计算机科学
COMPUTER SCIENCE

基于子图特征的节点排序算法

陈端兵, 杨志杰, 曾卓, 傅彦, 周俊临, 赵俊严

引用本文

陈端兵, 杨志杰, 曾卓, 傅彦, 周俊临, 赵俊严. 基于子图特征的节点排序算法[J]. 计算机科学, 2023, 50(11A): 230100122-7.

CHEN Duanbing, YANG Zhijie, ZENG Zhuo, FU Yan, ZHOU Junlin, ZHAO Junyan. Node Ranking Algorithm Based on Subgraph Features [J]. Computer Science, 2023, 50(11A): 230100122-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100122

基于子图特征的节点排序算法

陈端兵[1]　杨志杰[1]　曾　卓[1]　傅　彦[1]　周俊临[1]　赵俊严[2]
1 电子科技大学大数据研究中心　成都 611731
2 北京特种车辆研究所　北京 100072

摘　要　复杂网络理论已被广泛应用于各个领域,节点重要性排序研究是复杂网络领域的重要分支。复杂网络中节点重要性排序及重要节点挖掘对分析和理解复杂网络的结构与功能具有重要意义。众多学者针对复杂网络的关键节点识别和节点重要性排序问题进行了深入研究,取得了大量研究成果。但随着人工智能的发展和数据体量的飞速增长,复杂网络规模呈指数级增长,传统算法的准确性和泛化性已经无法满足现实需求。文中基于节点的二阶邻域信息,提出了一种基于子图特征的节点重要性排序(Subgraph Feature Extraction Rank,SFE Rank)的机器学习模型。利用二阶邻域信息,建立局部子图的含权邻接矩阵,通过矩阵特征分解提取能够有效反映节点局部特征信息的向量表征,在此基础上,建立机器学习模型,用于学习节点子图特征向量和节点重要性的关联关系。在9个真实网络上进行实验,结果表明,相比已有的节点重要性排序方法,所提方法具有更优的排序效果和更好的泛化性能。

关键词:复杂网络;重要节点;特征提取;子图特征;机器学习
中图法分类号　TP301

Node Ranking Algorithm Based on Subgraph Features

CHEN Duanbing[1],YANG Zhijie[1],ZENG Zhuo[1],FU Yan[1],ZHOU Junlin[1] and ZHAO Junyan[2]
1 Big Data Research Center,University of Electronic Science and Technology of China,Chengdu 611731,China
2 Beijing Special Vehicle Institute,Beijing 100072,China

Abstract　Complex network theory has been widely applied in various fields,and node ranking is an important branch in the complex networks. Node ranking and critical node mining are significant for analyzing and understanding the structure and function of complex networks. Many scholars have conducted in-depth researches on critical nodes identification and ranking in complex networks,and have achieved great success. However,with the development of artificial intelligence and rapid growth of data,the size of complex networks grows exponentially. The accuracy and generalization of traditional algorithms can no longer meet the real demand. A machine learning model on node ranking(subgraph feature extraction rank) based on subgraph features of the second-order neighborhood information of nodes is proposed in this paper. A weighted adjacency matrix of local subgraphs is established using the second-order neighborhood information firstly. Then,vector representations that can effectively reflect the local feature of nodes are extracted through matrix feature decomposition. Finally,a machine learning model is established to train the correlation between node's subgraph feature vector and its influence. Experimental results on nine real networks show that the proposed method has better performance and generalization compared with benchmark node ranking methods.

Keywords　Complex networks,Critical nodes,Feature extraction,Subgraph features,Machine learning

1　引言

二阶链路预测是一种基于历史数据的预测方法,指如何通过已知的网络节点以及网络结构等信息预测未来网络链路的状态,即尚未产生连边的两个节点之间产生链接的可能性[1-2]。

复杂网络中,不同节点的重要性通常也不同。重要节点是相比网络中其他节点,能更大程度地影响网络功能的特殊节点。近年来,节点重要性排序研究[3]受到了越来越广泛的关注,不仅因为其重大的理论研究意义,更因为其广泛的实用价值。如在传染病网络中,社交关系较复杂的人感染后能够迅速传播给大量其他易感者,在电力网络中,关键节点短路将导致大量线路断电。由此可见,挖掘网络中的相对重要节点[4]具有重要的应用价值。

节点重要性的一个重要影响因素是节点的邻域结构,根据信息传播的三度理论,二阶网络链路对节点重要性的影响较大。二阶链路预测在网络工作中有着广泛的应用,通过预测链路的状态,可以及时采取相应的措施,进行拥塞控制,以保证网络的稳定性和可靠性;其次通过预测链路的状态,可以提前发现链路的故障和异常情况,并采取相应的措施进行

