软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(9):2062−2077 [doi: 10.3724/SP.J.1001.2013.04400]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于多目标优化的网络社区发现方法*

黄发良[1], 张师超[2], 朱晓峰[2,3]

[1](福建师范大学 软件学院,福建 福州 350007)

[2](广西师范大学 计算机科学与信息工程学院,广西 桂林 541000)

[3](School of Information Technology and Electrical Engineering, University of Southern Queensland, Australia)

通讯作者: 黄发良, E-mail: faliang.huang@gmail.com

摘 要: 社区发现是复杂网络挖掘中的重要任务之一,在恐怖组织识别、蛋白质功能预测、舆情分析等方面具有重要的理论和应用价值.但是,现有的社区质量评判指标具有数据依赖性与耦合关联性,而且基于单一评判指标优化的网络社区发现算法有很大的局限性.针对这些问题,将网络社区发现问题形式化为多目标优化问题,提出了一种基于多目标粒子群优化的网络社区发现算法 MOCD-PSO,它选取模块度 *Q*、最小最大割 MinMax*Cut* 与轮廓(silhouette)这 3 个指标进行综合寻优.实验结果表明,MOCD-PSO 算法具有较好的收敛性,能够发现分布均匀且分散度较高的 Pareto 最优网络社区结构集,并且无论与单目标优化方法(GN 与 GA-Net)相比较,还是与多目标优化算法(MOGA-Net 与 SCAH-MOHSA)相比较,MOCD-PSO 算法都能在无先验信息的条件下挖掘出更高质量的网络社区.

关键词: 复杂网络;社区挖掘;多目标粒子群优化

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 黄发良,张师超,朱晓峰.基于多目标优化的网络社区发现方法.软件学报,2013,24(9):2062−2077. http://www.jos.org.cn/1000-9825/4400.htm

英文引用格式: Huang FL, Zhang SC, Zhu XF. Discovering network community based on multi-objective optimization. Ruan Jian Xue Bao/Journal of Software, 2013,24(9):2062−2077 (in Chinese). http://www.jos.org.cn/1000-9825/4400.htm

Discovering Network Community Based on Multi-Objective Optimization

HUANG Fa-Liang[1], ZHANG Shi-Chao[2], ZHU Xiao-Feng[2,3]

[1](Faculty of Software, Fujian Normal University, Fuzhou 350007, China)

[2](College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541000, China)

[3](School of Information Technology and Electrical Engineering, University of Southern Queensland, Australia)

Corresponding author: HUANG Fa-Liang, E-mail: faliang.huang@gmail.com

Abstract: Community discovery is an important task in mining complex networks, and has important theoretical and application value in the terrorist organization identification, protein function prediction, public opinion analysis, etc. However, existing metrics used to measure quality of network communities are data dependent and have coupling relations, and the community discovery algorithms based on optimizing just one metric have a lot of limitations. To address the issues, the task to discover network communities is formalized as a multi-objective optimization problem. An algorithm, MOCD-PSO, is used to discover network communities based on multi-objective particle swarm optimization, which constructs objective function with modularity *Q*, *MinMaxCut* and silhouette. The experimental results show that the proposed algorithm has good convergence and can find Pareto optimal network communities with relatively well uniform

* 基金项目: 国家自然科学基金(61170131, 61263035); 澳大利亚 ARC(DP0985456); 国家高技术研究发展计划(863)(2012AA011005); 国家重点基础研究发展计划(973)(2013CB329404); 教育部人文社会科学研究青年基金(12YJCZH074); 福建师范大学优秀青年骨干教师培养基金(fjsdjk2012082); 科学计算与智能信息处理广西高校重点实验室开放基金(GXSCIIP201212)

收稿时间: 2012-10-19; 修改时间: 2013-01-21; 定稿时间: 2013-03-22; jos 在线出版时间: 2013-04-27

CNKI 网络优先出版: 2013-04-27 15:35, http://www.cnki.net/kcms/detail/11.2560.TP.20130427.1535.001.html

and dispersive distribution. In addition, compared with the classical algorithms based on single objective optimization (GN, GA-Net) and multi-objective optimization (MOGA-Net,SCAH-MOHSA), the proposed algorithm requires no input parameters and can discover the higher-quality community structure in networks.

Key words: complex network; communities mining; multi-objective particle swarm optimization

现实世界中的许多复杂系统都可以用复杂网络加以表示,例如社会网络、生物网络、电力网络、Web 网络等.通过两个简单的映射,即对象到网络节点的映射与对象间关系到边的映射,可将复杂网络表示成图模型.复杂网络研究正在吸引着来自物理学、生物学、社会学与计算机科学等不同领域学者的关注.除了广为人知的小世界与无标度等特性外,复杂网络还具有极为重要的模块性,即复杂网络中隐含着丰富的社区结构模式.按照文献中对网络社区的描述,可以松散地将其定义为具有某种共同特征的相互连接的信息载体集合,例如,隶属于某个特定主题的 Web 页面集合,由具有某种共同兴趣爱好的微博者组成的微群,等等.从网络拓扑结构来看,一个网络社区就是一个网络图的稠密连通子图,在这个子图内的节点之间,连接密度高于子图内部节点与外部节点的连接密度.复杂网络社区发现的研究成果已被成功应用于诸如恐怖组织识别、蛋白质功能预测、舆情分析与处理等众多领域当中[1].

网络社区发现研究正在吸引着复杂网络研究者的广泛关注,近年来涌现出大量的方法,文献[2]对这些方法进行了系统分析并给出了初步的分类.从数据挖掘层面上看,网络社区发现的本质是基于网络链接的聚类学习,其目标是将网络节点集划分为多个内部链接紧密而外部链接稀疏的簇.从聚类学习的角度上讲,网络社区发现算法的质量很大程度上取决于网络社区结构质量评判指标的设计思路与优化策略.目前,网络社区发现算法中目标函数(网络社区结构质量评判指标)的优化求解策略大致可归纳为两类:基本启发式方法和超启发式方法.前者将复杂网络社区发现问题转化为预定义启发式规则的设计问题,根据各种社区质量评判指标的特征设计优化策略;后者利用各种超启发式算子在网络社区发现问题空间中对社区质量评判指标进行寻优.

社区质量评判指标的基本启发式方法可分为直接贪心法与间接贪心法.直接贪心法的思想非常简单,就是初始化网络为$|V|$个社区,反复迭代如下过程直到算法终止条件满足:计算各边相对于模块度的信息增益度,选取使社区质量评判指标值增量最大的边加入,从而实现社区合并.直接贪心法的代表算法就是模块度指标值 Q 的优化算法,原始的 Q 优化算法的时间复杂度为 $O((m+n)n)$或 $O(n^2)$[3].为了提高算法的效率与效果,提出了一系列的改进方法:Clauset 等人设计了数据结构 max-heaps 将算法时间复杂度降低到 $O(\text{mdlogn})$;Danon 等人提出对 Q 值增量进行规范化处理以发现与社区大小具有较大差异的社区结构[4];Wakita 等人提出利用合并比(consolidation ratio)对 Q 值增量进行加权来提高算法的扩展性[5];Blondel 等人提出在迭代合并的过程中允许多个社区合并而不仅仅是两个社区的合并[6].间接贪心法的基本思想是:将整个网络视为一个社区,然后循环如下过程直到社区质量评判指标值 Q 满足给定条件:选择具有某种特性的边并删除来实现网络社区的发现.该方法的选择策略主要包括:边介中性(betweenness)大者优先[7]、边聚集系数(clustering coefficient)小者优先[8]、边信息中心度(information centrality)大者优先[3]与边稳定系数(stability coefficient)大者优先[9].除了对边进行贪心的方法外,淦文燕等人提出了一种基于拓扑势的社区发现算法,将每个社区视为拓扑势场的局部高势区,通过对局部极大势值节点进行贪心寻优来实现网络的社区划分[10].

基本启发式方法的思想简单直观,容易实现,但是,该类方法需要借助先验知识定义递归终止条件,不具备自动识别网络社区总数的能力,这在很大程度上限制了此类优化方法在现实复杂网络社区发现中的应用.

为了克服基本启发式方法的不足,研究者们提出了一类用于优化社区质量评判指标的超启发式方法,主要包括基于单一目标的优化算法与基于多目标的优化算法.Tasgin 等人通过利用 GA(genetic algorithm)算法优化社区模块度 Q 函数来实现网络最优划分的近似[11].Pizzutiz 首先给出用于评判网络划分质量的社区分数(community score)的定义,然后运用 GA-Net 进行优化网络划分[12].考虑到社会网络的海量性,Lipczak 等人[13]提出一种基于社区足够小且社区数有限假设的 ACGA 算法:将一个社区编码为一个个体,根据社区质量潜在提高量来选择个体进行遗传操作.段晓东等人引入粒子群算法对网络进行迭代二分实现 Web 社区的发现[14].CDPSO 算法[15]采用基于节点邻居有序表的粒子编码方式,通过 PSO 全局搜索来挖掘社区.Gog 等人提出一种基于个体

信息共享机制的协同进化算法对网络社区结构进行寻优[16],结合局部搜索的 GA 变体算法 CCGA[17]与 LGA[18]通过优化社区质量评判指标 Q 来实现大规模复杂网络的社区发现.Zhu 与 Wang 提出挖掘网络社区的并行遗传算法 PGA[19].

