软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(7):1773−1789 [doi: 10.13328/j.cnki.jos.005114]

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

基于进化聚类的动态网络社团发现*

牛新征[1], 司伟钰[2], 佘 堃[3]

[1](电子科技大学 计算机科学与技术学院,四川 成都 611731)

[2](上海交通大学 电子信息与电气工程学院,上海 200240)

[3](电子科技大学 信息与软件工程学院,四川 成都 611731)

通讯作者: 牛新征, E-mail: xinzhengniu@uestc.edu.cn

摘 要: 社团的数目和时间平滑性的平衡因子一直是基于进化聚类的动态网络社团发现算法的最大的问题.提出一种基于标签的多目标优化的动态网络社团发现算法(LDMGA).借鉴多目标遗传算法思想,将进化聚类思想转换为多目标遗传算法优化问题,在保证当前时刻的聚类质量的同时,又能使当前聚类结果与前一时刻网络结构保持一致.该算法在初始化过程中加入标签传播算法,提高了初始个体的聚类质量.提出基于标签的变异算法,增强了算法的聚类效果和算法的收敛速度.同时,多目标遗传算法和标签算法的结合使算法可扩展性更强,运行时间随着节点或者边数目的增加呈线性增长.将该算法与目前的优秀算法在仿真数据集和真实数据集上进行对比实验,结果表明,该算法既有良好的聚类效果,又有良好的扩展性.

关键词: 进化聚类;标签传播;动态网络;社团发现

中图法分类号: TP181

中文引用格式: 牛新征,司伟钰,佘堃.基于进化聚类的动态网络社团发现.软件学报,2017,28(7):1773−1789. http://www.jos.org.cn/1000-9825/5114.htm

英文引用格式: Niu XZ, Si WY, She K. Evolutionary community detection in dynamic networks. Ruan Jian Xue Bao/Journal of Software, 2017,28(7):1773−1789(in Chinese). http://www.jos.org.cn/1000-9825/5114.htm

Evolutionary Community Detection in Dynamic Networks

NIU Xin-Zheng[1], SI Wei-Yu[2], SHE Kun[3]

[1](School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

[2](School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

[3](School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: The number of communities and temporal smoothness are always the main limitations in the field of evolutionary community detection for dynamic networks. In this paper, a new multi-objective approach based on the label propagation algorithm (LDMGA) is proposed. Employing the idea of multi-objective genetic algorithm, the evolutionary clustering algorithm is transformed into a multi-objective optimization problem, which not only improves the clustering quality, but also minimizes clustering drift from one time step to the successive one. Population initialization based on the label propagation algorithm improves the clustering quality of initial individuals. In addition, applying the label propagation algorithm to the mutation progress enhances the quality of clustering and the convergence rate. At the same time, the combination of the multi-objective genetic algorithm and the label propagation algorithm makes

* 基金项目: 国家自然科学基金(61300192); 国家科技支撑计划(2013BAH33F02); 中央高校基本科研业务费(ZYGX2014J052); 四川省科技支撑计划(2015GZ0096)

Foundation item: National Natural Science Foundation of China (61300192); National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2013BAH33F02); Fundamental Research Funds for the Central Universities (ZYGX2014J052); Science and Technology Support Program of Sichuan, China (2015GZ0096)

收稿时间: 2015-08-24; 修改时间: 2016-03-18; 采用时间: 2016-06-10; jos 在线出版时间: 2016-10-11

CNKI 网络优先出版: 2016-10-12 16:26:45, http://www.cnki.net/kcms/detail/11.2560.TP.20161012.1626.012.html

the algorithm more scalable, and the running time increases linearly with the increase of the number of nodes or edges. The experiment on the synthesized datasets and real datasets shows that the proposed algorithm consistently provides good clustering quality and scalability.

Key words: evolutionary clustering; label propagation; dynamic network; community discovery

近年来,相比于静态网络,社会网络的动态特性受到了广泛关注,大量动态网络被提了出来,比如信息交互网络及科学家合作网络[1]、社交网络等等.所有的复杂系统都具有某种动态特性,所以,将这些复杂网络模拟成动态网络是一种合理、有效的方式.目前,动态网络研究方法被广泛应用于科技网络、生物网络和社交网络中,用来发现和描述不同事物和个体之间的相互联系,比如生物分子之间的关系、合作关系等等,以及由多个体形成的社团结构,如朋友圈、生物圈,图 1 展示了具有 3 个社团结构的网络,节点连线的数值表示边的权重.对于大多数网络来说,它们的拓扑结构会随着时间发生明显的变化,如每个人的朋友圈会随着时间的变化而发生结构上的变化.

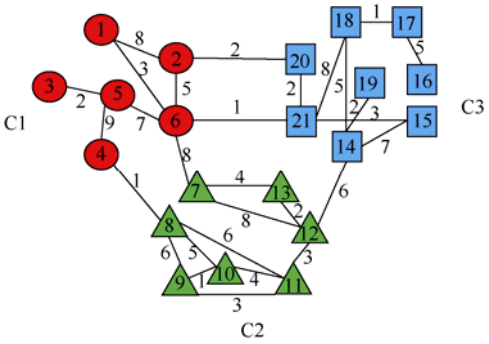


Fig.1 A network with three communities

图 1 具有 3 个社团结构的网络

动态网络中的社团结构发现日益成为一个重要的研究课题,并且,其在真实社会中有相当广泛的应用[2,3],比如信息影响力分析、客户推荐等等.动态网络的动态特性意味着,随着时间的变化,社团的结构也会发生相应的变化.在社团结构中,社团内部节点之间的联系比较紧密,在社团之间的节点联系比较稀疏.在文献[4,5]中,动态网络上的社团结构挖掘问题表述为,从一个离散的时间轴上观察网络中某些个体间的联系,即观察连续时间点上的子图.随着时间的推移,社团及其结构变化可以理解为这些子图上多数联系所形成的结构.

