

# 基于二阶共同邻节点的链路预测算法

王海龙<sup>1</sup>, 李晨璞<sup>2</sup>, 王浩森<sup>2</sup>

(1.河北建筑工程学院信息工程学院, 河北 张家口 075000;

2.河北建筑工程学院数理系, 河北 张家口 075000)

✉ xwanghl@126.com; lichenpu2005@126.com; 1145187343@qq.com



**摘要:**针对链路预测共同邻居算法(CN算法)预测精确度偏低且不适用于多类型网络的缺点,在CN算法的基础上,提出一种将共同邻居的概念扩展到二阶的链路预测算法,算法将待预测节点的邻节点分为3种类型,不同类型的邻居点被赋予不同的权重,建立基于二阶共同邻节点的链路预测算法(CN2算法)。以7个真实网络为例,通过计算分析AUC值(ROC曲线下的面积)测试算法预测精确度,将测试集数据划分为5%~50%的不同比例以测试算法的鲁棒性。综合得出CN2算法的链路预测准确性较CN算法在7个网络中平均提升9.9%左右,并且鲁棒性更优。

**关键词:**复杂网络;链路预测;共同邻居

**中图分类号:**TP393 **文献标志码:**A

## Link Prediction Algorithm Based on Second-Order Common Neighbor Nodes

WANG Hailong<sup>1</sup>, LI Chenpu<sup>2</sup>, WANG Haosen<sup>2</sup>

(1.School of Information Engineering, Hebei University of Architecture, Zhangjiakou 075000, China;

2.Department of Mathematics, Hebei University of Architecture, Zhangjiakou 075000, China)

✉ xwanghl@126.com; lichenpu2005@126.com; 1145187343@qq.com

**Abstract:** The link prediction Common Neighbor algorithm (CN algorithm) has the disadvantages of low prediction accuracy and being unsuitable for multi-type networks. Based on the CN algorithm, this paper proposes a link prediction algorithm that extends the concept of common neighbors to the second order. The algorithm divides the neighbor nodes of the nodes to be predicted into three types, and different types of neighbor nodes are assigned different weights to establish a link prediction algorithm based on Second-order Common Neighbor nodes (CN2 algorithm). Taking seven real networks as examples, the prediction accuracy of the algorithm is tested by calculating and analyzing the AUC (Area under the ROC Curve) value, and the robustness of the algorithm is tested by dividing the test set data into different proportions ranging from 5% to 50%. Comprehensive results show that the link prediction accuracy of the CN2 algorithm is improved by an average of about 9.9% compared to CN algorithm in the seven networks, and it has better robustness.

**Key words:** complex networks; link prediction; common neighbors

## 0 引言(Introduction)

链路预测的原理是通过分析现有网络的链路和节点等已知部分的信息寻找和判断网络中缺失的链路,并预测未来可能

出现的链路,这对更立体地研究复杂网络非常重要<sup>[1]</sup>。链路预测有着广泛的应用场景,研究者已在社交网络、交通信息网络、生物蛋白质作用网络、电力信息网络等领域开展了大量链路预

测的研究<sup>[2]</sup>。

链路预测算法主要分为三大种类:基于局部相似性的链路预测算法、基于路径的链路预测算法、基于随机游走的链路预测算法<sup>[3]</sup>。基于局部相似性的链路预测算法具有设计简单、计算复杂度相对较低的优点,其主要原理为若两个不直接相连的节点非常相似,比如相邻节点相同、节点的度相同或节点共同邻居相同,则两个节点就极有可能在未来某段时间产生链接。基于路径的链路预测算法是通过分析待预测节点对节点之间低路径、高阶路径甚至全路径进行设计的。基于随机游走的链路预测算法的设计原理是根据网络信息,从当前节点向其邻居节点转移,转移的选择由选用的算法决定,以此进行链路预测。

