计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800026

# 基于拓扑相似和 XGBoost 的复杂网络链路预测方法

龚追飞　魏传佳
浙江工业大学计算机科学与技术学院　杭州 310023

**摘　要**　为了提高复杂网络链路预测的性能,采用拓扑相似和 XGBoost 算法来完成复杂网络链路预测。利用复杂网络拓扑结构建立邻接矩阵,求解共同邻居集合,然后根据拓扑相似理论计算复杂网络相似得分函数,将各个时间窗的得分函数和权重参数作为输入,采用 XGBoost 算法实现复杂网络的链路预测。通过差异化设置 XGBoost 算法的两个正则化系数,测试其对链路预测准确率的影响,获取最优正则化系数,从而得到稳定的 XGBoost 链路预测模型。实验证明,时间窗数量设置合理的情况下,相比常用网络链路预测算法,基于拓扑相似和 XGBoost 算法的预测准确率优势明显,且预测时间性能和其他算法的差距较小,尤其适用于大规模的复杂网络链路预测。

**关键词**:复杂网络;链路预测;拓扑相似;XGBoost 算法;时间窗;正则化
**中图法分类号**　TP391

## Complex Network Link Prediction Method Based on Topology Similarity and XGBoost

GONG Zhui-fei and WEI Chuan-jia
College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China

**Abstract**　In order to improve the performance of complex network link prediction,topology similarity and XGBoost algorithm are used to complete link prediction in complex network. According to the topological structure of complex network,the adjacency matrix is established to solve the common neighbor set. Then the similarity score function of complex network is calculated according to the topological similarity theory. The score function and weight parameters of each time window are taken as input,and XGBoost algorithm is used to realize the link prediction of complex network. By setting two regularization coefficients of XGBoost algorithm through differentiation,the influence on link prediction accuracy is tested,and the optimal regularization coefficient is obtained,thus a stable XGBoost link prediction model is obtained. The experimental results show that,compared with the common network link prediction algorithms,the prediction accuracy based on topology similarity and XGBoost algorithm has obvious advantages,and the prediction time performance is smaller than other algorithms,especially suitable for large-scale complex network link prediction.

**Keywords**　Complex network,Link prediction,Topology similarity,XGBoost algorithm,Time window,Regularization

近年来,关于异构网络的网络拓扑分析及链路预测成为热点,根据网络历史状态,分析网络未来拓扑及链路,解析复杂网络的生成规律。根据节点的链路变化可以完成节点的有效聚类分析,并追溯节点的通信线路,分析节点在网络中的价值数据[1],从而实现对整个复杂网络的数据挖掘,这种从局部到整体的研究方式能有效提高复杂网络的研究深度。通过对网络的链路预测,可以有效分析网络拓扑结构,重构复杂网络的生成过程[2],挖掘复杂网络运行的内部机制和规律。当前的网络入侵防御、在线学习精准推荐、导航路线预测等都体现了链路预测的重要作用。

当前,关于复杂网络链路预测的方法较多,Wang 等[3]采用共同邻居(Common Neighbor,CN)算法进行复杂网络链路预测,主要思路是根据共同邻居结构计算节点间的相似性,结果显示出较高的预测准确率。Chen 等[4]采用基于局部路径(Local Path,LP)的距离矩阵计算复杂网络节点的距离,根据节点距离关系预测节点链路,从而有效去除重复路径,预测结果有一定的提升。Yang 等[5]提出基于网络嵌入与局部合力的社区划分算法。Wang 等[6]综述了特征分类的链路预测方法。但是,上述方法都是仅基于网络拓扑结构的改进方案,当复杂网络链路的规模较大时,其预测准确率均不够理想。

XGBoost 作为一种基于 Boosting 算法的集成强分类器,与基于网络拓扑结构的链路预测具有较强的互补性。Wu 等[7]提出了基于 Boost 集成学习分类器的链路预测优化算法。Mo[8]用信度函数评价节点的重要性。因此,本文尝试将

到稿日期:2020-08-04　返修日期:2020-09-21
基金项目:国家自然科学基金(61773348);浙江省自然科学基金(LY17F030016)
This work was supported by the National Natural Science Foundation of China(61773348) and Natural Science Foundation of Zhejiang Province,China(LY17F030016).
通信作者:龚追飞(793688937@qq.com)

