• 复杂性科学 •

基于特征工程的重要节点挖掘方法

潘 侃[1]，尹春林[1]，王 磊[2]，陈端兵[2,3*]

(1. 云南电网有限责任公司电力科学研究院 昆明 650217；2. 成都数之联科技有限公司 成都 610041；
3. 电子科技大学大数据研究中心 成都 611731)

【摘要】复杂网络中重要节点的挖掘对分析和治理现实复杂系统有着重要的指导意义。设计能反映节点重要性的有效计算方法，是高效准确挖掘重要节点的关键。该文基于节点的邻居信息，采用特征工程中的特征提取、特征重构等方法提取能有效反映节点局部结构的特征向量。利用局部特征向量，通过回归模型建立节点局部结构和重要性的关系模型。在 13 个真实网络上的实验结果表明，相比于已有的重要节点挖掘基准方法，该方法具有更优的性能。

关 键 词 复杂网络； 重要节点； 特征工程； 局部结构
中图分类号 TP301 文献标志码 A doi:10.12178/1001-0548.2021106

Identifying Critical Nodes Based on Feature Engineering

PAN Kan[1], YIN Chunlin[1], WANG Lei[2], and CHEN Duanbing[2,3*]

(1. Electric Power Research Institute, Yunnan Power Grid Co. Ltd. Kunming 650217; 2. Union Big Data Tech. Inc. Chengdu 610041;
3. Big Data Research Center, University of Electronic Science and Technology of China Chengdu 611731)

Abstract To mine important nodes in complex networks is very important for analyzing and governing real complex systems. Designing a good indicator that reflects the importance of nodes is a key issue on efficiently and accurately mining critical nodes. On the bases of the neighbor information of nodes, the features that can effectively reflect the local structure of nodes are extracted through feature extraction and reconstruction. The relational model between local structure and real importance of nodes is established by utilizing regression model based on the extracted features. The experimental results on 13 real networks show that the proposed method outperforms the benchmark methods of critical nodes identification.

Key words complex network; local structure; critical nodes; feature engineering

现实生活中的复杂系统 (如交通运输系统、生物系统) 可以很自然地用图表示，其中节点表示系统中的各个要素，边表示要素之间的关系[1]。复杂网络的研究逐渐从宏观层面深入微观层面[2]。节点作为系统中最小的元素，不同节点在系统中的地位是不同的。重要节点是指相比于网络中其他节点，能更大程度地影响网络功能的一些特殊节点。这种节点数量不多，但是其影响力却可以快速波及网络中大部分节点，如社交网络中权威账号的舆论引导，交通网路中重要路口堵塞导致交通系统瘫痪等。节点重要性排序[1] 和相对重要节点的挖掘[3-4] 对现实生活有着重要的指导意义。在网络分析中，节点的重要性通常用中心性[5] 来度量，其主要目的是为基础网络的每个节点分配一个实值，用于度量该节点对其他节点的影响力。目前已有不少成熟的节点中心性计算方法，主要分为两类[3]：1) 基于网络结构特征的指标和方法；2) 基于随机游走的指标和方法。

基于结构特征的指标和方法主要根据其他节点与已知节点之间的网络结构特征设计相对重要指标。这些方法通过捕捉节点之间的局部连边信息或路径信息，衡量节点的重要性。度中心性 (degree) 是最简单的中心性度量方法，主要利用网络节点的连边信息刻画节点的重要性。度中心性认为一个节点邻居数目越多，该节点影响力就越大。但若节点在网络中属于核心位置，即使它本身度很小，也有

收稿日期：2021－04－13；修回日期：2021－06－30
基金项目：国家自然科学基金 (61673085)
作者简介：潘侃 (1985－)，男，主要从事信息技术在电力系统中的应用方面的研究.
*通信作者：陈端兵，E-mail：dbchen@uestc.edu.cn

较高的影响力。基于此，文献 [6] 提出了基于 K-壳分解 (K-shell decomposition) 的 K-shell 中心性，该中心性将外围的节点层层剥去，使处于内层的节点拥有较高的影响力。还有一些基于路径的中心性计算方法，如节点的接近中心性 (closeness)[7] 考虑将节点与其他节点的测地距离之和的倒数作为节点重要性。而介数中心性 (betweenness)[8] 认为经过一个节点的最短路径越多，这个节点就越重要。受到介数中心性启发，流介数中心性 (flow betweenness)[9]、连通介数中心性 (communicability betweenness)[10]、随机游走介数中心性 (random walk betweenness)[11] 和路由介数中心性 (routing betweenness)[12] 相继被提出。除此以外，H-index[13]作为评价学者学术成就的权威方法，也能很自然地延伸到复杂网络的重要节点挖掘任务中。

上述方法能够很好地捕捉节点周围的局部结构信息。除此之外，很多学者采用基于路径和随机游走的方法，利用整个图的拓扑信息挖掘图中的重要节点。在不考虑时间开销的前提下，从初始节点出发将信息传播出去，当随机游走趋于稳定时，信息保留越多的节点越重要。特征向量中心性 (eigenvector) 传播时不仅考虑节点的邻居数目，也同时考虑每个邻居节点的重要性。另外，学者们还提出了 HITs[14]、LeaderRank[15]、PageRank[16]、Vote Rank[17] 等其他全局游走的方法。总体而言，这些基于全局游走的方法计算成本较高，不能有效地应用于超大规模网络。文献 [18] 考虑四阶邻居，提出了局部中心性方法 LocalRank，在时间复杂度和准确率之间找到了一个较好的平衡点。

