基于改进灰狼优化的复杂网络重要节点识别算法

顾秋阳[1,2]，吴宝[1,2]，孙兆洋[3]，池仁勇[1,2]

（1. 浙江工业大学管理学院，浙江 杭州 310023；2. 浙江工业大学中国中小企业研究院，浙江 杭州 310023；
3. 中国标准化研究院高新技术标准化研究所，北京 100191）

摘 要：近年来，如何识别影响力最大的重要节点已成为网络科学最前沿的热点方向。将复杂网络节点影响力最大化问题表述为一个优化问题，其成本函数表示为节点影响力及其间的距离，使用 Shannon 熵对节点影响力进行度量，并利用一种改进灰狼优化算法来解决此问题。最后，使用真实复杂网络数据集进行数值计算。结果表明，与现有算法相比，所提算法精度更高，且计算效率较高。

关键词：改进灰狼优化；复杂网络；重要节点识别；影响力最大化

中图分类号：TP181，TP391

文献标识码：A

DOI:10.11959/j.issn.1000−436x.2021088

Key node identification algorithm for complex network based on improved grey wolf optimization

GU Qiuyang[1,2], WU Bao[1,2], SUN Zhaoyang[3], CHI Renyong[1,2]

1. School of Management, Zhejiang University of Technology, Hangzhou 310023, China
2. China Institute for Small and Medium Enterprises, Zhejiang University of Technology, Hangzhou 310023, China
3. Institute of High and New Technology Standardization, China Institute of Standardization, Beijing 100191, China

Abstract: In recent years, how to select the most influential key node for identification has become the most cutting-edge hot direction in network science. Formulating the problem of maximizing the influence of complex network nodes as an optimization problem whose cost function was expressed as the influence of nodes and the distance between them, measures user influence using Shannon entropy, and solved this problem using an improved gray wolf optimization algorithm. Finally, numerical examples were performed with real complex network datasets. The experimental results show that the proposed algorithm is more accurate and computationally efficient than the existing method.

Keywords: improved grey wolf optimization, complex network, key node identification, the maximization of influence

1 引言

随着近年来信息技术的迅速发展，复杂网络节点间的交流、互动形式日趋多样化，由此产生海量的复杂网络数据，其中存在大量节点间的信息交互集。对上述数据进行挖掘，寻找具有影响力的节点，并基于此有效阻止负面信息传播、宣传正面信息、提升推荐效率已成为学术界关注的焦点之一。据中国互联网络信息中心（CNNIC）发布的第 47 次《中国互联网发展状况统计报告》，截至 2020 年 12 月，我国网民规模达 9.89 亿，互联网普及率达 70.4%，较 2020 年 3 月增长 5.9%；同期，我国网络新闻用户达 7.42 亿，占网民整体的 75.1%，社交应用使用率高达网民总量的 85.5%。2019 年，国家互联网信息办公室发布的《数据安全管理办法》中，对社交网络平台提出了新的要求。网络运营者应采取措施督促提醒用户对自己的网络行为负责、加强自律，对于节点通过网络转发给其他节点制作的信息，应

收稿日期：2020−11−02；**修回日期**：2021−03−24

基金项目：国家自然科学基金资助项目（No.71571162, No.71772164）；浙江省哲学社会科学重大课题基金资助项目（No.20YSXK02ZD）

Foundation Items: The National Natural Science Foundation of China (No.71571162, No.71772164), The Philosophy and Social Science Major Project of Zhejiang Province (No.20YSK02ZD)

自动标注信息制作者在该社交网络上的账户或节点标识。这往往会选择合适的节点群体进行信息传播，而社交网络庞大的节点群体会使病毒式传播成为在线社交网络中的最大受益者[1-4]。由于进行节点选取和控制的成本有限，因此只能在复杂网络中选择一组有限数量的种子节点，使其在复杂网络中传播信息，最终实现影响力最大化问题[5]。本文需解决的问题为如何有效识别最具影响力的节点，并将其加入种子节点集中，该问题又被称为节点影响力最大化（IM, influence maximization）问题[6-8]。尽管如文献[9-10]所示，在现有的节点影响力最大化算法中，节点的行为和兴趣已被视为信息传播过程的重要因素，但此类信息并非从真实复杂网络中获得的，故只有网络结构信息可用于识别影响力较大的节点。

目前，已有学者提出多种用于复杂网络重要节点识别的扩散模型[11-13]，可用于模拟信息扩散过程建模及种子节点影响力。由于节点影响力最大化问题的搜索范围往往很大，故评估所有可能的节点子集以定位最佳种子节点集为 NP-hard 问题[14]。而真实复杂网络的辐射范围很广，故利用影响力最大化模型来衡量节点影响力往往较费时。常用于识别影响力最大的节点集的方法是选择中心节点，可采用多种方法来度量复杂网络结构的中心度，如网络度[15]、介数中间性[16]、紧密度[17]、k-shell[18]及改进 k-shell[19]。但通过测量网络结构的中心度对影响力最大化节点进行识别通常不是解决此问题的最好方法。另外，有研究认为节点影响力最大化问题属于优化问题，并采用贪婪算法、启发式算法及元启发式算法等来解决该问题。Wang 等[20]提出了改进贪婪算法，以此求得扩散模型的最优解，尽管该算法的精度较高，但其计算过程十分复杂，故在大规模复杂网络中并不实用[21]。而 Bao 等[22]针对影响力最大化问题提出的启发式算法的时间复杂度和精度较低，但启发式算法得到的值在可达局部中最优[23]。

为识别一组接近最优的具有最大影响力的节点，首先需要对节点影响力进行度量，然后找到最有影响力的节点集。本文首先使用基于 Shannon 熵的指标来衡量节点影响力，并将该问题转化为优化问题，采用一种改进灰狼优化算法解决该问题[24]，并利用真实复杂网络数据集进行数值计算。

2　相关工作

本文基于图论对复杂网络图进行建模，其中，节点代表复杂网络节点，边代表节点间存在的关系。首先对一些经典的扩散模型以及用于影响力最大化问题的常见方法进行回顾。扩展模型常用于模拟真实世界中的传播过程，常用的扩散模型包括阈值模型[25]、级联模型[26]和传染病模型[27]。其中，独立级联模型（IC, independent cascading）为现今应用最广泛的扩散模型之一[26]，其中每个节点都处于活跃或非活跃模式中。为度量种子节点集 S 的影响力，将最初放置在种子节点集 S 中的节点定义为活跃节点，并将所有其他节点定义为非活跃节点。在每个时间段 t 中，时间段 $t-1$ 中被激活的每个节点都有一次机会激活其非活跃的邻节点（激活概率为 p），该过程会一直持续到该时间段内没有新节点为止。最后，在此过程迭代多次后得到的激活节点数量即为节点集 S 的影响范围。

有关节点影响力最大化问题，现有研究可分为两大类。第一类为识别节点影响力并对其进行排序。在多数相关研究中，根据网络结构和节点网络位置，可使用中心性度量法来衡量节点影响力。虽然中心性方法计算较简单，也能较好地确定节点活跃度[28]，但这类方法在识别 Top-k 影响力节点时精度不高。Bao 等[22]提出了一种优化种子节点集的方法，使其总体影响范围最大化，并在考虑节点影响力的同时，考虑了网络中的节点距离，以进一步提高重要节点识别算法的有效性。第二类节点影响力最大化方法在考虑网络可达性的基础上，将该问题转化为优化问题（如贪婪算法、启发式算法、元启发式算法等）。关于贪婪算法，Kempe 等[25]对含有 k 个最具影响力的种子节点集 S 进行识别，并迭代 k 次，每次迭代时使用扩散模型来估计该集合的影响力。