拓扑相似原理和 XGBoost 算法相结合进行复杂网络链路预测，旨在进一步提高链路预测准确率，并有效优化大规模复杂网络节点的链路预测时间性能。

## 1 拓扑相似分析

### 1.1 共同邻居结构

设无向图 $G_1=(V,E_1)$（见图 1(a)）有 $V$ 个顶点和 $E_1$ 条边，$G_2=(V,E_2)$（见图 1(b)）有 $V$ 个顶点和 $E_2$ 条边。

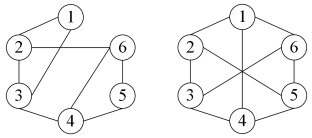


图 1 无向图结构

Fig. 1 Undirected graph structure

$\Gamma_1(i)$ 和 $\Gamma_2(i)$ 分别表示 $V_i$ 在图 1 中的邻居集合，顶点 $V_i$ 和 $V_j$ 共同邻居求解方法为[9]：

$$CN(i,j)=|\Gamma_1(i)\bigcap\Gamma_1(j)|+|\Gamma_1(i)\bigcap\Gamma_2(j)|+|\Gamma_2(i)\bigcap\Gamma_1(j)|+|\Gamma_2(i)\bigcap\Gamma_2(j)| \tag{1}$$

其中，$\Gamma_1(1)=\{2,3\}$，$\Gamma_1(2)=\{1,3,6\}$，$\Gamma_2(1)=\{2,4,6\}$ 和 $\Gamma_2(2)=\{1,3,5\}$。可以求解顶点 1 和 2 的共同邻居，$CN(1,2)=1+1+1+0=3$。

扩展到多图 $(G_1,G_2,\cdots,G_n)$ 中：

$$CN(i,j)=\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}|\Gamma_\alpha(i)\bigcap\Gamma_\beta(j)| \tag{2}$$

若节点 $v_k$，$v_k\in\Gamma_\alpha(i)\bigcap\Gamma_\beta(j)$，则 $v_k$ 是 $V_i$ 和 $V_j$ 在图 $G_\alpha$ 和 $G_\beta$ 上的共同邻居。

### 1.2 拓扑相似数学表示

设网络中 $V_i$ 和 $V_j$ 两点在 $t$ 时的相似程度为 $Sim_t(i,j)$，计算方法为：

$$Sim_t(i,j)=\sum_{\alpha=1}^{t}\sum_{\beta=1}^{t}K(\alpha,\beta)|\Gamma_\alpha(i)\bigcap\Gamma_\beta(j)| \tag{3}$$

$K(\alpha,\beta)$ 为核函数，根据时间可知离 $t$ 时刻越近，核函数越大；离 $t$ 时刻越远，核函数越小，采用指数衰减法。

$$Sim_t(i,j)=\sum_{\alpha=1}^{t}\sum_{\beta=1}^{t}\lambda^{2t-\alpha-\beta}|\Gamma_\alpha(i)\bigcap\Gamma_\beta(j)| \tag{4}$$

其中，$\lambda$ 表示每个时刻图对下一时刻的影响参数，如 $\lambda\in(0,1]$，$\alpha,\beta<t$。

相似性预测分值[10]为：

$$Score(i,j)=\sum_{\alpha=1}^{T}\sum_{\beta=1}^{T}\lambda^{2T-\alpha-\beta}|\Gamma_\alpha(i)\bigcap\Gamma_\beta(j)| \tag{5}$$

为了方便计算 $|\Gamma_\alpha(i)\bigcap\Gamma_\beta(j)|$ 的值，引入邻接矩阵方法，根据邻接矩阵 $\boldsymbol{A}=\{a_{ij}\}_{N*N}$，其中 $a_{ij}=\begin{cases}1, & (v_i,v_j)\in E\\ 0, & (v_i,v_j)\notin E\end{cases}$，对图 1(a)进行邻接矩阵建立，可以得到 $a_{12}=1$，$a_{13}=1$，$a_{21}=1$，$a_{23}=1$，$a_{26}=1$，$a_{31}=1$，$a_{32}=1$，$a_{34}=1$，$a_{43}=1$，$a_{45}=1$，$a_{46}=1$，$a_{54}=1$，$a_{56}=1$，$a_{62}=1$，$a_{64}=1$，$a_{65}=1$，其他节点的边为 0，建立邻接矩阵结构为：

$$\begin{bmatrix}0&1&1&0&0&0\\1&0&1&0&0&1\\1&1&0&1&0&0\\0&0&1&0&1&1\\0&0&0&1&0&1\\0&1&0&1&1&0\end{bmatrix}$$