虽然复杂网络中检测节点重要性的方法很多，但它们都试图找到能反映节点重要性的某种因素。但节点重要性之所以不同，是因为不同节点周围的结构是异质的[19]。因此，本文利用机器学习方法挖掘节点结构特征与节点重要性之间的关系。首先基于二步可达子图的节点信息，采用特征工程中的特征提取、特征重构方法，提出能描述节点周围信息的特征集合。再利用简单的线性回归模型 (linear regression model)[20]，学习节点局部结构与节点重要性之间的关系。在 13 个真实网络中，将训练所得模型与度中心性、介数中心性[8]、K-shell[6]、H-index[13] 和 DynamicRank[21] 中心性进行了比较。实验结果表明，本方法能更准确、更有效地识别出复杂网络中对信息传播影响较大的重要节点。

1　基于特征工程的重要节点挖掘方法

重要节点挖掘是网络攻击和信息传播及控制等领域中的核心问题之一。网络中的少数节点具有非常高的影响力。而造成网络中节点重要性差异的根本原因是节点周围的结构差异[19]。闭塞的局部结构会阻碍节点影响力的传播，而好的局部结构会促进信息在网络中传播。

本文研究主要针对无向无权图$G(V,E)$，其中$V=\{v_1,v_2,\cdots,v_n\}$是节点集合，$E=\{e_1,e_2,\cdots,e_m\}$是边集合，n 和 m 分别是节点数量和边数量。为了提取和重构节点邻居信息得到节点的局部结构特征，首先拓展两个邻居的定义。

定义 1　二阶邻居

若网络中节点 u 的一阶邻居定义为$\Gamma_1(u)$，那么节点 u 的二阶邻居可定义为：

$$\Gamma_2(u)=\{v|v\in\Gamma_1(x),x\in\Gamma_1(u),v\neq u\}$$

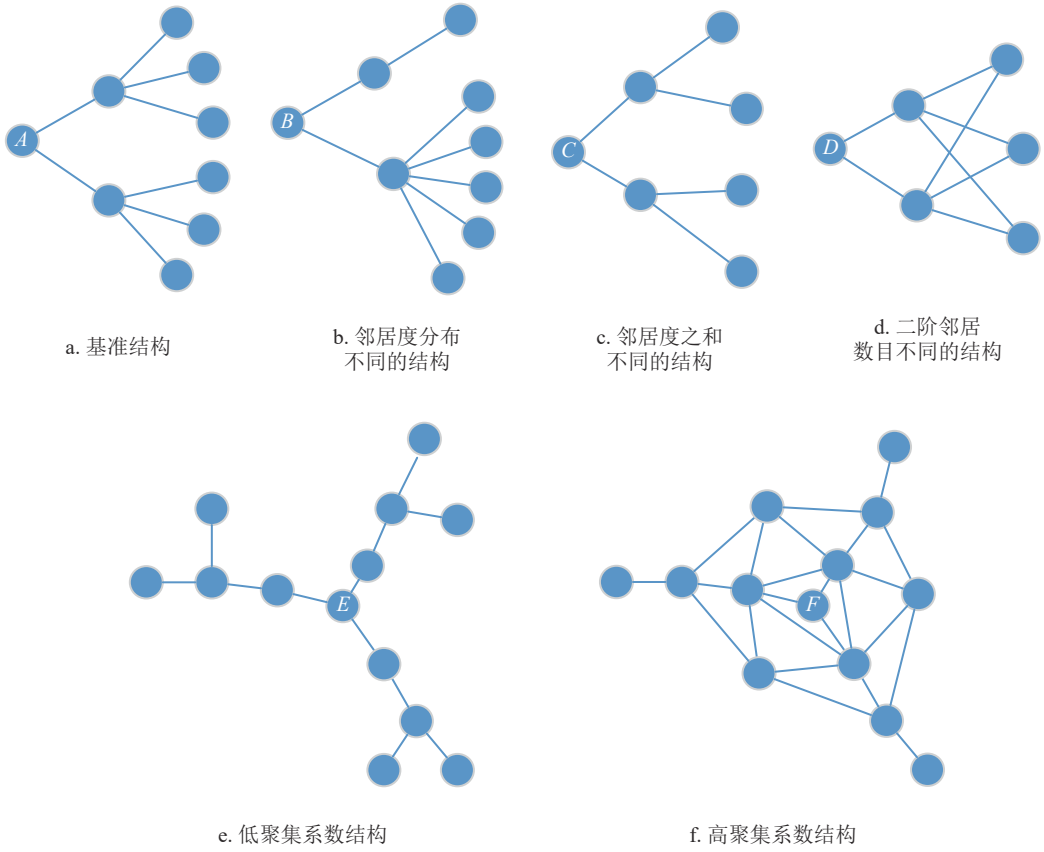
定义 2　二阶外联邻居

二阶外联邻居属于二阶邻居的子集，区别在于二阶外联邻居是二阶邻居与一阶邻居的差集，定义如下：

$$\tilde{\Gamma}_2(u)=\Gamma_2(u)-\Gamma_1(u)=\{v|v\in\Gamma_2(u)\wedge v\notin\Gamma_1(u)\}$$

