

基于神经网络的链路预测算法

潘永昊, 于洪涛, 刘树新

(国家数字交换系统工程技术研究中心, 河南 郑州 450002)

摘要: 针对当前基于网络拓扑结构相似性的链路预测算法普遍存在精确度较低且适应性不强的问题, 研究发现融合算法能够有效改善这些问题。提出了一种基于神经网络的融合链路预测算法, 主要通过神经网络对不同链路预测相似性指标进行融合。该算法使用神经网络对不同相似性指标的数值特征进行学习, 同时采用标准粒子群算法对神经网络进行了优化, 并通过优化学习后的神经网络模型计算出融合指标。多个真实网络数据集上实验表明, 该算法的预测精度明显高于融合之前的各项指标, 并且优于现有融合方法的精度。

关键词: 复杂网络; 链路预测; 神经网络; BP 算法

中图分类号: TN94; TP3; TP391

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2018049

Neural network-based link prediction algorithm

PAN Yonghao, YU Hongtao, LIU Shuxin

National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China

Abstract: To improve the difference existed in the link prediction accuracy and adaptability of different topology structure similarity based methods, a neural network-based link prediction algorithm, which fused similarity indices by neural network was proposed. The algorithm uses neural network to study the numerical characteristics of different similarity indices, and uses particle swarm optimization to optimize the neural network, and calculates the fusion index by the optimized neural network model. The experiment on the real network data set shows that the prediction accuracy of the algorithm is obviously higher than that before the fusion, and the accuracy is better than the existing methods.

Key words: complex network, link prediction, neural network, back propagation algorithm

1 引言

链路预测是复杂网络研究的一个重要部分, 是指通过网络中已知的节点以及结构信息, 预测网络中尚未产生连接的 2 个节点之间产生连接的可能性^[1]。这种预测包括对网络中已有但未观测到连接的预测, 以及网络中未来产生连接的预测。

链路预测的概念一经提出, 就获得了众多领域的广泛关注, 在很多领域得到了实际应用, 如指导生物蛋白质网络构建^[2]、分析社会网络结构^[3]、解决推荐系统中数据稀疏问题^[4]等。与此同时, 链路预测还具有重要的理论研究意义, 如网络结构与演化的研究^[5-6]等。经过了多年的研究和探索, 链路预测的算法逐步形成基于相似性的方法和基于似然估计的方法^[1]。在链路预测的

收稿日期: 2018-05-07; 修回日期: 2018-06-01

通信作者: 潘永昊, panyonghao2016@163.com

基金项目: 国家自然科学基金创新研究群体基金资助项目 (No.61521003), 国家自然科学基金资助项目 (No.61601513)

Foundation Items: The Innovative Research Groups of the National Natural Science Foundation of China (No.61521003), The National Natural Science Foundation of China (No.61601513)

研究过程中, 所研究的网络模型, 也从最初无权无向网络, 扩展到了含权网络^[7]、有向网络^[8]、异质网络^[9], 从静态模型发展到了时序模型^[10]。但不论模型如何复杂, 静态还是时序, 都是以无权无向网络为出发点, 其算法也是在无权无向网络算法思想的基础上不断改进。在无权无向网络的算法中, 也存在一系列矛盾: 1) 基于似然估计的方法最大的问题是计算复杂度高, 不适合在规模较大的网络中应用; 2) 基于相似性的方法具有运算复杂度低的特点, 但其运算效果会受到网络结构的影响, 在不同结构特点的网络中, 运算效果不够稳定, 不具有顽健性。

针对以上问题, 文献[1]中提出了基于不同的相似性算法进行融合的方法, 用以提高链路预测算法的稳定性和顽健性, 同时这种融合方法可以兼顾基于相似性的链路预测方法运算复杂度低的优点。本文借鉴这种融合方法^[1]的思想, 提出一种基于神经网络的链路预测算法 (NNLP, neural network-based link prediction algorithm), 对链路预测相似性指标融合问题进行研究。主要贡献如下: 1) 使用不同的链路预测相似性指标, 通过训练神经网络, 实现对不同相似性指标的非线性计算, 得出融合指标; 2) 在真实的数据集中进行测试, 并与基准方法进行比较, 在不同的真实网络数据集上都得到了理想的效果, 证明本文方法能够有效提高链路预测的准确性和顽健性。