根据邻接矩阵，式(5)可以表示为[11]：

$$\lambda^{2T-\alpha-\beta}|\Gamma_\alpha(i)\bigcap\Gamma_\beta(j)|=\sum_{k=1}^{n}e^{2T-\alpha-\beta}m_{ik}^{(\alpha)}m_{kj}^{(\beta)} \tag{6}$$

$$Score(i,j)=\sum_{\alpha=1}^{T}\sum_{\beta=1}^{T}\sum_{k=1}^{n}e^{2T-\alpha-\beta}m_{ik}^{(\alpha)}m_{kj}^{(\beta)} \tag{7}$$

## 2 基于 XGBoost 算法的链路预测

### 2.1 XGBoost 算法

对样本集 $D=\{(x_i,y_i)\}(|D|=n,x_i\in R^m,y_i\in R)$ 构建树型结构，$x_i$ 和 $y_i$ 分别为特征向量与标签，合并子树的训练结果得到[12]：

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i),f_k\in F \tag{8}$$

其中，$F=\{f(x)=\omega_{q(x)}\}(q:R^m\to T,\omega\in R^T)$ 表示所有子树，$q$ 为叶子节点，$T$ 为叶子总量，每棵树 $f_k$ 的 $q$ 所占权重为 $\omega$。通过各子树建立目标函数：

$$Obj(\Theta)=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{9}$$

其中，$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\|W\|^2$ 表示正则项[13]，$\gamma$ 和 $\lambda$ 均为正则参数，$l(y_i,\hat{y}_i)=\begin{cases}1, & y_i=\hat{y}_i\\ 0, & y_i\neq\hat{y}_i\end{cases}$。

在 Boosting 算法中，每次预测结果均与上一次预测结果相关，具体参考式(10)。

$$\begin{aligned}&\hat{y}_i^{(0)}=0\\&\hat{y}_i^{(1)}=f_1(x_i)=\hat{y}_i^{(0)}+f_1(x_i)\\&\hat{y}_i^{(2)}=f_1(x_i)+f_2(x_i)=\hat{y}_i^{(1)}+f_2(x_i)\\&\cdots\\&\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i)\end{aligned} \tag{10}$$

其中，$\hat{y}_i^{(t)}$ 表示第 $t$ 次的训练结果，将式(10)代入式(9)可得到：

$$\begin{aligned}Obj^{(t)}&=L^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{i=1}^{t}\Omega(f_i)\\&=\sum_{i=1}^{n}l((y_i,\hat{y}_i^{(t-1)})+f_i(x_i))+\Omega(f_t)+constant\end{aligned} \tag{11}$$

对式(11)进行泰勒展开[14]：

$$Obj^{(t)}=\sum_{i=1}^{n}[l(y_i,\hat{y}_i^{(t-1)})+g_if_t(x_i))+\frac{1}{2}h_if_t^2(x_i)]+\Omega(f_t)+constant \tag{12}$$

其中，$g_i=\partial_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$，$h_i=\partial^2_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$ (13)

优化 *constant* 项后得：

$$Obj^{(t)}=\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\Omega(f_t) \tag{14}$$

其中，$\Omega(f_t)$ 正则项可以变换为[15]：

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{15}$$

因为 $\gamma$ 和 $\lambda$ 为复杂度参数，$\gamma$ 决定 XGBoost 的树是否继续分叉，而 $\lambda$ 控制着正则化的权重，所以在设置参数值时需要根据实际情况而定。

对式(14)进一步进行泰勒展开得到[16-17]：

$$Obj=\sum_{i=1}^{n}[g_iw_{q(x_i)}+\frac{1}{2}h_iw_{q(x_i)}^2]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$$

$$=\sum_{j=1}^{T}[(\sum_{i\in I_j}g_i)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^2]+\gamma T \tag{16}$$

