原有的相似性.

在网络中,连边关系的局部聚集特性表现形式为所有的连边都比较紧密,聚集成一个簇,也可以称之为社区结构,而节点的局部聚集特性可以由聚类系数(clustering coefficient)[11]来表示. 用数学公式表达,其定义如下:

$$C_i = \frac{2T_i}{k_i(k_i-1)}. \tag{4}$$

其中:C_i 为节点的聚类系数;T_i 是 $\{e_{jk}: v_j, v_k \in L(i), e_{jk} \in E\}$ 中边的数目,节点 v_j, v_k 是节点 v_i 的邻居节点,$L(i)$ 是节点 v_i 的邻居节点集合,e_{jk} 是节点 v_i 的邻居节点之间的连边;k_i 表示节点 v_i 的度. 节点的聚类系数反映了节点的邻居节点之间相互连接的概率. C_i 取值在 0 与 1 之间,C_i 越接近 1,说明节点 v_i 的邻居们抱成一团,节点 v_i 的局部越紧密;C_i 越接近 0,说明节点 v_i 的邻居比较稀疏,整个结构接近树状.

定义 1(NCC 相似性指标) 用 S_{xy}^{NCC} 表示节点为 v_x 和 v_y 的 NCC(node clustering coefficient)相似性指标,定义为

$$S_{xy}^{\mathrm{NCC}} = \sum_z \frac{\sum_{j\in\Gamma(z)} \boldsymbol{a}_j}{k_z(k_z-1)/2}. \tag{5}$$

其中:$\Gamma(z)$ 表示节点 v_x 与 v_y 的共同邻居,$z \in \Gamma(x)\cap\Gamma(y)$;$k_z$ 是节点 v_z 的度,节点 v_j 是节点 v_z 的邻居节点,其邻接矩阵记为 $\boldsymbol{a}_j$. S_{xy}^{NCC} 的值越大,说明两节点越相似,则它们之间越有可能产生连边.

定义 2(NCCP 相似性指标) 用 S_{xy}^{NCCP} 表示节点为 v_x 和 v_y 的 NCCP(node clustering coefficient plus)相似性指标,定义为

$$S_{xy}^{\mathrm{NCCP}} = |\boldsymbol{\Theta}_{\mathrm{cn}}| \cdot \left(1 + \sum_z \frac{\sum_{j\in N_z} \boldsymbol{a}_j}{k_z(k_z-1)}\right). \tag{6}$$

其中,$|\boldsymbol{\Theta}_{\mathrm{cn}}|$ 表示对 CN 指标归一化处理,保证每个相似值平等对待,避免相似值过大,与事实相悖. 由 CN 指标计算公式 $S_{xy} = |\Gamma(x)\cap\Gamma(y)|$,其中 $\Gamma(x)$ 是节点 v_x 的邻居节点的集合,$z \in \Gamma(x)\cap\Gamma(y)$. 由此可知,CN 指标的相似性分值 $S\in[0,\infty)$,需要采用函数将 $S\in\mathbf{R}$($\mathbf{R}$ 为实数)映射到[0, 1],本文采用具有较好归一化效果的 Logistic 函数. 结合式(5)和式(6),可以得到适用于时态网络的节点相似性计算公式:

$$S_t^{\mathrm{NCC}}(v_x, v_y) = \left\{ \sum_z \frac{\sum_{j\in\Gamma(z)} a_j}{k_z(k_z-1)/2} \middle| t \right\}, \tag{7}$$

$$S_t^{\mathrm{NCCP}}(v_x, v_y) = \left\{ |\boldsymbol{\Theta}_{\mathrm{cn}}| \cdot \sum_z \frac{\sum_{j\in N_z} a_j}{k_z(k_z-1)} \middle| t \right\}. \tag{8}$$

其中:z 表示节点对的邻居节点,$z\in\Gamma(x)\cap\Gamma(y)$;$t$ 表示时刻,$t=1$ 表示第一个时刻,$t=1,2,\cdots,m$;$S_t^{\mathrm{NCC}}(v_x, v_y)$ 表示在第 t 时刻的网络结构中利用 NCC 指标计算节点对 (v_x, v_y) 的相似性值;$S_t^{\mathrm{NCCP}}(v_x, v_y)$ 表示在第 t 时刻的网络结构中利用 NCCP 指标计算节点对 (v_x, v_y) 的相似性值.