增量聚类[6]和进化聚类[7]是目前两种主流的研究动态网络社团发现的方法.增量聚类算法的基本思想为仅对第 1 个时间点的网络进行聚类分析,对在后面时间点的网络,根据网络缓慢变化特性,把前一时间点的聚类结果作为基础,同时利用当前时间点网络各方面特征对聚类结果进行局部调整,最后得到的聚类结果具有光滑性.代表算法有 IA-MCS[8],GraphScope[9]等.增量聚类的方法在某种程度上牺牲了聚类质量以获得较小的时间复杂度.而进化聚类的方法同时考虑了聚类质量和时间平滑性对于结果的影响,在保证聚类质量的同时,使聚类结果更加接近真实社团结构.

Lin 等人[10]提出 FacetNet 框架,该算法是目前最经典的进化聚类的算法.该框架采用随机块模型生成社团,并根据狄利克雷分布的概率模型分析社团的演化.他们利用 KL-divergence 算法定义快照质量和历史开销.将社团发现和社团演化融为一体,t 时刻的数据和历史社团结构同时影响 t 时刻的社团结构,所以,该算法得到的社团结构在抗噪性和合理性方面比较优越.在每一次的迭代中,将更新近似结构的值以降低历史开销,最终收敛到一个局部最优解.

Kim 等人[11]提出了基于微粒与密度的进化方法.该算法在初始时,将动态网络构建为一系列的粒子群(nano-communities),社团定义为粒子群中连接紧密的一个子集.然后,Kim 等人采用基于密度的聚类方法和开销嵌入技术(cost embedding technique)实现社团结构的时间平滑性.同时,该算法不依赖具体的聚类方法和节点间相似性的定义,进而提高了算法的效率.他们定义了相邻时刻节点之间的相似性连接,由这些连接将动态网络转换为一个完整的多部图,即 t 个时间点的动态网络对应一个 t 部图.这种算法解决了之前多数算法的两个问题:(1) 每个时刻社团数目不变;(2) 以迭代方式达到时间平滑而使效率降低.

但是,在以上的方法中,存在两个普遍的问题:(1) 社团数目的确定;(2) 用来实现时间平滑性的平衡因子的确定.FacetNet[10]算法只能用于发现固定社团数目的动态网络中,即在整个网络的时间序列上,社团数目不会发生改变.虽然 FacetNet 扩展版本弥补了之前一些问题,例如,可以处理增加节点和删除节点的情况,也可以处理相邻时刻社团数目改变的情况,但该算法本身也存在缺点.例如,需要很多次迭代才能使矩阵达到收敛.因此,FacetNet 扩展版不适合大规模的数据处理.而 Kim 等人的算法是基于密度的聚类算法,所以需要提前设置密

度参数.

另一个问题是,这些算法需要提前设置平衡因子来实现时间平滑性.Folino 等人[12]针对动态网络社团发现问题提出了一种基于多目标优化的进化聚类算法(DYNMOGA).其聚类的框架是多目标遗传算法,有效地平衡了时间开销和历史开销,并显著提高了聚类的质量,而且无需平衡因子,能够自动发现社团数目.通过大量实验,其结果表明,该算法在基于进化聚类的动态网络社团发现中,时间复杂度和聚类的精度方面明显优于经典的FacetNet 算法.但是,由于 DYNMOGA 算法采用遗传算法和基于图的编码方式,使得时间复杂度较高,不能很好地处理规模较大的网络.同时,在由于考虑网络的动态特性而采用的多目标优化方法中,个体产生的随机性较大,在增大种群数量和迭代次数以获取聚类质量的同时,牺牲了时间.DYNMOGA 算法采用基于邻接位置的编码方式,即每个节点用其邻居节点编号表示基因值,节点和基因值表示两个节点之间有边,同时说明两个节点属于同一个社团.解码过程就是划分出相应的社团以及确定社团的数目.基于邻接位置的编码在解码过程中需要耗费 $O(n\log n)$的时间.但是,从整个 DYNMOGA 算法的时间复杂度为 $O((gp\log p)\times(n\log n+m))$可以看出,解码时间在整个算法中占了很大一部分,影响了算法的效率.

因此,本文针对 DYNMOGA 算法中随机性较强和时间复杂度较高的问题,引入了标签传播的思想,结合网络的动态特性,即同时考虑时间开销和历史开销,并将其有效地应用到动态网络社团发现中,提出了基于标签传播的多目标优化的进化聚类算法 LDMGA.

本文的主要贡献包含以下 3 个方面.

(1) 初始化个体时,引入基于节点度的标签传播算法,使初始社团有一定的精度,提高了聚类的质量.

(2) 提出了基于标签传播的变异算法,在进一步提高聚类质量的同时,加快了算法的收敛速度.

(3) 在多目标遗传算法中结合标签传播算法,增强了算法可扩展性,算法运行时间随着节点或者边数目呈线性增长.

本文第 1 节给出多目标进化聚类模型的介绍.第 2 节给出基于标签传播算法的多目标优化的聚类算法 LDMGA 的介绍.第 3 节主要对 LDMGA 算法整个流程加以介绍并给出时间复杂度的相关分析.第 4 节对本文算法进行实验测试与分析.第 5 节总结全文.

1 基于多目标优化的进化聚类模型

Chakrabarti 等人[7]提出了进化聚类的思想.该思想框架认为,短时间内网络的变化可能是由于噪声引起的,所以,在时间序列上,社团的变化具有时间平滑性.在对每个时刻的网络进行社团发现时,需要对两个相互冲突的准则进行考察:第一,使当前社团聚类结果尽量准确地反映当前时刻的网络结构.第二,与上一时刻的聚类结果相比,当前时刻聚类结果不能变化剧烈.因此,Chakrabarti 等人提出了用来考察两个相互冲突的准则的概念:快照质量(snapshot quality,简称 ST)和历史开销(temporal cost,简称 TC).快照质量用来衡量当前聚类结果 C_t 在当前网络结构 G_t 下的聚类质量,而历史开销用来衡量当前时刻聚类结果 C_t 与前一时刻聚类结果 C_{t-1} 的相似性.所以,同时满足快照质量最大和历史开销最小的聚类结果被认为是当前时刻最优的聚类结果.引入平衡因子来衡量这两个准则的影响程度,聚类质量可以用公式(1)来描述.

$$cost=\alpha\cdot SC+(1-\alpha)\cdot TC \tag{1}$$

在这个公式中,α是平衡因子,其值由用户自定义.当 α =1 的时候,结果只考虑聚类质量.当 α =0 时,结果为最接近前一时刻的聚类结果.当α值介于 0 和 1 之间时,可以控制两个准则,以便达到最佳平衡点,找到最优聚类结果.