## 1 链路预测研究(Link prediction research)

近年来,基于随机游走的链路预测算法吸引了许多学者的关注。例如,JORDAN等<sup>[4]</sup>提出机器学习中的集成学习可以很好地做到对算法取长补短,然后引入链路预测科研领域,结合多种特征进行系统地分析和研究。GUNS等<sup>[5]</sup>通过集成学习中的随机森林算法,用非洲、中东等地区一些国家的城市疾病网络进行仿真实验,其预测精确度优于单个链路预测算法。LIU等<sup>[6]</sup>提出了改进后的图嵌入随机游走的链路预测算法,可以增强网络的拓扑特性,提升算法的预测精确度,但大多需要提前对数据进行预处理和训练,预测成本较高。

在基于路径的链路预测算法中,ZHOU等<sup>[7]</sup>在共同邻居算法的基础上考虑了更深的一个层次——三阶路径,并提出在某种特定的网络中,三阶路径算法的预测精确度要高于二阶路径算法的预测精确度。张齐<sup>[8]</sup>考虑将信息熵引入链路预测算法中以提提高路径预测算法的精确度,其中包括最短路径的信息熵,提出了基于路径熵的链路预测算法(如PE算法等)。MOUTINHO等<sup>[9]</sup>提出了基于路径的链路预测量子算法,算法使用受控的量子行走对偶数和奇数路径进行预测分数的编码。基于机器学习和路径的链路预测算法提高了预测精确度,但是两类算法的时间复杂度都比较高,不适用于大型网络。

在基于局部相似性的链路预测算法的研究中,NEWMAN<sup>[10]</sup>提出在同一个网络中有更多共同合作者的两个节点在未来产生链接的可能性更大。ZHOU等<sup>[11]</sup>采用每个节点将单位资源平均分配给其相邻节点为权重,提出了基于资源分配的链路预测算法。YANG等<sup>[12]</sup>在研究CN算法<sup>[13]</sup>的基础上,引入引力场理论,提高了链路的预测精确度。考虑到预测精确度和预测成本,在CN算法的基础上,本文将共同邻居的概念进行了扩展,把这些节点分为3种不同的类型,不同类型的邻居点被赋予不同的权重,提出二阶共同邻节点的链路预测算法(CN2算法),并将其应用到7个真实网络进行实验,实验结果显示本文提出的算法具有较高的预测精确度。

## 2 算法介绍(Algorithm introduction)

### 2.1 问题描述

首先对真实网络数据 $G(V, E)$ 进行网络定义,其中 $V$ 为网络中的节点集合, $E$ 为网络中的连边集合。网络中的节点总数为 $N$ 。通常用 $S_{xy}$ 表示链路预测算法中两个节点相似度预测的量值, $S_{xy}$ 的值越大,两个节点间产生连边的概率越高。

### 2.2 链路预测经典算法

表1为下文所用符号及其含义。

表1 符号及其含义

Tab.1 Symbol and its meaning

| 符号                         | 含义                            |
|----------------------------|-------------------------------|
| $\Gamma(x) \cap \Gamma(y)$ | 节点 $x$ 和节点 $y$ 的共同邻居节点集合      |
| $k_z$                      | 节点 $z$ 的度                     |
| $\alpha$                   | 大于0且小于1,是控制邻居节点权重的参数          |
| $ paths_{x,y}^{<l>} $      | 连接节点 $x, y$ 的路径中,长度为 $l$ 的路径数 |

表2为几种算法的公式定义。

表2 公式定义

Tab.2 Formula definition

| 算法                                      | 公式定义                                                                               |
|-----------------------------------------|------------------------------------------------------------------------------------|
| CN                                      | $S_{xy}^{CN} =  \Gamma(x) \cap \Gamma(y) $                                         |
| AA(Adamic-Adar) <sup>[14]</sup>         | $S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}$           |
| RA(Resource Allocation) <sup>[15]</sup> | $S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$                |
| PA(Priority Allocation) <sup>[16]</sup> | $S_{xy}^{PA} = k_x k_y$                                                            |
| Salton <sup>[15]</sup>                  | $S_{xy}^{Salton} = \frac{ \Gamma(x) \cap \Gamma(y) }{\sqrt{k_x k_y}}$              |
| Jaccard <sup>[16]</sup>                 | $S_{xy}^{Jaccard} = \frac{ \Gamma(x) \cap \Gamma(y) }{ \Gamma(x) \cup \Gamma(y) }$ |
| Katz <sup>[17]</sup>                    | $S_{xy}^{Katz} = \sum_{l=1}^{\infty} \alpha^l \cdot  paths_{x,y}^{<l>} $           |