本文将网络时间快照计算得到的相似性值序列 $S=\{S_t(v_x, v_y)\mid t=1,2,\cdots,m\}$ 看作是一组动态数列,为了使模型简单易于计算,降低算法的计算复杂度,采用线性回归(linear regression, LR)预测模型[12]作为基本的回归预测模型,该模型计算效率高、性能良好,且模型的预测范围较小,预测值在[0,1]之间. 在文献[13]中作者将科学家合作网络划分成网络时间序列,然后利用有监督和无监督两种方法进行链路预测实验,结果表明在同等指标下,无监督预测的线性回归模型(LR)表现较好.

将前 $m-1$ 层网络时间快照中节点对之间的相似性序列 $S=\{S_t(v_x, v_y)\mid t=1,2,\cdots,m-1\}$ 作为训练集数据,$S_m(v_x, v_y)$ 作为预测目标,建立与时间 t 之间的函数. 假设 $\hat{S}_t(v_x, v_y)$ 是第 t 时刻模型计算的节点对 (v_x, v_y) 之间的相似性值,那么回归方程表示为

$$\hat{S}_t(v_x, v_y) = a + b\cdot t, \tag{9}$$

结合式(7)和式(8),利用一元线性回归分析法可以计算最佳参数 a 和 b 的值,最终得到在第 t 时刻的相似性计算公式为

$$S_t^{\mathrm{NCC}}(v_x, v_y) = a_{\mathrm{NCC}} + b_{\mathrm{NCC}}\cdot t, \tag{10}$$

$$S_t^{\mathrm{NCCP}}(v_x, v_y) = a_{\mathrm{NCC+}} + b_{\mathrm{NCC+}}\cdot t. \tag{11}$$

2.2　算法过程描述

本算法的核心思想是:将网络划分为多个时间快照,然后利用所有快照的历史信息来预测未来时刻的网络状态. 时态网络链路预测算法过

程为

输入:时态网络 $G(V,E,T)$

输出:算法评价指标值

1)读取网络 $G(V,E,T)$,获取网络中所有可能出现的边.

2)将时态网络划分为 m 个单层网络,构建一系列网络时间快照$\{G_t \mid t=1,2,\cdots,m\}$.

3)根据式(7)、式(8)分别计算前 $m-1$ 个时刻的节点对之间的相似性值,构建相似性值时间序列.

4)选择对比度量指标[NCC, NCCP, CN, JC, AA, RA],计算在其他相似性指标下的节点对的相似性值.

5)根据得到的前 $m-1$ 个时刻的分数值序列训练得到预测模型,采用式(10)、式(11)计算第 m 时刻节点间的相似性值.

6)计算衡量算法的多个评价指标,如 AUC(接受者操作特性曲线下方的面积)、精确度 P、召回率 R 和 $F1$ 指标等.

2.3 复杂度分析

CN 指标[14]的时间复杂度与节点的度有关,假设网络节点数为 N,整个网络的平均度为 k,则计算共同邻居的时间复杂度为$O(k)$,则 CN 算法的时间复杂度为 $O(N^2k)$. 基于共同邻居的 JC, AA,RA 算法[15]与 CN 算法有类似的计算过程,因此它们有相同的时间复杂度. 基于随机游走的 SimRank 算法[16]的时间复杂度为 $O(Nk^l)$,其中 l 是随机游走的步数. 本文所提出的两种相似性度量方法 NCC 和 NCCP 需要计算节点的聚类系数,进行链路预测过程的时间复杂度为 $O(N^2k)$. 以上算法的时间复杂度比较如表 1 所示.

表 1 经典算法的时间复杂度比较

Table 1 Time complexity comparison of the classic algorithms

指标公式	复杂度
$S_{xy}^{\mathrm{SR}}=C\dfrac{\sum_{v_z\in\Gamma(x)}\sum_{v'z\in\Gamma(y)}S_{zz'}^{\mathrm{SR}}}{k_xk_y}$,$C$ 为衰减参数	$O(Nk)$
$S_{xy}^{\mathrm{CN}}=\lvert\Gamma(x)\cap\Gamma(y)\rvert$	$O(N^2k)$
$S_{xy}^{\mathrm{JC}}=\dfrac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$	$O(N^2k)$
$S_{xy}^{\mathrm{AA}}=\sum_{z\in\lvert\Gamma(x)\cap\Gamma(y)\rvert}\dfrac{1}{\lg k_z}$	$O(N^2k)$
$S_{xy}^{\mathrm{RA}}=\sum_{z\in\lvert\Gamma(x)\cap\Gamma(y)\rvert}\dfrac{1}{k_z}$	$O(N^2k)$

由表 1 可以看出,SR 算法时间复杂度最低,是 $O(Nk)$,因此效率最高. 但由于其属于全局迭代算法,包含随机游走过程,因此实验结果并不稳定. 本文提出算法的时间复杂度与其他同类算法相同,都是 $O(N^2k)$,虽然略高于 SR 算法,但因为其不存在随机过程,保证了结果的稳定性,因此具有更好的适用性.