1.1 动态网络和社团

本文用 $G_t=(V_t,E_t)$表示在时刻 t 的网络,V_t 是网络 G_t 中节点的集合,E_t 则是网络 G_t 中边的集合.如果 G_t 是一个带权重的网络,那么节点之间的边有不同的权重值.一个有 T 个时间点的网络可以被描述成一个网络序列,$G=\{G_1,G_2,\ldots,G_T\}$.

社团是动态复杂网络的一个普遍的特性.在同一个社团内的节点间边的连接密度高,社团之间的节点间边的连接密度低.如果在时间点 t 的网络中有 tk 个社团,那么网络 G_t 可以被描述成 $C_t=\{C_{t1},C_{t2},\ldots,C_{tk}\}$.$C_{tp}$ 表示第 p 个社团,并且 $C_{tp}\cap C_{tq}=\varnothing, p,q\in\{1,2,...,k\}$. 网络序列 $G=\{G_1,G_2,\ldots,G_T\}$发现的社团结果表示为 $C=\{C_1,C_2,\ldots,C_T\}$.

1.2 多目标优化问题定义

定义 1. 对于一个静态网络 G_t,多目标优化问题可以定义为

$$\left.\begin{aligned}\min F(C_t)=(f_1(C_t),f_2(C_t),...,f_h(C_t))\\ C_t\in Q, Q=\{C_{t1},C_{t2},...,C_{tm}\}\end{aligned}\right\} \tag{2}$$

多目标优化问题的解是使目标函数 $F(C_t)\in R^h$(h 为目标函数的个数)中的各个分函数 $f_i(i=1,2,...,h)$取得最小值的 C_t.每个分函数就是一个独立的目标函数,用来衡量聚类的效果.以上是极小化问题的定义,对于极大化问题的定义与上述定义相似.由于 F 是一个目标函数的矢量,而且所有分函数同时进行优化,所以,在多目标优化问题中,解是一个最优解的集合,而不是唯一的.

1.3 多目标优化最优解

多目标优化问题解的一大特征是至少存在一个目标优于其他所有的解,具有这样特征的解就称为 Pareto 最优解、非劣最优解集或非支配最优解集.多目标优化算法的目标就是构造非支配解集,不断地寻找最优的非支配解集,直到找到最优解集.

定义 2. Pareto 最优解或非支配最优解:若认为 $C^*\in\Omega$是最优解(即 Pareto optimal solution),则对$\forall C\in\Omega$,满足下列条件:

$$\cap_{i\in I}(f_i(C)\geq f_i(C^*)) \tag{3}$$

$I=\{1,2,\ldots,h\}$,h 为目标函数个数或者且至少存在一个 $j\in I$,使

$$f_j(C)>f_j(C^*) \tag{4}$$

定义 3. 支配关系.

对于两个解 $C_1\in\Omega$和 $C_2\in\Omega$,支配关系定义如下.

(1) 若$\forall k\in\{1,2,...,h\}$,$f_k(C_1)\leq f_k(C_2)$;(2) $\exists l\in\{1,2,...,h\}$,使 $f_l(C_1)<f_l(C_2)$,则称 C_1 支配 C_2,表示为 $C_1<C_2$.

非支配集(non-dominated set)即是由满足上述条件的 C_1 构成的集合,其中,所有满足上述条件的 C_1 构成的集合即为最大非支配集.非支配集在通常情况下被认为是最大非支配集.

定义 4. 最优边界.

最优解总是在目标函数搜索区的边界线或者面上(Pareto front),这样的边界线或者面称为最优边界.对所有目标函数而言,Pareto 最优解集中的解相互之间是不可以比较的.换句话说,当同时考虑所有目标时,这些解是目标函数搜索空间中最优的解,没有更优的解.

2 基于标签传播算法的多目标优化的聚类算法 LDMGA

2.1 基于多目标优化的聚类算法框架

LDMGA 算法的基本框架是遗传算法.遗传算法[13]是一类可适应的搜索方法.在遗传算法中,多目标优化问题的解集被定义为种群中的最优个体,每个个体代表一种可能解.种群中个体的数量表示种群的规模.每一个个体是多个基因的集合,可以理解为某种基因的排列组合,不同的组合方式决定了个体的社团结构.因此,算法在初始时需要将个体实现从表现型到基因型的映像,即编码工作.个体在连续的后代中得到不断进化.子代个体一般这样产生:将两个父代个体进行交叉操作,继承父代中基因结构,然后,通过变异操作使父代个体的基因产生突变,以生成更好的结构.在每一代中,都要进行个体的适应值的计算,即目标函数的计算,有较高适应度的个体将被选择进入下一代的迭代.经过数代之后,新生代中个体会趋于满足某种给定的条件,即有好的社团结构,最终的个体被认为是对所有目标函数优化的最优解或接近最优解.LDMGA 算法基本框架如图 2 所示.

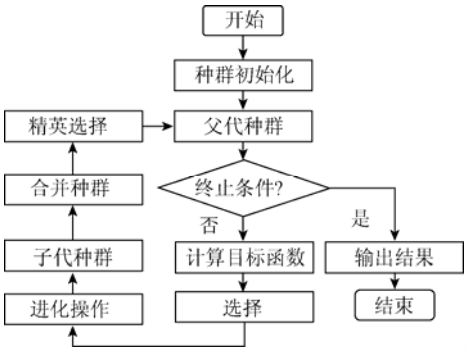
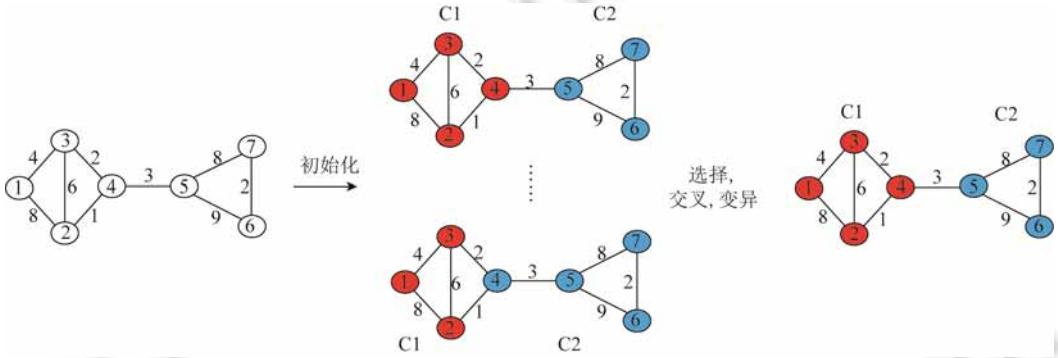