CN算法和Jaccard算法主要考虑了共同邻居的数量对节点相似度的影响,AA算法、RA算法和Salton算法在此基础上又考虑了共同邻居节点度对节点相似度的影响。AA算法中两个节点的共同邻节点对其相似性的影响与共同邻节点度的对数值成反比。RA算法考虑了资源平均分配的影响,认为两个节点的共同邻节点对其相似性的影响与共同邻节点的度值成反比。PA算法没有考虑共同邻居节点,只考虑了待预测节点对两个节点自身度值的影响。Katz算法考虑了两个节点之间全部路径对相似性的影响,此算法的相似性的预测精确度较高,但计算复杂度会随着路径长度的增加而快速提升。

### 2.3 基于二阶共同邻节点的链路预测算法

在数据爆炸式增长的时代,大型社交和信息网络需要具备快速的处理和分析能力,这时基于局部节点信息的链路预测算法因具有较低的时效性而备受研究人员的关注。本文在CN算法的基础上将共同邻居的概念扩展到二阶邻节点,然后把把这些节点分为3类,不同类型的邻节点被赋予不同的权重,建立二阶共同邻节点链路预测算法,将 $S_{xy}$ 定义为

$$S_{xy}^{CN2} = N_I + \alpha N_{II} + \alpha^2 N_{III} \quad (1)$$

CN2算法的核心思想是将待预测两个节点( $X$ 和 $Y$ )的二阶邻节点进行分类,第一类节点( $N_I$ )是待测的两个节点 $X$ 和 $Y$ 的共同邻节点,第二类节点( $N_{II}$ )是与其中一个节点(如 $X$ )的距离为1的节点,与另一个节点(如 $Y$ )的距离为2的节点,

第三类节点( $N_{III}$ )是与  $X$  和  $Y$  两个节点的距离均为 2 的节点。CN2 算法对二阶邻节点集合范围内的节点进行分类,考虑到算法的时间复杂度,此算法只研究距离不大于 2 的邻居节点,CN2 算法弥补了 CN 算法中预测信息较少的缺陷。

图 1 是 CN2 算法的一个示例网络,通过示例网络可进一步理解 CN2 算法建立的思想。图 1 中, $X$ 、 $Y$  是待预测节点对,节点 1 和节点 2 为  $X$  和  $Y$  的第一类共同邻节点,即  $N_I$ ,坐标形式为(1,1),其中第一个坐标表示邻节点距离  $X$  的步长,第二个坐标表示邻节点距离  $Y$  的步长;节点 3、4、6 和 7 为第二类共同邻节点,即  $N_{II}$ ,坐标形式为(1,2)或(2,1);节点 5 为第三类共同邻节点,即  $N_{III}$ ,坐标形式为(2,2)。经典的 CN 算法只考虑了待预测节点对  $X$  和  $Y$  的第一类共同邻节点,预测信息较少,本文将  $X$  和  $Y$  节点对的共同邻节点拓展到二阶,按照距离将其分为 3 类并赋予不同的权重,以提高链路预测的精确度。

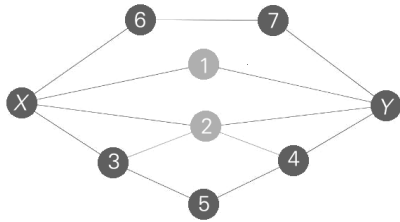


图 1 示例网络  
Fig. 1 Example network

3 实验结果与分析(Experimental results and analysis)

在分析阶段,对网络数据中的所有连边进行数据集的划分。每次随机从网络连边集中抽取 90% 的连边作为训练集,剩余部分为测试集。每种网络独立进行 100 次重复实验,取结果的平均值作为算法的最终预测效果。

3.1 实验数据

为验证本文提出算法的有效性和精确性,本文在网络数据集网站(<https://networkrepository.com>)中选取了 7 个不同的领域,对结构和大小均不相同的实际网络进行实验,分别为线虫的神经网络(C. elegans)、科学家合作网络(NS)、食物链网络(FWFB)、海豚网络(Dolphin)、柏林传染病网络(Dublin)、政治家博客网络(PB)、电力网络(Power)。表 3 为 7 个网络的具体拓扑特征。