尽管上述这些基于单一目标优化算法[12−19]具有较好的时间效率且能够挖掘出满足某种特定目标的网络社区结构,但是,实际应用中的网络社区发现问题常需要兼顾多个目标,且这些目标可能是相互冲突的.显然,基于单一目标优化的社区发现方法无法满足这样的应用需求.因此,基于多目标优化的社区发现开始受到关注.公茂果等人提出一种用于网络社区发现的基于数学规划方法与进化算法相结合的多目标优化算法,同时对内部链接密度与外部链接密度进行优化[20].多目标优化算法(NNIA-Net[21],MOGA-Net[22],MOHSA[23]与 SCAH-MOHSA[24])都是选取社区评分(community score)与社区适应度(community fitness)作为优化目标来实现网络社区的挖掘,不同的是所采用的超启发式方法:NNIA-Net 运用免疫算法,MOGA-Net 运用 GA 算法,MOHSA 自适应混合多目标和谐搜索算法,SCAH-MOHSA 在 MOHSA 基础上添加一个谱聚类的预处理算子.

与单一目标优化算法相比较,这些多目标优化算法能够对各种社区质量评判指标进行综合考量,可以发现更多更高质量的网络社区结构.由于基于多目标优化的社区发现研究刚刚起步,还存在着一些不足,比如,当前的算法都是假设网络社区质量评判指标之间是可能不相一致的,而没有对假设是否成立给出理论证明或者实验验证,也没有研究网络社区质量评判指标之间的关系性质.另外,现有的基于多目标优化的社区发现方法几乎都是基于遗传算法,将多目标粒子群优化算法用于社区发现的研究报告尚未见到,而相关研究表明多目标粒子群优化算法具有优良的全局寻优能力[25].

本文首先从实验的角度验证了网络社区质量评判指标耦合关联性与数据依赖性的存在性,由此推导出进行多目标优化的社区发现的必要性,接着对多目标优化网络社区的问题进行形式化描述,然后提出了一种基于多目标粒子群优化的网络社区发现算法 MOCD-PSO,该方法通过同时优化多个网络社区质量评判指标产生 Pareto 最优社区划分集合,用户可以根据特定需要从中选择最满意的社区结构.实验结果表明,无论与单目标优化方法(GN 与 GA-Net)比较,还是与多目标优化算法(MOGA-Net 与 SCAH-MOHSA)相比较,MOCD-PSO 算法总能在无先验信息的条件下挖掘出更高质量的网络社区.

本文第 1 节对现有网络社区质量评判指标进行简介.第 2 节通过实验分析得出社区质量评判指标具有数据依赖性与耦合关联性,并对两种性质进行形式化描述.第 3 节给出与多目标优化网络社区相关的形式定义.第 4 节给出 MOCD-PSO 算法的详细描述及性能分析.第 5 节给出真实网络数据与人工网络数据的测试与比较结果.第 6 节对全文进行总结.

1 相关工作

由于“关于什么是社区”缺乏严格统一的定义,研究者们就从不同的应用场景与理论角度入手,定义出各种不同形式的社区,而根据社区的松散定义,可以把这些形形色色的社区归为 3 大类:具有最大内部链接密度的连通子图,具有最小外部链接密度的连通子图,同时具有最大内部链接密度与最小外部链接密度的连通子图.研究人员为了从网络中挖掘出各自定义的不同社区,提出了多种用于测度网络社区质量的量化指标.本节在给出网络社区结构形式定义(定义 1)的基础上,列出了常用的社区质量评判指标(表 2),表 1 是各种度量指标的相关符号.更多的社区质量评判指标见文献[26].

当前,使用最广泛的社区质量评判指标就是 Newman 与 Girvan 提出的 Q 值函数[27],它是通过比较网络子图与其对应随机图零模型的连接密度来定义的,Q 值越大,则网络社区质量越高.然而该指标存在着粒度有限的缺点[28],Li 等人[29]将其改进为模块度密度指标 Q_{Li}.评判指标 MinMax*Cut*[30]试图在最大化社区内节点相似度的同时最小化社区间节点相似度,MinMax*Cut* 值越小,则网络社区质量越高.评判指标 *silhouette*[31]是借鉴于图形显示中的同一簇内的像素点值相似的观点来度量网络社区质量,该指标中的 $a_i = avg_{j\in V_l}(A_{ij}), i\in V_l$ 表示社区 V_l 内节点 i 与社区 V_l 中其他节点的平均相似度,$b_i = \max_{V_k\in V}\left(\frac{1}{|V_k|}\sum_{j\in V_k}A_{ij}\right), i\in V_l, l\neq k$ 表示社区 V_l 内节点 i 与其他

社区的最大平均相似度,*silhouette* 指标值越大,则网络社区质量越高.

定义 1(网络社区结构). 给定无向网络 $G=(V,E)$,网络节点集合为 V,网络边集合为 $E=\{e=(u,v)|u\in V,v\in V\}$,$G$ 用一个大小为$|V|\times|V|$的矩阵 A 来表示,若边 $e=(i,j)\in E$,则 $A_{ij}=1$;否则,$A_{ij}=0$.网络社区结构就是网络节点集合 V 的一个 m 划分方案 $P=(V_1,V_2,\ldots,V_m)$,其中,V_i 必须满足 4 个条件:$V_i\subseteq V$,$V_i\neq\varnothing(i=1,2,\ldots,m)$,$\bigcup_{i=1}^{m}V_i=V$ 与 $V_i\cap V_j=\varnothing(i\neq j)$.

Table 1 Symbols and notations

表 1 相关符号

Symbol	Remark
$d_j^{in}=\sum_{j\neq k,k\in V_i}A_{jk}$	社区 V_i 内的节点 j 的内部度
$d_j^{out}=\sum_{j\neq k,k\in \overline{V_i}}A_{jk}$	社区 V_i 内的节点 j 的外部度
$d(j,V_k)=\sum_{i\in V_k}A_{ji}$	社区 V_i 内的节点 j 与社区 V_k 的关联度
$d^{in}(V_i)=\sum_{j\neq k,\,j\in V_i,k\in V_i}A_{jk}$	社区 V_i 内部度
$d^{out}(V_i)=\sum_{j\neq k,\,j\in V_i,k\in \overline{V_i}}A_{jk}$	社区 V_i 外部度
$d(V_i,V_j)=\sum_{i'\in V_i,\,j'\in V_j}A_{i'j'}$	社区 V_i 与社区 V_j 的关联度

Table 2 Common metrics to measure network community goodness

表 2 常用的社区质量评判指标

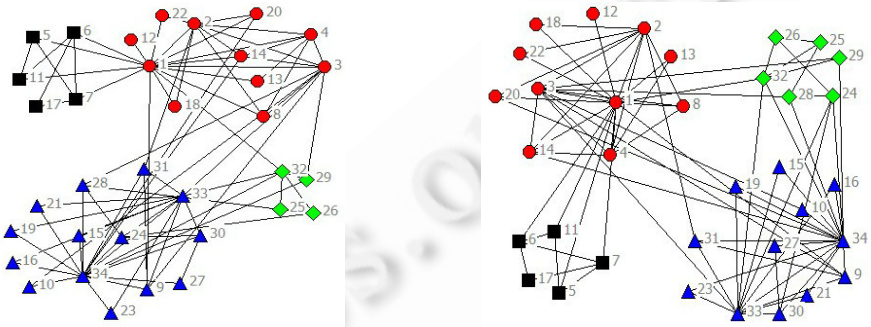
Metrics	Formula
Q	$Q=\sum_{i=1}^{m}\left(\frac{d^{in}(V_i)}{d^{in}(V)}-\left(\frac{d^{out}(V_i)}{d^{in}(V)}\right)^2\right)$
Q_{Li}	$Q_{Li}=\sum_{i=1}^{m}\left(\frac{d^{in}(V_i)-d^{out}(V_i)}{\lvert V_i\rvert}\right)$
MinMax*Cut*	$MMC=\sum_{i=1}^{m}\frac{d^{out}(V_i)}{d^{in}(V_i)}$
silhouette	$GS=\frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{\lvert V_j\rvert}\sum_{v_i\in V_j}\frac{a_i-b_i}{\max(b_i,a_i)}\right)$
ductance[32]	$\sum_{i=1}^{m}\frac{d^{out}(V_i)}{2d^{in}(V_i)+d^{out}(V_i)}$
Expansion[33]	$\sum_{i=1}^{m}\frac{d^{out}(V_i)}{\lvert V_i\rvert}$
NCut[34]	$\sum_{i=1}^{m}\frac{d^{out}(V_i)}{2d^{in}(V_i)+d^{out}(V_i)}+\frac{d^{out}(V_i)}{2(\lvert E\rvert-d^{in}(V_i))+d^{out}(V_i)}$
community score[23]	$\sum_{i=1}^{k}\left(\sum_{j,k\in V_i}A_{jk}\right)\left(\frac{1}{\lvert V_i\rvert}\sum_{j\in V_i}\left(\frac{k_j^{in}(V_i)}{\lvert V_i\rvert}\right)^r\right)$
community fitness[23]	$\sum_{i=1}^{k}\sum_{j\in V_i}\frac{k_j^{in}(V_i)}{(k_j^{in}(V_i)+k_j^{out}(V_i))^{\alpha}}$

2 社区质量评判指标的性质

社区质量评判指标值有最大化最优,亦有最小化最优,二者可相互转化.不失一般性,本文仅讨论最大化最优的情形.下面,我们首先通过实验法来探讨网络社区质量评判指标耦合关联性与数据依赖性的存在性,然后给出对应的形式定义.