为了继续求解式(16)，设定：

$$G_j=\sum_{i\in I_j}g_i \tag{17}$$

$$H_j=\sum_{i\in I_j}h_i \tag{18}$$

根据式(17)和式(18)，可将式(16)转化为：

$$Obj^{(t)}=\sum_{j=1}^{M}[G_jw_j+\frac{1}{2}(H_j+\lambda)w_j^2]+\gamma M \tag{19}$$

根据式(19)可获得第 $j$ 个叶子的最优值 $w_j^*$ 和目标函数的最优解 $obj^*$。

$$w_j^*=-\frac{G_j}{H_j+\lambda} \tag{20}$$

$$obj^*=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{21}$$

求得所有叶子的最终权重之后，获得最优决策树结构，确定稳定的 XGBoost 复杂网络链路模型。

### 2.2 基于拓扑相似和 XGBoost 的链路预测流程

根据 1.2 节和 2.1 节的拓扑相似和 XGBoost 算法，将拓扑相似得分函数作为 XGBoost 求解的目标函数，以每个时间窗的拓扑相似得分数据和权重参数作为输入变量，经过 XGBoost 多次训练后获得网络的预测结果，结合测试集判断链路预测性能，当预测准确率达到设定的阈值时，输出链路预测结果，基于拓扑相似和 XGBoost 算法的复杂网络链路预测流程如图 2 所示。

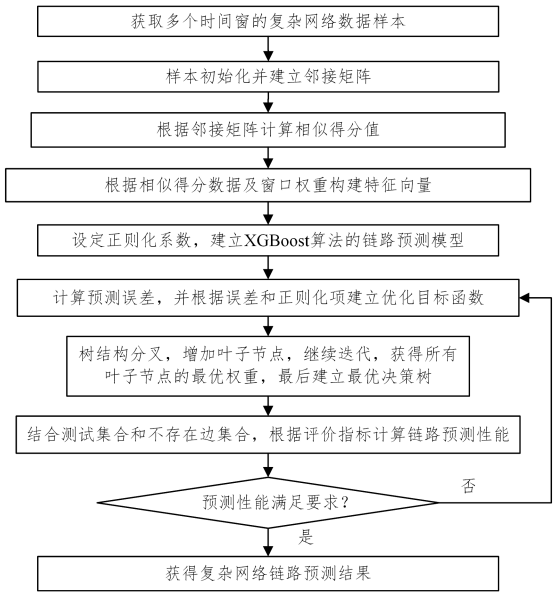


图 2 基于拓扑相似和 XGBoost 的链路预测流程

Fig. 2 Link prediction flow based on topology similarity and XGBoost

本文的主要工作创新是将网络拓扑结构分为多个分区，并对各个分区构建弱学习器，然后利用 XGBoost 集成学习方法进行融合。XGBoost 的求解主要是通过决策树来预测分析复杂网络链路，而决策树的叶子权重将决定决策树的具体结构，通过式(20)求解每片叶子的最优权重，获得最优决策树，根据决策树模型对输入的复杂网络特征向量进行预测分析，得到复杂网络的链路预测结果。在求解过程中，需要进行必要的正则化处理，以便能够得到稳定的 XGBoost 链路预测模型。

## 3 实例仿真

为了验证拓扑相似和 XGBoost 算法在复杂网络链路预测中的性能，采用 Matlab 进行实例仿真。仿真数据来源于斯坦福 SNAP 数据集的 Social networks 中的样本，大部分为各主流社交平台和搜索平台的网络数据，数据集中共有 22 470 个复杂网络节点，有 171 002 条边。

首先，通过差异化设置 $\gamma$ 和 $\lambda$ 值，验证不同 $\gamma$ 和 $\lambda$ 值的链路预测准确率情况来确定适合本实例的正则化参数；其次，验证不同时间段的历史网络数据对预测未来链路的影响，确定合适的时间窗口数；最后，将常用复杂链路预测算法与本文算法进行对比，比较各算法在复杂网络链路预测方面的性能。

### 3.1 正则化参数的链路预测影响

图 3 给出了 XGBoost 算法不同 $\gamma$ 和 $\lambda$ 值情况下的链路预测准确率。由图 3 可知，$\gamma$ 和 $\lambda$ 的取值对复杂网络链路预测准确率的影响较敏感，当 $\gamma=0.2$ 和 $\lambda=1$ 时，准确率处于最高点；当 $\gamma=0.83$ 和 $\lambda=0.1$ 时，准确率处于最低点。因此，本文 XGBoost 算法选择的正则化参数为 $\gamma=0.2$，$\lambda=1$。