为改善节点影响力最大化问题优化过程中存在的时间复杂度问题，研究者提出了一系列启发式算法，用于选择节点影响力最大化集合。启发式算法通过使用递归法给每个节点分配一个分数，并选择分数最高的 top-k 节点作为种子节点集。Chen 等[29]提出了 2 种方法用于选择种子节点集，首先以每个节点度作为影响力指标，并以 k 次迭代选择种子节点集。在每个步骤中，都可将影响力最大的节点添加到种子节点集中。如节点 v 加入种子节点集，则

其邻节点影响力将会随之降低，且被选为种子节点的概率也会降低。Wang 等[30]所提度惩罚法尝试同样方法来选择种子节点集，将节点 v 选为种子节点，就需惩罚二跳内的所有邻节点，并降低被选为种子节点的概率。该方法可有效减少网络中种子节点的重叠现象，并增加节点分散性。Guo 等[31]提出了基于距离图对具有适当距离的节点进行识别和分类，以使各组中的节点对间距离都大于阈值。Bao 等[22]提出了启发式聚类算法，确定每对节点间的相似度，并选择各集群中影响力最大的节点作为种子节点，将节点影响最大化问题建模为多目标优化问题。

元启发式算法通常首先定义适应度函数，将节点影响力最大化问题建模为优化问题，并应用各种改进优化算法对该问题进行解决。Jiang 等[32]将种子节点集 S 的影响定义为预期扩散值（EDV, expected diffusion value）。Gong 等[21]引入改进 EDV 值，并使用粒子群优化（PSO, particle swarm optimization）算法来优化目标函数。Sun[33]用复制对称平均场理论来解决节点影响力最大化问题。本文首先定义了适应度函数，然后利用灰狼优化算法提出了一种适应度函数对重要节点识别方法进行优化。

3 算法设计

3.1 预备知识

本文首先利用熵的概念定义了适应度函数，并将节点影响力最大化问题设计为优化问题，随后利用改进灰狼优化算法解决该问题。根据 Shannon 熵值，如 $X \in \{x_1, x_2, \cdots, x_n\}$，且 p_i 为目标 x_i 被选择的概率，其中 $\sum_{i=1}^{n} p_i = 1$，可根据式(1)计算 X 的熵。

$$B(X) = -\sum_{i=1}^{n} p_i \ln(p_i) \tag{1}$$

灰狼优化（GWO, gray wolf optimization）算法[24]是一种基于种群的演化算法，其灵感来源于灰狼的狩猎行为。如图 1 所示，灰狼遵循严格的社会等级制度，在社会中主要分为 α、β、γ、δ 这 4 种。α 狼、β 狼及 γ 狼在狩猎过程中主要负责攻击，而 δ 狼在团队中扮演替罪羊的角色。根据不同狼的特征，可将 GWO 算法建模如下。首先，随机生成一组解，使得每个解都与一头狼相对应，以表示其位置。其中，适应性排名第一的为 α 狼，适应性排名第二和第三的分别为 β 狼和 γ 狼，剩余则为 δ 狼。该算法根据 α、β、γ 狼的位置进行空间搜索，以找到猎物的位置（即最优解）。上述 3 种狼可以预测猎物位置，而 δ 狼则根据另外 3 种狼的位置来改变自身位置，从而找到离猎物更近的位置，该算法尝试在迭代过程中找到最优解。在每次迭代过程中，δ 狼都会试图根据其他狼群的位置来改变自身位置。在每次迭代 t 中，δ 狼都会根据迭代次数 $t+1$ 确定自身的新位置，即 $\boldsymbol{X}_i(t+1)$，具体如式(2)所示。

$$\boldsymbol{X}_i(t+1) = \frac{\boldsymbol{Y}_1 + \boldsymbol{Y}_2 + \boldsymbol{Y}_3}{3} \tag{2}$$

其中，每只 δ 狼都会试图根据 $\boldsymbol{Y}_1$、$\boldsymbol{Y}_2$、$\boldsymbol{Y}_3$ 来确定更优位置，$\boldsymbol{Y} \in [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3]$ 值可根据狼的当前位置和 α 狼的位置计算得到，具体如式(3)所示。

$$\boldsymbol{Y}_1 = \boldsymbol{X}_\alpha - \boldsymbol{X}_1 \boldsymbol{D}_\alpha \tag{3}$$

其中，$\boldsymbol{D}_\alpha$ 的计算式为

$$\boldsymbol{D}_\alpha = |\boldsymbol{C}_1 \boldsymbol{X}_\alpha - \boldsymbol{X}_i(t)| \tag{4}$$

其中，t 表示当前迭代次数，$\boldsymbol{X}_i(t)$ 表示本次迭代中狼 i 的位置，而 $\boldsymbol{X}_\alpha$ 表示 α 狼的位置。结合 β 狼和 γ 狼的当前位置，利用式(5)可计算得到 $\boldsymbol{Y}_2$ 和 $\boldsymbol{Y}_3$ 的值。

$$\{\boldsymbol{Y}_2, \boldsymbol{Y}_3\} = \{\boldsymbol{X}_\beta - \boldsymbol{A}_2 \boldsymbol{D}_\beta, \boldsymbol{X}_\gamma - \boldsymbol{A}_3 \boldsymbol{D}_\gamma\} \tag{5}$$

其中，$\boldsymbol{D}_\beta$ 和 $\boldsymbol{D}_\gamma$ 的值可由式(6)计算得到。

$$\{\boldsymbol{D}_\beta, \boldsymbol{D}_\gamma\} = \{|\boldsymbol{C}_2 \boldsymbol{X}_\beta - \boldsymbol{X}_i(t)|, |\boldsymbol{C}_3 \boldsymbol{X}_\gamma - \boldsymbol{X}_i(t)|\} \tag{6}$$

其中，$\boldsymbol{X}_\beta$ 和 $\boldsymbol{X}_\gamma$ 分别表示 β 狼和 γ 狼的位置，$\boldsymbol{A}$ 和 $\boldsymbol{C}$ 分别表示如式(7)所示的系数向量。

$$\{\boldsymbol{A}, \boldsymbol{C}\} = \{2a\boldsymbol{r}_1 - a, 2\boldsymbol{r}_2\} \tag{7}$$

其中，$\boldsymbol{r}_1$ 和 $\boldsymbol{r}_2$ 表示区间 $[0,1]$ 内的随机值，a 值为迭代过程中从 2 线性趋近于 0 的控制参数。a 值在时段 t 中的计算方法如式(8)所示。

$$a = 2 - \frac{2t}{\max_t} \tag{8}$$

其中，$\max_t$ 表示重要节点识别算法的迭代次数。在迭代过程中，狼群间的位置矛盾减少，算法逐渐收敛，最后得到 α 狼最优解，并将其作为最优解决方案。经典灰狼优化算法的执行过程如算法 1 所示[25]，其中 n 表示种群规模。

算法 1 灰狼优化算法

输入 狼 $\alpha, \beta, \gamma, \delta$ 的位置

输出 集合 $\{x_\alpha, x_\beta, x_\gamma, x_\delta\}$

1)初始化每只狼的位置为 $x_i(i\in[1,2,\cdots,n])$
2)初始化参数 $\alpha,\beta,\gamma,\delta$
3)计算每只狼的适应度值
4)定义 x_α、x_β、x_γ 分别为排序 1、2 和 3 的狼
5)将其他狼定义为 δ
6) while 时间小于迭代周期
7)　for each δ 狼为 x_i do
8)　　更新 x_i 的位置
9)　end for
10)　更新参数 a,A,C
11)　计算每只狼 δ 的适应值
12)　更新集合 $\{x_\alpha,x_\beta,x_\gamma\}$
13) end while
14) return 集合 $\{x_\alpha,x_\beta,x_\gamma,x_\delta\}$