Fig.2 Workflow of the LDMGA

图 2 LDMGA 算法基本框架

LDMGA 算法过程示意图如图 3 所示,节点连线的数值表示边的权重.图 3(a)表示初始输入网络.在算法初始化过程中,将网络划分出不同的社团结构,生成具有不同社团结构的个体,如图 3(b)所示.然后,初始种群经过选择、交叉、变异的过程,最终选出最优的社团结构,如图 3(c)所示.



(a) (b) (c)

Fig.3 Schematic diagram of LDMGA process

图 3 LDMGA 算法过程示意图

2.2 个体的编码与解码

目前在社团结构发现算法中,主要有字符串编码方式[14,15]和基于图的编码方式[16−18].与图的编码方式相比,字符串编码方式在表示社团结构方面更加直观和高效,所以本文采用字符串编码方式.一个网络任意的一种划分被称为一个个体,包含 n 个基因 $ge_1,ge_2,\ldots,ge_n$,n 是节点的数量.每个基因都对应一个值 j.这些基因构成了网络,并且每个值 j 对应第 i 个基因 ge_i,并且 j 表示第 i 个基因 ge_i 所属社团标签,意味着有相同标签的基因属于同一个社团.图 4 展示了一个网络划分和相应的编码.

在字符串表示法中,网络节点所属社团仅仅是一个标识符,也就是说,不同个体中拥有相同标签的节点未必属于同一个社团.如图 4 所示的网络可能的两个个体(1,1,1,2,2,2,2)和(2,2,2,2,3,3,3),第 1 个个体中存在两个社团{1,2,3}和{4,5,6,7},

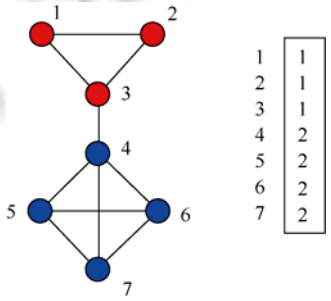


Fig.4 A network of 7 nodes partitioned in two communities {1,2,3} and {4,5,6,7} and the corresponding representation

图 4 7 个节点网络被划分成两个社团{1,2,3}和{4,5,6,7}以及相应编码表示

第 2 个个体存在两个社团{1,2,3,4}和{5,6,7}.如果这两个个体之间进行交叉操作,那么交叉结果可能是(2,2,2,2,2,2,2),则整个社团结构被破坏掉.所以,本文给出的解码过程为针对任意个体,其初始标签为 $L(i)(i=1,2,\ldots,n)$.设 $L(1)=1$,如果 $L(2)=L(1)$,则 $L(2)=1$.如果 $L(2)\neq L(1)$,则 $L(2)=2$,以此类推.如果 $S(k)\neq S(j)(j=1,2,\ldots,k-1)$,并且如果此时 $p-1$ 是当前最大社团标签值,那么 $S(k)=p$.在整个解码过程中,个体(1,1,1,2,2,2,2)和(2,2,2,2,3, 3,3)解码为(1,1,1,2,2,2,2)和(1,1,1,1,2,2,2),然后再进行交叉操作,便于保留社团结构.在解码后,节点的标签表示节点所属社团编号.

2.3 基于标签传播的初始化算法

在生成初始种群时,增强初始种群的社团结构和初始种群的多样性可以提高算法的效率.为了达到这个目标,本文借鉴标签传播算法的思想生成初始种群.

基于图的半监督学习的标签传播算法(LPA)是由 Zhu 等人[19]提出来的,LPA[19]算法采用的基本方法就是用已经标记的节点标签信息去预测其他未标记节点的标签信息.LPA 通过节点之间标签的传递进行分类,它并不受限于数据的分布形状,只要是空间分布上同一类的数据,标签传播算法都能将它们分到同一类中,算法简单,时间复杂度低,聚类的效果好,可扩展性好.Raghava 等人[20]首次提出将 LPA 应用于社团发现,该算法简称为 RAK 算法.在 RAK 算法中,首先,每个节点被赋予一个唯一的标签,并且每个节点有若干个邻居;在每次迭代中,每个节点根据其邻居节点标签的情况不断更新为其多数邻居的标签,当节点的标签不再发生变化时,算法结束.最后,根据每个节点的标签划分出相应的社团.RAK 算法的主要步骤如下.

(1) 对于网络 $G=(V,E)$,$\forall x\in V$,算法初始赋予任意节点 x 一个唯一的标签值 Lx,Lx 表示节点 x 所在的社团编号.

(2) 根据节点 x 邻居集 $N(x)$的标签情况,节点 x 不断迭代更新自己的标签值 Lx 为多数邻居的标签值.在迭代更新过程中,如果有不止一个可选的标签,则随机选择其中一个邻居的标签更新为节点 x 的新标签值.在经过 k 次迭代之后,每个节点的标签变化趋于稳定.

(3) $\forall x,y\in V$,如果存在两个节点标签值相同,即 $Lx=Ly$,那么就认为节点 x 和节点 y 属于同一个社团,生成社团划分.

RAK 算法的时间复杂度为 $O(km)$,k 表示算法的迭代次数,m 表示网络的边数.由于 RAK 算法的时间复杂度几乎为线性,所以在处理各种大规模数据时拥有很好的效率.但是,从以上描述的步骤中可以看出,RAK 算法存在一些随机因素,比如在节点标签更新时,当有不止一个可选标签时,随机选择一个标签来更新.在这些可选标签中,不同的标签会导致每次迭代产生的社团结构有一定的差异.所以,RAK 算法的这些随机因素会使算法结构不够稳定.算法在处理大规模数据时,这样的情况会更加明显和频繁.