表3 7 个网络的拓扑特征

Tab.3 Topological characteristics statistics of 7 networks

| 网络         | V     | E      | C     | K      | D     |
|------------|-------|--------|-------|--------|-------|
| C. elegans | 297   | 2 345  | 0.311 | 15.791 | 0.053 |
| NS         | 379   | 914    | 0.741 | 4.82   | 0.013 |
| FWFB       | 128   | 2 075  | 0.344 | 32.422 | 0.259 |
| Dolphin    | 62    | 159    | 0.259 | 5.129  | 0.084 |
| Dublin     | 410   | 2 765  | 0.456 | 13.488 | 0.033 |
| PB         | 1 222 | 16 714 | 0.320 | 27.355 | 0.022 |
| Power      | 4 941 | 6 594  | 0.080 | 2.669  | 0.001 |

注:V 代表网络的节点数,E 代表网络的边数,C 代表网络的平均聚类系数,K 代表网络的平均度,D 代表网络密度。

3.2 评价指标

采用评价指标 AUC(ROC 曲线下面积)<sup>[18]</sup> 测试算法的预

测精确度,原理为在网络测试集中随机选择一条连边,其概率大于在网络中不存在的连边中随机选择一条连边的概率。例如,有  $n_1$  次选中在网络测试集中连边的概率大于不存在连边的概率,有  $n_2$  次选中在网络测试集中连边的概率小于或等于不存在连边的概率,AUC 公式定义为

$$AUC = \frac{n_1 + 0.5n_2}{n} \quad (2)$$

3.3 实验步骤

- (1)对真实的网络数据进行预处理,得到用于实验的稀疏矩阵。
- (2)采用随机抽样的方式对实验数据集进行划分。
- (3)将 CN2 算法和其他经典预测算法进行对比,计算每个算法的 AUC 平均值,并进行比较。
- (4)多次重复实验获得平均预测值,根据实验结果明确每一个真实网络中  $\alpha$  的最优值。
- (5)划分不同比例的测试集,检验算法的预测鲁棒性。
- (6)对算法的预测精确度和鲁棒性进行总结。

3.4 实验结果

在 CN2 算法中, $\alpha$  是可以调节的权重参数。为了得到最优的预测效果,本文设置  $\alpha$  值为 0.001~1,将其分为 3 段。第一段为 0.001~0.01,每次递增 0.001。第二段为 0.01~0.1,每次递增 0.01。第三段为 0.1~1,每次递增 0.1。对 7 个真实网络进行链路预测实验,以 AUC 为评价指标,获得每个网络中最优的  $\alpha$  值。AUC 随  $\alpha$  变化的趋势如图 2 所示。

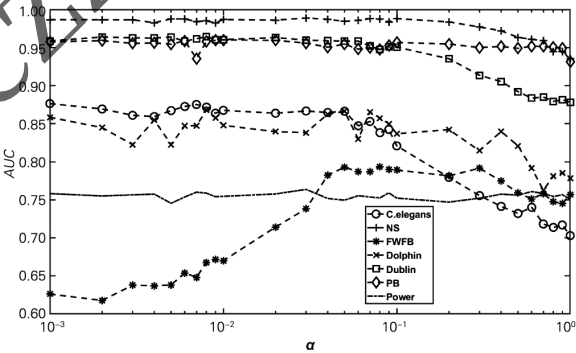


图 2 AUC 随  $\alpha$  变化的趋势

Fig. 2 AUC variation trend with  $\alpha$

从图 2 可以看出,当  $\alpha$  取 0.008 时,CN2 算法对大部分网络的预测精确度较好,在 FWFB 网络中且当  $\alpha$  取 0.04 时的预测精确度最好。所以,在后续的计算过程中,在 FWFB 网络中采用 CN2 算法进行预测时, $\alpha$  取 0.04,其他网络中的  $\alpha$  取 0.008。将 CN2 算法分别与 CN 算法、AA 算法、PA 算法、RA 算法、Salton 算法、Jaccard 算法和 Katz 算法在上述 7 个网络中进行对比,得到的 AUC 值如表 4 所示。