从局部角度考虑，节点的度以及节点邻居的度最能反映节点的局部结构特征。除此以外，现有的中心性算法中，H-index 和 K-shell 也是能较好反映节点重要程度的中心性指标。然而这些中心性指标对节点周围复杂多样的局部结构还是很难刻画。

度中心性可以广泛地概括简单图中重要节点的规律，一般来说，节点的邻居越多，影响力越大。现实网络中节点的局部结构非常复杂，单独用某一种复杂网络指标无法准确地刻画节点周围的结构信息。如图 1a～1d 中，节点 A、B、C、D 具有不同的局部结构，相应的中心节点的影响力也有差异，使用传统的中心性方法无法准确区分这 4 个节点的真实重要性。如采用度中心计算时，A、B、C、D 属于同一类型节点 ($d_A=d_B=d_C=d_D=2$)。而 H-index 无法判断节点 A、C、D($h_A=h_C=h_D=2$)。另外 K-shell 中心性也无法判断 A、B 和 C、D ($k_A=k_B=1, k_C=k_D=2$)。可以看出，传统方法在节点重要性分析中还属于粗粒度方法，对于不同的微观局部结构有时很难区分。



a. 基准结构　b. 邻居度分布不同的结构　c. 邻居度之和不同的结构　d. 二阶邻居数目不同的结构

e. 低聚集系数结构　f. 高聚集系数结构

图 1　复杂网络中节点的局部结构示例

1.1　特征提取

由于传统的基于中心性的方法不能很好地刻画节点的局部结构，特别是对于二阶邻居结构信息的刻画过于粗糙。因此本文主要以节点的邻居信息为基础，提取和重组能刻画节点局部结构的特征。

1.1.1　一阶邻居特征

从一阶邻居开始，一般而言，度越大，信息越有可能传播出去，因此，节点的度是刻画信息传播能力的一个重要特征。除此以外，一阶邻居度的分布一定程度上反映了节点二阶邻居的结构信息。如图 1 中，虽然节点 A、B、C、D 的度都为 2，但是它们的一阶邻居度分布相差却很大。特别地，A 的一阶邻居度分布为 [4,4]，而 B 的分布是 [2,6]。显然，A 的一阶邻居度的分布更加均衡，而 B 的邻居度分配不均衡。由于这两个一阶邻居度的分布对应的局部结构不同，导致节点的影响力也不同。在低感染率下，邻居度分布越均匀，信息往外传播能力越强。若度分布极度不均衡，在图 1b 中，若度为 6 的节点没有被感染，B 节点的传播能力会大打折扣。

为了描述邻居度的分布均衡性，本文引入国际通用的，用以衡量一个国家或地区居民收入差距的常用指标：基尼系数 (Gini coefficient)，基尼系数最大为 1，最小等于 0。系数越大说明该分布越不均匀，系数越接近 0 表明收入分配越是趋向平等。对给定的序列$x=[x_1,x_2,\cdots,x_n]$，该序列数据平均值为μ，可采用下式直接计算序列的基尼系数：

$$\mathrm{Gini}(x)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|}{2n^2\mu} \tag{1}$$

如图 1 所示，B 的一阶邻居度的差距很大，而 A 的一阶邻居相对平衡。给定节点 u，其一阶邻居为$\Gamma_1(u)$，一阶邻居度的集合为$D_1(u)=\{d_v|v\in\Gamma_1(u)\}$。为了刻画节点 u 的一阶邻居度分布的平衡度，定义节点 u 的一阶邻居度的基尼系数：

$$\mathrm{Gini}(D_1(u))=\frac{\sum_{v_i\in\Gamma_1(u)}\sum_{v_j\in\Gamma_1(u)}|d_{v_i}-d_{v_j}|}{2n\sum_{v\in\Gamma_1(u)}d_v} \tag{2}$$

然而，只有基尼系数还不能完全反映节点一阶邻居局部结构。如图 1 中 A 和 C，一阶邻居的基尼系数都为 0 且中心节点的度都为 2，仅靠这两个特征还不能很好区分相同度节点重要性的差异，有时小度节点甚至比大度节点具有更高的传播影响力。

为了体现这种差异性，引入特征 2 区分这种情况，特征 2 为一阶邻居度之和，定义如下：

$$\mathrm{SUM}(D_1(u)) = \sum_{d_i \in D_1(u)} d_i \tag{3}$$

1.1.2　二阶邻居特征

有时仅用一阶邻居的特征还不能很好地刻画节点周围的局部特征，如图 1 中的节点 A 和 D，$\mathrm{Gini}(D_1(A)) = \mathrm{Gini}(D_1(D)) = 0$ 且 $\mathrm{SUM}(D_1(A)) = \mathrm{SUM}(D_1(D)) = 0$，仅从这两个角度还是无法区别 A、D 两种局部结构。针对上述情况，本文将二阶邻居数目作为特征，记为$\mathrm{Len}(\Gamma_2(u))$，其中$\mathrm{Len}(\Gamma_2(A)) = 6$，$\mathrm{Len}(\Gamma_2(D)) = 3$。