实验 1. 检验社区质量评判指标的耦合关联性.实验数据:Karate 网络;算法框架:CDPSO[15];算法框架内嵌的社区质量评判指标:模块度 *Q*,*silhouette* 值 *GS*.

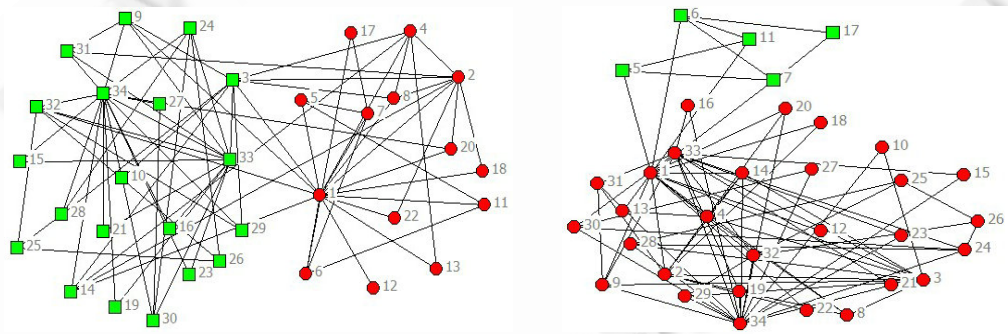
算法(CDPSO+*Q*)是以社区质量评判指标模块度 *Q* 为单一目标的网络社区结构优化算法,而算法(CDPSO+*GS*)是以社区质量评判指标 *silhouette* 值为单一目标的网络社区结构优化算法.从图 1 中可以看出:在迭代次数为 91 时,*Q* 与 *GS* 分别为 *Q*=0.415,*GS*=0.789;当迭代次数为 105 时,*Q* 与 *GS* 分别为 *Q*=0.42,*GS*=0.777.这说明在优化模块度 *Q* 的迭代过程中,*silhouette* 值并没有得到同步优化.类似地,由图 2 中可知:在迭代次数为 85 时,*GS* 与 *Q* 分别为 *GS*=0.861,*Q*=0.335;当迭代次数为 96 时,*GS* 与 *Q* 分别为 *Q*=0.133,*GS*=0.952.这说明在优化 *silhouette* 值的迭代过程中,模块度 *Q* 也并没有得到同步优化.



(a) (*T*=91,*Q*=0.415,*GS*=0.789) (b) (*T*=105,*Q*=0.42,*GS*=0.777)

Fig.1 Network community structure generated in the iteration process of algorithm (CDPSO+*Q*)

图 1 算法(CDPSO+*Q*)迭代过程中的网络社区结构图



(a) (*T*=85,*GS*=0.861,*Q*=0.335) (b) (*T*=96,*Q*=0.133,*GS*=0.952)

Fig.2 Network community structure generated in the iteration process of algorithm (CDPSO+*GS*)

图 2 算法(CDPSO+*GS*)迭代过程中的网络社区结构图

实验 2. 社区质量评判指标的数据依赖性实验.实验数据:运用 LFR 网络数据生成器[35]构造的 6 个人工网络 SynNet_1~SynNet_6;算法框架:CDPSO;算法框架内嵌的社区质量评判指标:模块度 *Q*,*silhouette* 值 *GS*.采用不同评判指标优化策略得到的网络社区结构的 *F*-Measure[36]值见表 3.由 *F*-Measure 定义可知,网络社区结构对应的 *F*-Measure 值越大,则表明其质量越高.从表 3 可以看出,相比较模块度 *GS*,以指标 *Q* 为优化目标的算法能更好地发现 SynNet_2,SynNet_3,SynNet_4 与 SynNet_6 中的隐含社区结构,类似地,相比较模块度 *Q*,以 *GS* 指标为优化目标的算法能够挖掘出 SynNet_1 与 SynNet_5 中的更优质社区结构.由此实验可知两种不同的评判指标在不同网络数据集的优化性能表现不一样.

网络社区质量评判指标耦合关联性与数据依赖性的形式定义如下:

给定网络 $G=(V,E),|V|=N$,社区划分函数 $clusterer:V\to P,|P|=2^N$,社区质量评判指标向量 $F=(F_1,F_2,\ldots,F_k)$,其中,$F_i:p\to R(p\in P,i=1,\ldots,k)$.

性质 1(耦合关联性). 对于网络 G,在社区划分算法 clusterer 的第 t_m 次与第 t_n 次迭代($m>n$)时,社区质量评判指标 F_i 有 $F_i(t_m)>F_i(t_n)$,$\exists F_j(t_m)<F_j(t_n)\wedge j\neq i$,称 F_i 与 F_j 存在耦合关联性.

性质 2(数据依赖性). 对于网络 G_1,其先验社区结构为 P_1^{true},聚类有效性函数 CV(诸如 F-Measure,Clustering Error),社区划分函数 clusterer 分别采用最优化评判指标 F_i 与 F_j 对网络 G_1 进行划分,获得社区结构为 P_1^i 与 P_1^j,$F\text{-}Measure(P_1^{\text{true}},P_1^i)>F\text{-}Measure(P_1^{\text{true}},P_1^j)$,记为 $F_i \overset{G_1}{\succ} F_j$.而对具有先验社区结构 P_2^{true} 的网络图 G_2,有 $F_j \overset{G_2}{\succ} F_i$,社区质量评判指标 F_i 与 F_j 具有数据依赖性.

Table 3 Data dependency of modularity metrics Q and *silhouette*

表 3 模块度 Q 与 GS 的数据依赖性

Dataset	(CDPSO+Q)	CDPSO+GS
SynNet_1	0.72	0.73
SynNet_2	0.75	0.70
SynNet_3	0.73	0.72
SynNet_4	0.74	0.72
SynNet_5	0.72	0.75
SynNet_6	0.73	0.71

理论上,可以根据先验的权系数将不同的评判指标合成一个评判指标,进而采用单目标优化策略.但由于性质 1 的存在,即各评判指标之间的耦合关联性,使得这种做法很难保证得到最优解.

从实验 2 可以看出,对于某个具体的网络而言,的确存在着某个最优的评判指标,但这个最优评判指标是因网络而异,而且无法事先得知.因此,性质 2 的存在说明了只选取某种质量评判指标进行网络社区发现是具有局限性的.

3 多目标优化网络社区的形式描述

性质 1 与性质 2 说明了采用多目标优化方法解决网络社区发现问题的必要性,为了更好地理解与描述多目标优化网络社区的问题,本节从 Pareto 最优[37]的角度对此问题进行形式化描述.

定义 2(Pareto 支配关系). 对于由社区划分函数 $clusterer$:$V\rightarrow P$ 发现的蕴含于网络 $G=(V,E)$中的两种不同社区结构 $P_1,P_2\in P$,社区结构 P_1 Pareto 支配社区结构 P_2 当且仅当下式成立,并记为 $P_1\succ P_2$:

$$\forall i\in(1,2,\ldots,n):F_i(P_1)\geqslant F_i(P_2) \tag{1}$$

定义 3(Pareto 相等关系). 类似 Pareto 支配关系,社区结构 P_1 与社区结构 P_2 Pareto 相等当且仅当式(2)成立,并记为 $P_1=P_2$:

$$\forall i\in(1,2,\ldots,n):F_i(P_1)=F_i(P_2) \tag{2}$$

定义 4(Pareto 不可明辨关系). 类似 Pareto 支配关系,社区结构 P_1 与社区结构 P_2 满足 Pareto 不可明辨关系当且仅当式(3)成立,并记为 $P_1\lozenge P_2$:

$$\neg(P_1\succ P_2)\wedge\neg(P_2\succ P_1) \tag{3}$$

定义 5(Pareto 最优社区结构). 若网络社区划分方案集合 P 中的某一划分方案 P^*被称为 Pareto 最优社区结构,当且仅当如下条件成立:

$$\neg\exists P_i\in P:P_i\succ P^* \tag{4}$$

定义 6(Pareto 最优社区结构集). 所有 Pareto 最优社区结构组成的集合 $PS=\{P^*|\neg\exists P_i\in P:P_i\succ P^*\}$称为 Pareto 最优社区结构集.