由于标签传播算法的良好的聚类质量和接近线性的时间,本文借鉴 RAK 标签传播算法的思想.在生成初始个体时,标签传播算法能够产生有一定社团结构的个体.同时,节点的度数越大,那么这个节点对周围社团造成的影响也就越大,因为它可以作为更多节点的邻居节点来影响这些节点的标签传播过程.所以,为了使初始个体有好的社团结构,我们选择从度大的节点进行标签更新,下面的实验结果表明从度大的节点进行标签更新会提高算法聚类的质量.

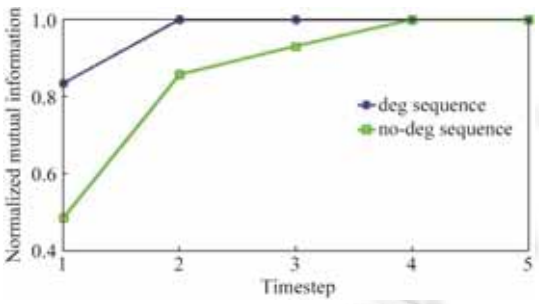


Fig.5 NMI values of the results of community detection of deg sequence and no-deg sequence

图 5 按照度大小顺序和序号顺序进行社团发现的 NMI 值

本文采用 SYN-VAR(z=5),SYN-VAR 网络中有 256 个节点,被分成 4 个社团,每个社团有 64 个节点.通过从每个社团中随机选择 8 个节点并且生成一个新的 32 个节点社团,该实验生成 5 个连续网络.社团中平均每个节点的度数被设置为社团大小的一半.此外,在每个时间点,随机删除 16 个节点,同时增加 16 个新节点.数据集由 DYNMOGA 算法[12]作者提供.NMI 值用来评价算法运行结果和真实社团结构的相似性(详见第 3.1 节).图 5 是按照度大小顺序(deg sequence)和序号顺序(no-deg sequence)进行社团发现的结果 NMI 的值.从图

5 可以看出,从度大的节点进行标签更新会提高算法的聚类质量.

在初始化过程中,为网络图中每一个节点分配唯一的标签,标签代表了节点所属的社团编号.为了增强标签传播算法的稳定性,标签更新的顺序是从度大的节点开始,采用异步更新,并且只进行一次标签更新过程.基于标签传播的初始化算法如算法 1 所示.

算法 1. 基于标签传播初始化算法.

输入:种群数量 p,图 G_t 的邻居集 N_t 和度 D_t;

输出:初始解 $g=\{g_t^1,g_t^2,...,g_t^p\}$.

1. **For** i=1 **to** p
2. $g^i=[g_1^i,g_2^i,\ldots,g_n^i], i\in\{1,2,...,p\}, g_j^i=j, j\in\{1,2,...,n\}, n$ 为节点数
3. 任意个体$\forall g^i\in g, i\in\{1,2,...,p\}$随机生成序列 $X=[x_1,x_2,...,x_n], x_i\in N_t(i)$
4. Sort(D_t),从度大的节点开始更新,同时采用异步更新策略,更新一次

   异步更新: $g^i(t)=f(g_1^i(t),...,g_h^i(t),g_{h+1}^i(t-1),...,g_l^i(t-1)), t\in\{1,2,...,k\}, k$ 为迭代次数
5. **End For**
6. **Return** p 个初始种群,定义时间点 t 的初始种群为 $g_t=\{g_t^1,g_t^2,...,g_t^p\}$

这里定义的节点的度,指的是一个节点连接的边的数目.如果是无权图,则节点的度就是节点连接的边的数目;如果是有权图,这个度就是节点连接的边的权重之和.在更新节点的过程中,有同步更新方式和异步更新方式.文献[20]给出的实验结果表明,异步更新方式比同步更新方式的结果更加稳定,但是比同步更新方式要更新的次数更多.本文采用异步更新策略,这一策略是指节点 x 在 t 次更新过程中的标签依据于第 t 次更新中已经更新过的标签的节点和第 t 次中没有更新标签的节点.

2.4 单路交叉策略

由于本算法的解码过程一定程度上解决了不同个体间标签不相容的状况,并且,为了进一步保留良好的社团结构,本文采用 2007 年 Tasgin 等人[21]提出的单路交叉策略.交叉策略见表 1,S 表示节点的编号.表 1 给出一个有 6 个节点的网络,假设随机选中节点 2,在个体 A 中节点 2 所在的社团节点{1,2,6}标签将传播到 B 中节点{1,2,6},即更新个体 B 中节点{1,2,6}标签为 A 中节点 2 的标签 1,这样就得到交叉后的新个体 C.这种交叉操作可以将 A 中的社团结构信息传播给 B.在本文算法中,每次交叉操作执行两次,一次 A 传播给 B,一次 B 传播给 A.

Table 1 One-Way crossover

表 1 单路交叉策略

S(节点)		A(源)		B(目标)		C(新)
1		1	→	1	→	1
2	→	1	→	2	→	1
3		2		2		2
4		3		1		1
5		2		2		2
6		1	→	1	→	1

2.5 基于标签传播算法的变异策略

对变异算法的改进是本文的一个创新点,本文没有采取以往经典的变异算法,而是将变异过程改为标签传播的一个中间过程,进一步提高了个体的聚类质量,加快了算法的收敛速度.在变异操作中,只进行一次更新的标签传播.与基于标签传播的初始化不同的有两点,第 1 点是更新的个体是交叉算法后的个体,第 2 点是在传播方向上,从节点编号为 1 的节点开始标签更新,增强了随机性.这样,变异过程便实现了一个遗传算法和标签传播算法的嵌套.基于标签传播算法的变异策略如算法 2 所示.

算法 2. 基于标签传播的变异算法.

输入:选择后的个体 $g=\{g^1,g^2,...,g^M\}$,种群数量 M,图 G_t 的邻居集 N_t;

输出:交叉后的个体 $g=\{g^{1'},g^{2'},...,g^{M'}\}$.