在对一阶邻居的规模和分布进行分析后，将基尼系数和规模作为二阶邻居的特征。但与一阶邻居不同的是，一阶邻居与二阶邻居会出现重叠邻居的情况。如图 1f 中的 F 节点，其周围很多一阶邻居之间存在连接。在获取二阶邻居时，很多一阶邻居还会被判定为二阶邻居。重叠的邻居越多，节点聚集系数越大，节点的影响力在局部区域内能充分地传播，但这种结构会导致信息很难再往外传播[22]。如图 1 所示，在邻居节点数目一致的情况下，E 节点往外传播的能力大于 F 节点。因此中心节点的二阶外联邻居$\tilde{\Gamma}_2(u)$度的分布和规模反映了信息从中心节点向外传播的能力。基于此，本文提取二阶外联邻居度的基尼系数和 SUM 值作为节点的局部结构特征。

表 1　节点局部结构特征度量

序号	名称	计算公式
1	自身度	d_v
2	一阶邻居的度之和	$\sum_{v\in\Gamma_1(u)} d_v$
3	一阶邻居度的基尼系数	$\dfrac{\sum_{v_i\in\Gamma_1(u)}\sum_{v_j\in\Gamma_1(u)}\lvert d_{v_i}-d_{v_j}\rvert}{2n\sum_{v\in\Gamma_1(u)} d_v}$
4	二阶邻居总数	$\mathrm{Len}(\Gamma_2(u))$
5	二阶邻居度之和	$\sum_{v\in\Gamma_2(u)} d_v$
6	二阶邻居度的基尼系数	$\dfrac{\sum_{v_i\in\Gamma_2(u)}\sum_{v_j\in\Gamma_2(u)}\lvert d_{v_i}-d_{v_j}\rvert}{2n\sum_{v\in\Gamma_2(u)} d_v}$
7	二阶外联邻居度之和	$\sum_{v\in\tilde{\Gamma}_2(u)} d_v$
8	二阶外联邻居度的基尼系数	$\dfrac{\sum_{v_i\in\tilde{\Gamma}_2(u)}\sum_{v_j\in\tilde{\Gamma}_2(u)}\lvert d_{v_i}-d_{v_j}\rvert}{2n\sum_{v\in\tilde{\Gamma}_2(u)} d_v}$

至此，本文从局部结构的规模和平衡性两个角度，针对一阶、二阶邻居，提取了共 8 个特征，具体计算方法和描述总结在表 1 中。除上述特征外，还有其他类型的特征对排序结果也有影响，如邻居度的最大值、平均值、方差等。这些特征都会对节点重要性判断带来影响，本文仅作为一种算法思路，通过重构二阶邻居内的度信息，得到刻画节点邻居结构最主要的 8 个特征用于节点重要性排序。

1.2　节点重要性学习建模

节点的重要性与节点周围的局部结构有着紧密的关系。本文根据表 1 列出的特征，采用线性回归 (linear regression) 模型对节点局部特征与节点重要性关系进行建模。定义一个线性回归函数$f: x \to s$，将节点的结构特征映射为节点的相对重要性，具体可表示为：

$$f_w(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x} + b \tag{4}$$

式中，$\boldsymbol{w}$ 为特征向量的权重向量；$\boldsymbol{x}$是特征向量；b 是误差项。

节点的数字化结构特征提取过程如图 2 所示，首先提取出节点 v 的二步可达闭环子图。接着，根据表 1，计算得到 8 个描述节点局部结构信息的特征，组成特征向量$\boldsymbol{x}_v = [x_v^1, x_v^2, \cdots, x_v^8]$。为了将提取出的特征用于不同网络的训练和预测，对每一维度的特征进行线性归一化处理，从而得到节点 v 归一化后的特征向量$\boldsymbol{x}_v = [x_v^1/x_{\max}^1, x_v^2/x_{\max}^2, \cdots, x_v^8/x_{\max}^8]$。

而为了获取节点真实的重要性，目前主要采用基于传播动力学的 SIR 模型进行仿真得到。具体地，在每个时间步，每个已感染节点与其邻居进行接触，每个易感邻居以概率β被感染，然后，每个已感染节点以概率λ恢复，为简单起见，本文将λ设置为 1。为了量化目标节点 v 的传播影响，以 v 作为唯一感染的种子开始向外传播，当不再有任何已感节点时，传播过程结束。此时，恢复的节点数记为R_v，假设节点 v 的三步可达子图的节点规模为$R_{\max}^v$。此时，选取$R_v/R_{\max}^v$用于衡量节点 v 的重要性。

如果感染概率β很小，受到感染的节点数量也很少，信息几乎传播不出去。当β值很高时，几乎感染网络中所有节点，因此过大或过小的β，通过仿真得到的节点重要性没有太大差异，对节点重要性评估没有实质性意义。本文根据非均匀平均场理论[23-25]，使用真实网络的平均度信息计算 SIR 模型中的传染阈值[26] $\beta_c \approx \langle k\rangle/(\langle k^2\rangle - \langle k\rangle)$，用此阈值作为感染概率进行仿真获得节点的真实重要性。为了消除波动的影响，本文进行 1 000 次独立仿真，取平均值 s_v 作为节点 v 的真实重要性。