定义 7(Pareto 前沿). Pareto 最优社区结构集的社区所对应的质量评判指标向量组成的集合,称为 Pareto 前沿.

定义 8(网络社区发现). 给定网络 $G=(V,E)$,F 为用户指定的目标函数集合,V 所对应的 k 划分的网络社区结构集合 $\boldsymbol{P}$,k 值可由基于整数编码的各种进化聚类算法(例如本文提出的 MOCD-PSO)决定或用户给出.网络社区

发现的过程就是寻找使 F 函数达到最优化的划分 $P^* = \arg\max_{P_i \in P} F(P_i)$.

根据上述定义,网络社区结构划分的多目标优化问题可用如下数学模型加以描述:

$$\left.\begin{aligned} \max F(X) &= (F_1(X), F_2(X), ..., F_n(X)) \\ \text{s.t.} \quad g_j(X) &\leqslant 0 \ \ (j = 1, 2, ..., p) \\ h_k(X) &= 0 \ \ (k = 1, 2, ..., q) \end{aligned}\right\} \tag{5}$$

其中,X 为给定网络的某种社区划分方案.X 的具体表示形式由进化算法的个体编码方式而定,为目标向量,目标函数 $F_i(X)$(i=1,…,n)为第 i 种社区质量评判指标,$g_j(X)$与 $h_k(X)$为约束函数,借此约束函数可以指定算法发现满足某些特定条件的网络社区结构.

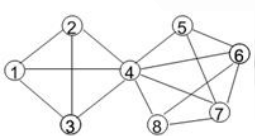
4 MOCD-PSO 算法

由性质 1 与性质 2 可知,设计基于优化多种社区质量评判指标的社区挖掘算法可以较好地克服以单一评判指标为优化目标的网络社区发现算法所具有的不足,因此,本节尝试设计一种基于多目标离散粒子群优化的网络社区发现算法 MOCD-PSO,该算法是在多目标演化优化的 NSGA-II 算法框架下,采用基于 PSO 的繁殖策略,选取 3 种代表性的评判指标(Q,MinMaxCut 与 GS)进行网络社区结构优化.值得指出的是,该算法的设计思想具有很好的推广性,可以对算法框架、选择更新策略、繁殖策略、个体编码策略与社区质量评判指标进行合理组合,从而形成一个基于多目标优化的网络社区发现算法家族,MOCD-PSO 是此家族的一分子而已.当前,代表性的算法框架[38]主要包括 NSGA-II、MOEA/D、基于偏好的 MOEAs、基于指示器(indicator)的 MOEAs、Hybrid MOEAs 与 Memetic MOEAs;选择更新策略主要有基于个体指标排序的方法(轮盘赌选择法、随机遍历抽样法、锦标赛选择法等)与基于群体指标的方法;个体繁殖策略[38]主要有基于差分进化的方法、基于免疫的方法、基于 PSO 的方法、基于概率模型的方法与基于模拟退火的方法;个体编码方法[38]有二进制编码、整数编码与实数编码.

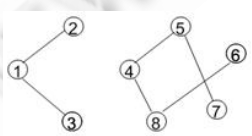
4.1 粒子编码

粒子编码采用 CDPSO 中的基于节点邻居有序表的编码方法,其基本思想是:首先对图中所有节点进行编号,然后对每个节点的邻居根据其编号进行排序形成邻居有序表,在初始化或粒子位置更新阶段生成新粒子时,确保该个体的合法性.以图 3(a)中的网络为例,首先建立各节点的邻居有序表(如图 3(d)所示),根据此表可以对网络社区结构(图 3(b))进行个体编码,结果如图 3(c)所示.该编码过程比较简单,下面以图 3(b)中的节点 1 为例加以说明:对于图 3(b)中的节点 1,其与节点 2 相连接,由图 3(d)中的中心节点 1 的邻居有序表可知,节点 2 是该有序表中的第 1 个元素,故在图 3(c)中,粒子的第 Dim(Dim=1)维的值为 1;类似地,可以得到粒子其他 Dim 所对应的值.该编码方式有如下 3 个优势:

1) 避免非法粒子的产生,能减少多目标优化问题中的约束限制条件数;
2) 自动确定社区数;
3) 避免基于二值编码的迭代二划分策略[14]所遭遇的容易陷入局部最优划分的处境.



(a) 网络



(b) 社区结构

Dim	1	2	3	4	5	6	7	8
Pos	1	1	1	4	3	4	2	1

(c) CDPSO 的粒子编码

中心节点	邻居有序节点列表							
1	2	3	4					
2	1	3	4					
3	1	2	4					
4	1	2	3	5	6	7	8	
5	4	6	7	8				
6	4	5	7	8				
7	4	5	6	8				
8	4	6	7					

(d) 邻居有序表

Fig.3 Encoding particle based on ordered neighbor list

图 3 基于节点邻居有序表的粒子编码

4.2 粒子更新策略

粒子群算法的提出,主要受启发于这样的生物学研究结果:鸟群在进行觅食时会记忆并共享自己所发现的最优觅食路径,通过这种共享信息,鸟群能够更快地找到更优质的食物.该算法在求解最优化问题时,将每个优化问题的解都对应着搜索空间中的一只鸟,或称为粒子,所有的粒子都可以由被优化的函数计算出其适应值;同时,每个粒子还具有相应的速度决定其飞翔的方向和距离.也就是说,整个优化过程是通过对粒子的位置与速度不断更新来实现的.

粒子的速度更新方法有很多,基本上可以概括为如下公式(6):

$$V_i(t+1)=wV_i(t)+c_1\times rand(\cdot)\times(P_i-X_i(t))+c_2\times rand(\cdot)\times(P_g-X_i(t)) \tag{6}$$

其中,$X_i=(x_{i1},x_{i2},\ldots,x_{id})$与 $V_i=(v_{i1},v_{i2},\ldots,v_{id})$分别是粒子 i 的位置与速度,t 为进化代数,w 为惯性系数,c_1 与 c_2 为学习因子,$rand(\cdot)$为均匀分布在[0,1]之间的随机数,$P_i=(p_{i1},p_{i2},\ldots,p_{id})$是粒子 i 的历史最优位置,$P_g=(p_{g1},p_{g2},\ldots,p_{gd})$是粒子 i 当前所处邻域中的最优粒子位置.公式(6)的作用是对在问题解空间中飞行的粒子 P_i 的速度进行调整,即通过记忆自身迄今为止搜索到的最优位置 $P_i=(p_{i1},p_{i2},\ldots,p_{id})$来实现粒子的自学习和通过感知整个粒子群迄今为止搜索到的最优位置 $P_g=(p_{g1},p_{g2},\ldots,p_{gd})$来实现群体信息共享.由于 P_g 是粒子群中具有最佳适应度的粒子,亦称为 leader,即 P_{leader},其主要作用是引导粒子群向包含潜在最优解的解区域方向飞行.粒子邻域拓扑结构决定 P_{leader} 的身份.例如:当粒子邻域拓扑结构为环形时,$P_{leader}=P_{lbest}$;当粒子邻域拓扑结构为全连接时,$P_{leader}=P_{gbest}$;当粒子邻域拓扑结构为星形时,$P_{leader}=P_{focal}$.MOCD-PSO 采用全连接的粒子邻域拓扑结构.

传统离散 PSO 的粒子位置更新策略可形式化为公式(7),MOCD-PSO 对此进行改进为如下公式(8)、公式(9):

$$x_i(t+1)=\begin{cases}1, & \text{if } \rho<sig(v_{ij}(t+1))\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

$$x_{ij}(t+1)=\begin{cases}k, & \text{if } \rho<sig(v_{ij}(t+1)) \ \&\ \deg(v_j)>1\\ x_{ij}(t), & \text{otherwise}\end{cases} \tag{8}$$

$$sig(v_{ij})=\left|\frac{1-\exp(-v_{ij})}{1+\exp(-v_{ij})}\right| \tag{9}$$

其中,k 为除当前连接邻居外的任意随机的邻居节点,即 $k=ceil(rand\times\deg(v_j))$ & $k\neq x_{ij}(t)$;$ceil$ 为上取整函数;$\deg(v_j)$表示节点 v_j 的度(在图 G 中与节点 v_j 关联的边数);ρ为预定阈值.公式(8)的含义可以简单描述为:将通过 S 型函数对粒子速度进行映射,若此映射值大于预定阈值,则将此粒子的位置向量的第 i 个分量赋值为该粒子的与当前分量值不同的邻居节点.基于第 4.1 节粒子编码方法的位置更新公式(8)可使粒子具有更强的搜索能力,而公式(9)是 S 型函数 $sig(v_{ij}(t+1))=1/[1+\exp(-v_{ij}(t+1))]$的修正,其目的是为了提高算法的收敛性,同时,将 v_{ij} 的取值限制在区间[−4,4]内,以防止函数 sig 饱和.