2 相关工作

文献[1]中介绍了几种基于相似性的链路预测方法, 包括基于局部相似性的方法, 如 common neighbors (CN)^[11]、adamic-adar (AA)^[12]、resource allocation index (RA)^[13]、preferential attachment Index (PA)^[14]、extended resource allocation index^[15], 以及基于全局路径相似性的方法, 最常用的是 Katz 指标^[16]。Liu 等^[17]在研究中通过增加节点对之间包含的局部拓扑信息, 提出了基于节点间拓扑加权的相似性链路预测方法。文献[18-19]通过实验对比了基于相似性的链路预测指标性能, 在不同的实际网络中, 不同的结构相似性指标表现各有优劣, 性能差异较大, 顽健性较差。文献[20]研究发现一些算法输出的结果中部分正确结果具

有互补性。同时在文献[1]在展望中提出可以通过合适的方法, 对不同的相似性指标进行融合, 使融合以后的指标能够适应不同的实际网络, 表现出更高的精确度和更好的顽健性。

在融合算法方面, 研究者开展过一些相关研究, 但是总体来说相对较少。文献[18]中采用 3 种不同的 OWA (ordered weighted averaging) 算子对 9 种基于局部信息的结构相似性指标进行了融合。文献[20]非常巧妙地把链路预测问题定义为了二分类问题, 并提出一种基于 Adaboost 的链路预测算法, 借用 Boosting 方法通过错误反馈提升弱学习算法得到强学习算法的思想, 并利用算法互补性原则选取若干链路预测算法作为弱分类器, 基于 AdaBoost 算法提出并实现了一个新的算法。文献[21]则提出一种基于 choquet 模糊积分的链路预测算法, 引入数学中模糊测度和模糊积分的概念, 在考虑不同指标的重要性和指标交互作用的基础上, 对指标进行融合。通过已有的相关研究可以证实, 融合结构相似性指标的方法的确能够提高链路预测的效果, 使融合后的算法能够克服各个单一指标只适用于与其结构特性接近网络的问题。

3 算法描述

针对任意无向网络 $G = (V, E)$, 其中, V 为节点的集合, E 为连边的集合。令 U 为 $\frac{N(N-1)}{2}$ 个节点对组成的全集 (N 为节点的个数)。给定一种链路预测的方法, 为每对没有连边的节点对 v_x 、 v_y 赋予一个分数值 s_{xy} 。这个分数值可以理解作为一种接近性, 它与两节点的链接概率正相关。将所有未连接节点对按照该分数值从大到小排序, 排在最前面的节点对相互连接的概率最大。不同的链路预测算代表对网络不同结构特征的提取方式, 在本文提出的 NNLP 算法中, 通过神经网络对不同的结构特征进行融合计算, 使融合计算结果中能够充分包含多种结构特征。

3.1 训练集说明

对网络中的所有节点, 两两组成一个节点对, 2 个节点之间有连边的称为有连边节点对, 没有连边的称为未连边节点对, 很显然, 未连边节点

对的数量远大于有连边节点对。从有连边节点对集合中, 随机抽取一定数量的节点对, 再从未连边节点对集合中抽取同样数量的节点对组成训练集 D , $D = \{(s_{x_1 y_1}, t_{x_1 y_1}), (s_{x_2 y_2}, t_{x_2 y_2}), \dots, (s_{x_n y_n}, t_{x_n y_n})\}$, 其中, 输入向量 $s_{x_i y_i} = \{s_{x_i y_i}^{(1)}, s_{x_i y_i}^{(2)}, \dots, s_{x_i y_i}^{(n)}\}$, 表示选定的用于融合计算的 d 个相似性指标对节点对 v_{x_i} 、 v_{y_i} 计算的相似性数值的归一化数值, 训练集的输出为标量 $t_{x_i y_i}$, 如果节点对 v_{x_i} 、 v_{y_i} 之间有连边, 则 $t_{x_i y_i} = 1$, 否则 $t_{x_i y_i} = 0$ 。

3.2 使用粒子群算法确定神经网络的初始参数

使用多种相似性指标对神经网络进行训练, 各个相似性指标之间具有很强的相关性, 直接用于训练神经网络很容易过拟合, 因此, 首先采用标准粒子群 (PSO, particle swarm optimization) 算法优化神经网络的初始阈值和权值。

粒子群算法源于对鸟类捕食行为的模拟^[23]。构造 I 个 N 维向量 $p = \{p^1, p^2, \dots, p^n, \dots, p^N\}$, 每个 N 维向量代表一个粒子, p^n 表示向量的第 n 个值。粒子以一定规则在解空间中飞行, 以寻找最优解。解的优劣由粒子群算法中的适应函数进行衡量, 粒子群算法的目标是使适应度函数值最大, 此时对应的解最优。粒子根据式(1)更新自身速度和位置。

$$\begin{aligned} v_i^{j+1} &= w_i v_i^j + c_1 r_1 (pbest_i^j - p_i^j) + c_2 r_2 (gbest^j - p_i^j) \\ p_i^{j+1} &= p_i^j + v_i^{j+1} \end{aligned} \quad (1)$$