本文使用无向图 $G=(V,E)$ 对复杂网络图进行建模，其中，$V=\{v_1,v_2,\cdots,v_{|V|}\}$ 表示复杂网络节点集，E 表示节点间的关系边集。边 $e_{i,j}\in E$ 表示 2 个节点 v_i 和 v_j 间的距离，且这 2 个节点为邻节点。$N_i\subset V$ 表示节点 v_i 的邻居，$d_i=|N_i|$ 表示节点 v_i 的影响力。N_i^2 表示节点 v_i 的二跳邻节点（即邻节点的邻节点）。节点影响力最大化问题的目标是识别具有 k 个节点的集合 S，以启动传播过程，从而使传播影响最大化，即激活节点数量最大化。在本文的其余部分中，S' 表示种子节点集 S 中节点的邻节点或二跳邻节点的节点集合。

3.2　适应度函数

本文选择节点作为种子节点集中的节点，其中每个节点都具有较高影响力，且选择的集合尽可能地覆盖整个网络，以保证信息传播最大化。在传播过程中，节点 $v_j\in S'$ 被激活，即接受信息，可使用式(9)进行计算。

$$I(v_j)=\sum_{e_{i,k}\in E,e_{kj}\in E,v_i\in S}p_{i,j}+p_{i,k}p_{k,j} \tag{9}$$

其中，$p_{i,j}$ 表示信息从 v_i 传播到 v_j 的概率。式(9)可用来计算节点 v_j 收到其邻节点发送信息的概率，且其邻节点都为 S 集成员。如果节点 v_j 在 S 集中没有邻节点，则可将其视为零集合。而式(9)等号右边第二项可用来计算节点 $\boldsymbol{v}_j$ 收到其二跳邻节点发送信息的概率，且其二跳邻节点也为 S 集中的成员。这两部分都根据概率规则进行求和，在传播过程中不同节点不会产生相同影响。如果影响力较大的节点接收到该信息，则其可能会传播至复杂网络的其他领域，各节点 $v_j\in S'$ 值。本文所提基于改进灰狼优化的重要节点识别算法中，适应度函数为 $\sum_{v_j\in S'}w(v_j)$。使用本文所提 IGW-CNI 算法寻找种子节点，以使 S' 集中的节点数量最大化，且其中所有节点都具有较大影响力。在运用求和运算符后，将 S' 集中有限的具有较高价值的节点和无价值节点相组合，使适应度函数计算得到的值实现最大化。在此情况下，如果无法将信息传播给有影响力的节点，会导致所选节点的影响力大幅降低。在适应度函数中使用熵来解决此问题，使节点在 S' 集中的均衡节点能更有效提升集 S 的影响力。另一方面，熵值随着 S 值的增大而增大，故需进行归一化。集 S 的适应度熵值可定义为

$$E(S)=-\sum_{v_j\in S'}\frac{I(v_j)d_j}{\sum_{v_j\in S'}w(v_j)}\ln\left(\frac{I(v_j)d_j}{\sum_{v_j\in S'}w(v_j)}\right) \tag{10}$$

3.3　改进灰狼优化算法

参考文献[34]的结论可知，度为 1 的节点被选为种子节点的概率非常低。在许多真实复杂网络中，大部分节点只存在一个邻节点。为有效降低本文所提 IGW-CNI 算法的时间复杂度，仅将影响力大于 1 的节点选为候选种子节点，并将这些节点表示为 $V'=\{v_1',v',\cdots,v'_{|V'|}\}$。而每只狼（即每个解）都存在 2 个属性，即位置和对应的种子节点集。狼 i 的位置可表示为带有节点 $|V'|$ 的向量 $\boldsymbol{x}_i$，其中第 j 个节点表示节点 v_j' 被选为种子节点的概率。狼 i 对应的种子集为 S_i，其中包含 k 个节点，且其值在 X_i 中最高。

算法 2 给出了本文所提 IGW-CNI 算法的具体流程。首先，随机生成 n 个主要解，并利用随机位置函数生成每个随机解；其次，利用式(10)来计算解的适应度值，并根据解的适应度值来选择确定 α、β、γ 狼。进行 $\max_t$ 次迭代操作，以最大化适应度。在每次迭代过程中，更新 δ 的位置，并给出更新后的位置。为每只狼 i 确定种子集 S_i，而后计算每个子集 S_i 的适应度值，并根据适应度值，将最佳解更新 α、β、γ 狼；随机更新 β、γ 狼的位置，避免产生局部最优解。

算法 2 本文所提 IGW-CNI 算法

输入 图 $G=(V,E)$，种子集规模 k，群体规模 n，迭代周期 $\max_t$

输出 集合 $\{x_\alpha,x_\beta,x_\gamma,x_\delta\}$，不同参数下的种子集 S

1) 初始化参数 $a,A,C|V'|$

2) 在随机位置初始化 n，并确定对应于每个 $\boldsymbol{X}_i$ 的 S_i

3) 计算对于每个 S_i 的适应值

4) 并基于此选择最优解 $\boldsymbol{X}_\alpha,\boldsymbol{X}_\beta,\boldsymbol{X}_\gamma$，并令剩余解为 $\boldsymbol{X}_\delta$

5) while 时间小于迭代周期

6) for each δ do

7) 更新 $\boldsymbol{X}_i$ 的位置，并确定 S_i

8) end for

9) 更新参数 a,A,C

10) 对于 S_i 计算适应值

11) 更新集合 $\{x_\alpha,x_\beta,x_\gamma\}$

12) end while

13) return 集合 $\{x_\alpha,x_\beta,x_\gamma,x_\delta\}$

在本文所提 IGW-CNI 算法中，利用随机位置函数生成基本解 i（狼 i 的基本位置 $\boldsymbol{X}_i$ 及其对应的种子节点集 S_i），该函数的运行机制如文献[13]所提算法所示。其中，给每个节点 $v_j\in V'$ 赋随机值 $\boldsymbol{X}_{i,j}$，并将该随机值作为该节点被选为种子节点的概率，且该概率与节点的影响力成正比。在本文所提 IGW-CNI 算法中，根据 α、β、γ 狼的位置来更新 δ 狼的位置，以期获得更优位置。故基于此设置了更新位置函数，该函数进程如文献[12]所提算法所示。其中，为更新节点 v_j 在狼 i，即 $\boldsymbol{X}_{i,j}$ 中的概率，参考了该节点为 α、β、γ 的位置 $\boldsymbol{X}_i$，可计算 $\boldsymbol{Y}_1$、$\boldsymbol{Y}_2$、$\boldsymbol{Y}_3$ 的值，并相应计算 $\boldsymbol{X}_{i,j}(t+1)$ 的值。通过计算每个节点 $v_j\in V'$ 的 $\boldsymbol{X}_{i,j}(t+1)$ 值，将迭代周期 $t+1$ 内狼 i 的位置更新为 $\boldsymbol{X}_i(t+1)$，并更新 $\boldsymbol{X}_i$ 中的 $\boldsymbol{X}_{i,j}$，以对每个 $\boldsymbol{X}_{i,j}(j\in1,\cdots,|V'|)$，都在多次迭代中更新狼 i 的位置，重复此计算可更新每次迭代中狼的位置。