基金项目:国家自然科学基金重点项目(T2293771);教育部哲学社会科学研究重大课题(21JZD055)
This work was supported by the Major Program of National Natural Science Foundation of China(T2293771) and Key Research Project of Philosophy and Social Sciences of the Ministry of Education(21JZD055).
通信作者:陈端兵(dbchen@uestc.edu.cn)

修复，以保证网络的可靠性；另外，二阶链路预测可以实现自适应网络控制，根据预测的链路状态，自动调整网络参数和拓扑结构，以达到最佳的网络性能和资源利用效率。但是在网络链路预测中存在诸多难点：(1)网络链路状态数据量庞大，同时需要处理多个链路的状态，因此需要处理大量的数据，并进行有效的数据压缩和处理，以保证预测效率；(2)网络链路状态数据存在不确定性和噪声，需要进行有效的数据清洗和预处理，以获得高质量的预测结果；(3)二阶链路预测需要考虑多个变量之间的关系，因此需要采用复杂的算法进行预测，算法的设计和实现难度较大。

复杂网络中节点重要性排序算法大致分为3类：1)基于节点邻居的排序方法，如度中心性、k 壳分解法[6]、H-index[7]等；2)基于路径的排序方法，如接近中心性[8]、介数中心性[9]等；3)基于特征向量的排序方法，如 PageRank[10]、LeaderRank[11]等。在社会网络分析中，节点的重要性也称为“中心性”，其主要观点是节点的重要性等价于该节点与其他节点的连接使其具有的显著性。而度中心性认为一个节点的邻居数目越多，这个节点对整个网络功能的影响越大。但是度中心性仅仅考虑了节点的邻居数目，忽略了节点的位置，倘若一个节点处于网络核心位置，此时即使该节点的度较小，往往也具有较高的影响力。针对这个问题，文献[6]提出了基于 K-壳分解(K-shell Decomposition)的 K-shell 中心性，该方法的基本思想是将外围的节点层层剥去，每一个节点将属于一个特定的壳，壳中所有节点具有相同的 K-shell 值，而 K-shell 值越高的节点具有越高的影响力。文献[12]提出了一种基于特殊邻居节点数目的重要性排序方法。还有一些基于路径排序的方法，如介数中心性[9](Betweenness Centrality)认为任意两节点之间的最短路径经过一个节点的次数越多，这个节点就越重要。在介数中心性基础上，又陆续提出了流介数中心性[13](Flow Betweenness Centrality)、连通介数中心性[14](Communicability Betweenness Centrality)和随机游走介数中心性[15](Random Walk Betweenness Sentrality)等。基于特征向量的方法在考虑邻居节点的数量同时还考虑了节点质量对节点重要性的影响，如 Page Rank[10]和 Leader Rank[11]算法通过模拟用户在网络上浏览网页的过程，使得节点的分值沿着网页之间的访问路径增加。也有一些文献提出了基于网络结构的节点重要性评估方法，如基于复杂网络社区划分的节点重要性排序方法[16]、同时考虑节点自身重要性和其他所有节点影响力的方法[17]、Dynamic Rank 中心性[18]和基于特征工程的重要节点挖掘方法[19]，这类方法的基本思想是通过节点邻域结构特征来计算节点的重要性。另外，文献[20]还通过融合多种方法来实现节点的重要性排序。这些算法基本都是尝试去找到能够反映节点重要性的某种因素或者特征，模型效果通常依赖于特定的网络结构，在某些类型网络上表现较好的算法在其他类型网络可能表现较差。

近年来，学者们对基于机器学习和深度学习的重要节点挖掘模型也开展了大量研究。Yu 等[21]以节点局部结构为特征输入，提出了一种基于图卷积网络的重要节点挖掘深度学习模型。Munoz 等[22]提出了一种处理贷款产品的双层方法，用于挖掘高质量潜在金融产品客户，包括两个分类器：一个用于模拟贷款意向；另一个用于建模贷款资格。Rezaei 等[23]基于数据驱动的机器学习模型，利用网络中少量节点的重要性建立模型，对网络中剩下的节点重要性进行预测排序，此方法对每一个网络都需要单独建立模型，泛化性较弱。Xie 等[24]对 K-shell 网络进行改进，提出了一种新的算法 KBKNR，通过获取每个节点的 K 壳(K-Shell，Ks)值，来衡量复杂网络全局结构的影响。Gu 等[25]在 Leader Rank 算法的基础上利用节点相似度来衡量节点间的相互作用，提出了 SRank 算法来进行重要节点排序。Liu 等[26]在 H 指数和 Ha 指数的基础上，提出了一种 HHa 节点中心性算法。此外，文献[27]还通过奖励矩阵和强化学习来识别关键节点。文献[28]提出了一种路径感知图神经网络(PAGNN)，主要探索了在链路预测任务中如何将节点间的复杂网络结构学习集成到传统的 GNN 计算模型中，从而提高链路预测的准确率。文献[29]针对机会网络的多维链路属性和网络结构动态变化的特点，提出了基于网络表示学习的链路预测方法。文献[30]从特征谱视角出发，分析了复杂网络中链路可预测性与网络拓扑结构之间的关系。目前针对复杂网络关键节点的研究大部分是通过拓扑结构信息来进行重要性排序[6-7,9,21,25,31,33-36]，但是也有部分学者通过网络信息熵[26,32]来计算节点重要性。以上这些研究本质上是通过各种方式来获取复杂网络更多的特征信息，如结构信息、特征谱信息、计算网络信息熵，以及通过 GNN 获得更深层次的特征信息。但是，这些方法存在一定的限制，如需要全局信息进行计算、提取的特征不够丰富以及过于依赖深度学习。而在实际应用中，可能存在无法获取节点全局信息，如新冠疫情中较难获取到完整的传播链、传播网，也存在大规模网络的处理速度慢、需要的计算资源高等问题。