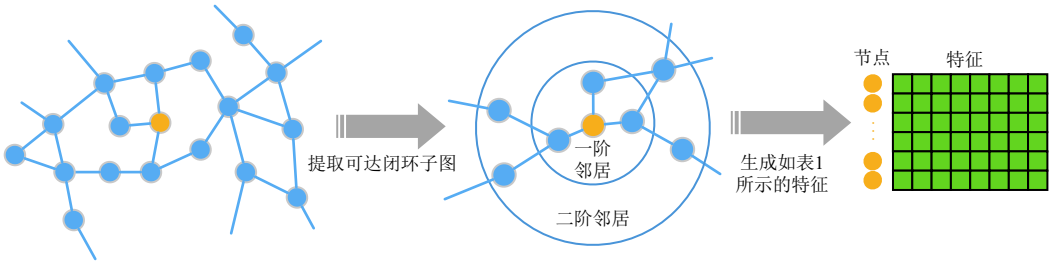


图 2 节点局部特征生成示例

至此，在获得了节点 v 的归一化结构特征$\boldsymbol{x}_v$和真实重要性 s_v 后，采用线性回归模型，选取均方误差 (mean squared error, MSE) 建立目标函数以学习节点局部结构特征与真实重要性之间的关系：

$$\min J = s - f(\boldsymbol{x}) \tag{5}$$

为了获得模型最优的回归系数，本文采用 Adam 优化器[27] 优化目标函数。

1.3 模型训练

本文用 LastFM[28] 作为训练网络对节点重要性挖掘模型进行训练学习。LastFM 是一个 2020 年 3 月从公共 API 收集的社交网络，节点代表亚洲的用户账号，边代表它们之间相互关注的关系，其节点规模为 7 624，边数量为 27 806，最大度为 216。首先从 LastFM网络中提取节点的特征向量；同时，以 LastFM 网络中每个节点为初始感染节点，进行 1 000 次独立的 SIR 传播仿真，将 1 000 次的平均 s_v 作为每个节点的标签；最后，将标签和特征向量作为训练集输入线性回归模型，训练得到节点重要性度量模型，用于预测其他网络中每个节点的重要性。

2 实验与讨论

本文用 13 个不同类型的真实网络对本文提出的方法进行测试，并和度中心性、介数中心性、K-shell 中心性、H-index 中心性和 DynamicRank 中心性进行对比。

2.1 评估指标

为了评估各方法是否能准确发现网络中的重要节点，采用 Kendall Tau 系数[29] 评估节点重要性预测排序与真实排序的相关性。给定两个有序序列，每个序列关联图中$|V|$个节点的重要程度，$X=(x_1, x_2,\cdots,x_{|V|})$，$Y=(y_1,y_2,\cdots,y_{|V|})$。从图中选取两个节点$i, j$, 若$x_i>x_j$且$y_i>y_j$，或者$x_i<x_j$且$y_i<y_j$，这两个节点是和谐的。而如果$x_i>x_j$且$y_i<y_j$，或者$x_i<x_j$且$y_i>y_j$，那么两个节点是不和谐的。除此以外，若$x_i=x_j$或$y_i=y_j$，这两个节点不属于任何一种。对网络中$\frac{1}{2}|V|(|V|-1)$节点对进行比较，可得到两个序列的 Kendall Tau 系数：

$$\tau=\frac{2(n_+-n_-)}{|V|(|V|-1)} \tag{6}$$

式中，n_+是和谐节点对的数目；n_-是不和谐节点对的数目。这个系数值在 [−1,1] 的范围内。若 X 和 Y 不相关，那么τ趋近于 0。相反 Kendall Tau 系数越大，说明两个序列相关性强，排序结果越一致。

2.2 数据集

表 2 13 个真实网络的基本特征数据

网络	n	m	$\langle k\rangle$	$k_{\max}$	$\langle c\rangle$
Router	5 022	6 258	2.490	106	0.010 0
Grid	4 941	6 594	2.669	17	0.103 0
CM	27 519	116 181	3.030	202	0.630 0
Stelzl	1 702	3 155	3.700	95	0.006 0
Vidal	3 023	6 149	4.100	129	0.065 8
Sex	16 730	39 044	4.700	305	0
NS	379	914	4.820	34	0.370 0
Figeys	2 239	6 432	5.700	314	0.039 9
Email	1 133	5 451	9.620	71	0.110 0
Hamster	2 426	16 631	13.700	273	0.537 6
Jazz	198	2 742	27.690	91	0.520 0
Polblog	1 224	19 025	31.08	467	0.226 0
USAir	1 574	28 236	35.880	596	0.380 0

本文采用的 13 个真实网络中，包括了规模较小的网络 (如 Jazz)，也有规模较大的网络 (如 Cond-Mat, CM)，其平均度的范围为 2~35。其中，1) Jazz 是爵士乐手之间的协作网络，每条边表示两个乐手在一个乐队中一起演奏；2) NetScience(NS) 是发表关于复杂网络主题论文的科学家之间的合作者网络；3) Email 是 Rovirai Virgili 大学成员之间的电子邮件交换网络；4) Sex 是研究男女性伙伴的网络；5) Polblog 是 2004 年美国大选中博客之间的超链接形成的网络；6) USAir 是 2010 年美国机场之间的航空网络；7) Router 是由 Rocketfuel 项目收集的互联