4 数值算例

4.1 数据说明与参数设置

为证明本文所提 IGW-CNI 算法的有效性，本文使用 4 个真实复杂网络数据集进行数值模拟，以便观察算法对现实情况的适应度（实验中使用的真实复杂网络数据集特征如表 1 所示）。其中，激活率为节点激活概率，根据网络图的稀疏性进行设置，并基于网络节点度和二阶度的平均值进行计算。本文使用 Python 软件以近期的热点事件“HUAWEI event”和“华为事件”的 30 个热点评论节点作为初始节点，分别爬取 Facebook、Twitter、新浪微博和豆瓣的复杂网络节点数据集作为仿真实验的基础数据集（爬取时间为 2020 年 4 月 23 日至 2020 年 11 月 16 日）。本文将每个节点作为一个节点，使用节点间的边界表示节点间的关系。本文选择了 10 个具有较强影响力的节点及其邻节点列表作为复杂网络初始节点，以此生成了简单复杂网络。实验在 MATLAB 2017b 环境下实施，并在 Windows10 操作系统的服务器（Intel Xeon 处理器（34 GHz）和 32 GB 内存）上进行。

为评估本文所提 IGW-CNI 重要节点识别算法的性能，将其结果与现有较成熟的重要节点识别算法进行比较，其中包括 Node2vec[17]、PageRank 法[28]、度递减搜索策略（DDSE, degree descending search strategy）[23]和模拟退火 EDV 值（SADV, simulated annealing EDV）[35]。由于实验中 IGW-CNI 重要节点识别算法预测列表在每次运行时的结果都可能不同，故设置评估结果为迭代 100 次运行后的平均值，运行的平均标准差为

表 1 复杂网络数据集说明

网络序号	网络名称	类型	节点数量/个	节点边界数量/个	平均度	节点平均路径	聚类系数	激活率
1	Facebook	无向	35 847	332 822	32.98	6.59	0.696	0.04
2	Twitter	无向	48 372	439 804	59.08	4.76	0.513	0.07
3	新浪微博	无向	38 790	653 738	56.49	8.60	0.639	0.05
4	豆瓣	无向	41 556	837 194	54.20	6.49	0.651	0.05

1.524。首先，本文对所提 IGW-CNI 算法参数对适应度函数值的影响进行分析，得到了节点规模和迭代周期的最优值。为分析迭代周期的影响，设迭代周期的取值范围为 [0,50] 。在适应度函数值为 $[10,20,30,40,50]$ 的情况下，各网络数据集中的实验结果如图 1 所示。由图 1 可知，当迭代周期超过 20 时，适应度值没有显著增加，故在后续实验中将 $\max_t$ 设为 20。

本文还对节点规模与适应度值间的关系进行研究，设迭代周期的取值范围为[0,50]，并计算当种群适应度函数值为 $\{10,20,30,40,50\}$ 时不同节点规模下的适应度值，结果如图 2 所示。由图 2 可知，适应度值随着节点规模的增大而增大，但当节点规模大于 20 时，适应度值没有显著增加，故将节点规模固定为 20。

4.2　实验结果

本节对所提算法的收敛速度进行了研究，并研究了优化过程中的狼群运动，计算了 2 次连续迭代 t 和 $t+1$ 中狼 i 位置间的欧氏距离。而所有类似的狼在连续 2 次迭代 t 和 $t+1$ 中的平均移动（AM, average movement）为

$$\mathrm{AM}(t,t+1)=\frac{\sum_{i=1}^{n}\mathrm{movement}(i,t,t+1)}{n} \tag{11}$$

其中，n 表示节点规模。实验结果如图 3 所示。由图 3 可知，在初始迭代中的平均运动增加，随着迭代次数的增加，平均移动次数收敛。尽管在最终迭代中存在振荡行为，避免了算法陷入局部最优解，但所提算法最终会呈收敛趋势。

本文对所提 IGW-CNI 算法与其他重要节点识别算法的性能进行比较。为此，将种子节点集的规模定义为[0, 50]，并对种子节点集的影响力进行评价，相关结果如图 4 所示。由图 4 可知，与其他算法相比，本文所提 IGW-CNI 算法具有较好表现，且在种子节点数量相同的情况下影响力较大，这是由于随着种子节点集规模的增加，种子节点间的距离也发挥重要作用，故 IGW-CNI 算法比其他算法所选择的种子节点集具有更高的影响力，即本文所提算法在不同种子集规模下都优于其他算法。随着种子集规模的扩大，本文所提算法得到的种子集影响力的增长速度快于其他算法，这是由于 IGW-CNI 算法利用改进灰狼优化算法来进行种子集估计并利用随机位置函数生成基本解，因此得到的种子集规模较其他经典算法会具有更快的增长速度。

本文进一步改变 IC 模型的激活率以分析其对算法性能的影响。为此，将 k 值设为 30，激活率值设置为区间 [0,0.2] ，间隔为 0.005。结果如图 5 所示，随着 p 值的增大，节点集的影响力也随之增加。本文所提 IGW-CNI 算法的性能在大多数情形下较其他算法为最优，且随着 p 值的增大，其优势更加突出。这是由于节点影响力随着 p 值的增大而增大，而在适应度函数中应用熵函数可使 IGW-CNI 算法选择重叠较少的种子集，故本文所提 IGW-CNI 算法在激活率较高时方面优于其他算法，即本文所提 IGW-CNI 算法在不同 p 值下都优于其他算法，且随着激活率 p 值的增大，本文所提算法得到种子集的影响力的增长速度快于其他算法。

本文使用如式(12)所示精度计算方法进一步进行算法比较。

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{12}$$

其中，真正例（TP, true positive）表示正确预测链接的数量，真负例（TN, true negative）表示正确的未预测链接的数量，假正例（FP, false positive）表示错误预测链接的数量，假负例（FN, false negative）表示错误的未预测链接的数量。

算法精度比较结果如表 2 所示。由表 2 可知，本文所提 IGW-CNI 算法在不同数据集中都优于其他重要节点识别算法，这是由于本文将影响最大化问题建模为具有成本函数的优化问题，并采用改进灰狼优化算法加以解决，有效提升了算法精度。在 Facebook 数据集中，PageRank 与 DDSE 法的性能仅次于本文所提 IGW-CNI 法，这是由于在聚类系数较高、激活率较低的情况下，PageRank 与 DDSE 法都考虑了入链数量的变化，故计算精度相对较高；在 Twitter 数据集中，SADV 法的性能仅次于本文所提 IGW-CNI 法，这是由于在平均度与激活率较高的情况下模拟退火算法可更好地加速收敛过程，因此可以得到相对较优的计算精度结果；在新浪微博数据集中，N2V 和 CN 法的性能仅次于本文所提 IGW-CNI 法，这是由于在平均路径较大的情况下，基于网络图最短路径进行邻节点度计算可有效提升算法计算精度。