表 4 7 个不同网络中各算法预测的 AUC 值(训练集比例为 90%)

Tab.4 The AUC values predicted by various algorithms in 7 different networks (with a training set ratio of 90%)

| 网络         | AUC 值 |       |       |       |        |         |       |
|------------|-------|-------|-------|-------|--------|---------|-------|
|            | CN    | AA    | RA    | PA    | Salton | Jaccard | Katz  |
| C. elegans | 0.847 | 0.864 | 0.867 | 0.755 | 0.797  | 0.790   | 0.864 |
| NS         | 0.979 | 0.983 | 0.983 | 0.655 | 0.978  | 0.977   | 0.986 |
| FWFB       | 0.605 | 0.606 | 0.608 | 0.731 | 0.526  | 0.526   | 0.676 |

0.876  
0.988  
0.793



续表

| 网络      | AUC 值 |       |       |       |        |         |              |              |
|---------|-------|-------|-------|-------|--------|---------|--------------|--------------|
|         | CN    | AA    | RA    | PA    | Salton | Jaccard | Katz         | CN2          |
| Dolphin | 0.809 | 0.815 | 0.813 | 0.656 | 0.811  | 0.813   | 0.849        | <b>0.858</b> |
| Dublin  | 0.938 | 0.942 | 0.944 | 0.692 | 0.943  | 0.941   | 0.963        | <b>0.964</b> |
| Polblog | 0.923 | 0.926 | 0.927 | 0.909 | 0.878  | 0.876   | 0.935        | <b>0.960</b> |
| Power   | 0.625 | 0.624 | 0.624 | 0.579 | 0.624  | 0.625   | <b>0.963</b> | 0.758        |

由表 4 可知,在 C. elegans、NS、FWFB、Dolphin、Dublin 和 Polblog 六个网络中,与其他七个算法相比,CN2 算法预测得到的 AUC 值均最大,特别是在 FWFB 网络中,AUC 值有较大幅度的提高。但在 Power 网络中,由于网络中心性比较高,每个网络节点相差不大,因此 CN2 算法的预测精确度没有达到最好的效果。从整体的实验结果来看,CN2 算法达到了较好的预测精确度。