同时，现有的大部分方法更多地依赖于具体的网络结构，从而导致重要节点排序算法只能够适用于某一类特定或者某几类网络，泛化性以及鲁棒性较差。基于以上问题，本文提出了一种基于机器学习的重要节点排序算法，与已有工作不同的是，本文通过节点的局部结构采用机器学习模型自动提取到节点特征，以此来度量节点的重要性。本文通过建立节点重要性和节点局部结构特征之间的关联关系，构建了一种不需要节点全局信息而泛化性能更好以及鲁棒性更强的节点重要性排序模型。首先通过节点二阶邻接矩阵的特征分解提取能够反映此节点局部特征的向量表征，之后通过机器学习模型学习节点的局部特征与节点重要性之间的关联关系，模型的核心是构建能反映节点重要性的节点向量表征，利用少量网络建模，使其应用于各类网络的节点重要性排序。在9个真实网络中，将本文模型 SFE Rank 得到的结果与度中心性、介数中心性[9]、K-shell[6]、H-index[7]、DynamicRank 中心性[18]、HHa[26]、SRank[25]以及基于特征工程的中心性[19]的结果进行了比较。实验结果表明，本文方法能更准确、更有效地识别复杂网络中高影响力节点，排序模型具有更好的泛化性能。

2 基于子图特征的节点排序方法

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，其中 n 和 m 分别代表节点数量和边的数量。为了提取节点子图的结构特征，首先给出邻居的定义。

定义1(二阶邻居) 如果 $\Gamma_1(u)$ 为网络中节点 u 的一阶邻居，那么节点 u 的二阶邻居可以定义为：

$$\Gamma_2(u)=\{v|v\in\Gamma_1(x)\backslash\Gamma_1(u),x\in\Gamma_1(u),v\neq u\} \quad (1)$$

类似地，节点 u 的三阶邻居可定义为：

$$\Gamma_3(u)=\{v|v\in\Gamma_1(x)\backslash(\Gamma_1(u)\cup\Gamma_2(u)),x\in\Gamma_2(u)\} \quad (2)$$

2.1 特征提取

根据信息传播的一般规律，距离目标节点越近的节点或边，对信息传播做出的贡献越大，它们的权值越高，而距离较远的则权值较低。下文给出点到目标节点距离的定义。

定义 2（点到目标节点 u 的距离） 定义点 u 相对于点 u 的距离为 $\lambda(u,u)=0$，对不同于点 u 的点 v，如果 v 是 u 的 d 阶邻居，那么点 v 到 u 的距离 $\lambda(v,u)$ 就等于 d。

定义 3（边权） 边权与该边关联的两个节点到目标节点 u 的距离相关，且一阶邻居之间的连边、一阶邻居和二阶邻居之间的连边以及二阶邻居之间的连边对传播做出的贡献不一样，具体地，定义边 (a,b) 相对于点 u 的权值为：

$$\omega(a,b)=\frac{1}{2^{\lambda(a,u)}\cdot 2^{\lambda(b,u)}} \quad (3)$$

其中，$a,b\in\Gamma_1(u)\cup\Gamma_2(u)\cup\Gamma_3(u)\cup\{u\}$，$a\neq b$。

定义 4（点权） 对于目标节点 u 的邻域节点 $v\in\Gamma_1(u)\cup\Gamma_2(u)\cup\{u\}$，其点权定义为：

$$\theta(v)=\sum_{(v,w)\in E}\omega(v,w) \quad (4)$$

根据信息传播的三度理论，信息在网络中从节点 u 向外传播时，节点 u 对其三阶邻居的影响就已经很小了，本文在计算 u 的影响力时，仅利用节点 u 及其一阶和二阶邻居构建邻域子图，并依据边权和点权获得其含权邻接矩阵 $adjMatrix$。为了统一特征数目，将每一个节点 u 的邻接矩阵都转为大小为 $k\times k$ 的矩阵，若邻接矩阵规模小于 $k\times k$，则在原矩阵中添加值全为 0 的若干行和若干列，将其补全为一个 $k\times k$ 的矩阵，若邻接矩阵规模大于 $k\times k$，则按如下规则去掉子图中的部分节点：