其中, v_i^j 和 p_i^j 分别为第 i 个粒子在第 j 步时的速度和位置, $pbest_i^j$ 表示第 i 个粒子搜寻到的历史最优解, $gbest^j$ 表示整个群体内所有粒子在第 j 步搜索到的最优解, 粒子根据这 2 种最优解进行位置和速度更新。 w 是惯性因子, 能够平衡算法的全局搜索能力和局部搜索能力。文献[23]指出, 惯性权重应在迭代过程中逐渐减少, 这样可以使 PSO 算法在初始阶段有较强的全局搜索能力, 而在后续阶段有较强的局部搜索能力, 本文采用文献[23]给出的建议值, 即设 w 的最大值为 0.9, 最小值为 0.4。 c_1 、 c_2 是加速常量, 分别表示粒子向自己的历史经验和种群中最优粒子学习的程度, 它们也起着平衡算法局部搜索能力和全局搜索能力的作用, 一般取值为 2。 r_1 、 r_2 是 (0,1) 之间的随机数。

采用粒子群算法优化神经网络初始参数时, 向量 p 表示神经网络的所有参数, 即神经网络的连接权值和阈值分别用向量中的一个分量表示。定义神经网络, 训练集 D 训练神经网络, 把预测输出和期望输出之间的误差绝对值和作为适应度函数 F , 即

$$F = \sum_{i=0}^n abs(t_i - o_i) \quad (2)$$

其中, n 为训练集中训练样例的个数, t_i 为第 i 个训练样例的期望输出, o_i 为第 i 个训练样例的预测输出。

迭代完成以后, 最优解 $gbest$ 即为优化以后的神经网络参数。本文中粒子搜索最优解的终止条件有两个: 一是达到最大迭代次数, 二是迭代次数满足限制条件以后, 粒子找到的最优解在迭代中没有大的变化。为了防止算法过早地陷入局部最优解, 本文中将粒子的取值空间平均分为 N 份, 然后在每份取值空间中随机地生成一定数量的粒子, 以增大初始化时种群的多样性。

3.3 训练神经网络

使用 3.2 节中通过粒子群算法优化以后的连接权值和阈值作为神经网络的初始参数, 使用误差逆传播算法对神经网络进行训练^[24]。采用 Sigmoid 函数作为神经元的激活函数 $f(\bullet)$, 神经网络由 d 个输入神经元、 q 个隐层神经元、1 个输出神经元组成, 隐层第 h 个神经元的阈值用 γ_h 表示, 输出层神经元用 θ 表示, 输入层第 i 个神经元到隐层第 h 个神经元的连接权值用 v_{ih} 表示, 隐层第 h 个神经元到输出层神经元的连接权值用 w_h 表示。

对于训练集 D 中的训练样例 (s_k, t_k) , 经过神经网络计算以后的实际输出为 $\hat{t}_k$, 即

$$\hat{t}_k = f(\beta - \theta) \quad (3)$$

则当前神经网络在训练样例 (s_k, t_k) 处的均方误差为

$$E_k = \frac{1}{2} (\hat{t}_k - t_k)^2 \quad (4)$$

通过 BP 迭代学习算法, 对神经网络中的所有参数, 即连接权值和阈值进行更新估计, 对于任意参数的更新估计表达式为

$$v = v + \Delta v \quad (5)$$

BP 算法基于梯度下降 (gradient descent) 策略, 以负梯度方向对参数进行调整, 以隐含层到输出层的连接权 w_h 为例推导出, 对于本轮迭代中的误差 E_k , 给定学习率 η , 有

$$\Delta w_h = -\eta \frac{\partial E_k}{\partial w_h} \quad (6)$$

其中, w_h 先影响到输出层神经元的输入值 β , 再影响到其输出值 $\hat{y}_k$, 然后影响到 E_k , 有

$$\frac{\partial E_k}{\partial w_h} = \frac{\partial E_k}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial \beta} \frac{\partial \beta}{\partial w_h} \quad (7)$$

由 β 的定义, 得到

$$\frac{\partial \beta}{\partial w_h} = b_h \quad (8)$$

由 Sigmoid 函数的性质得到

$$f'(x) = f(x)(1 - f(x)) \quad (9)$$

由式(3)和式(4)得到

$$g_j = -\frac{\partial E_k}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial \beta_j} = -(\hat{y}_j^k - y_j^k) f'(\beta_j - \theta_j) = \hat{y}_j^k (1 - \hat{y}_j^k) (y_j^k - \hat{y}_j^k) \quad (10)$$