1. **For** i=1 **to** M
2. 任意个体$\forall g^i \in g, i \in \{1,2,...,M\}$随机生成序列 $X=[x_1,x_2,...,x_n], x_i \in N_t(i)$
3. 从节点编号为 1 的节点开始更新,同时采用异步更新策略,更新一次

   异步更新: $g^i(t)=f(g_1^i(t),...,g_h^i(t),g_{h+1}^i(t-1),...,g_l^i(t-1)), t \in \{1,2,...,k\}$, k 为迭代次数
4. **End For**
5. **Return** M 个交叉个体 $g=\{g^{1'},g^{2'},...,g^{M'}\}$

3 LDMGA 算法流程

3.1 目标函数

目标函数:如前面公式(1)所述,本算法着重通过优化快照质量 SC 和历史开销 TC 来达到最终优化 $cost$ 的目的.因为快照质量 SC 衡量在 t 时刻社团结构的好坏,所以需要一个目标函数来最大化每个社团中的边的数量,最小化社团之间边的数量.所以,本文采用在社团结构发现领域广泛采用的标准-模块度 Q[22].

网络 $G_t=(V_t,E_t)$在时间 t 上有 n 个节点和 m 条边,其社团结构记为 $C=\{C_1,C_2,\ldots,C_k\}$,k 为社团数量.l_s 表示社团 C_s 中所有节点间的边的数目,d_s 表示社团中节点度数之和.

模块度 Q[22]的定义如下:

$$Q=\sum_{s=1}^{k}\left[\frac{l_s}{m}-\left(\frac{d_s}{2m}\right)^2\right] \tag{5}$$

在模块度 Q 的公式中,第 1 部分表示一个社团中边的概率,第 2 部分表示如果边随机分配,没有考虑社团结构,那么边在整个网络中的概率值.

第 2 个目标函数必须最小化历史开销 TC,本文采用标准化互信息 Normalized Mutual Information(NMI)[23],用一个矩阵来衡量时间 t 的社团结构和前一个时刻的社团结构的相似性.假设一个网络两个划分 $A=\{A_1,A_2,\ldots,A_a\}$和 $B=\{B_1,B_2,...,B_b\}$,C 是一个矩阵,其中的元素 C_{ij} 是同时在社团 $A_i \in A$ 和社团 $B_i \in B$ 中节点的数量.NMI[23]的定义如下:

$$NMI(A,B)=\frac{-2\sum_{t=1}^{C_A}\sum_{j=1}^{C_B}C_{ij}\log(C_{ij}N/C_{i.}C_{.j})}{\sum_{t=1}^{C_A}C_{i.}\log(C_{i.}/N)+\sum_{t=1}^{C_B}C_{.j}\log(C_{.j}/N)} \tag{6}$$

在这个公式中,C_A 表示划分 A 中社团的数量,C_B 表示划分 B 中社团的数量,$C_{i.}$表示矩阵 C 中行的和,$C_{.j}$ 表示矩阵 C 中列的和,N 是节点的数量.如果 $A=B$,则 $NMI(A,B)=1$.如果 A 和 B 完全不同,则 $NMI(A,B)=0$.所以,本文的第 2 个目标就是在时间 t 最大化 $NMI(C_t,C_{t-1})$.

3.2 LDMGA 算法流程

给定一个动态网络图序列 $G=\{G_1,G_2,\ldots,G_T\}$.LDMGA 算法首先发现网络 G_1 的划分,对网络进行单目标优化算法,采用轮盘赌注选择算法,只计算和优化第 1 个目标函数值,即 Q 的值,然后通过运行基于标签的遗传算法得到时刻 T=1 的网络社团划分.当时刻 T>1 时,多目标优化的遗传算法首先采用基于标签初始化算法生成一群个体,计算两个目标函数值,进行非支配排序给每个个体划分一个 $rank$.采用精英保留策略,选择 $rank$ 低的个体,通过交叉变异产生新的个体,子代与父代混合并进行非支配排序,选择较优个体进入下一次的迭代.算法经过固定次数的迭代后结束,同时,算法返回一组解,这些解都是 Pareto front.每个解对应两个目标函数之间不同的平衡点,并且每个网络的划分包含不同数目的社团.这些解已经是非支配解中满足快照质量和历史开销的解,在这些解中,本文选择社团结构最好的解,即选择模块度值最大的划分作为最后时间 t 返回的结果.LDMGA 算法如算法 3 所示.

算法 3. LDMGA(G,T).

输入:图序列 $G=\{G_1,G_2,\ldots,G_T\}$,时间点数 T;

输出:每个网络上的社团划分 $C_t=\{C_{t1},C_{t2},\ldots,C_{tk}\}$.

1. 基于标签的初始化算法,得到初始解 p 个 $g_t=\{g_t^1,g_t^2,\ldots,g_t^p\}$
2. $\forall g^i\in g,i\in\{1,2,\ldots,p\}$,解码得到社团 $C_t=\{C_{t1},C_{t2},\ldots,C_{tk}\}$,$k$ 为社团数量
3. 计算两个目标函数 Q,NMI
4. t=1 时,选择轮盘赌注算法,只优化第 1 个目标函数
5. **For** t=2 **to** T
6. **While** 终止条件不满足 **do**
7. 对每一个个体进行非支配排序,每个个体分配等级 *Rank*
8. 选择最优的个体生成子代
9. 对子代个体加以进化,即进行单路交叉操作和基于标签的变异操作
10. 子代和父代进行非支配排序,为每一个个体分配等级 *Rank*
11. 精英保留,选择等级低的个体进入下一代
12. **End while**
13. **Return** Q 值最大的个体作为解 $C_t=\{C_{t1},C_{t2},\ldots,C_{tk}\}$
14. **End for**

3.3 LDMGA 算法时间复杂度分析

在LDMGA算法中,基于标签的初始化算法时间为$O(m)$,m表示边的数量.在每次迭代过程中,算法解码的时间为 $O(n)$,单路交叉策略时间为 $O(n)$,基于标签的变异时间为 $O(m)$.计算 Q 值时,对于每个节点 i 和 di 个邻居,时间复杂度是 $O(m)$,m 是边的数量.对于 NMI 的计算,在文献[24]中证明 NMI 值可以在 $O(n)$的时间中被有效计算.在遗传算法中种群数量是 p,迭代次数为 g,则 LDMGA 算法时间复杂度为 $O(gp\log p\times(n+m))$.