1）优先去掉二阶邻居中的节点，若邻接矩阵规模还大于 $k\times k$，再从一阶邻居中选择节点去掉。

2）同一层中去节点时，优先去除点权小的节点，使其得到规模为 $k\times k$ 的矩阵。

之后将得到的 $k\times k$ 矩阵进行 SVD 分解，并将其特征值形成的向量作为节点 u 的特征。对于一个秩为 r 的矩阵 $A_{k\times k}$，必然存在 $k\times k$ 的正交矩阵 $U_{k\times k}$，$V_{k\times k}$ 和 $S_{k\times k}$，使得：

$$A_{k\times k}=U_{k\times k}S_{k\times k}V_{k\times k}^{\mathrm{T}}=U_{k\times k}\begin{pmatrix}D_{r\times r}0\\0\quad 0\end{pmatrix}_{k\times k}V_{k\times k}^{\mathrm{T}} \quad (5)$$

其中，$D_{r\times r}$ 为对角矩阵：

$$D_{r\times r}=\begin{pmatrix}\sqrt{b_1}&&&\\&\sqrt{b_2}&&\\&&\ddots&\\&&&\sqrt{b_r}\end{pmatrix}_{r\times r} \quad (6)$$

其主对角线上的值 $\sqrt{b_1},\sqrt{b_2},\cdots,\sqrt{b_r}$ 满足从大到小的排列，称为矩阵 $A_{k\times k}$ 的正奇异值，而 $\sqrt{b_1},\sqrt{b_2},\cdots,\sqrt{b_r},0_{r+1},0_{r+2},\cdots,0_k$ 称为矩阵 $A_{k\times k}$ 的奇异值，也是最终提取到的子图特征。

算法 1 子图特征提取算法（Subgraph Feature Extraction Algorithm）

输入：节点 u 的二阶邻居子图矩阵 Matrix

输出：节点 u 的子图特征

1. n·n←Matrix.size;

2. for Point v In SubGraph do

3. $\theta(v)=\sum_{(v,w)\in E}\omega(v,w)$;

4. 将矩阵 Matrix 的行列按照 θ 进行降序排序；

5. if n>k then

6. 将矩阵 Matrix 裁剪为 k·k；

7. else

8. 将矩阵用 0 补到 k·k；

9. $A_{k\cdot k}=U_{k\cdot k}\cdot S_{k\cdot k}\cdot V_{k\cdot k}^{T}=U_{k\cdot k}\cdot\begin{pmatrix}D_{r\cdot r}&0\\0&0\end{pmatrix}$;

10. 得到 $D_{r\cdot r}=\begin{pmatrix}\sqrt{b_1}&&&\\&\sqrt{b_2}&&\\&&\ddots&\\&&&\sqrt{b_r}\end{pmatrix}$;

11. feature=$\sqrt{b_1},\cdots,\sqrt{b_r},0_{r+1},\cdots,0_k$;

12. 返回 feature.

2.2 节点重要性学习模型

对于节点 u，节点的重要性和节点周围的局部结构有着紧密的关系。本文根据式(4)和式(5)的矩阵分解得到描述节点 u 的特征向量表征 $\boldsymbol{t}=[\sqrt{b_1},\sqrt{b_2},\cdots,\sqrt{b_r},0_{r+1},0_{r+2},\cdots,0_k]^{\mathrm{T}}$，采用线性回归模型对节点局部特征与节点重要性进行建模。定义一个线性回归函数 $f:\boldsymbol{t}\to s$，将节点局部特征向量映射为节点的相对重要性，具体可以表示为：

$$f_u(\boldsymbol{t})=\boldsymbol{w}\cdot\boldsymbol{t}+b \quad (7)$$

其中，$\boldsymbol{w}$ 为特征向量的权重向量；b 是误差项。为了获取节点真实的重要性，本文通过基于传播动力学的 SIR 模型进行仿真。具体地，每个已感染的节点 u 与易感邻居节点 v 接触时，v 将以概率 β 被感染，即从易感状态变为感染状态。同时，u 将以概率 γ 恢复，即从感染状态变为恢复状态。为简单起见，本文中 γ 设置为 1，即传播 1 次之后立即变为恢复状态。为了量化目标节点 u 的传播影响，以节点 u 为唯一的传播源开始向外传播，当整个子图中不存在任何的感染节点时，认为在该子图中传播达到稳态，记此时恢复状态的节点数目为 R_u，即节点 u 的重要性为 R_u。