将式(10)和式(8)代入式(7), 再代入式(6), 得出 BP 算法中对于 w_h 的更新公式。

$$\Delta w_h = \eta g b_h \quad (11)$$

由类似原理可以计算出

$$\Delta \theta = -\eta g \quad (12)$$

$$\Delta v_{ih} = \eta e_h s_i \quad (13)$$

$$\Delta \gamma_h = -\eta e_h \quad (14)$$

其中

$$e_h = -\frac{\partial E_k}{\partial b_h} \frac{\partial b_h}{\partial \alpha_h} = -\frac{\partial E_k}{\partial \beta} \frac{\partial \beta}{\partial \alpha_h} f'(\alpha_h - \gamma_h) = w_h g f'(\alpha_h - \gamma_h) = b_h (1 - b_h) w_h g \quad (15)$$

根据上述过程, 不断迭代更新连接权值和阈值, 最终目标是 minimized 训练集 D 上的累积误差。

$$E = \frac{1}{m} \sum_{k=1}^m E_k \quad (16)$$

最终确定神经网络的参数, 即连接权值和阈值。

3.4 计算融合指标

对于节点对 v_x 、 v_y , 选定的 d 个相似性指标

计算归一化数值为 $s_{xy}^{(1)}, s_{xy}^{(2)}, \dots, s_{xy}^{(n)}$, 把神经网络的输出记为融合以后的融合指标 s_{xy} 。即节点对 v_x 、 v_y 之间连边相似性融合指标为

$$\begin{aligned} s_{xy} &= f(\beta - \theta) = \\ &= f\left(\sum_{h=1}^q w_h b_h - \theta\right) = \\ &= f\left(\sum_{h=1}^q w_h f(\alpha_h - \gamma_h) - \theta\right) = \\ &= f\left(\sum_{h=1}^q w_h f\left(\sum_{i=1}^d v_{ih} s_{xy}^{(i)} - \gamma_h\right) - \theta\right) \end{aligned} \quad (17)$$

按照此方法, 计算出网络中未连边节点对的连边相似性融合指标的数值, 并按照这个数值, 从大到小进行排列, 排在最前面的节点对之间存在或未来产生连边的概率最大。整个计算过程如算法 1 所示。

算法 1

输入 网络 G , 不同的单一链路预测算法 $A = \{A_1, A_2, A_3, \dots\}$

输出 NNLP 算法计算的融合指标

1) 随机抽取连边集合 E 和未连边集合 $U - E$ 中 10% 的节点对构成训练集 D 。

2) 使用集合 A 中的算法计算节点对相似性指标。

3) 使用 POS 算法初始化神经网络参数 $\gamma_h, \theta, v_{ih}, w_h$ 。

4) 使用训练集 D 中的节点对在步骤 2) 中计算得到的相似性指标训练神经网络。

5) 使用步骤 4) 中训练的神经网络计算融合指标。

6) 返回计算得出的融合指标。

4 实验与分析

4.1 实验数据集

为了验证所提链路预测算法的有效性, 本文在 6 个真实网络数据集上进行链路预测对比实验, 分别如下。

1) 政治家博客网络 (PB), 原数据是有向网络, 节点为博客网页, 边表示网页之间的超链接。

2) 科学家合作网络 (NS), 网络中的节点是在网络科学领域发表过论文的科学家, 连边表示科学家的合作关系。

3) 爵士音乐家合作网 (Jazz), 网络中的节点为爵士音乐家, 连边表示音乐家的合作关系。

4) 线虫的神经网络 (CE), 节点表示线虫的神经元, 边表示神经元突触。

5) 美国航空网络 (USAir), 网络中的每个节点对应一个机场, 连边表示 2 个机场之间有直飞的航线。

6) 蛋白质相互作用网络 (Yeast), 网络中的每个节点表示一种蛋白质, 边表示蛋白质间的相互作用关系。

以上 6 个网络的基本结构参数如表 1 所示。其中, $|V|$ 表示网络中节点的个数, $|E|$ 表示边的数量, $\langle k \rangle$ 表示网络的平均度, $\langle d \rangle$ 表示网络平均距离, C 表示网络簇系数, r 表示结合系数, H 表示度的分布熵。