3 实验分析

3.1 相似性度量方法对比实验

本文选取不同领域的 6 个真实网络数据集:空手道俱乐部网络(Karate)、海豚社会关系网络(Dolphins)、911 恐怖袭击网络(911data)、美国政治书籍网络(Polbooks)、美国大学生足球俱乐部(Footballs)和科学家合作网络(Scientists). 6 个真实网络数据集的统计特征见表 2.

本文采用随机抽样法划分网络数据集,测试集的比例默认设定为 10% . 选择传统的 4 个链路预测方法作为对比算法,分别是基于共同邻居的 CN 指标、Jaccard 指标(JC)、AA 指标、RA 指标. 循环重复实验多次,采用评价指标 AUC 的平均值作为算法的评估结果,如表 3 所示.

表 2 网络的统计特征

Table 2 Statistic characteristics of networks

数据集	$\lvert V\rvert$	$\lvert E\rvert$	k	$\lvert C\rvert$	$<c>$	$\lvert D\rvert$
Karate	34	78	4.588	0.571	2.408	2.218
Dolphins	62	159	5.129	0.259	3.064	3.357
911data	69	159	4.609	0.470	1.758	3.215
Polbooks	105	411	8.400	0.488	5.260	3.079
Footballs	115	613	10.661	0.403	10.231	2.508
Scientists	1589	2742	4.597	0.638	0.079	5.989

注:$\lvert V\rvert$ 表示网络的节点数;$\lvert E\rvert$ 表示网络中的连边数;k 表示网络的平均度;$\lvert C\rvert$ 为网络的平均聚类系数;$<c>$ 表示网络的平均连通程度;$\lvert D\rvert$ 表示网络的平均最短路径的距离.

表 3 不同网络数据集的 AUC 值

Table 3 AUC for different network datasets

数据集	CN	JC	AA	RA	NCC	NCCP
Karate	0.689	0.599	**0.724**	**0.733**	0.693	0.705
Dolphins	0.762	0.762	**0.764**	**0.764**	0.763	**0.766**
911data	0.851	0.844	0.858	**0.865**	**0.864**	0.861
Polbooks	0.867	0.863	0.871	**0.872**	0.870	**0.874**
Footballs	**0.848**	**0.860**	**0.848**	**0.848**	0.847	**0.848**
Scientists	0.930	0.930	0.930	0.930	0.930	**0.931**

注:表中黑体加下划线标注的是最大值,黑体标注的是次大值.

由表 3 可知,对于 Karate,Dolphins 和 911data 这 3 个较小规模的网络,AA 和 RA 指标具有较高

的预测精确度，所提出的 NCC 和 NCCP 指标表现也比较优异. 在 Polbooks 网络数据集中，NCCP 和 RA 指标表现显著，且平均 AUC 值达到了 0.873. 在 Scientists 网络中 AUC 值达到了 0.931，预测精确度高，NCCP 指标预测效果最好. 实验结果表明，网络邻居节点的聚类系数可以提高预测精确度.

为了更加清晰地展现 NCC 和 NCCP 指标的性能，做了以下显著性检验：本文采用皮尔逊相关系数进行检验，首先将 NCC 和 NCCP 指标与 CN，JC，AA，RA 指标进行对比计算，分别得到相应指标的假设机率（p），然后把分别得到的 p 加和取平均得到 NCC 指标的 p 为 0.000 6，NCCP 指标的 p 为 0.000 8，均远小于 0.05. 所以本实验效果较显著.

为了验证 NCCP 相似性指标对 CN 指标的增强效果，本文利用 AUC 值的对比情况来刻画，实验得到如表 4 所示数据. 由后两列计算结果可以看出，在添加邻居节点的聚类系数后构建的 NCCP 指标比 CN 指标预测效果有了普遍的提高，说明 NCCP 有一定的增强效果.