4.3 Leader选择策略

作为多目标离散 PSO 算法,MOCD-PSO 会在迭代优化的过程中形成多个非支配解(nondominated),即相互之间不满足 Pareto 支配关系的网络社区划分方案集合 $NonSet$,这提出了一个如何在粒子更新时进行 leader 选择的问题.MOCD-PSO 运用基于核密度估计 leader 选择机制进行 leader 选择(如图 4 所示).为简化说明,令 $|objectives|=2$, $|NonSet|=10$,即,优化目标数为 2,算法当前迭代过程中共有 10 种相互之间不满足 Pareto 支配关系的网络社区划分方案,分布在由目标 1 与目标 2 构成的平面上.对于 $NonSet$ 中的每一个点 x,都存在一个以该点为中心的 r 邻域 $Neighbor(x,r)$,计算该中心点到在其邻域内其他点的距离 $dist$ 的平均值,选择具有最大平均值的中心点作为 leaders.若存在多个这样的 leader,则从中选取具有最多 r 邻居者为 leaders;若此 leaders 含有多个 leader,则从中随机选取一个.该过程可以形式化为算法 leaderSelection.若令粒子的 r 邻域邻居数 $\widetilde{Nb}$,则有该算法的时间复杂度为 $O(\widetilde{Nb}\times|NonSet|)$.

算法. leaderSelection.

输入:满足非支配关系的网络社区划分方案集合 $Nonset$,邻域半径 r;

输出:粒子飞行的 leader.

Step 1: $leaders = \arg\max_{x \in NonSet} \frac{\sum_{y \in Neighbor(x,r)} dist(x,y)}{|Neighbor(x,r)|}$

Step 2: 若$|leaders|>1$,则 $leaders = \arg\max_{x \in leaders} |Neighbor(x,r)|$

Step 3: 若$|leaders|>1$,则 $leaders=randomSelec(leaders)$

Step 4: return leaders

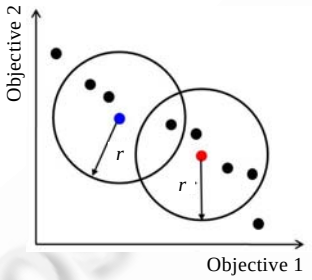


Fig.4 Leader selection based on kernel density estimation

图 4 基于核密度估计的 leader 选择机制

4.4 Pareto最优社区结构集更新策略

研究表明,一个好的多目标优化算法应该能产生尽可能多且分布尽可能均匀平滑的高精确极优解[39].现有大多数多目标优化算法是借助外部存档(external archive,简写 EA)存储 Pareto 最优解,基本思想是:通过对算法当前迭代过程中产生的最优解与 EA 中的元素进行 Pareto 关系的比较来实现 Pareto 最优解的更新,即:当 *EA* 中的元素数目达到其最大容量 MAX*P* 时,直接将新生的非被支配解插入;而当 EA 已满时,随机决定是否用新生的非被支配解代替 EA 中的某个随机元素.显然,这样无法很好地达成一个优质多目标优化算法的期望:使极优解的分布尽可能均匀.在 EA 已达其最大容量且当前迭代过程产生的网络社区划分方案 *CurBestP* 满足如公式(10)所示条件时,会面临这样一个问题:如何从(|*CurBestP*|+MAX*P*)候选 Pareto 最优社区结构集合中选择出 MAX*P* 个具有最大均匀分布特性的 Pareto 最优社区结构.一种朴素的策略就是遍历候选 Pareto 最优社区结构集合的所有大小为 MAX*P* 的组合情况,令 *EA*′为其中一种组合情况,再利用 KL 散度[19] $D_{KL}(EA'\|Y)=\sum_i EA'(i)\ln\frac{EA'(i)}{Y(i)}$ (*Y* 为 *N* 维均匀分布,*N* 为粒子编码向量的长度)来度量 *EA*′的均匀性.然而,高达 $\binom{|CurBestP|+\text{MAX}P}{\text{MAX}P}$ 的组合数使得这种朴素策略在计算上是 prohibitive 的.

值得说明是,由于在多目标优化算法中每个网络社区结构都对应到网络社区结构质量评判指标空间中的一个点,我们希望 MOCD-PSO 算法在寻优过程中找到最优网络社区结构集合所对应的点集合分布更加均匀,因为这样可以加强算法的寻优能力.而在信息论中,KL 散度通常被用来度量两个不同分布之间的差异,故我们采用 KL 散度来度量 *EA* 中的点分布均匀性.

$$\forall P_i \forall P_j (P_i \in EA \wedge P_j \in CurBestP \wedge P_i \lozenge P_j) \tag{10}$$

基于上述分析,我们提出一种启发式策略来实现 Pareto 最优社区结构集更新,给 *EA* 中粒子赋以年龄属性,用以标识其从进入 *EA* 到当前迭代所经历的代数.若简单地实现给 *EA* 中的每个粒子配置一个年龄跟踪器,这一方面会增加空间开销,同时还会加大计算复杂度.由于我们只对粒子的年龄大小顺序感兴趣,而无需知道粒子的精确年龄,因此在具体实现中,我们巧妙地将 *EA* 表示成队列 Q_{PS},这既节省空间又提高了计算效率.该策略可形式化为算法 UpdatePS.

算法. UpdatePS.

输入:外部存档 *EA*,当前迭代过程产生的网络社区划分方案 *CurBestP*;

输出:更新后的 *EA*.

Step 1: 将 *CurBestP* 中的每个网络社区划分方案 P_i 分别与队列 Q_{PS} 中的每个粒子对应的社区划分方案 P_j 进行 Pareto 关系比较,若 $P_i \succ P_j$,则对 Q_{PS} 执行 P_i 入队与 P_j 删除操作,同时将删除 *CurBestP* 中的 P_i;

Step 2: 将 *CurBestP* 中的每个网络社区划分方案 P_i 与队列 Q_{PS} 中的每个粒子对应的社区划分方案 P_j 进行 Pareto 关系比较,若$(P_i \diamond P_j) \vee ((P_i=P_j) \wedge (P_i$ 与 P_j 为不同的社区划分方案)),则:

Step 2-1: 若队列 Q_{PS} 未满,则对 Q_{PS} 执行 P_i 入队操作;否则,执行 Step 2-2;

Step 2-2: 对 Q_{PS} 执行 P_i 入队与 P_j 删除操作;

Step 3: 输出更新后的 *EA*.

从上述步骤不难看出,算法的时间复杂度为 $O(\text{MAX}P \times |CurBestP|)$,若将其置于多目标优化的过程中,则其时间复杂度为 $O(t \times ObjNum \times \text{MAX}P \times |CurBestP|)$,*ObjNum* 为目标数,*t* 为 MOCD-PSO 算法的迭代次数.

4.5 目标函数

MOCD-PSO 算法选择最为常见的 3 个社区质量评判指标(Q,MaxMin*Cut* 与 *GS*)作为目标函数.结合第 3 节,构造网络社区结构划分的多目标优化问题(公式(11)~公式(14)).

$$\text{Max}F(P)=(F_1(P),F_2(P),F_3(P)) \tag{11}$$

$$F_1(P)=Q(P) \text{ s.t. } -1/2 \leqslant F_1(P) \leqslant 1 \tag{12}$$

$$F_2(P)=\frac{1}{1+\text{MinMax}Cut(P)}, \text{ s.t. } 0 \leqslant F_2(P) \leqslant 1 \tag{13}$$

$$F_3(P)=GS(P) \text{ s.t. } -1 \leqslant F_3(P) \leqslant 1 \tag{14}$$

4.6 算法描述与分析

MOCD-PSO 算法的具体步骤描述如下:

Step 1: 设置粒子群规模、粒子位置和速度的范围与维度、粒子群惯性因子、邻域半径以及外部存档最大容量;

Step 2: 建立网络各节点的邻居节点编号表;

Step 3: 采用基于节点邻居有序表的编码方法初始化粒子群;

Step 4: 计算粒子适应度向量 $F(P)$;

Step 5: 进行粒子的 Pareto 支配关系比较;

Step 6: 调用 UpdatePS 算法更新 Pareto 最优社区结构集;

Step 7: 调用 leaderSelection 算法选择粒子飞行的 leader;

Step 8: 根据公式(6)、公式(8)与公式(9)对粒子的位置和速度进行更新;

Step 9: 重复 Step 4~Step 7,直至算法迭代次数达到用户指定的最大迭代次数值,输出全部 Pareto 最优解集元素所对应的网络社区结构.