在 SIR 传播过程中，如果感染概率 β 很低，那么信息几乎传播不出去，反之，若感染概率 β 很高，那么整个网络中的节点都会被感染，因此过大或者过小的 β 得到的传播影响范围不具区分性，各种排序方法都将失效。本文利用平均场理论，根据真实网络的平均度及度平方平均值计算 SIR 模型中感染概率的阈值 $\beta_c\approx\langle k\rangle/(\langle k^2\rangle-\langle k\rangle)$，依据这个阈值进行仿真获得每个节点的真实重要性。为了消除随机波动的影响，本文采用 100 次独立仿真实验结果的平均值 s_u 作为节点 u 的真实重要性。

至此，获得了节点 u 的特征 $\boldsymbol{t}[\sqrt{b_1},\sqrt{b_2},\cdots,\sqrt{b_r},0_{r+1},0_{r+2},\cdots,0_k]^{\mathrm{T}}$ 和真实重要性 s_u，采用回归模型，选取均方误差（Mean Square Error，MSE）作为优化目标，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：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \quad (8)$$

其中，$\hat{y}_i$ 为第 i 个节点的预测值，y_i 为第 i 个节点的真实值（本文采用 SIR 仿真得到）。

2.3 模型训练

本文选用 Twin 作为训练网络对节点重要性排序模型

进行训练学习。Twin是一个2017年10月从维基百科收集的孪生城市关系连接网络，节点代表的是城市，边代表的是孪生关系，其节点数量为14274，边数量为20573，最大度为99。首先从Twin网络中提取节点的特征向量；同时将Twin中的每个节点当做传播源，进行100次独立的SIR传播仿真，将得到的平均值 s_u 作为每个节点的标签值；最后，将特征向量和标签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。

为了获得模型最优的回归系数，本文采用Adam优化器优化目标函数。并且因为用来训练的网络信息比较丰富，节点种类繁多，为了能够获得相对丰富的局部信息和较快的训练推理速度，因此特征数量 k 设置为了100。

3 实验和讨论

本文用9个不同类型的真实网络对本文方法进行测试，并将通过本文方法得到的结果与度中心性、介数中心性[9]、K-shell中心性[6]、H-index中心性[7]、基于特征工程的中心性[19]和DynamicRank[18]中心性所得结果进行了对比。

3.1 评估指标

为了评估各方法挖掘网络中重要节点的能力，采用了Kendall Tau系数来评估节点重要性预测排序 $X=(x_1, x_2, \cdots, x_n)$ 与真实排序 $Y=(y_1, y_2, \cdots, y_n)$ 的相关性。对序列中任意两个位置 i, j，如果 $x_i>x_j$ 且 $y_i>y_j$ 或者 $x_i<x_j$ 且 $y_i<y_j$ 时，则认为序列中这两个数对是和谐的，和谐数对数目记为 n_+。而如果 $x_i>x_j$ 且 $y_i<y_j$，或者 $x_i<x_j$ 且 $y_i>y_j$ 时，则认为序列中这两个数对是不和谐的，不和谐数对数目记为 n_-。如果存在 $x_i=x_j$ 或 $y_i=y_j$，则这个数对不属于两者中的任何一种。对网络中 $n(n-1)/2$ 个序对进行比较，可以得到两个序列的Kendall Tau系数：

$$\tau=\frac{2(n_+-n_-)}{n(n-1)} \tag{9}$$

这个系数值在[-1,1]的范围内。如果序列 X 和序列 Y 不相关，那么 τ 趋近于0。如果两个序列为负相关则为负值，且负相关性越强，越接近于-1，反之如果系数值越接近1，则说明两个序列的相关性越强，排序结果越一致。

3.2 数据集

为了保证本文方法的普适性，本文采用的9个真实网络的规模、平均度及聚集系数等参数不尽相同：包含了规模较大的网络（如Cond-Mat，CM），也包含了规模较小的网络（如Jazz），这9个真实网络的平均度介于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关系网络；3）Email是Rovirai Virgili大学成员之间电子邮件交换网络；4）Sex是研究男女性伙伴的网络；5）Polblogs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）Cond-Mat（CM）是1995－1999年arXiv出版物的科学家合作网络；8）Grid是美国西部的某电力网络；9）Hamster是一个包含网站用户之间的友谊和家庭关系的网络。以上数据集可以从网站（http://konect.cc/networks/）获得，这9个真实网络的基本特征数据如表1所列，其中 n 是节点数目，m 是边数目，$\langle k\rangle$ 是所有节点的平均度，σ 是所有节点的聚集系数的方差，$\langle c\rangle$ 是所有节点的平均聚集系数。