4 实验结果与分析

在进化聚类的算法中,最经典的是 Lin 等人[10]提出的 FacetNet 框架和 Kim 等人[11]提出的基于微粒与密度的进化聚类方法.Folino 等人[12]提出了 DYNMOGA 算法,通过对比这两种算法发现,无论是聚类的效果,还是时间平滑性,DYNMOGA 算法都要优于前面两种算法.本节对所提出的 LDMGA 算法和 DYNMOGA 算法、FacetNet 算法以及 Kim 等人提出的算法进行综合测评.

4.1 实验设置

LDMGA 算法、DYNMOGA 算法和 FacetNet 算法均用 Matlab 实现,其中,DYNMOGA 算法和 FacetNet 算法的源代码由作者提供.本文实验的硬件环境为 CPU 2.3GHz 的 Intel(R) Core(TM) i3-2350M,内存 4G;软件环境为 Windows 7:Matlab R2008b.

4.1.1 参数设置

LDMGA 算法采用遗传算法,所以参数的选择在进化算法中很重要.在文献[25]中,作者已经证明对于普遍问题很难找到好的参数.所以,本文采用 trial-and-error 方法,通过改变交叉和变异概率进行实验,实验的数据集采用 Girvan 和 Newman 等人[26]提出的标准生成.本文生成的数据集网络由 128 个节点组成,分成 4 个社团,每个社团 32 个节点.每个节点有固定的平均度数 *avgDegree*=16,并且每个节点有 z=5 条边连接所属社团以外的节点.

从图 6 可以看出,标准化互信息 NMI 值没有表现出明显的变化.鉴于一般采用高的交叉概率,低的变异概率,所以,设置交叉概率为 0.8,变异概率为 0.2.群体数量为 100,迭代次数为 100,结果为运行一次的值.DYNMOGA[12]算法,参考作者在文中的参数设定,实验设置参数为:交叉概率 0.8,变异概率 0.2,群体数量 100,迭代次数为 100,结果为运行 50 次取平均值.FacetNet 算法设置 *alpha*=0.8.

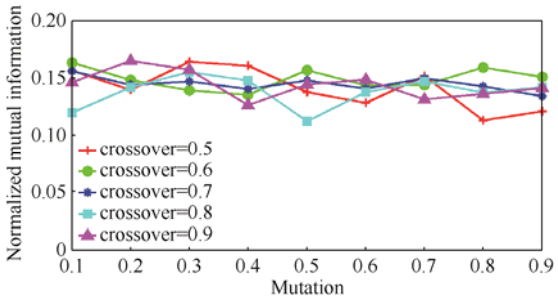


Fig.6 Normalized mutual information for different combinations of crossover and mutation rates

图 6 不同的交叉概率和变异概率上的标准化互信息的值

4.1.2 评价指标

本文使用两个验证函数来评价结果的质量,NMI 和 Error Rate[27],NMI 函数前面已经介绍过.在 Error Rate 的计算中,首先建立一个标识矩阵 Z,Z 为 $n\times k$ 的矩阵,n 为节点的数量,k 为社团数量,另外还有一个相似标识矩阵 G,表示真实社团.Error Rate 定义为$||ZZ^T-GG^T||$,表示社团 Z 和真实社团 G 在社团结构上的距离.

4.2 数据集#1

数据集#1 考虑一些重要的事件标志动态网络的演化[28,29].本文采用 Greene 等人[29]生成数据集的方法生成 4 个数据集,总共 20 个时间点.参数的设置为 1 000 个节点,每个节点的平均度数为 15,最大度数为 50,混合参数为 0.2,即社团之间边的概率.

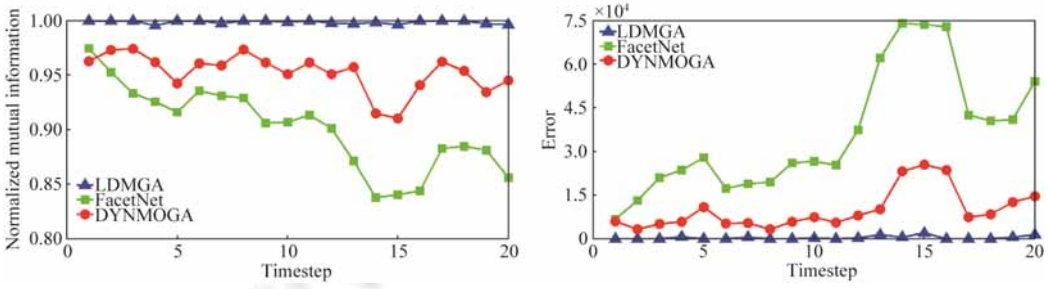
数据 Merging and splitting:在每个时间点上,10%的社团被分裂,10%的社团被选择,并进行合并.

数据 Expansion and contraction:随机选择 10%的社团进行扩张或收缩社团大小,比例是 25%.当扩张时,新的节点被随机地从其他社团中进行选择.

数据 Intermittent communities:10%的社团从第 1 个时间点开始被隐藏.

数据 Birth and death:从第 2 个时间点,10%的新社团被创建,通过从存在的社团中移动节点,随机移除 10%现有的社团.

从图 7(a)、图 7(b)的实验结果来看,LDMGA 算法得到的 NMI 值和 Error 值明显比 DYNMOGA 算法和 FacetNet 算法得到的结果要好.在社团的分裂合并过程中,加入了标签的 LDMGA 算法能够准确地发现社团结构,所以,NMI 的值一直接近 1,而 DYNMOGA 算法的 NMI 值一直处于 0.95 左右.从图 7(a)可以看出,FacetNet 算法的 NMI 值一直处于下降的趋势,且下降幅度明显.由于 LDMGA 算法发现的社团中几乎没有分错社团的节点,所以其错误率接近于 0,而 DYNMOGA 算法的错误率平均值在 7 000 左右.FacetNet 算法的错误率呈上升趋势(如图 7(b)所示),所以 FacetNet 算法并不适合该动态网络的社团发现.从实验数据分析得知,DYNMOGA 算法中,存在较多节点分错了社团,或者是存在较小的社团没有合并.