MOCD-PSO 算法步骤可划分为两大块:第 1 块是 Step 1~Step 3,主要负责初始化算法执行所需要的相关参数与数据结构;第 2 块是 Step 4~Step 9,主要通过粒子群在解空间中协作飞行来实现多个目标函数的寻优,若迭代寻优的结束条件满足,则输出 Pareto 最优解对应的网络社区,算法结束.

假定原网络的节点数为 *n*,由用户指定的最大迭代次数为 *t*,由用户指定的粒子数为 *k*,则算法的复杂度可以估计如下:

首先分析第 1 部分:

Step 1 是算法相关参数的设置,为常量复杂度 $O(1)$;

对于 Step 2,令网络的节点平均度数为 d,又由于真实网络的 d 往往是一个很小的常量,建立各节点的邻居有序表的复杂度为 $O(d\times n\times \log d)$,故有复杂度 $O(n)$;

Step 3 是粒子的初始化,则其复杂度为 $O(n\times k)$.

故第 1 部分的复杂度为 $O(n\times k)$.

然后分析第 2 部分:

Step 4 计算粒子适应度向量,若令一次迭代产生 r 个社区,社区的平均大小为[n/r],则社区内链接度的计算复杂度为 $O(r\times(n/r)^2)=O(n^2/r)$,社区间链接度的计算复杂度为 $O(r\times(n/r)^2\times(r-1))\approx O(n^2)$.故 Step 4 的复杂度约为 $O(n^2)$.

Step 5 是进行粒子的 Pareto 支配关系比较,需要耗费时间为 $O(k^2)$;

Step 6 是调用 UpdatePS 算法,由于此处 *ObjNum* 为常量 3,故其时间复杂度为

$$O(n^2\times ObjNum\times \mathrm{MAX}P\times|CurBestP|)=O(n^2\times \mathrm{MAX}P\times|CurBestP|);$$

Step 7 调用时间复杂度为 leaderSelection 算法,由第 4.3 节得知,其时间复杂度为 $O(\widetilde{Nb}\times|NonSet|)$,由于 $\widetilde{Nb}\ll k$ 且$|NonSet|\leqslant k$,故该步骤在最坏情形下的复杂度为 $O((nk)^2)$;

Step 8 的计算复杂度为 $O(n\times k)$;

Step 9 是停止条件(迭代次数)判定,为常量复杂度 $O(1)$.

故算法第 2 部分的时间复杂度为 $O(t\times n^2\times\max(k^2,\mathrm{MAX}P\times|CurBestP|))$.

综合上述分析可知,MOCD-PSO 算法的复杂度为 $O(t\times n^2\times\max(k^2,\mathrm{MAX}P\times|CurBestP|))$.

由于 MOCD-PSO 算法是建立在多目标粒子群优化算法的框架之上,故算法收敛性的相关理论证明见文献[40].

5 实验与分析

本文实验的硬件环境是:CPU 为 3.4 GHz 的 Intel(R) Core(TM) i7-2600,内存 4G;软件环境为 Windows 7+Matlab R2010.

5.1 数据集

我们在 3 个真实网络与 3 个人工网络上对 MOCD-PSO 算法进行实验分析.3 个真实网络分别是:第 1 个是 Karate 网络,该网络是美国一所大学中的空手道俱乐部成员间的相互社会关系,共有 34 个节点,78 条连接边;第 2 个是 Dolphins 网络,共有 62 个节点、159 条边;第 3 个是 HLM 网络,该网络包含 77 个节点,121 条边,主要依据名著《红楼梦》中的血缘关系与典型的媒介关系构造 5 个家族(宁国府、荣国府、王府、史府和薛府)之间的社会关系网络.3 个人工网络是由 LFR 网络数据生成器[35]模拟生成的.实验网络的特征描述见表 4.

Table 4 Tested networks

表 4 实验中的网络

Networks	Nodes	Edges
Karate	34	78
Dolphins	62	159
HLM	77	121
SynNet_1	100	276
SynNet_2	150	448
SynNet_3	200	523

5.2 收敛性分析

为了评价 MOCD-PSO 算法的收敛性,我们引入世代距离(generation distance,简称 GD)作为评价标准,GD 指标定义为公式(15).GD 越小,说明求得的 Pareto 最优解集越逼近全局最优解集,即算法收敛性也越好.

$$GD=\left(\frac{1}{|PS|}\sum_{i=1}^{|PS|}d_i^2\right)^{\frac{1}{2}} \tag{15}$$

其中,$|PS|$为 Pareto 最优解集中解的数目,d_i 是第 i 个 Pareto 最优解到全局最优解的最小欧氏距离.表 5 给出了 MOCD-PSO 算法对 6 个实验网络分别运行 50 次所得的 GD 指标的均值(mean)与方差(var).由表 5 可以看出,不论在真实网络上,还是在人工网络上,MOCD-PSO 算法都表现出了较好的收敛性.

Table 5 Index GD of tested networks

表 5 各种不同测试网络上的 GD 指标

Tested networks	Mean	Var
Karate	1.2699e−4	2.7850e−6
Dolphins	5.5761e−4	3.4842e−6
HLM	1.0398e−3	6.3236e−4
SynNet_1	4.2176e−3	6.5574e−4
SynNet_2	4.8538e−3	3.9223e−4
SynNet_3	6.9483e−3	7.4313e−4

5.3 最优解分布

本节从 Pareto 最优解的分布均匀性与分散度两个方面对 MOCD-PSO 算法生成的 Pareto 最优社区结构集进行分析.首先利用 *SP* 指标描述 Pareto 最优解在目标空间上的分布均匀性,令网络节点数是 r, F_i^k 表示第 i 个粒子的第 k 维,*SP* 指标可定义为公式(16)~公式(18),*SP* 越小,则意味着 Pareto 最优解集分布越均匀.

$$SP=\sqrt{\frac{1}{|PS|-1}\sum_{i=1}^{|PS|}(\overline{d}-d_i)^2} \tag{16}$$

$$\overline{d}=\frac{1}{|PS|}\sum_{i=1}^{|PS|}d_i \tag{17}$$

$$d_i=\min_{j=1,2,\ldots,|PS|}\left(\sum_{k=1}^{r}|F_i^k-F_j^k|\right) \tag{18}$$

我们把 Pareto 最优社区结构集的传统更新策略与本文的启发式更新策略进行对比,表 6 给出了 MOCD-PSO 算法对 6 个实验网络分别运行 50 次所得的 SP 指标的均值(mean)与方差(var).

Table 6 Distribution uniformity of Pareto optimal community structures

表 6 Pareto 最优社区结构集的分布均匀性

Tested networks	Traditional approach		Our approach	
	Mean	Var	Mean	Var
Karate	3.2358e−3	4.9335e−4	1.8687e−5	3.8156e−6
Dolphins	6.3874e−3	1.2977e−4	2.7603e−5	1.6261e−6
HLM	5.8527e−3	2.5510e−3	8.9090e−4	5.0596e−5
SynNet_1	2.5751e−3	1.4929−3	2.4352e−4	1.9660e−4
SynNet_2	4.7329e−3	3.5166e−4	2.8584e−3	3.1045e−4
SynNet_3	3.3712e−3	3.1122e−4	2.6218e−4	1.6565e−5

从表 6 可以看出,本文的启发式更新策略能使 Pareto 最优社区结构集的解分布得更加均匀.然后运用 Pareto 最优解集的统计量全距(range)(公式(19))进行其分散度分析,从表 7 可以看出,本文的启发式更新策略能使 Pareto 最优社区结构集的解分布更加分散,并且最优解的个数与网络节点数、网络模块度相关.

$$width(F_i(P))=\max\{F_i(P)|P\in PS\}-\min\{F_i(P)|P\in PS\} \tag{19}$$

5.4 有效性分析

为了测评 MOCD-PSO 算法的有效性,我们采用规范化互信息(normalized mutual information,简称 NMI)[41] 的评价标准来衡量 MOCD-PSO 算法的计算结果社区与网络真实社区相一致的程度(公式(20)),该值越大,意味着计算结果社区与网络真实社区相吻合的程度越高.将 MOCD-PSO 算法与 4 个代表性社区发现算法(GN[27],

GA-Net,MOGA-Net 与 SCAH-MOHSA)进行比较分析,其中,前二者为单目标优化算法,后二者为多目标优化算法.表 8 给出了各种算法在 6 个实验网络上分别运行 50 次所得的 *NMI* 的均值与方差.