表1 9个真实网络的基本特征数据

Table 1 Basic feature data of nine real networks

网络	n	m	$\langle k\rangle$	σ	$\langle c\rangle$
Grid	4941	6594	2.669	0.050	0.1030
CM	27519	116181	3.030	0.136	0.6300
Sex	16730	39044	4.700	0	0
NS	379	914	4.820	0.119	0.3700
Email	1133	5451	9.620	0.058	0.1100
Hamster	2426	16631	13.700	0.135	0.5376
Jazz	198	2742	27.690	0.041	0.5200
Polblogs	1224	19025	31.08	0.060	0.2260
USAir	1574	28236	35.880	0.128	0.3800

3.3 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的所有节点进行100次SIR传播仿真，其中传播概率 $\beta=1.5\beta_c$，将100次的SIR仿真结果平均值记为 s_u，作为测试网络节点的真实影响力。根据节点的子图特征（子图邻接矩阵大小取100×100），通过线性回归模型预测得到影响力的预测值，再通过计算所有节点影响力的预测值和真实值两个序列的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2所列。从表2可以看出，本文提出的方法在大多数网络上的表现都优于对照实验中的其他方法。

表2 不同方法与SIR模型仿真结果的Kendall Tau相关性系数

Table 2 Kendall Tau correlation coefficients of simulation results between different methods and SIR models

网络	度中心性	介数中心性[9]	h-index[7]	k-shell[6]	HHa[26]	KBKNR[25]	Dynamic Rank[18]	特征工程[19]	SFE Rank
Email	0.7780	0.6318	0.8057	0.5341	0.829	—	0.8861	0.8674	**0.9014**
Hamster	0.7208	0.5728	0.7216	0.2437	—	—	0.8702	0.8510	**0.8784**
Jazz	0.7696	0.4540	0.7964	0.4832	0.587	0.7764	0.8215	0.6876	**0.8659**
Polblogs	0.8213	0.6659	0.8385	0.4498	—	0.8152	0.8529	0.7543	**0.8560**
Grid	0.5766	0.4183	0.5972	0.1111	0.542	0.5067	0.8029	0.7827	**0.8064**
USAir	0.7269	0.5434	0.7534	0.5049	0.834	—	**0.8485**	0.6197	0.8276
NS	0.7125	0.3730	0.6970	0.2934	0.713	0.7741	0.8900	0.8747	**0.9157**
CM	0.6649	0.3758	0.7031	0.3794	—	—	0.8377	0.7969	**0.8547**
Sex	0.6045	0.5315	0.6358	0.1290	—	—	**0.8115**	0.7874	0.7922

通过表1和表2可以发现，当数据集的聚集度方差过低时，本文方法的结果并不是很好，如Sex数据集。在Sex数据集中，聚集系数方差和平均值均为0，且对4个网络提取到的特征通过热力图进行观察对比，发现sex网络中每个点通过子图提取到的特征基本一致，而其他网络中节点的特征分布差异较大，如图1所示。对于线性回归而言，特征高度相似时预测结果也会相近，此时得到的结果大量地集中在某个值附近，从而导致节点重要性排序出现较大

偏差。同时，为了验证本文方法的鲁棒性，本文在不同感染概率下对模型效果进行了分析。设置 $\beta=c\beta_c$，选用不同的 c 值用于分析不同感染概率对重要节点挖掘的影响。如图 2 所示，在不同感染概率 $\beta=\beta_c$，$1.5\beta_c$，$2\beta_c$，$2.5\beta_c$ 下，本文通过子图特征的方法能够很好地描述节点在网络中的重要性，在不同结构的网络中都取得了较好的效果。图 2 所示的结果表明，虽然基于子图特征的方法在训练时依赖于感染概率，但是训练得到的重要性评估模型在预测过程中对感染概率并不敏感，适用于不同感染概率下节点重要性的挖掘。