网路由器拓扑网络；8) Cond-Mat(CM) 是 1995 年−1999 年 arXiv 出版物的科学家合作网络；9) Grid 是美国西部的某电力网络；10) Figeys、Stelzl 和 Vidal 是 3 个蛋白质−蛋白质相互作用网络；11) Hamster 是一个包含网站用户之间的友谊和家庭关系的网络。以上数据集可从网站 (http://konect.cc/networks/) 获得，这 13 个真实网络的详细特征如表 2 所示，其中，n 是节点数目，m 是边数目，$\langle k\rangle$表示所有节点的平均度，$k_{\max}$代表节点的最大度，所有节点的平均聚集系数为$\langle c\rangle$。

2.3　实验及分析

为了检测模型预测的准确性，本文首先对测试网络中每个节点作 1 000 次 SIR 传播仿真，将 1 000 次的平均s_v作为测试网络节点的真实影响力。再根据节点影响力的预测值和真实值的 Kendall Tau 系数评价模型的预测效果。本文方法和其他基准方法的对比结果如表 3 所示。

从表 3 可以看出，本文提出的方法在大部分网络中表现非常好，13 个网络中有 10 个网络都好于对比方法，尤其在 NS 网络中，相比于表现第二好的 DynamicRank 中心性方法，相关系数提升了 0.2456。在平均度比较高 (平均度大于 20) 的网络中，由于训练集中缺少类似的大度点的局部结构，无法学习到大度节点的重要性，极大影响了模型的判断，如在 Polblog 网络中，最大度为 467，远高于训练网络的最大度 216。另一方面，平均度反映网络中常见的局部结构。如训练网络 LastFM 的平均度为 7.294，虽然也存在度为 20 的局部结构，但这种结构在训练网络中并不常见，转换得到的训练集会极不平衡。模型对度为 20 的局部结构无法充分学习，因此模型在度大于 20 的网络表现也就较差。

表 3　不同方法与 SIR 模型仿真结果的 Kendall Tau 相关性系数

网络	H-index	介数中心性	K-shell	度中心性	DynamicRank中心性	本文方法
Router	0.453 1	0.207 9	0.454 0	0.434 2	0.641 1	**0.837 7**
Grid	0.422 0	0.390 0	0.326 0	0.440 6	0.575 0	**0.782 7**
CM	0.592 6	0.304 8	0.577 4	0.561 7	0.602 3	**0.796 9**
Stelzl	0.530 2	0.440 0	0.523 1	0.496 9	0.702 9	**0.891 7**
Vidal	0.587 8	0.491 1	0.588 3	0.555 3	0.730 2	**0.906 2**
Sex	0.526 1	0.444 4	0.534 6	0.496 3	0.739 9	**0.870 4**
NS	0.515 0	0.181 0	0.502 0	0.540 8	0.632 0	**0.877 6**
Figeys	0.670 0	0.585 0	0.668 0	0.652 8	0.793 1	**0.828 7**
Email	0.802 5	0.682 5	0.769 5	0.792 4	0.825 4	**0.925 2**
Hamster	0.734 0	0.579 0	0.717 0	0.742 5	0.778 5	**0.858 2**
Jazz	**0.893 0**	0.502 0	0.759 0	0.884 1	**0.893 0**	0.581 4
Polblog	0.930 7	0.713 5	0.911 2	0.926 9	**0.944 0**	0.754 3
USAir	0.808 1	0.541 9	0.801 4	0.785 7	**0.828 7**	0.619 7

为了验证本文学习模型的鲁棒性，本文在不同感染概率下对模型效果进行了分析。设置 $\beta = c\beta_c$，选取不同 c 值用于分析选取不同传染概率对重要节点挖掘的影响。

如图 3 所示，在不同感染概率 β=β_c、1.5β_c、2β_c、2.5β_c 下，本文利用特征工程的方法提出的特征能够很好地描述节点在网络中的重要性。在不同的感染概率下，本文方法依旧能在低平均度的网络中取得最好的效果。图 3 的结果表明，虽然基于特征工程的方法在训练时依赖于感染概率，但训练得到的重要性评估模型对感染概率并不敏感，适用于对不同感染概率下，节点重要性的挖掘。

进一步，为了验证这 8 个特征的有效性，本文在不同网络上选取不同特征组合进行实验分析。

1) 在 Figeys 网络中去除特征 1 后，算法排序结果和实际仿真排序结果的 Kendall Tau 相关性系数从 0.83 下降到 0.77。

2) 在 NS 网络中去除特征 1 后，Kendall Tau 相关性系数从 0.879 下降至 0.872，若去除特征 2，Kendall Tau 相关性系数下降更为明显，降至 0.861。

3) 若在 Grid 网络中去除特征 2，Kendall Tau 相关性系数从 0.775 下降至 0.728。若再进一步去除特征 7，Kendall Tau 相关性系数大幅降低至 0.688。