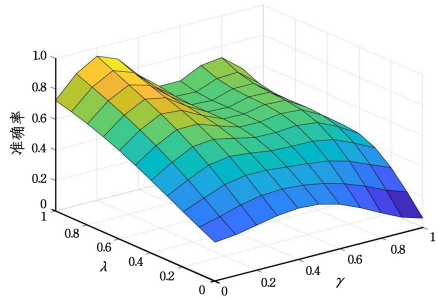


图 3 不同正则化参数的链路预测准确率

Fig. 3 Link prediction accuracy with different regularization parameters

### 3.2 不同时间窗的预测准确率

为了预测当前复杂网络的链路，需要根据历史时间窗口的网络拓扑来预测未来链路的情况，选择合适的历史时间段对复杂链路的网络预测性能有重要影响。为了验证不同时间窗口对复杂网络链路预测的性能，差异化设置时间窗口，选择 1 000 个节点，预测结果如表 1 所列。

表 1 不同时间窗的链路预测准确率

Table 1 Link prediction accuracy of different time windows

| 时间窗口($T$) | 预测准确率/% | 预测时间/s |
|---|---|---|
| 1 | 79.317 | 51.023 |
| 2 | 80.441 | 53.378 |
| 3 | 84.719 | 57.114 |
| 4 | 87.374 | 61.167 |
| 5 | 88.267 | 66.834 |
| 6 | 93.109 | 69.324 |
| 7 | 88.214 | 70.103 |
| 8 | 81.337 | 71.816 |
| 9 | 77.665 | 73.735 |
| 10 | 64.812 | 75.621 |

从表 1 可以看出，网络链路预测准确率随着时间窗口的增加而先增后减，当时间窗口数量达到 6 时，网络预测准确率达到最高，为 93.109%。当时间窗口数较少时，XGBoost 获

得的有效网络数据过少，因此预测准确率不高；当时间窗口数增加到 10 时，网络链路预测准确率只有 64.812%，这是因为过多的历史数据影响了当前网络链路的判断，导致准确率下降。在预测时间方面，随着时间窗口数的增多，需要处理的网络数据样本也增多，因此耗时缓慢增加。故将时间窗口数设置为 6，更适合本文复杂网络链路预测。

### 3.3 不同算法的预测性能

#### 3.3.1 预测准确率

选取训练及测试样本量分别为 1 000 和 300，采用 Matlab 软件，分别对共同邻居(CN)、局部路径(LP)、AdaBoost 常用链路预测算法[18]及本文算法进行预测准确率的仿真，结果如图 4 所示。

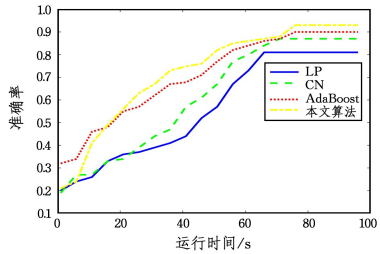


图 4 不同算法的预测准确率

Fig. 4 Prediction accuracy of different algorithms

从图 4 可以看出，复杂网络链路预测准确率随着预测时间的增加而提升，当运算时间达到 66 s 左右时，LP 算法开始收敛，CN 算法大约在 72 s 时开始收敛，AdaBoost 和本文算法的预测准确率大约在 78 s 时趋于稳定。当算法稳定时，本文算法的网络链路预测准确率最高，为 0.93，LP 算法的预测准确率最差。

#### 3.3.2 预测时间性能

下文对 4 种算法的预测时间性能进行仿真，仿真结果如图 5 所示。

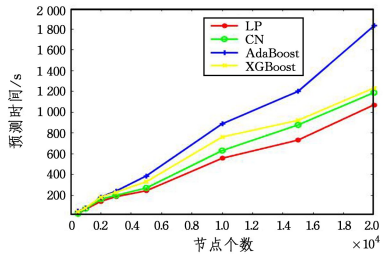


图 5 不同算法的预测时间

Fig. 5 Prediction time of different algorithms

从图 5 可以看出，在预测时间性能方面 LP 算法表现最优，AdaBoost 算法表现最差，AdaBoost 和 XGBoost 都需要建立树形结构并多次迭代求最优解，因此耗时较长。当需要预测的复杂网络节点个数较多时，本文算法和 CN 算法的链路预测消耗时间非常接近，与 LP 算法的预测时间也相差较小。