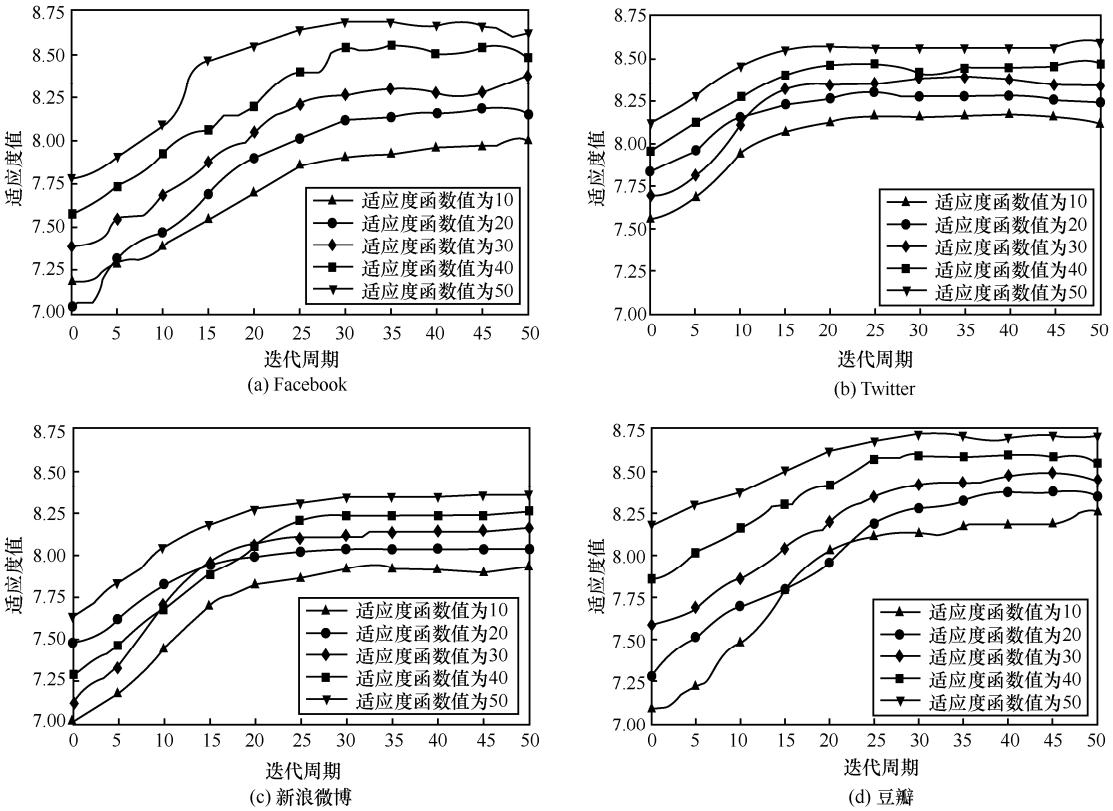

图 1 不同迭代周期下的适应度值

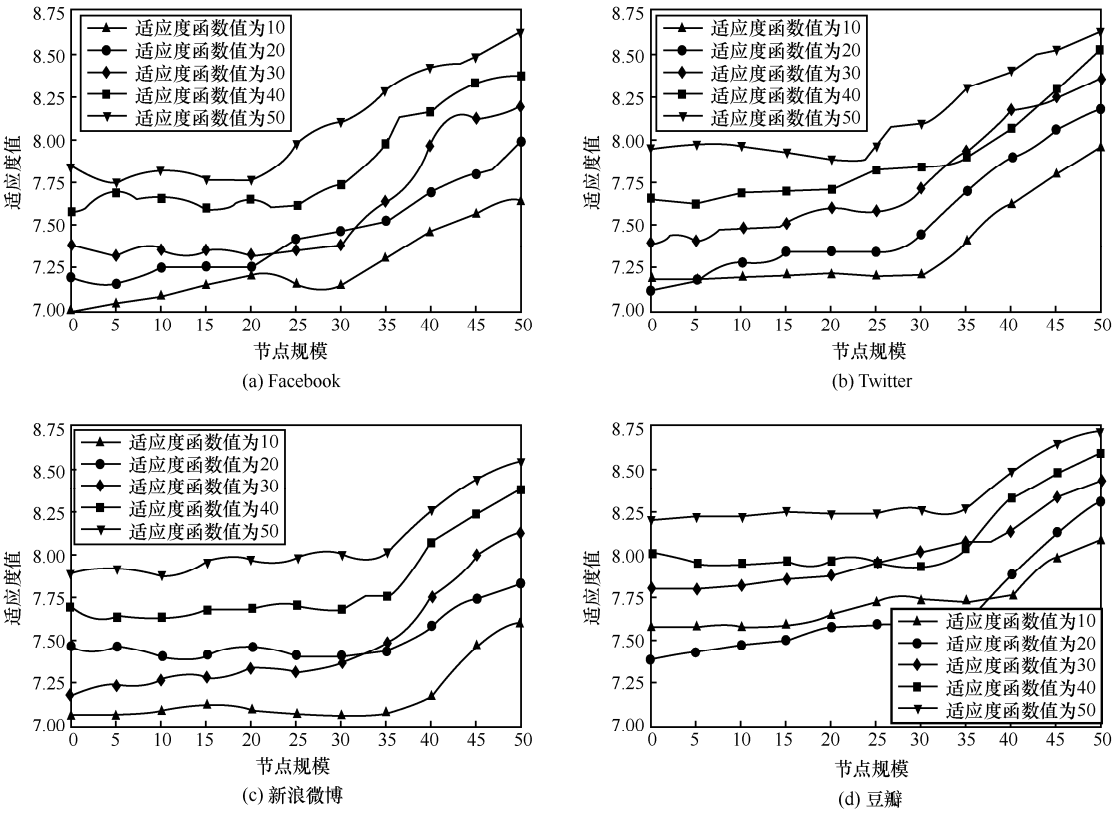

图 2 不同节点规模下的适应度值

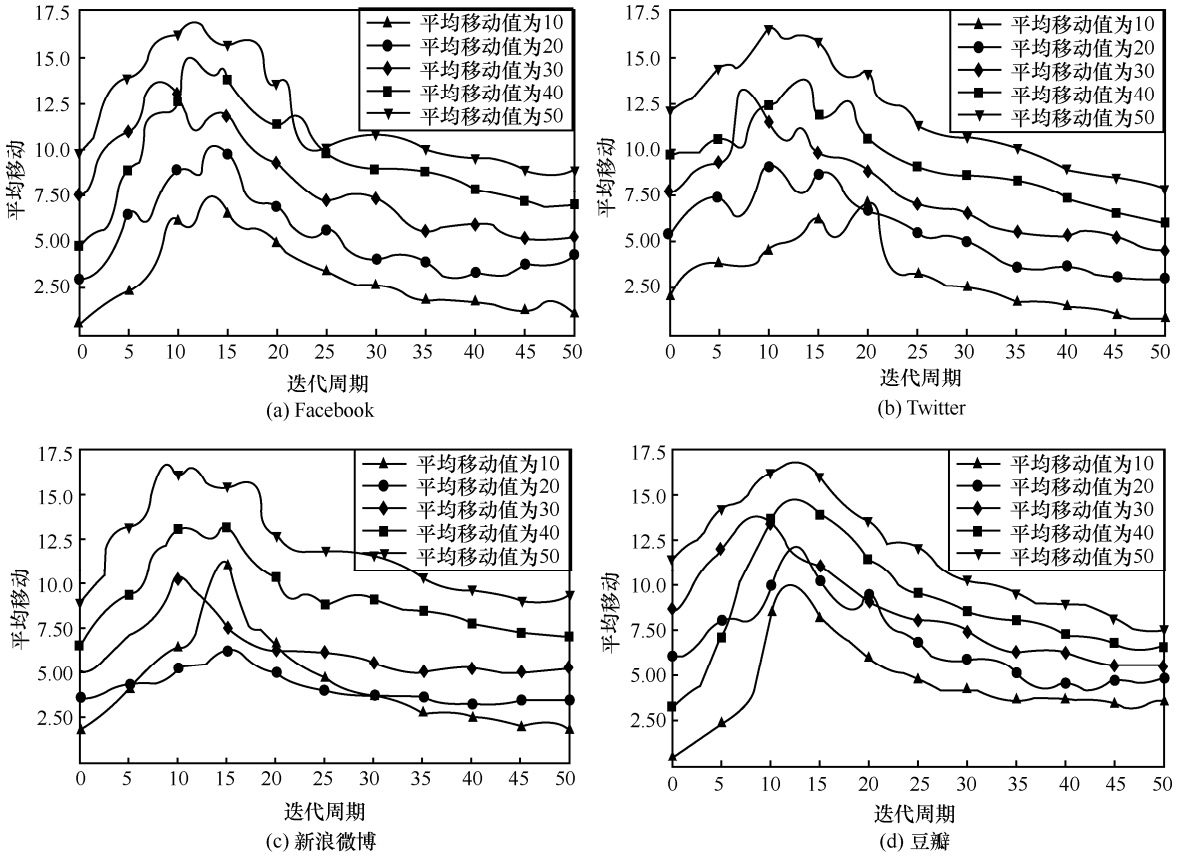

图 3　不同复杂网络中平均移动分析结果

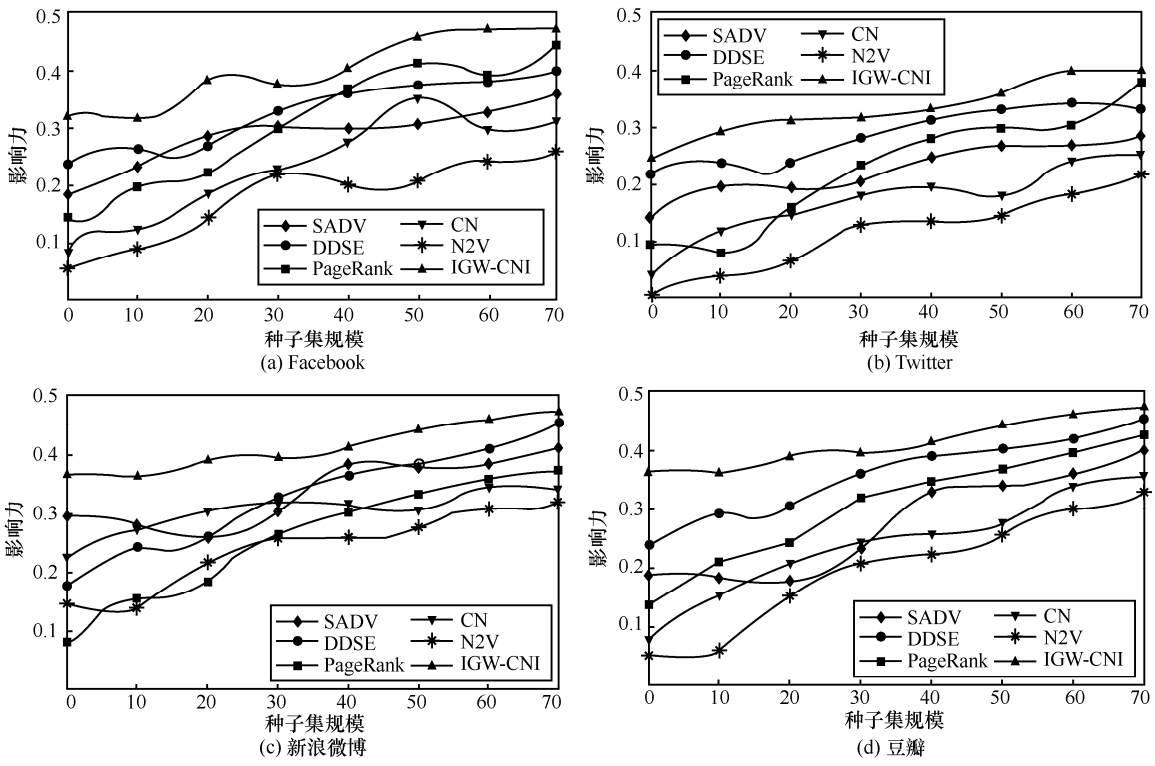

图 4　不同复杂网络中种子集规模对影响力的算法比较结果

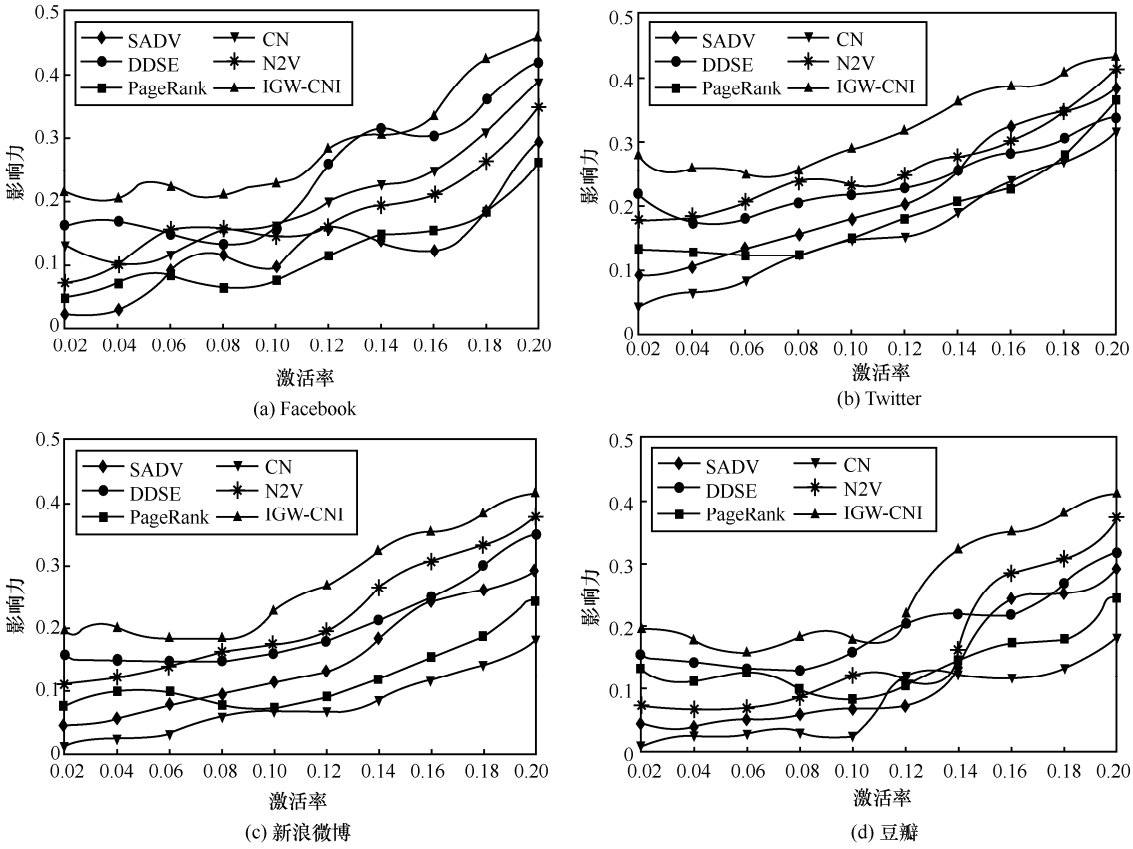

图 5　不同复杂网络中激活率对影响力的算法比较结果

表 2　　**算法精度比较结果**

数据集	N2V	CN	PageRank	SADV	DDSE	IGW-CNI
Facebook	0.728 1	0.752 0	0.831 4	0.743 3	0.856 5	**0.963 4**
Twitter	0.741 2	0.728 7	0.532 8	0.840 3	0.751 1	**0.929 3**
新浪微博	0.724 6	0.816 2	0.705 5	0.728 1	0.722 5	**0.983 2**
豆瓣	0.745 6	0.821 4	0.837 1	0.643 5	0.785 6	**0.981 3**