4) 在 Stelzl 网络中，若同时去除特征 3 和 7 时，Kendall Tau 相关性系数从 0.89 大幅下降至 0.79。

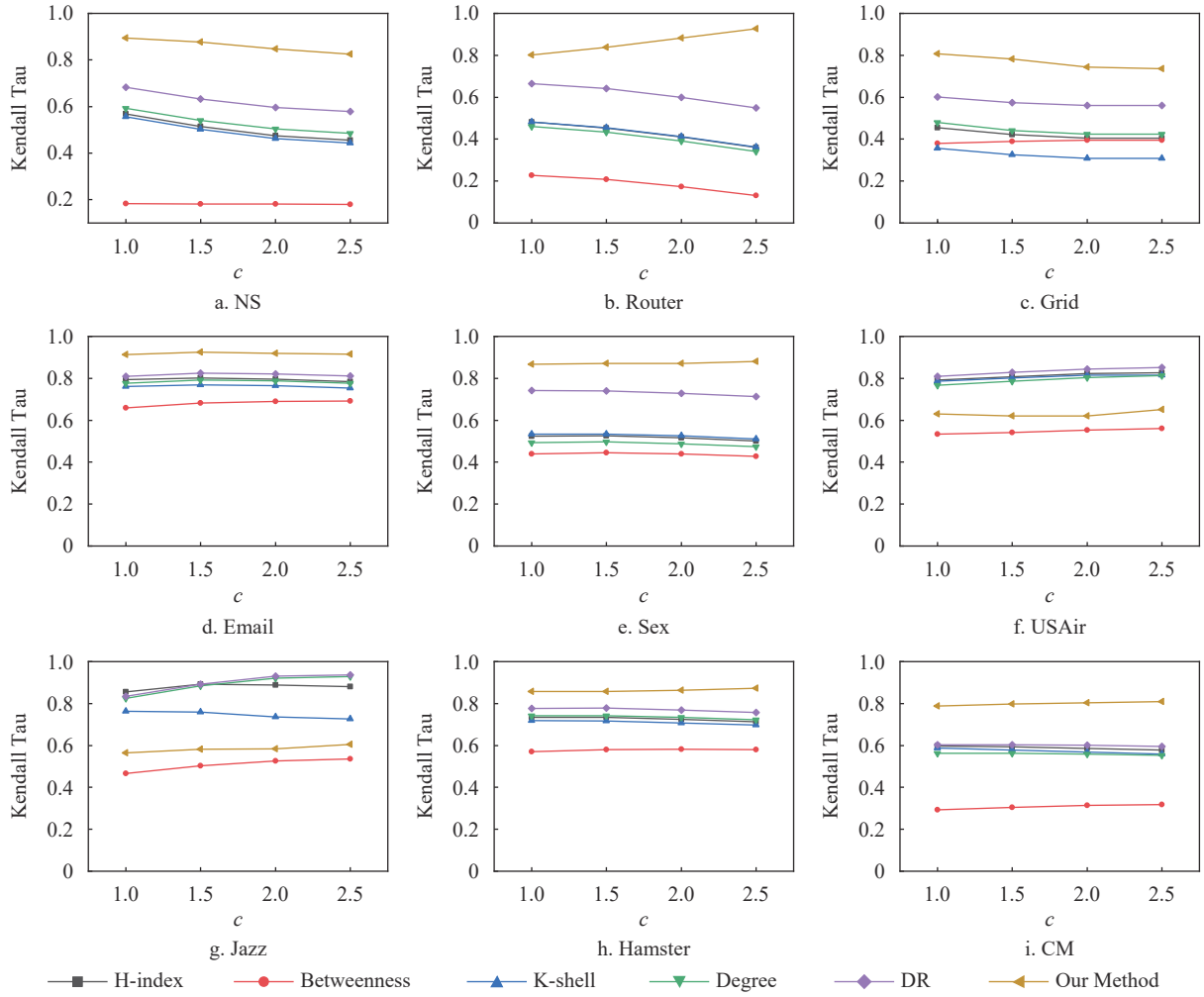


图 3 本文方法与其他基准方法在各网络中不同感染概率下的 Kendall Tau 相关性系数对比

从上面的分析可以看出，8 个特征在不同网络的重要节点排序上相互补充和促进，去掉某个或某组特征，对节点重要性研判将带来直接影响。而完整的 8 个特征，模型更稳定，也能更准确地判断网络中节点的重要性。

同时，根据信息传播理论，节点对三阶邻居以外的影响已经很小，更高阶的邻居信息趋于同质化[30]。为了验证更高阶邻居对模型的影响，根据特征 4-8，拓展三阶邻居的特征 9-13(三阶邻居的度之和、三阶邻居数目、三阶邻居度的基尼系数、三阶外联邻居度之和、三阶外联邻居度的基尼系数)。选取 email 作为测试网络，发现 8 个特征训练所得模型的排序结果与仿真结果的 Kendall Tau 相关性系数为 0.925，而 13 个特征的相关系数为 0.927，提升并不明显。实验表明，选取二阶邻居以内的信息已足够。

3 结 束 语

本文利用特征工程方法对节点的邻居信息进行提取和重构，提取更能反映节点局部结构的特征向量。根据节点的局部结构特征信息，建立了用于挖掘网络中重要节点的机器学习模型。用 13 个实际网络对本文所提方法的有效性进行了测试，并和典型的基准方法进行了对比。实验结果表明，本文提出的机器学习模型能有效地挖掘网络中的重要节点，13 个网络中有 10 个网络的效果显著地优于已有方法。由于本文方法一定程度上依赖于训练网络的局部结构，对于训练数据中出现较少的局部结构，由于训练不充分，在测试时表现出的效果整体欠佳。在未来的研究中，一方面是构建更加丰富多样的训练集，另一方面，需提取更为丰富的局部特征，提升模型的预测能力。近年来，随着深度学习的发展，尤其是图神经网络的研究深入，如何利用图神经网络训练泛化性能更好的复杂网络局部结构特征的表达模型[31]，从而提高重要节点识别的准确率也是一个重要的研究方向。

参 考 文 献

[1] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59(13): 1175-1197.
REN X L, LYU L Y. Review of ranking nodes in complex networks[J]. Science Bulletin, 2014, 59(13): 1175-1197.