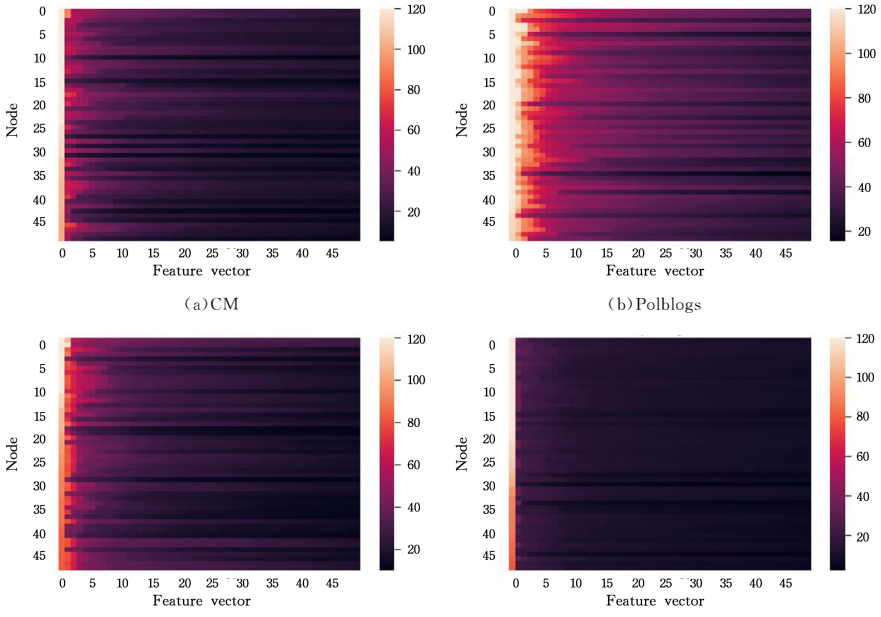

图 1　4 个真实网络中任意 100 个点的特征热力图

Fig. 1　Heat map of any 100 points' feature in four real networks

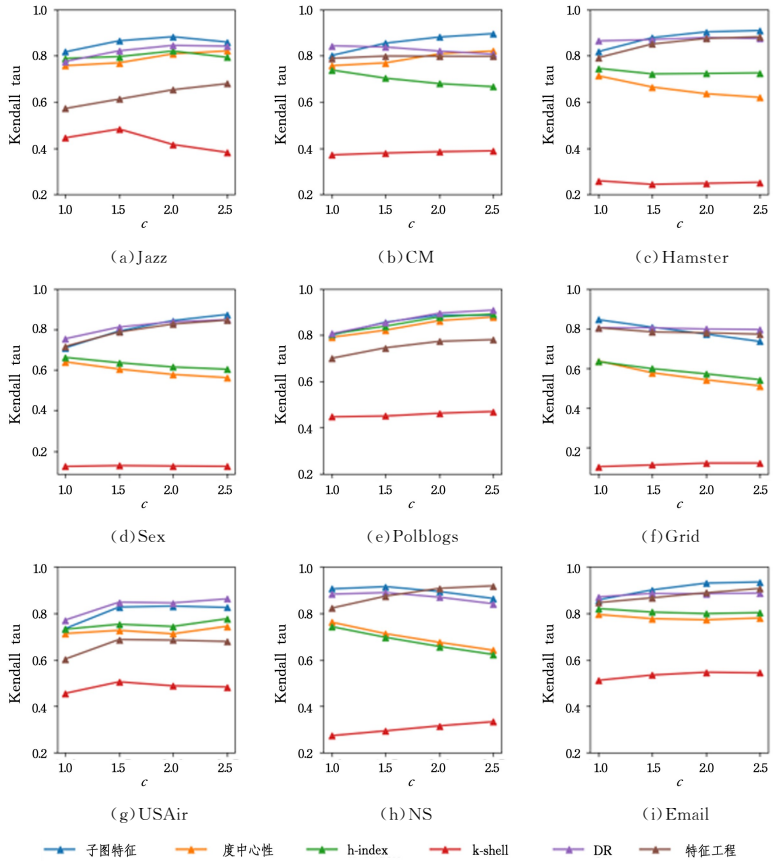

图 2　本文方法与其他基准方法在不同感染概率下 Kendall Tau 相关系数的对比

Fig. 2　Comparison of Kendall Tau correlation coefficients for different infection probabilities between our method and other benchmark methods

结束语 本文通过对节点邻域子图的含权邻接矩阵进行特征分解，来得到能够反映节点局部结构的特征向量。根据节点两步可达子图的加权邻接矩阵进行奇异值分解之后得到的特征向量和SIR仿真传播规模，建立了用于复杂网络中节点重要性排序的机器学习模型。在9个真实网络上对本文方法进行了测试，并与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，9个网络中有7个网络的效果优于其他基准方法，另外2个未达到最优值的测试网络也仅仅略低于最优值。未来的研究方向中，一方面可以研究如何解决在本文实验中因为平均聚集度和聚集度方差较低而导致的效果下降，另一方面是可以尝试更优的边权计算方式。同时，目前通过子图提取特征的方式及线性回归模型虽然获得了较好的效果，但是线性回归模型不能充分利用图特征向量，使用局部邻接矩阵也是固定大小，后续如何更为充分地利用图特征向量，利用图卷积建立节点 u 的 k 阶邻居内所有节点的压缩向量表征并将其作为节点 u 的特征，从而建立更为通用和广泛的重要节点排序模型，也是一个很重要的研究方向。

参 考 文 献

[1] LÜ L, ZHOU T. Link prediction in complex networks: A survey [J]. Physica A: statistical mechanics and its applications, 2011, 390(6): 1150-1170.