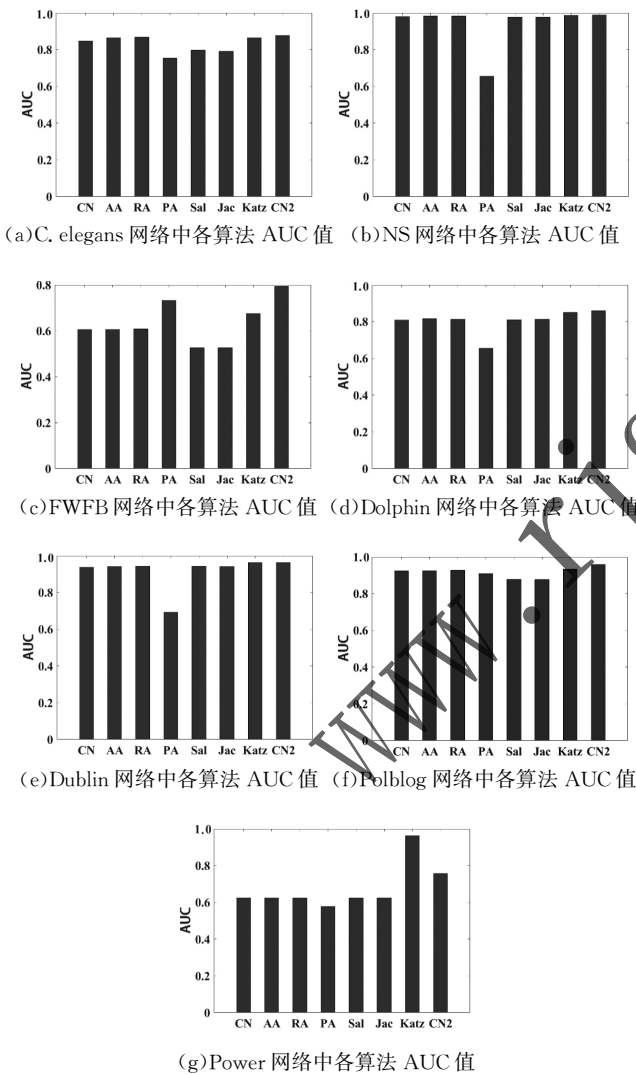


图 3 8 种链路预测算法在 7 个网络中的 AUC 值直方图

Fig. 3 AUC histograms of 8 link prediction algorithms in 7 networks

为了更直观地展示算法之间的差别,图 3 显示了 8 种链路预测算法在 7 个网络中的 AUC 值直方图。结合表 4 中的数据对不同网络的 AUC 值进行分析发现,7 个实验网络中,CN2 算法的 AUC 值在大部分情况下处于最高值,特别是在 FWFB 网络中,CN2 算法的 AUC 值比 CN 算法的 AUC 值高出 31 百分

点左右。此外,该算法在不同的网络中预测精确度都很高,在 7 个不同类型的网络中,将 CN 算法的 AUC 值与 CN2 算法的 AUC 值进行比较得出,CN2 算法精确度平均提升 9.9 个百分点左右,说明该算法具有一定的适用普遍性,是一种较好的链路预测算法。

### 3.5 鲁棒性分析

为了更全面地对 CN2 算法进行测试,本文将网络数据集进行不同比例的划分,比较在不同测试集比例下,CN2 算法与其他 7 个预测算法预测精确度的差异,并对算法进行鲁棒性测试。测试集比例(Probe Size)从 5% 增加到 50%,每次增加 5 个百分点。对于 Power 网络,由于划分测试集太大会导致网络连通性降低,因此在 Power 网络中将测试集的最大比例划分定为 0.2。实验中对所有网络,在每一种训练集比例下独立重复 100 次,取 AUC 值均值作为最终的结果。在不同测试集比例划分下,7 个网络中 8 种链路预测算法的 AUC 值如图 4 所示。

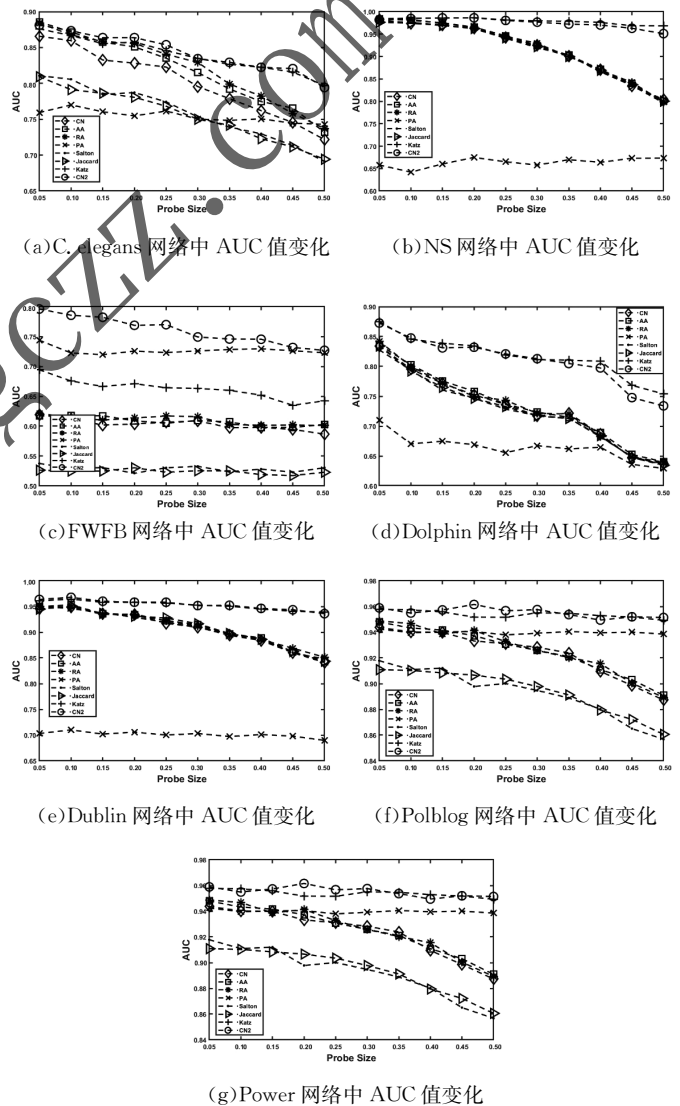


图 4 在不同测试集比例划分下 7 个网络中 8 种链路

预测算法的 AUC 值

Fig. 4 AUC values of 8 link prediction algorithms in 7 networks under different test set proportion divisions