4.2 度量指标

在实验中选用文献[1]中给出的 AUC (area under the receiver operating characteristic curve) 指标作为链路预测算法精度的评价标准。在衡量链路预测算法预测精确度时, AUC 是指在网络中随机抽取存在连边的节点对的链路预测评分高于随机抽取不存在连边的节点对的链路预测评分的概率。在实验中, 去除真实网络中边的权值、方向以及自环。本节在保证训练集连通的前提下, 采用随机抽取的方式, 把每一个数据集的连边集合 E 中 90% 的连边作为训练集 E^T , 剩余 10% 作为测试集 E^P 。其算法实现为, 分别在测试集 E^P 和未连边节点对集合 $U - E$ 中随机选取一个节点对, 并对比其评分高低, 在 n 次相互独立的比较中, 测试集 E^P 中选取的节点对评分高的次数记为 n' , 评分相等的次数记为 n'' , 则 AUC 值为

$$AUC = \frac{n' + 0.5n''}{n} \quad (18)$$

精确度 (precision) 指标表示预测正确的连边在测试集中所占的比例。定义 L 为预测的连边数, m 为预测正确的连边数, 则 precision 为

$$precision = \frac{m}{L} \quad (19)$$

4.3 神经网络的参数优化与训练

选择要融合的相似性指标是一个基础, 决定了本算法的结果精确性。从这点出发, 应该考虑结果精确度高、重叠少的指标。为了能进一步用此算法分析挖掘网络结构和指标之间、指标和指标的关系, 取 CN、PA、AA、Katz 4 种含义较为明确的基准指标进行融合。使用神经网络对指标进行融合, 实质上就是把不同指标进行非线性叠加, 使其发挥各自的优势, 使融合后的算法具有更加广泛的适用性。

神经网络的优点在于它的算法具有学习能力, 能够在训练过程中通过每一轮迭代后的输出误差优化神经网络中的参数, 使神经网络对各个指标的融合效果达到最佳。但在实验中注意到, 把不同的相似性指标作为节点之间有无连边的分类属性, 本身就具有很大的相关性, 这就使在训练神经网络时很容易出现过拟合。在实验中, 对同一个网络进行多次计算, 得出的计算结果时好时坏, 偏差较大, 计算结果很不稳定。本文使用粒子群算法对神经网络初始参数进行优化, 防止过拟合现象出现。通过对神经网络的初始参数进行优化后, 计算结果具有很好的收敛性, 使算法在多次计算的结果能够稳定一致。

实验中发现, 数量过大的训练集并不能够有效提高实验算法的预测效果, 反而会降低算法的时间复杂度, 并且引起过拟合。通过大量实验总结认为, 在确定训练集的时候, 选择从

表 1 静态网络数据集拓扑特征参数

网络	$ V $	$ E $	$\langle k \rangle$	$\langle d \rangle$	C	r	H
NS	1461	2742	3.75	5.82	0.878	0.461	1.85
PB	1222	16717	27.36	2.51	0.360	-0.221	2.970
Jazz	198	2742	27.7	2.24	0.633	0.02	1.4
CE	453	2025	8.94	2.66	0.655	-0.225	4.49
USAir	332	2128	12.81	2.74	0.749	-0.208	3.36
Yeast	2 370	10 904	9.2	5.16	0.378	0.469	3.35

连边集合中随机抽取 10% 的连边数目，再从未连边的节点对集合中随机抽取同样数量的节点对，组成训练集，既能够达到良好的预测精确度，又能降低算法的时间复杂度，提高运算效率。

4.4 算法有效性验证

本文使用 NNLP 算法对 CN、AA、PA、Katz 这 4 种相似性指标进行融合实验，计算 6 种真实网络数据集中未连接节点对的连边可能性值。本节使用 AUC 值精确衡量各个算法的预测精度。在每种数据集上分别单独计算 10 次，每次计算前重新划分训练集和测试集，随机从训练集中抽取节点对组成神经网络的训练集，取 10 次计算的平均值作为未连边节点对的数值，可以得到各个算法在 6 种不同数据集上的 AUC 值，如表 2 所示，每个数据集中最优值用粗体标出，Katz 指标的参数固定为 0.01。

如表 2 所示，本文中提出的 NNLP 算法在 6 种真实网络数据集中的表现都优于其他指标。这 6 种真实网络来自不同的领域，具有不同的网络拓扑特征，因此说明 NNLP 算法在不同种类的网络中

可以有比较好的表现，NNLP 算法可以有效提高结构相似性指标的顽健性。

表 3 中数据与表 2 中结果一致，在精确度指标中，NNLP 算法依然表现最佳。

通过分析 NNLP 算法和 4 种结构相似性指标可以发现，4 种指标是基于不同的连边机理提出的，因此，它们在不同的网络中表现不同。例如，PA 指标在 Jazz 网络中表现优秀，而在 NS 网络中表现非常一般；Katz 指标虽然考虑了全局路径信息，但在所有网络中的表现也不是最好，它在 USAir 网络中的表现与 AA 指标有一定差距；而 NNLP 算法合理融合了 4 种指标，因此在所有网络中都有比较好的表现。值得注意的是，NNLP 算法提供的是一种融合指标框架，它也可以用于融合其他的相似性指标以获得更好的表现，本章不再选取其他指标进行比较。