注：加粗字体表示此算法在此数据集下相对最优。

4.3　统计检验

本文在此对测试算法性能的统计显著性差异进行进一步分析。为取得统计显著性，进行了 Friedman 检验，使用 Bonferroni 法来校正实验结果[36]，并设置可信度分数为 $\alpha=0.05$，这表明如果 $p<0.05$，则存在差异显著。表 3 中，本文所提 IGW-CNI 算法将成本函数表示为节点影响力及其间的距离，并使用 Shannon 熵对节点影响力进行度量的优势较其他经典算法会充分体现。Friedman 检验对 AUC 的观察检验值 F_f 为 54.142，均大于相应的 χ^2 值（即 $\chi^2(\alpha_c,D_f)$）。当置信区间 $\alpha=0.05$、自由度 $D_f=8$ 时，χ^2 值为 15.51，故拒绝零假设。

4.4　大数据集实验

为检验本文所提 IGW-CNI 算法在极限条件下的适用性，本节使用规模更大的 MovieLens 20M、DatingT 和 Netflix 等复杂网络数据集对上述重要节点识别算法进行对比实验，以进一步分析本文所提算法在大数据集中的适用性（其中 MovieLens 20M 和 Netflix 数据集常用于推荐算法研究，且呈现包含节点和目标 2 种节点的二部图网络结构，但本文所提算法更加重视在网络环境下的应用，故在上述 2 种数据集中只采用节点以进行大数据集实验。这也从另一角度对本文所提 IGW-CNI 重要节点识别算法在二部网络图中的适

表 3　　对 AUROC 值的 Friedman 统计检验结果

数据集	N2V	CN	PageRank	SADV	DDSE	IGW-CNI
Facebook	0.948 1	0.952 0	0.931 4	0.943 3	0.873 2	**0.971 6**
Twitter	0.941 2	0.928 7	0.532 8	0.940 3	0.810 2	**0.964 8**
新浪微博	0.921 5	0.916 2	0.605 5	0.928 1	0.903 8	**0.994 4**
豆瓣	0.705 7	0.738 9	0.459 4	0.820 9	0.783 3	**0.953 9**

注：加粗字体表示此算法在此参数条件下相对最优。

用性进行分析，进一步证实了本文所提算法的普适性）。其中，MovieLens20M 被国内外学者认为是最稳定的基准数据集之一，其包含 139 428 个节点对 27 493 部影片的 20 351 902 条评分；DatingT 数据集包括 136 739 名男性和 179 332 名女性组成的 18 293 016 条交友记录；Netflix 数据集中包括 480 189 个节点对 17 770 部电影的 100 480 507 条评分记录。图 6 给出了同等参数条件下上述 3 种大数据集中的算法比较结果。由图 6 可知，在体量更大的数据集中的实验结果基本与图 5 中的实验结果一致。此外，本文还使用上述 3 种大数据集作为对照组重复进行了精度与效率等方面的实验，结论与前文一致，故在此不再赘述。

4.5　时间复杂度

最后，本文对所提 IGW-CNI 算法的计算时间复杂度效率进行了分析，并对算法的平均运行时间进行了比较，结果如表 4 所示。由表 4 可知，本文所提 IGW-CNI 法的效率高于其他算法，且获得的种子节点集的影响力性能更优。

5　结束语

复杂网络中的重要节点识别对电子商务、网络营销等多个领域都有很大影响。在预算有限的情况下，网络营销所面临的一个主要挑战是识别少数具有影响力的重要节点集，即种子节点集，以期通过将信息传递给种子节点集在复杂网络中获得较大影响力，故可将此问题转化为节点影响最大化问题。本文将影响最大化问题建模为具有成本函数的优化问题，并采用改进灰狼优化算法解决此问题。最后基于真实复杂网络数据集进行比较实验，结果表明，本文所提 IGW-CNI 算法优于其他影响力最大化算法，不仅效率更高，计算时间也短。

尽管本文已提出了上述具有重要意义的发现，但还具有一定局限性，其中一些可能会为未来的进

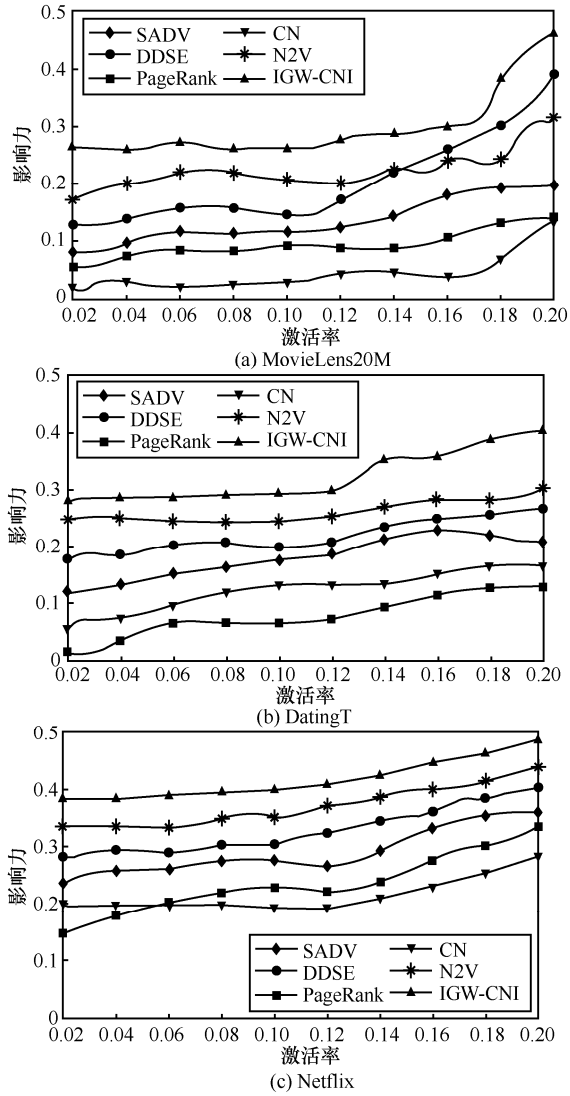

图 6　不同大规模复杂网络中激活率对影响力的算法比较结果

一步研究指明方向。首先，本文所提改进灰狼优化算法为基于种群的优化算法，仍可能存在局部最小值，可考虑引入随机梯度等方法避免出现局部最小值。其次，可结合社会化调查法和计算实验等研究框架对所提重要节点识别算法进行进一步佐证。最后，可尝试使用如自编码器等深度学习方法进一步提升重要节点识别算法的效率和精度。

表 4　　时间复杂度实验结果

数据集名称	N2V	CN	PageRank	SADV	DDSE	IGW-CNI
Facebook	3 132.14	2 513.41	1 445.14	8 593.19	1 933.40	**528.40**
Twitter	3 415.65	2 735.25	1 856.65	7 459.02	2 418.94	**450.83**
新浪微博	3 242.15	2 541.58	1 814.46	8 069.14	1 415.76	**391.56**
豆瓣	3 154.89	2 419.28	1 375.96	6 495.59	3 128.48	**543.59**

注：加粗字体表示此算法法在此参数条件下相对最优。

参考文献：

[1] KIMURA M, SAITO K, NAKANO R, et al. Extracting influential nodes on a social network for information diffusion[J]. Data Mining and Knowledge Discovery, 2009, 20(1): 70-97.

[2] SHEIKHAHMADI A, NEMATBAKHSH M A, ZAREIE A. Identification of influential users by neighbors in online social networks[J]. Physica A: Statistical Mechanics and Its Applications, 2017, 486: 517-534.