综合而言，在复杂网络链路预测方面，本文算法的预测准确性最好，LP 算法的预测时间性能最优；当需要预测的网络节点数量较多时，本文算法能保证高准确率，且其预测时间与 LP 和 CN 也非常接近。

**结束语** 采用拓扑相似与 XGBoost 算法的复杂网络链路预测，利用拓扑相似得分函数和最优决策树模型，通过输入网络节点各时间窗的网络拓扑数据及时间窗权重值，来获得网络节点的链路预测结果且预测准确率高。后续研究将考虑对标准 XGBoost 算法进行优化，提高 XGBoost 算法的链路预测效率。

## 参 考 文 献

[1] LI H，MA X P，SHI J，et al. Research on trust transfer based recommendation model in complex network environment[J]. Acta Automatica Sinica，2018，44(2)：363-376.

[2] XU X K，XU S，ZHU Y X，et al. Link predictability in complex networks[J]. Complex Systems and Complexity Science，2014，11(1)：41-47.

[3] WANG K，LIU S X，YU H T，et al. Complex network link prediction algorithm based on common neighbor validity[J]. Journal of University of Electronic Science and Technology of China，2019，48(3)：114-121.

[4] CHEN B，ZHU W，LIU Y. Algorithm for complex network diameter based on distance matrix[J]. Journal of Systems Engineering & Electronics，2018，29(2)：118-124.

[5] YANG X H，WANG C. Community Detection Algorithm in Complex Network Based on Network Embedding and Local Resultant Force[J]. Computer Science，2021，48(4)：229-236.

[6] WANG H，LE Z C，GONG X，et al. Review of Link Prediction Methods Based on Feature Classification[J]. Computer Science，2020，47(8)：302-312.

[7] WU Z F，LIANG Q，LIU Q，et al. link prediction optimization algorithm based on AdaBoost[J]. Journal of Communications，2014(3)：116-123.

[8] MO H M. Application of Belief Function to Identification of Node Influence in Complex Networks[J]. Journal of Chongqing Technology and Business University(Natural Science Edition)，2018，35(1)：71-78.

[9] WANG X，CHEN X，QIAN F L，et al. Node similarity link prediction algorithm based on common neighbor contribution[J]. Data Acquisition and Processing，2018，33(5)：900-910.

[10] GUO W Y，LIU H Y，SUN Q，et al. Topological similarity measurement of contour lines using tree edit distance[J]. Journal of Surveying and Mapping Science and Technology，2019(1)：79-85.

[11] FU L D，WEI H，LI D，et al. Community dividing algorithm based on similarity of common neighbor nodes[J]. Journal of Computer Applications，2019，39(7)：2024-2029.

[12] GÓMEZ-RÍOS A，LUENGO J，HERRERA F. A Study on the Noise Label Influence in Boosting Algorithms：AdaBoost，GBM and XGBoost[C]// International Conference on Hybrid Artificial Intelligence Systems. 2017.

[13] ZHANG Y，YAO Y G. Research on network intrusion detection based on xgboost algorithm[J]. Information Network Security，2018(9)：102-105.

[14] LI Z S，LIU Z G. Feature selection algorithm based on xgboost

[J]. Journal of Communications, 2019(10): 101-108.
[15] ZHANG X Y, TANG K. Traffic identification method based on xgboost algorithm and domain name information screening[J]. Electronic Design Engineering, 2019(6): 177-182, 187.
[16] CUI Y P, SHI K X, HU J W. Research of Webshell Detection Method Based on XGBoost Algorithm[J]. Computer Science, 2018, 45(0z1): 375-379.
[17] SU B J, ZHOU Y P, LIANG X G. Emotion recognition model of e-commerce review text based on xgboost algorithm[J]. Internet of Things Technology, 2018, 8(1): 54-57.
[18] ZHANG Y X, FENG Y X. Overview of link prediction methods and development[J]. TT & C Technology, 2019, 38(2): 8-12.

**GONG Zhui-fei**, born in 1977, postgraduate, Ph. D, lecturer, senior engineer. Her main research interests include complex network and link prediction.