融合各种结构相似性指标的研究仍处于起步阶段，相关文献还比较少；本文选择与 He 等^[18]提出的一种基于 OWA 算子的链路预测算法进行对比，称为 LPE_{OWA} 算法，该算法采用 MEM (maximum entropy method)、MVM (minimum

表 2 NNLP 算法和各结构相似性指标在数据集中的 AUC 值

算法	AUC 值					
	NS	PB	USAir	CE	Jazz	Yeast
CN	0.9657	0.9225	0.9581	0.8248	0.9514	0.9082
PA	0.7400	0.9093	0.9260	0.7507	0.7674	0.8649
AA	0.9658	0.9238	0.9689	0.8408	0.9591	0.9084
Katz	0.9800	0.9329	0.9506	0.8493	0.9409	0.9697
NNLP	0.9998	0.9453	0.9807	0.9024	0.9766	0.9814

表 3 NNLP 算法和各结构相似性指标在数据集中的 precision 值

算法	precision 值					
	NS	PB	USAir	CE	Jazz	Yeast
CN	0.495	0.178	0.389	0.101	0.507	0.151
PA	0.039	0.066	0.328	0.062	0.135	0.095
AA	0.508	0.177	0.404	0.103	0.522	0.175
Katz	0.478	0.176	0.381	0.105	0.446	0.256
NNLP	0.516	0.185	0.413	0.134	0.543	0.265

variance method) 和 CSM (chi-square method) 3 种 OWA 算子分别给 9 种基于局部信息的结构相似性指标赋予融合权重。在 WSDP98、ChesLower、C96 和 B97 这 4 个真实社会网络上的实验表明, 该算法能在一定程度上提高结构相似性指标的预测精度和顽健性; 其中, 与 MVM 算子和 CSM 算子相比, 基于 MEM 算子的 LPE_{OWA} 算法在大部分数据集中可以取得更高的预测精度。基于上述分析, 本章采用 MEM 算子对 CN、AA、PA 和 Katz 这 4 种结构相似性指标进行融合, 并在 6 种真实网络数据集上与 NNLP 算法进行对比。

LPE_{OWA} 算法的主要参数是优化权重, 记为 α , 文献[18]采用了 15 个不同的 α 值 (0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 和 0.98) 进行实验。实验结果表明, 在不同的网络数据集中, 最高的预测结果对应的 α 值不同。因此, 本文也采用这 15 个 α 值进行实验, 并取最好的预测结果与 NNLP 算法进行对比。按 20%~80% 的比例随机划分训练集和测试集, 在各个数据集上分别单独计算 10 次后取其平均值, 得到实验结果如表 4 所示。

如表 4 所示, 在 6 个真实网络数据集中, LPE_{OWA} 算法和 NNLP 算法的预测精度都优于单个结构相似性指标, 而 NNLP 算法的预测精度又都高于 LPE_{OWA} 算法。在许多数据集中, LPE_{OWA} 算法的预测精度与表现最优的单个指标接近, 说明 LPE_{OWA} 算法对预测精度的提升能力有限。在 Jazz 网络和 PB 网络中, NNLP 算法和 LPE_{OWA} 算法的表现接近, 但在 NS、USAir、CE 和 Yeast 网络中, NNLP 算法明显优于 LPE_{OWA} 算法, 这是因为 OWA 算子在融合各结构相似性指标时,

只考虑到了各指标分数值在大小顺序上的重要性, 并没有考虑到各指标之间存在的交互作用。

4.5 融合局部相似性指标后与全局指标的比较

在本文选取的 4 种结构相似性指标中, CN、PA、AA 属于局部结构相似性指标, Katz 指标属于全局结构相似性指标。前 3 种指标时间复杂度较低, 但预测精度一般没有全局指标高, Katz 指标的精度一般较高, 但时间复杂度也较高。如表 4 所示, Katz 指标在 NS、PB、Yeast 等大多数网络中的表现要优于其他 3 种只考虑了局部信息的结构相似性指标, 甚至在很多网络中的表现与 NNLP 算法接近。从表 4 中也可以看出, 在许多数据集中, Katz 指标的模糊密度大于其他 3 种局部结构相似性指标, 说明考虑了全局信息的 Katz 指标在融合中发挥了重要作用。基于上述分析, 本文中采用神经网络来融合局部相似性指标, 比较其与 Katz 指标的优劣。将融合了 3 种局部结构相似性指标的算法称为 NNLP₃ 算法, 表 5 列出了 Katz 指标、NNLP₃ 算法和 NNLP 算法的 AUC 值。