由图4可以看到,在C. elegans、NS、FWFB、Dolphin、Dublin和Polblog网络中,随着测试集比例的增大,CN2算法的AUC值均高于其他算法的AUC值,并且其AUC值随测试集比例的增大而下降较缓慢。在Power网络中,虽然CN2算法的AUC值低于Katz算法的AUC值,但是比其他算法的AUC值均高,因此综合评测表明,CN2算法具有较高的预测鲁棒性。

#### 4 结论(Conclusion)

本文提出了一种基于二阶共同邻节点的链路预测算法——CN2算法,它是在CN算法的基础上对共同邻节点进行了扩展,增加了更多待预测节点对的周围信息,使算法获得了更好的预测精确度。对不同的邻节点进行分类以达到更好的区分待测节点对的相似性。对7个不同类型的网络进行仿真测试,将CN2算法与7个不同的链路预测算法进行实验对比发现,CN2算法具有较好的预测精确度。在7个不同网络中精确度平均提升约9.9百分点。通过改变测试集比例的划分测试算法的鲁棒性发现,CN2算法具有较好的鲁棒性,综合得出CN2算法具有较好的链路预测性能。

#### 参考文献(References)

- [1] 周涛,张子柯,陈关荣,等. 复杂网络研究的机遇与挑战[J]. 电子科技大学学报,2014,43(1):1-5.
- [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.
- [3] 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.
- [4] JORDAN M I, MITCHELL T M. Machine learning: trends, perspectives, and prospects[J]. Science, 2015, 349(6245):255-260.
- [5] GUNS R, ROUSSEAU R. Recommending research collaborations using link prediction and random forest classifiers[J]. Scientometrics, 2014, 101(2):1461-1473.
- [6] LIU Z, ZHANG Q M, LÜ L Y, et al. Link prediction in complex networks: a local naïve Bayes model[J]. Institute of physics, 2011, 96(4):48007.
- [7] ZHOU T, LEE Y L, WANG G N. Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms[J]. Physica A: statistical mechanics and its applications, 2021, 564:125532.
- [8] 张齐. 基于熵的复杂网络结构特性研究[D]. 重庆:西南大学,2017.
- [9] MOUTINHO J P, MELO A, COUTINHO B, et al. Quantum link prediction in complex networks[J]. Physical review A, 2023, 107(3):032605.
- [10] NEWMAN M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2):167-256.
- [11] ZHOU T, LÜ L Y, ZHANG Y C. Predicting missing links via local information[J]. The European physical journal B, 2009, 71(4):623-630.
- [12] YANG Y L, YE Z L, ZHAO H X, et al. A novel link prediction framework based on gravitational field[J]. Data science and engineering, 2023, 8(1):47-60.
- [13] RAPOPORT A. Spread of information through a population with socio-structural bias: II. various models with partial transitivity[J]. The bulletin of mathematical biophysics, 1953, 15(4):535-546.
- [14] ADAMIC L A, ADAR E. Friends and neighbors on the web[J]. Social networks, 2003, 25(3):211-230.
- [15] SALTON G, FOX E A, WU H. Extended boolean information retrieval[J]. Communications of the ACM, 1983, 26(11):1022-1036.
- [16] JACCARD P. Etude comparative de distribution florale dans une portion des alpes et jura[J]. Bulletin de la société vaudoise des sciences naturelles, 1901, 37(142):547-579.
- [17] LV L Y, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. Physical review E, 2009, 80(4):122-131.
- [18] 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.

#### 作者简介:

王海龙(1998-),男,硕士生。研究领域:复杂网络,链路预测。  
李晨璞(1979-),男,博士,副教授。研究领域:复杂网络理论及应用。  
王浩森(1987-),男,博士,副教授。研究领域:神经网络算法及应用。