表 4　NCCP 指标的增强效果

Table 4　Enhancement effect of NCCP index

数据集	NCC	NCCP	CN	α/%	β/%
Karate	0.693	0.705	0.689	2.322	1.732
Dolphins	0.751	0.766	0.762	0.525	1.997
911data	0.838	0.858	0.851	0.823	2.387
Polbooks	0.866	0.870	0.867	0.346	0.462
Footballs	0.843	0.846	0.848	−0.236	0.356
Scientists	0.874	0.930	0.930	0	6.407

注：α 表示 NCCP 比 CN 增强的部分与 CN 的百分比；β 表示 NCCP 比 NCC 增强的部分与 NCC 的百分比.

根据表 4 结果可知，NCC 指标与 NCCP 指标在整体上优于 CN 指标.

3.2　时态网络链路预测实验

为了验证时态网络的链路预测算法的效果，本文使用 Email－Eu－core temporal network 时态网络数据集[5]. 该网络是根据欧洲某大型研究机构内部成员的电子邮件往来关系所构建的，对所有接收和发出的邮件信息内容作匿名处理. 数据集不包含成员和其他地方地区的通信邮件，仅限于机构内部核心成员之间的通信，完整的数据集包含了来自 4 个部门成员之间的所有电子邮件，时间跨度为 802 天. 本文所使用的网络数据集来自第三个部门（Dept3），该数据集有 89 个节点、12 216条时态边，转化为静态网络则有 1 506 条边，时间跨度为 802 天. 节点代表机构内部的部门成员，每条连边代表他们之间有一次邮件往来，数据集中每条数据 (u,v,t) 表示在时间 t 从用户 u 向用户 v 发送了电子邮件.

利用精确度 P、召回率 R 和 $F1$ 指标对预测结果进行评价，如果按天划分时态网络，那么每日的邮件数量即代表每层网络的连边数目，得到如表 5 所示结果.

表 5　按日划分时态网络预测结果的评价指标值

Table 5　Evaluation indicators of temporal network prediction results by dividing days

评价指标	CN	JC	RA	AA	NCC	NCCP
P	0.060	0.056	0.061	0.066	0.071	0.060
R	0.458	0.417	0.5	0.458	0.542	0.458
$F1$	0.106	0.098	0.107	0.107	0.126	0.106

由表 5 可知，NCC 指标的预测结果明显优于其他指标，说明基于邻居节点聚类的度量指标可以提高链路预测精确度. 表 5 中的 P 值都普遍偏低，究其原因，一方面时态网络数据集本身较小，另一方面网络划分的层数过多. 按天划分时态网络导致每层的连边都特别稀疏，信息很零散，使所有指标的预测效果都偏低.

如果按月划分时态网络，得到如表 6 所示结果，NCC 指标预测结果仍然优于其他指标，显示其有效性；且 3 个评价指标值比按天划分的网络的评价指标值都有所提高，说明时态网络链路预测的精确度还与网络划分的层数有关系，当网络划分过细时，网络分辨率很高，则预测效果不理想.

表 6　按月划分时态网络预测结果的评价指标值

Table 6　Evaluation indicators of temporal network prediction results by dividing months

评价指标	CN	JC	RA	AA	NCC	NCCP
P	0.217	0.207	0.208	0.216	0.218	0.216
R	0.692	0.646	0.636	0.682	0.723	0.708
$F1$	0.331	0.314	0.313	0.328	0.335	0.331

选取目前公认的比较好的 RA 指标结合 ARIMA 模型作为基线算法，与本文提出的 NCC 指标结合线性回归模型进行比较，使用按月划分的网络数据集，随着层数的增加得到的 AUC 结果如图 2 所示.

由图 2 可看出，除了层数为 4 和 13 的时候本文提出的方法低于 ARIMA_RA 方法，其他情况均好于 ARIMA_RA 方法.

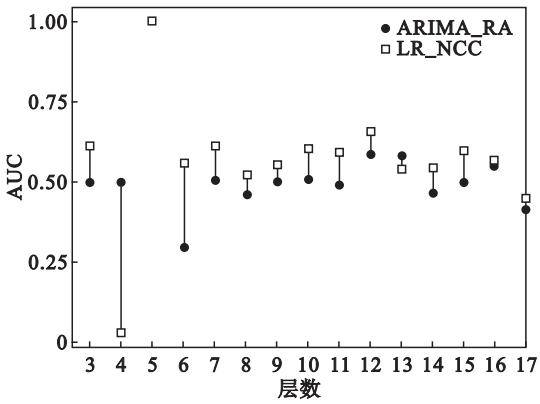

图 2 本文提出的方法与基线方法的对比
Fig. 2 Comparison of the proposed method with the baseline method

综合以上实验结果,本文所提出的 NCC 指标在对时态网络的链路进行预测时优于其他相似性指标,说明在考虑邻居节点的聚集情况后,更贴近真实网络中人们交流协作的过程,因此预测效果更好. 并且,从以上两种划分形式可以看出,时态网络的预测效率还与时间的划分和时态网络模型有密切关系.