从表 5 可以看出, NNLP 算法在除 Jazz 网络外的其他 5 个真实网络数据集集中的表现优于 Katz 指标和 NNLP₃ 算法, 在表 6 中, NNLP 算法在 6 个真实网络数据集中表现均优于 Katz 指标和 NNLP₃ 算法, 在大多数网络中, NNLP₃ 算法也优于考虑了全局信息的 Katz 指标。但在 PB 和 Yeast 网络中, Katz 指标要优于 NNLP₃ 算法, 这是因为 CN、PA、AA 指标在 PB 和 Yeast 网络中的表现明显低于 Katz 指标, 使神经网络在对 3 种局部结构相似性指标进行融合后预测精度有所提升, 但提升效果受到这 3 种指标预测精度的限制。在 NS 指标中, CN、PA、AA 指标也与 Katz 指标有较

表 4 NNLP 算法、 LPE_{OWA} 算法和各指标在数据集集中的 AUC 值

算法	AUC 值					
	NS	PB	UsAir	CE	Jazz	Yeast
CN	0.969 1	0.919 1	0.950 7	0.829 7	0.952 5	0.909 8
PA	0.737 6	0.911 6	0.918 3	0.752 8	0.766 5	0.866 4
AA	0.969 6	0.922 4	0.956 7	0.845 8	0.946 4	0.911 2
Katz	0.980 8	0.930 4	0.950 9	0.855 6	0.938 7	0.970 6
LPE_{OWA}	0.981 9	0.941 7	0.966 7	0.867 9	0.960 8	0.972 9
NNLP	0.999 8	0.945 3	0.980 7	0.902 4	0.975 3	0.981 4

表 5 Katz 指标、NNLP₃ 算法和 NNLP 算法在数据集中的 AUC 值

算法	AUC 值					
	NS	PB	USAir	CE	Jazz	Yeast
Katz	0.980 0	0.932 9	0.947 2	0.859 7	0.942 8	0.974 1
NNLP ₃	0.988 0	0.929 8	0.978 5	0.870 7	0.977 8	0.919 0
NNLP	0.999 8	0.945 3	0.980 7	0.902 4	0.975 3	0.981 2

表 6 Katz 指标、NNLP₃ 算法和 NNLP 算法在数据集中的 precision 值

算法	precision 值					
	NS	PB	USAir	CE	Jazz	Yeast
Katz	0.478	0.177	0.381	0.105	0.446	0.256
NNLP ₃	0.511	0.138	0.389	0.109	0.531	0.143
NNLP	0.516	0.185	0.413	0.134	0.543	0.266

大的差距，但融合后的预测效果却好于 Katz 指标，说明 CN、PA、AA 指标在 NS 网络的预测计算中，相互之间具有更加良好的互补性，而在 PB 和 Yeast 网络中互补性不够明显。在 Jazz 网络中，NNLP₃ 算法的结果要优于 NNLP 算法，说明融合相似性指标的方法，并不是选择的指标越多，效果越好。

5 结束语

本文提出的 NNLP 算法通过神经网络对不同链路预测指标进行了融合，主要利用指标在预测结果上的互补性，提高了链路预测的精度，同时兼顾了算法的时间复杂度。NNLP 算法在计算中针对不同的网络数据训练神经网络参数，然后计算融合指标，针对不同的网络具有较强的适应性。通过本文的算法推导以及实验，验证了融合算法能够提升链路预测的精度，同时也发现，融合指标并不是越多越好。下一步研究主要关注 2 方面问题：1) 融合方法中如何确定融合指标的种类和数量；2) 通过探讨各个算法之间的互补性，进一步分析网络特性，并探索新的算法。

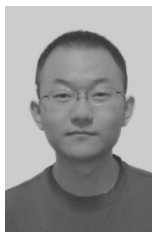
参考文献：

[1] LÜ L, ZHOU T. Link prediction in complex networks: a survey[J]. Physica a: Statistical Mechanics and its Applications, 2011, 390(6):