Table 7 Dispersiveness of Pareto optimal community structures

表 7 Pareto 最优社区结构集的分散度

Tested networks	Traditional approach				Our approach			
	# of solutions	Width of the range			# of solutions	Width of the range		
		$F_1(P)$	$F_2(P)$	$F_3(P)$		$F_1(P)$	$F_2(P)$	$F_3(P)$
Karate	12	0.389 7	0.321 1	0.406 7	21	0.398 9	0.333 2	0.475 5
Dolphins	16	0.403 9	0.360 5	0.544 3	36	0.423 9	0.370 7	0.637 6
HLM	55	0.575 2	0.448 6	0.589 2	85	0.587 4	0.483 8	0.592 2
SynNet_1	97	0.649 1	0.502 5	0.759 9	100	0.686 8	0.512 5	0.817 3
SynNet_2	100	0.622 5	0.537 7	0.748 1	100	0.644 3	0.554 0	0.779 7
SynNet_3	100	0.587 0	0.522 3	0.689 3	100	0.607 0	0.534 9	0.713 2

$$NMI(P^{res},P^{\text{true}})=\frac{-2\sum_{V_m\in P^{res}}\sum_{V_n\in P^{\text{true}}}\frac{|V_m\cap V_n|}{|V|}\log\left(\frac{|V\|V_m\cap V_n|}{|V_m\|V_n|}\right)}{\sum_{V_m\in P^{res}}\frac{|V_m|}{|V|}\log\left(\frac{|V_m|}{|V|}\right)+\sum_{V_n\in P^{\text{true}}}\frac{|V_n|}{|V|}\log\left(\frac{|V_n|}{|V|}\right)} \tag{20}$$

Table 8 Quality comparison of network communities

表 8 网络社区质量的比较

Tested networks	GN	GA-Net	MOGA-Net		SCAH-MOHSA		MOCD-PSO	
			Max_*Avg*	Max_*Std*	Max_*Avg*	Max_*Std*	Max_*Avg*	Max_*Std*
Karate	0.692	0.863	**1**	**0**	**1**	**0**	**1**	**0**
Dolphins	0.573	0.885	**1**	**0**	0.932	0.062	**1**	**0**
HLM	0.654	0.897	0.927	0.038	0.916	0.043	**0.935**	**0.028**
SynNet_1	0.582	0.779	0.814	0.047	0.823	0.051	**0.831**	**0.034**
SynNet_2	0.613	0.813	0.852	0.049	0.848	0.057	**0.857**	**0.031**
SynNet_3	0.647	0.855	0.876	0.054	0.869	0.061	**0.882**	**0.032**

从表 8 可以看出,与单目标算法 GN 与 GA-Net 相比较,MOCD-PSO 算法发现的社区结构质量要远远高于这二者的结果社区结构质量,在诸如 Karate 与 Dophins 之类的社区结构模块性较强的网络中,能够准确发现网络的真实社区结构.与多目标优化算法 MOGA-Net 与 SCAH-MOHSA 相比较,不仅从均值 Min_*Avg*(公式(21))意义上看,MOCD-PSO 算法展现出明显的优势,而且从方差 Max_*Std*(公式(22))意义上看,MOCD-PSO 算法的最佳结果社区结构的 *NMI* 的分布更加均匀,因而,我们的方法具有更强的鲁棒性.

$$\text{Max\_}Avg=\frac{1}{RUNS}\sum_{run=1}^{RUNS}\max_{P^{res}\in PS_{run}}(NMI(P^{res},P^{\text{true}})) \tag{21}$$

$$\text{Max\_}Std=std\{NMI_{run}|run=1,2,\ldots,RUNS\} \tag{22}$$

P^{res} 表示待比较算法生成的第 *res* 个结果社区结构;P^{true} 表示网络的真实社区结构;V_m 表示 P^{res} 中的第 *m* 个社区;V_n 表示 P^{true} 中的第 *n* 个社区;PS_{run} 表示算法在第 *run* 次实验中计算所得的网络社区结构集合;NMI_{run} 表示第 *run* 次实验中计算所得的所有网络社区结构的 *NMI* 的最大值;*RUNS* 表示实验的次数,这里取值为 50 (*RUNS*=50).

5.5 时间效率分析

为了评价 MOCD-PSO 算法的时间性能,我们将 MOCD-PSO 算法与上述两种基于多目标优化的社区发现算法(MOGA-Net 与 SCAH-MOHSA)进行比较分析.为了更好地对这 3 种超启发式的随机搜索算法进行时间性能比较,我们将最大迭代次数的循环终止条件修改为网络社区结构的 Max_*Avg* 大于某个指定阈值.值得指出的是,我们没有选择 GN 算法与 GA-Net 作为比较对象,主要出于这样的考虑:一般来说,基于单目标优化的社区发现算法在时间性能上要优于基于多目标优化的社区发现算法,由于基于多目标优化的社区发现算法需要多个目标函数值,而且目标函数估值是优化算法中的重要耗时部件.图 5 给出了 3 种算法在 6 个实验网络上分别运

行 50 次所得的耗费时间的均值.从图 5 可以看出:在小规模的真实网络 Karate 与 Dolphins 中,MOCD-PSO 的耗时与其他两种算法的耗时相当;在网络 HLM 与 SynNet_1 上,MOCD-PSO 逐步体现出时间效率优势,但优势相对微弱;而在规模更大的网络 SynNet_2 与 SynNet_3 上,这种时间效率优势就变得显著.由此不难得出这样的结论:与其他两个基于多目标优化的社区发现算法(MOGA-Net 与 SCAH-MOHSA)相比较,随着网络规模的增加,3 种算法的耗时都会增加,但 MOCD-PSO 耗时增长速度要远低于其他两种算法,即 MOCD-PSO 具有更好的可扩展性.

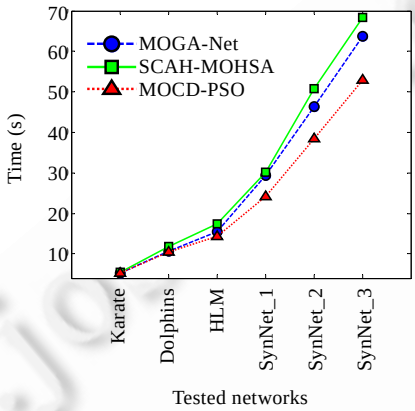


Fig.5 Time performance analysis of MOCD-PSO

图 5 MOCD-PSO 的时间性能分析

6 结束语

复杂网络社区发现研究对互联网文化安全与信息个性化服务等多个方面都有着重要的意义,当前,绝大多数基于生物启发优化的复杂网络社区发现算法都是以某种单一的社区结构质量测度目标来驱动的,而社区质量评判指标的多样性使得网络社区结构分析的实践人员在面对众多的评判指标难以决策,并且现实的复杂网络隐含的真实社区结构不是通过单一测度指标能完全反映出来的.本文通过大量的实验发现了社区质量评判指标的数据依赖性与耦合关联性,分析了这两种性质导致基于单一评判指标优化的网络社区发现算法所具有的局限性.因此,本文进一步将复杂网络社区发现问题形式化为多目标优化问题,同时提出了基于多目标粒子群优化的网络社区发现算法 MOCD-PSO,在此算法中,我们提出了新的 Pareto 最优网络社区结构集更新策略并予以巧妙地实现,选取模块度指标 Q、最小最大割指标 MinMax*Cut* 与轮廓(silhouette)指标来构建 MOCD-PSO 的优化.实验结果表明,MOCD-PSO 算法具有较好的收敛性,能发现分布均匀且分散度较高的 Pareto 最优网络社区结构集,并且无论与单目标优化方法(GN 与 GA-Net)相比较,还是与多目标优化算法(MOGA-Net 与 SCAH-MOHSA)相比较,MOCD-PSO 算法所发现的 Pareto 最优网络社区结构与网络内在社区结构更相吻合.

References:

[1] Fortunato S. Community detection in graphs. Physics Reports, 2010,486(3-5):75−174. [doi: 10.1016/j.physrep.2009.11.002]

[2] Coscia M, Giannotti F, Pedreschi D. A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining, 2011,4(5):512−546. [doi: 10.1002/sam.10133]

[3] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physics Review E, 2004,70(6):066111. [doi: 10.1103/PhysRevE.70.066111]

[4] Danon L, Díaz-Guilera A, Arenas A. The effect of size heterogeneity on community identification in complex networks. Journal of Statistical Mechanics, 2006,11:P11010. [doi: 10.1088/1742-5468/2006/11/P11010]

[5] Wakita K, Tsurumi T. Finding community structure in mega-scale social networks. In: Akerkar R, Maret P, Vercouter L, eds. Proc. of the WWW 2007. New York: ACM Press, 2007. 1275−1276. [doi: 10.1145/1242572.1242805]

[6] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008,10:P10008. [doi: 10.1088/1742-5468/2008/10/P10008]

[7] Girvan M, Newman MEJ. Modularity and community structure in networks. Proc. of the National Academy of the Sciences of the United States of America, 2006,103(23):8577−8582. [doi: 10.1073/pnas.0601602103]

[8] Zhang P, Wang J, Li X, Li M, Di Z, Fan Y. The clustering coefficient and community structure of bipartite networks. Physica A: Statistical Mechanics and its Applications, 2008,387(27):6869−6875. [doi: 10.1016/j.physa.2008.09.006]

[9] Lin YF, Wang TY, Tang R, Zhou YW, Huang HK. An effective model and algorithm for community detection in social networks. Journal of Computer Research and Development, 2012,49(2):337−345 (in Chinese with English abstract).