将网络划分为 m 个时间快照,然后利用所有快照的历史信息来预测未来时刻的网络状态是本文方法的研究思路. 然而,在现实网络中,不同时刻网络状态对预测未来时刻网络状态所贡献的重要性是不同的,例如信息传播过程中,最近时间节点的信息最重要,历史时间中的过时信息的影响力不大. 为了在本次时态网络链路预测中验证该思想,选择预测时刻(即第 m 层)的前 n 层网络信息进行预测. 实验中 n 取 1,2,4,7,9,10,得到如表 7,表 8 所示的预测结果.

表 7 预测结果精确度评价指标值
Table 7 Evaluation indicators of precision of prediction results

层数	NCC	NCCP	CN	JC	AA	RA
1	0.448	0.448	0.448	0.448	0.448	0.448
2	0.358	0.345	0.363	0.355	0.374	0.368
4	0.366	0.364	0.364	0.358	0.369	0.357
7	0.300	0.302	0.301	0.286	0.299	0.290
9	0.290	0.284	0.286	0.277	0.286	0.290
10	0.274	0.269	0.272	0.264	0.268	0.268

由表 7 和表 8 中数据可知,$n=1$ 时所有预测算法的精确度都相同;当 $n=2$ 时,两个评价指标在所有算法中的值都开始减小,此时 AA 指标预测效果最好;当 n 逐渐增大时,两个评价指标都普遍呈减小的趋势,这说明越贴近预测时间的网络结构对最终结果的影响越大. 而且,在 n 增大的过程中,本文所提出的 NCC 和 NCCP 指标对时态网络的链路预测效果逐渐开始显示其优越性.

表 8 预测结果 $F1$ 值评价指标值
Table 8 Evaluation indicators of $F1$-value results

层数	NCC	NCCP	CN	JC	AA	RA
1	0.485	0.485	0.485	0.485	0.485	0.485
2	0.425	0.406	0.440	0.428	0.450	0.439
4	0.448	0.449	0.449	0.440	0.452	0.436
7	0.404	0.409	0.409	0.387	0.405	0.390
9	0.403	0.396	0.401	0.381	0.398	0.398
10	0.389	0.378	0.388	0.374	0.377	0.374

4 结 语

本文将时态网络划分为一系列时间快照序列,利用所提出的度量指标计算每一层网络中的节点对相似性,构建节点对相似性时间序列,然后,结合时间序列回归模型预测节点对未来的相似性. 实验结果表明,利用邻居节点的聚类信息可以提高预测精度,利用真实邮件网络数据集验证了所提出的指标的预测效果优越性,并且实验结果证明越接近预测时间的网络结构对预测结果影响越大.

参考文献:

[1] Das K, Sinha S K. Identification and analysis of future user interactions using some link prediction methods in social networks [C]//Data, Engineering and Applications. Singapore: Springer, 2019: 83 – 94.

[2] Liu Z, Zhang Q M, Lyu L Y, et al. Link prediction in complex networks: a local naïve Bayes model [J]. *Europhysics Letters*, 2011, 96(4): 48007.

[3] Holme P. Modern temporal network theory: a colloquium [J]. *European Physical Journal B*, 2015, 88(9): 1 – 30.

[4] Tang J K. Temporal network metrics and their application to real world networks [D]. Cambridge: University of Cambridge, 2012.

[5] Paranjape A, Benson A R, Leskovec J. Motifs in temporal networks [C]//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. New York, 2017: 601 – 610.

[6] Lei K, Qin M, Bai B, et al. GCN-GAN: a non-linear temporal link prediction model for weighted dynamic networks [C]// IEEE INFOCOM 2019—IEEE Conference on Computer Communications. Paris, 2019: 388 – 396.

[7] Yasami Y, Safaei F. A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks [J]. *Physica A: Statistical Mechanics and Its Applications*, 2018, 492: 2166 – 2197.

(下转第 43 页)