[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.
LYU L Y. Link prediction in complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.

[3] 朱军芳, 陈端兵, 周涛, 等. 网络科学中相对重要节点挖掘方法综述[J]. 电子科技大学学报, 2019, 48(4): 595-603.
ZHU J F, CHEN D B, ZHOU T, et al. A survey on mining relatively important nodes in network science[J]. Journal of University of Electronic Science and Technology of China, 2019, 48(4): 595-603.

[4] 赫南, 李德毅, 淦文燕, 等. 复杂网络中重要性节点发掘综述[J]. 计算机科学, 2007, 34(12): 1-5.
HE N, LI D Y, GAN W Y, et al. Mining vital nodes in complex networks[J]. Computer Science, 2007, 34(12): 1-5.

[5] LYU L Y, CHEN D, REN X L, et al. Vital nodes identification in complex networks[J]. Physics Reports, 2016, 650: 1-63.

[6] KITSAK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6: 888-893.

[7] FREEMAN L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1978, 1(3): 215-239.

[8] FREEMAN L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1): 35-41.

[9] FREEMAN L C, BORGATTI S P, WHITE D R. Centrality in valued graphs: A measure of betweenness based on network flow[J]. Social Networks, 1991, 13(2): 141-154.

[10] ESTRADA E, HIGHAM D J, HATANO N. Communicability betweenness in complex networks[J]. Physica A, 2009, 388(5): 764-774.

[11] NEWMAN M. A measure of betweenness centrality based on random walks[J]. Social Networks, 2005, 27(1): 39-54.

[12] DOLEV S, ELOVICI Y, PUZIS R. Routing betweenness centrality[J]. Journal of the ACM, 2010, 57(4): 1-27.

[13] LYU L Y, ZHOU T, ZHANG Q M, et al. The H-index of a network node and its relation to degree and coreness[J]. Nature Communications, 2016, 7(1): 10168.

[14] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[15] LYU L Y, ZHANG Y C, YEUNG C H, et al. Leaders in social networks, the delicious case[J]. PLoS ONE, 2011, 6: e21202.

[16] BRIN S, PAGE L. The Anatomy of a Large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30: 107-117.

[17] ZHANG J X, CHEN D B, DONG Q, et al. Identifying a set of influential spreaders in complex networks[J]. Scientific Reports, 2016, 6(1): 27823.

[18] CHEN D, LYU L Y, SHANG M S, et al. Identifying influential nodes in complex networks[J]. Physica A, 2012, 391: 1777-1787.

[19] MOREENO Y, PASTOR-SATORRAS R, VESPIGNANI A. Epidemic outbreaks in complex heterogeneous networks[J]. The European Physical Journal B, 2002, 26(4): 521-529.

[20] COHEN J, COHEN P, WEST S G, et al. Applied multiple regression/correlation analysis for the behavioral sciences[M]. London: Routledge, 2013.

[21] CHEN D B, SUN H L, TANG Q, et al. Identifying influential spreaders in complex networks by propagation probability dynamics[J]. Chaos, 2019, 29(3): 033120.

[22] CHEN D B, GAO H, LYU L Y, et al. Identifying influential nodes in large-scale directed networks: The role of clustering[J]. PLoS ONE, 2013, 8(10): e77455.

[23] NEWMAN M. Spread of epidemic disease on networks[J]. Physical Review E, 2002, 66(1): 016128.

[24] CASTELLANO C, PASTOR-SATORRAS R. Thresholds for epidemic spreading in networks[J]. Physical Review Letters, 2010, 105(21): 218701.

[25] COHEN R, EREZ K, BEN-AVRAHAM D, et al. Resilience of the internet to random breakdowns[J]. Physical Review Letters, 2000, 85(21): 4626-4628.

[26] SHU P, WANG W, TANG M, et al. Numerical identification of epidemic thresholds for susceptible-infected-recovered model on finite-size networks[J]. Chaos, 2015, 25(6): 063104.

[27] KINGMA D, BA J. Adam: A method for stochastic optimization[EB/OL]. (2014-12-22). https://arxiv.org/abs/1412.6980.

[28] ROZEMBERCZKI B, SARKAR R. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. [S.l.]: ACM, 2020: 1325-1334.

[29] KNIGHT W R. A computer method for calculating Kendall ’s Tau with ungrouped data[J]. Journal of the American Statistical Association, 1966, 61(314): 436-439.

[30] CHRISTAKIS N A, FOWLER J H. Social contagion theory: Examining dynamic social networks and human behavior[J]. Statistics in Medicine, 2013, 32(4): 556-577.

[31] YU E Y, WANG Y P, FU Y, et al. Identifying critical nodes in complex networks via graph convolutional networks[J]. Knowledge-Based Systems, 2020, 198: 105893.

编　辑　蒋　晓