[3] 邓琨, 李文平, 余法红, 等. 基于多核心标签传播的复杂网络重叠社区识别方法[J]. 通信学报, 2017, 38(2): 53-66.
DENG K, LI W P, YU F H, et al. Overlapping community detection in complex networks based on multi kernel label propagation[J]. Journal on Communications, 2017, 38(2): 53-66.

[4] ZAREIE A, SHEIKHAHMADI A, JALILI M. Identification of influential users in social networks based on users' interest[J]. Information Sciences, 2019, 493: 217-231.

[5] ZAREIE A, SHEIKHAHMADI A, KHAMFOROOSH K. Influence maximization in social networks based on TOPSIS[J]. Expert Systems With Applications, 2018, 108: 96-107.

[6] LU F, ZHANG W K, SHAO L W, et al. Scalable influence maximization under independent cascade model[J]. Journal of Network and Computer Applications, 2017, 86: 15-23.

[7] SHEIKHAHMADI A, NEMATBAKHSH M A, SHOKROLLAHI A. Improving detection of influential nodes in complex networks[J]. Physica A: Statistical Mechanics and Its Applications, 2015, 436: 833-845.

[8] 刘露, 胡封晔, 牛亮, 等. 异质网络中基于节点影响力的相似度度量方法[J]. 电子学报, 2019, 47(9): 1929-1936.
LIU L, HU F Y, NIU L, et al. Node influence based similarity measure method in heterogeneous network[J]. Acta Electronica Sinica, 2019, 47(9): 1929-1936.

[9] ZAREIE A, SHEIKHAHMADI A, JALILI M. Influential node ranking in social networks based on neighborhood diversity[J]. Future Generation Computer Systems, 2019, 94: 120-129.

[10] 韩忠明, 陈炎, 李梦琪, 等. 一种有效的基于三角结构的复杂网络节点影响力度量模型[J]. 物理学报, 2016, 65(16): 289-300.
HAN Z M, CHEN Y, LI M Q, et al. An efficient node influence metric based on triangle in complex networks[J]. Acta Physica Sinica, 2016, 65(16): 289-300.

[11] HUANG C Y, LEE C L, WEN T H, et al. A computer virus spreading model based on resource limitations and interaction costs[J]. Journal of Systems and Software, 2013, 86(3): 801-808.

[12] JALILI M, PERC M. Information cascades in complex networks[J]. Journal of Complex Networks, 2017, 5(5): 665-693.

[13] NOWZARI C, PRECIADO V M, PAPPAS G J. Analysis and control of epidemics: a survey of spreading processes on complex networks[J]. IEEE Control Systems Magazine, 2016, 36(1): 26-46.

[14] 韩忠明, 陈炎, 刘雯, 等. 社会网络节点影响力分析研究[J]. 软件学报, 2017, 28(1): 84-104.
HAN Z M, CHEN Y, LIU W, et al. Research on node influence analysis in social networks[J]. Journal of Software, 2017, 28(1): 84-104.

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

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

[17] SABIDUSSI G. The centrality of a graph[J]. Psychometrika, 1966, 31(4): 581-603.

[18] KITSAK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11): 888-893.

[19] 于冬梅, 高雷阜, 赵世杰. 考虑凸形障碍的应急设施选址与资源分配决策研究[J]. 系统工程理论与实践, 2019, 39(5): 1178-1188.
YU D M, GAO L F, ZHAO S J. Emergency facility location-allocation problem with convex barriers[J]. Systems Engineering-Theory & Practice, 2019, 39(5): 1178-1188.

[20] WANG Y, CONG G, SONG G J, et al. Community-based greedy algorithm for mining top-K influential nodes in mobile social networks[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and Data Mining. New York: ACM Press, 2010: 1039-1048.

[21] GONG M G, YAN J N, SHEN B, et al. Influence maximization in social networks based on discrete particle swarm optimization[J]. Information Sciences, 2016, 367/368: 600-614.

[22] BAO Z K, LIU J G, ZHANG H F. Identifying multiple influential spreaders by a heuristic clustering algorithm[J]. Physics Letters A, 2017, 381(11): 976-983.

[23] CUI L Z, HU H X, YU S, et al. DDSE: a novel evolutionary algorithm based on degree-descending search strategy for influence maximization in social networks[J]. Journal of Network and Computer Applications, 2018, 103: 119-130.

[24] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer[J]. Advances in Engineering Software, 2014, 69: 46-61.

[25] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 137-146.

[26] GOLDENBERG J. LIBAI B. MULLER E. Using complex systems analysis to advance marketing theory development: modeling heterogeneity effects on new product growth through stochastic cellular automata, academy of marketing science review 9[J]. Monthly Labor Review, 2001, 31(3): 8-11.

[27] PENG S C, YU S, YANG A M. Smartphone malware and its propaga-

tion modeling: a survey[J]. IEEE Communications Surveys & Tutorials, 2014, 16(2): 925-941.

[28] 李理, 单而芳. 图上博弈的 Page-Shapley 值[J]. 系统工程理论与实践, 2019, 39(11): 2771-2783.
LI L, SHAN E F. The Page-Shapley values for graph games[J]. Systems Engineering-Theory & Practice, 2019, 39(11): 2771-2783.

[29] CHEN W, WANG Y J, YANG S Y. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 1-9.

[30] WANG X J, SU Y Y, ZHAO C L, et al. Effective identification of multiple influential spreaders by DegreePunishment[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 461: 238-247.

[31] GUO L, LIN J H, GUO Q, et al. Identifying multiple influential spreaders in term of the distance-based coloring[J]. Physics Letters A, 2016, 380(7/8): 837-842.

[32] JIANG Q, SONG G, CONG G, et al. Simulated annealing based influence maximization in social networks[C]//Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2011: 127-132.

[33] SUN Y F. Optimal selection of nodes to propagate influence on networks[J]. The European Physical Journal B, 2016, 89(11): 253.

[34] BORRERO J S, PROKOPYEV O A, KROKHMAL P. Optimization of cascading processes in arbitrary networks with stochastic interactions[J]. IEEE Transactions on Network Science and Engineering, 2019, 6(4): 773-787.

[35] 於志勇, 陈基杰, 郭昆, 等. 基于影响力与种子扩展的重叠社区发现[J]. 电子学报, 2019, 47(1): 153-160.
YU Z Y, CHEN J J, GUO K, et al. Overlapping community detection based on influence and seeds extension[J]. Acta Electronica Sinica, 2019, 47(1): 153-160.

[36] 王炯滔, 金明, 李有明, 等. 基于 Friedman 检验的非参数协作频谱感知方法[J]. 电子与信息学报, 2014, 36(1): 61-66.
WANG J T, JIN M, LI Y M, et al. Nonparametric cooperative spectrum sensing algorithm based on Friedman test[J]. Journal of Electronics & Information Technology, 2014, 36(1): 61-66.

[作者简介]

顾秋阳（1995− ），男，浙江杭州人，浙江工业大学博士生，主要研究方向为智能信息处理、数据挖掘、中小企业高质量发展等。

吴宝（1979− ），男，浙江金华人，博士，浙江工业大学研究员、博士生导师，主要研究方向为复杂网络链路预测、金融信用风险控制与中小企业发展。

孙兆洋（1979− ），女，北京人，博士，中国标准化研究院副研究员，主要研究方向为智能信息处理与数据挖掘。

池仁勇（1959− ），男，浙江温州人，博士，浙江工业大学教授、博士生导师，主要研究方向为复杂网络链路预测、中小企业智能信息管理与创新创业。