1150-1170.
[2] CANNISTRACI C V, ALANISLOBATO G, RAVASI T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks[J]. Scientific Reports, 2015 ,3 (4): 1613-1613.
[3] KOSSINETIS G. Effects of missing data in social network[J]. Social Networks, 2006 , 28 (3): 247-268.
[4] ESSLIMAMI I, BRUN A, BOYER A. Densifying a behavioral recommender system by social networks link prediction methods[J]. Social Network Analysis and Mining, 2011,1(3): 159-172.
[5] 刘树新, 季新生, 刘彩霞, 等. 一种信息传播促进网络增长的网络演化模型[J]. 物理学报, 2014, 63(15).
LIU S X, JI X S, LIU C X, et al. A complex network evolution model for network growth promoted by information transmission[J]. Acta Physica Sinica, 2014, 63(15).
[6] 刘树新, 季新生, 刘彩霞, 等. 局部拓扑信息耦合促进网络演化[J]. 电子与信息学报, 2016, 38 (9): 2180-2187.
LIU S X, JI X S, LIU C X, et al. Information coupling of local topology promoting the network evolution [J]. Journal of Electronics & Information Technology, 2016, 38(9): 2180-2187.
[7] ZHOU C, MOTTER A E, KURTHS J. Universality in the synchronization of weighted random networks[J]. Physical Review Letters, 2006, 96 (3): 034101.
[8] FIRTUNATO S. Community detection in graph[J]. Physics Reports, 2010, 486 (3): 75-174.
[9] 黄立威, 李德毅, 马于涛, 等. 一种基于元路径的异质信息网络链路预测模型[J]. 计算机学报, 2014, 37 (4): 848-858.
HUANG L W, LI D Y, MA Y T, et al. A meta path-based link prediction model for heterogeneous information networks[J]. Chinese Journal of Computers 2014 ,37 (4): 848-858.
[10] 王守辉, 于洪涛, 黄瑞阳, 等. 基于模体演化的时序链路预测方

- 法[J]. 自动化学报, 2016, 42 (5): 735-745.
- WANG S H, YU H T, HUANG R Y, et al. A temporal link prediction method based on motif evolution[J]. Acta Automatica Sinica, 2016, 42 (5): 735-745.
- [11] MITZENMACHER M. A brief history of generative models for power law and lognormal distributions[J]. Internet Mathematics, 2004, 1 (2): 226-251.
- [12] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25 (3): 211-230.
- [13] OU Q, JIN Y D, ZHOU T, et al. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks[J]. Physical Review E, 2007, 75 (2).
- [14] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286 (5439): 509-512.
- [15] 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.
- [16] KATZ L. A new status index derived from sociometric index[J]. Psychometrika, 1953, 18 (1): 39-43.
- [17] LIU S, JI X, LIU C, et al. Similarity indices based on link weight assignment for link prediction of unweighted complex networks[J]. International Journal of Modern Physics B, 2017, 31 (2): 412-1054.
- [18] HE Y, LIU J N K, HU Y, et al. OWA operator based link prediction ensemble for social network[J]. Expert Systems with Applications, 2015, 42 (1): 21-50.
- [19] GAO F, MUSIAL K, COOPER C, et al. Link prediction methods and their accuracy for different social networks and network metrics[J]. Scientific Programming, 2014, 501: 172879.
- [20] 吴祖峰, 梁棋, 刘峤, 等. 基于 AdaBoost 的链路预测优化算法[J]. 通信学报, 2014, 35 (3): 116-123.
- WU Z F, LIANG Q, LIU Q, et al. Modified link prediction algorithm based on AdaBoost [J]. Journal on Communications, 2014, 35(3): 116-123.
- [21] YU H T, WANG S H, MA Q Q. Link prediction algorithm based on the Choquet fuzzy integral[J]. Journal on Communications, 2016, 20 (4): 809-824.
- [22] 吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社, 2013: 46-47.
- LYU L Y, ZHOU T. Link prediction[M]. Beijing: High Education Press, 2013.
- [23] BRITS R, ENGELBRECHT A P, et al. A niching particle swarm optimizer[C]//The 4th Asia-Pacific conference on simulated evolution and learning, Singapore, 2002, 692-696.
- [24] HECHT-NIELSEN R. Theory of the backpropagation neural network[C]//International Joint Conference on Neural Networks, 1989, 1(1): 593-605.

[作者简介]



潘永昊 (1992-), 男, 甘肃金昌人, 国家数字交换系统工程技术研究中心硕士生, 主要研究方向为复杂网络、链路预测。



于洪涛 (1970-), 男, 辽宁丹东人, 博士, 国家数字交换系统工程技术研究中心研究员, 主要研究方向为网络大数据分析处理。



刘树新 (1987-), 男, 山东潍坊人, 博士, 国家数字交换系统工程技术研究中心助理研究员, 主要研究方向为复杂网络、网络信息挖掘。