[10] Gan WY, He N, Li DY, Wang JM. Community discovery method in networks based on topological potential. Ruan Jian Xue Bao/Journal of Software, 2009,20(8):2241−2254 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3318.htm [doi: 10.3724/SP.J.1001.2009.03318]

[11] Tasgin M, Herdagdelen A, Bingol H. Community detection in complex networks using genetic algorithms. arXiv:0711.0491.

[12] Pizzuti C. GA-NET: A genetic algorithm for community detection in social networks. In: Rudolph G, Jansen T, Lucas SM, Poloni C, BeumeProc N, eds. Proc. of the PPSN 2008. Heidelberg: Springer-Verlag, 2008. 1081−1090. [doi: 10.1007/978-3-540-87700-4_107]

[13] Lipczak M, Milios E. Agglomerative genetic algorithm for clustering in social networks. In: Raidl GR, Alba E, Bacardit J, Beyer HG, Birattari M, Blum C, *et al*., eds. Proc. of the GECCO 2009. New York: ACM Press, 2009. 1243−1250. [doi: 10.1145/1569901.1570068]

[14] Duan XD, Wang CR, Liu XD, Lin YP. Web community detection model using particle swarm optimization. In: Michalewicz Z, Reynolds RG, eds. Proc. of the CEC 2008. Los Alamitos: IEEE Press, 2008. 1074−1079. [doi: 10.1109/CEC.2008.4630930]

[15] Huang FL, Xiao NF. Particle-Swarm-Optimization algorithm to discover network community. Control Theory & Applications, 2011, 28(9):1135−1140 (in Chinese with English abstract).

[16] Gog A, Dumitrescu D, Hirsbrunner B. Community detection in complex networks using collaborative evolutionary algorithms. In: Costa FA, Rocha L, *et al*., eds. Proc. of the ECAL 2007. Heidelberg: Springer-Verlag, 2007. 886−894. [doi: 10.1007/978-3-540-74913-4_89]

[17] He DX, Zhou X, Wang Z, Zhou CG, Wang Z, Jin D. Community mining in complex networks—Clustering combination based genetic algorithm. ACTA AUTOMATICA SINICA, 2010,36(8):1160−1170 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2010.01160]

[18] Jin D, Liu J, Yang B, He DX, Liu DY. Genetic algorithm with local search for community detection in large-scale complex networks. ACTA AUTOMATICA SINICA, 2011,37(7):873−882 (in Chinese with English abstract).

[19] Zhu XL, Wang B. Community mining in complex network based on parallel genetic algorithm. In: Lin YC, Shieh CS, Liao BY, *et al*., eds. Proc. of the ICGEC 2010. Shenzhen: IEEE Press, 2010. 325−328. [doi: 10.1109/ICGEC.2010.87]

[20] Gong MG, Ma LJ, Zhang QF, Jiao LC. Community detection in networks by using multiobjective evolutionary algorithm with decomposition. Physica A, 2012,391:4050−4060. [doi: 10.1016/j.physa.2012.03.021]

[21] Gong MG, Hou T, Fu B, Jiao LC. A non-dominated neighbor immune algorithm for community detection in networks. In: Krasnogor N, Lanzi PL, eds. Proc. of the GECCO 2011. New York: ACM Press, 2011. 1627−1634. [doi: 10.1145/2001576.2001796]

[22] Pizzuti C. A multiobjective genetic algorithm to find communities in complex networks. IEEE Trans. on Evolutionary Computation, 2012,16(3):418−430. [doi: 10.1109/TEVC.2011.2161090]

[23] Amiri B, Hossain L, Crawford JW. An efficient multiobjective evolutionary algorithm for community detection in social networks. In: Smith A, ed. Proc. of the CEC 2011. Piscataway: IEEE Press, 2011. 2193−2199. [doi: 10.1109/CEC.2011.5949886]

[24] Li YY, Chen J, Liu RC, Wu JS. A spectral clustering-based adaptive hybrid multi-objective harmony search algorithm for community detection. In: Abbass A, Essam D, Sarker R, eds. Proc. of the CEC 2012. Brisbane: IEEE Press, 2012. 1−8. [doi: 10.1109/CEC.2012.6253013]

[25] Zhou AM, Qu BY, Li H, Zhou SZ, Suganthan PN, Zhang QF. Multiobjective evolutionary algorithms: A survey of the state of the art. Swarm and Evolutionary Computation, 2011,1(1):32−49. [doi: 10.1016/j.swevo.2011.03.001]

[26] Leskovec J, Lang KJ, Mahoney MW. Empirical comparison of algorithms for network community detection. In: Rappa M, Jones P, Freire J, Chakrabarti S, eds. Proc. of the WWW 2010. New York: ACM Press, 2010. 631−640. [doi: 10.1145/1772690.1772755]

[27] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113. [doi: 10.1103/PhysRevE.69.026113]

[28] Fortunato S, Barthélemy M. Resolution limit in community detection. Proc. of the National Academy of the Sciences of the United States of America, 2007,104:36−41. [doi: 10.1073/pnas.0605965104]

[29] Li ZP, Zhang SH, Wang RS, Zhang XS, Chen LN. Quantitative function for community detection. Physical Review E, 2008,77(3):036109. [doi: 10.1103/PhysRevE.77.036109]

[30] Demir GN, Uyar AS, Oguducu SG. Graph-Based sequence clustering through multiobjective evolutionary algorithms for Web recommender systems. In: Lipson H, ed. Proc. of the GECCO 2007. New York: ACM Press, 2007. 1943−1950. [doi: 10.1145/1276958.1277346]

[31] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987,20(1):53−65. [doi: 10.1016/0377-0427(87)90125-7]

[32] Kannan R, Vempala S, Vetta A. On clusterings: Good, bad and spectral. Journal of the ACM, 2004,51(3):497−515. [doi: 10.1145/990308.990313]

[33] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. Proc. of the National Academy of the Sciences of the United States of America, 2004,101(9):2658−2663. [doi: 10.1073/pnas.0400054101]

[34] Shi JB, Malik J. Normalized cuts and image segmentation. IEEE TPAMI, 2000,22(8):888−905. [doi: 10.1109/34.868688]

[35] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4):046110. [doi: 10.1103/PhysRevE.78.046110]

[36] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. In: Simoff SJ, Zaane OR, eds. Proc. of the KDD Workshop on Text Mining 2000. New York: ACM Press, 2000. 525−526.

[37] Kung HT, Luccio F, Preparata FP. On finding the maxima of a set of vectors. Journal of the ACM, 1975,22(4):469−476. [doi: 10.1145/321906.321910]

[38] Hruschka ER, Campello RJGB, Freitas AA, de Carvalho ACPLFD. A survey of evolutionary algorithms for clustering. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2009,39(2):133−155. [doi: 10.1109/TSMCC.2008.2007252]

[39] Zitzler E, Deb K, Thiele L. Comparison of multiobjective evolutionary algorithms: Empirical results. Evolutionary Computation, 2000,8(2):173−195. [doi: 10.1162/106365600568202]

[40] Reyes-Sierra M, Coello CAC. Multi-Objective particle swarm optimizers: A survey of the state-of-the-art. Int'l Journal of Computational Intelligence Research, 2006,2(3):287−308.

[41] Nicosia V, Mangioni G, Carchiolo V, Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities. Journal of Statistical Mechanics: Theory and Experiment, 2008,2009(3):1−22. [doi: 10.1088/1742-5468/2009/03/P03024]

附中文参考文献:

[9] 林友芳,王天宇,唐锐,周元炜,黄厚宽.一种有效的社会网络社区发现模型和算法.计算机研究与发展,2012,49(2):337−345.

[10] 淦文燕,赫南,李德毅,王建民.一种基于拓扑势的网络社区发现方法.软件学报,2009,20(8):2241−2254. http://www.jos.org.cn/1000-9825/3318.htm

[15] 黄发良,肖南峰.网络社区发现的粒子群优化算法.控制理论与应用,2011,28(9):1135−1140.

[17] 何东晓,周栩,王佐,周春光,王喆,金弟.复杂网络社区挖掘-基于聚类融合的遗传算法.自动化学报,2010,36(8):1160−1170. [doi: 10.3724/SP.J.1004.2010.01160]

[18] 金弟,刘杰,杨博,何东晓,刘大有.局部搜索与遗传算法结合的大规模复杂网络社区探测.自动化学报,2011,37(7):873−882.

黄发良(1975－),男,湖南永州人,博士,讲师,主要研究领域为数据挖掘,智能计算.
E-mail: faliang.huang@gmail.com

朱晓峰(1973－),男,博士生,主要研究领域为数据挖掘,图像检索.
E-mail: seanzhuxf@gmail.com

张师超(1962－),男,博士,教授,博士生导师,主要研究领域为数据挖掘,人工智能.
E-mail: zhangsc@mailbox.gxnu.edu.cn