[2] LU L Y, ZHOU T. Link Prediction[M]. Beijing: Higher Education Press, 2013.

[3] REN X L, LU L Y. Review of ranking nodes in complex networks[J]. Science Bulletin, 2014, 59(13): 1175-1197.

[4] 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.

[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.

[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] LÜ L, 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): 1-7.

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

[9] FREEMAN L C. A Set of Measures of Centrality Based on Betweenness[J]. Sociometry, 1977, 40(1): 35-41.

[10] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN systems, 1998, 30(1-7): 107-117.

[11] LÜ L, ZHANG Y C, YEUNG C H, et al. Leaders in social networks, the delicious case[J]. PLoS One, 2011, 6(6): e21202.

[12] ZHAO N, LI J, WANG J, et al. Relatively important nodes mining method based on neighbor layer diffuse[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(1): 121-126.

[13] 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.

[14] ESTRADA E, HIGHAM D J, HATANO N. Communicability betweenness in complex networks[J]. Physica A Statistical Mechanics & Its Applications, 2009, 388(5): 764-774.

[15] NEWMAN M. A measure of betweenness centrality based on random walks[J]. Social Networks, 2003, 27(1): 39-54.

[16] WANG A. Research on node importance evaluation method based on complex network structure features[D]. Beijing: People's Public Security University of China, 2020.

[17] ZHAO J, WANG Y, DENG Y. Identifying influential nodes in complex networks from global perspective[J]. Chaos, Solitons & Fractals, 2020, 133: 109637.

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

[19] PAN K, YIN C L, WANG L, et al. Identifying critical nodes based on feature engineering[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(6): 930-937.

[20] XUE Y, DENG Y. Entailment for intuitionistic fuzzy sets based on generalized belief structures[J]. International Journal of Intelligent Systems, 2020, 35(6): 963-982.

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

[22] MUNOZ J, REZAEI A A, JALILI M, et al. Deep learning based bi-level approach for proactive loan prospecting[J]. Expert Systems with Applications, 2021, 185: 115607.

[23] REZAEI A A, MUNOZ J, JALILI M, et al. Machine learning-based approach for vital node identification in complex networks [J]. Expert Systems with Applications, 2023, 214: 119086.

[24] XIE L X, SUN H H, YANG H Y, et al. Key node recognition in complex networks based on the K-shell method[J]. J Tsinghua Univ(Sci & Technol), 2022, 62(5): 849-861.

[25] GU Y R, ZHU Z Y. Node ranking in complex networks based on LeaderRank and modes similarity[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(2): 441-448.

[26] LIU J C, MA T C, YUE M L. HHa Centrality Algorithm: A Node Centrality Algorithm Based on the H-Index and Ha-Index [J]. Library and Information Service, 2021, 65(20): 92-100.

[27] WANG L M. Identification of vital nodes in complex networks based on deep reinforcement learning[D]. Bengbu: Anhui University of Finance and Economics, 2020.

[28] YANG S, HU B, ZHANG Z, et al. Inductive link prediction with interactive structure learning on attributed graph[C]// Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, 2021: 383-398.

[29] LIU L L, SONG X Y, CHEN Y B. Link prediction in opportunistic networks based on networks representation learning[J]. Journal of Beijing University of Posts and Telecommunications, 2022, 45(4): 64-69.

[30] TAN S Y,QI M Z,WU J,et al. Link predictability of complex network from spectrum perspective[J]. Acta Physica Sinica,2020,69(8):188-197.

[31] HUANG H B,YANG L M,WANG J X,et al. Identification technique of essential nodes in protein networks based on combined parameters[J]. Acta Automatica Sinica,2008,34(11):1388-1395.

[32] HU G,NIU Q,XU L P,et al. The model to analyses of node importance order structure evolution based on network hyperlink information entropy[J]. Acta Electronica Sinica,2022,50(11):2638-2644.

[33] ZHENG W P,WU Z K,YANG G. A novel algorithm for identifying critical nodes in networks based on local centrality[J]. Journal of Computer Research and Development,2019,56(9):1872-1880.

[34] LUO J,YAN G H,ZHANG M,et al. Research on node importance fused multi-information for multi-relational social networks[J]. Journal of Computer Research and Development,2020,57(5):954-970.

[35] GAO C,JIANG S H,WANG Z,et al. A novel method to identify influential stations based on dynamic passenger flows[J]. Scientia Sinica Informationis,2021,51(9):1490-1506.

[36] ZHOU M Y,WU X Y,CAO Y,et al. A novel method to identify multiple influential nodes in complex networks[J]. Scientia Sinica Informationis,2019,49(10):1333-1342.

CHEN Duanbing, born in 1971, Ph.D, professor. His main research interests include big data mining and complex networks.