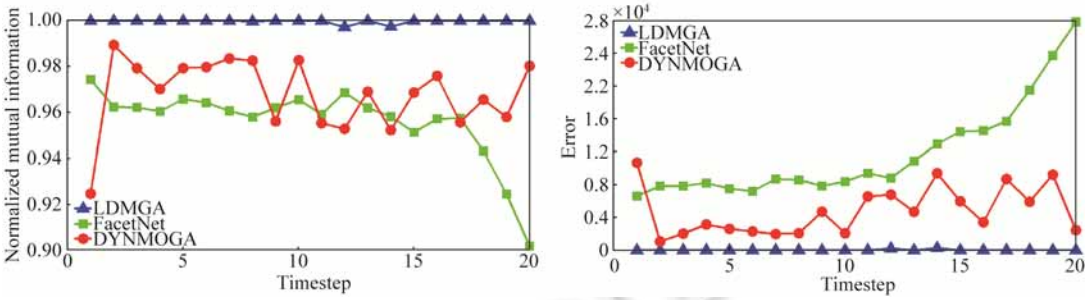


(a) 标准化互信息 (b) 错误率

Fig.7 Merging and splitting

图 7 数据集 Merging and splitting

从图 8(a)、图 8(b)可以看出,3 种算法运行结果和在数据集 Merging and splitting 上的结果相似,只是 FacetNet 算法和 DYNMOGA 算法的结果差距在缩小.在社团进行扩张或收缩的过程中(如图 8(a)所示),LDMGA 算法发现的社团明显比 DYNMOGA 算法和 FacetNnet 算法得到的社团结构更加接近真实的社团结构.从 NMI 的结果一直为 1 或者接近 1,错误率接近 0 或者为 0 可以看出,LDMGA 算法非常适用于该种类型的网络.而 DYNMOGA 算法的错误率平均值在 5 000 左右,NMI 值在 0.97 左右,说明存在部分节点分错社团.而 FacetNet 算法在该种动态网络上,性能在下降,当时间点为 17 时,性能下降明显.

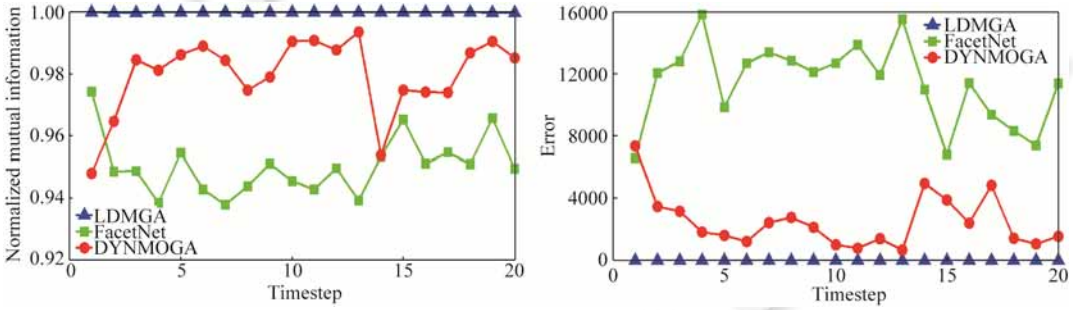


(a) 标准化互信息 (b) 错误率

Fig.8 Expansion and contraction

图 8 数据集 Expansion and contraction

从图 9(a)、图 9(b)可以看出,20 个时间点上,LDMGA 算法的 NMI 值都为 1,Error 值为 0.这意味着,在每个时间点存在社团隐藏的情况下,LDMGA 算法依然能够准确发现这种网络类型的社团结构.虽然 DYNMOGA 算法随着时间点的增长其 NMI 值不断接近 1(如图 9(a)所示),但是,LDMGA 算法发现的社团结构更好,相比下来,FacetNet 算法的 NMI 值一直在 0.95 左右,而错误率较高(如图 9(b)所示),明显低于 LDMGA 算法的性能.

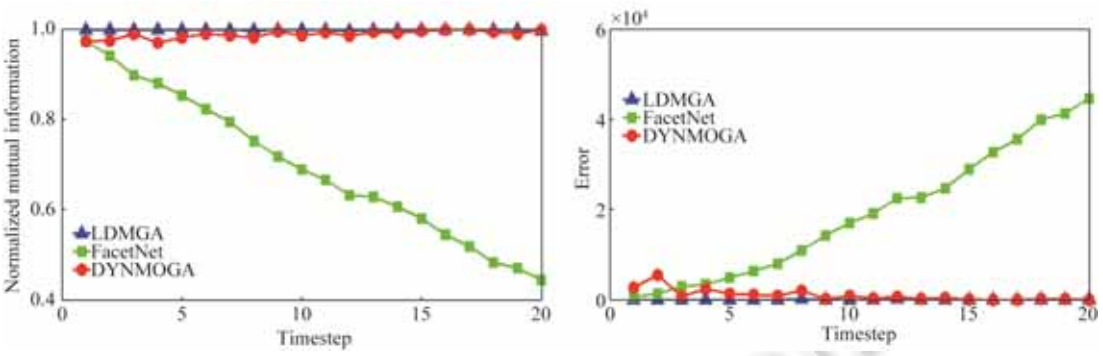


(a) 标准化互信息 (b) 错误率

Fig.9 Intermittent communities

图 9 数据集 Intermittent communities

从图 10(a)、图 10(b)可以看出,DYNMOGA 算法的结果虽然与 LDMGA 算法的结果相近,但是 LDMGA 算法的运行结果比 DYNMOGA 算法的运行结果要好.在新社团产生和旧社团消失的网络中,从 NMI 的结果一直为 1 或者接近 1(如图 10(a)所示)可以看出,LDMGA 算法发现的社团结构比 DYNMOGA 算法得到的社团结构更加接近真实的社团结构.尤其是从 Error 值可以看出,DYNMOGA 算法的错误率在 3 000 左右,而 LDMGA 算法的错误率接近 0.这说明,LDMGA 算法非常适用于该种类型的网络.而 FacetNnet 算法的性能随着社团的产生和消失,NMI 值在不断地下降,Error 值在不断地增大.



(a) 标准化互信息 (b) 错误率

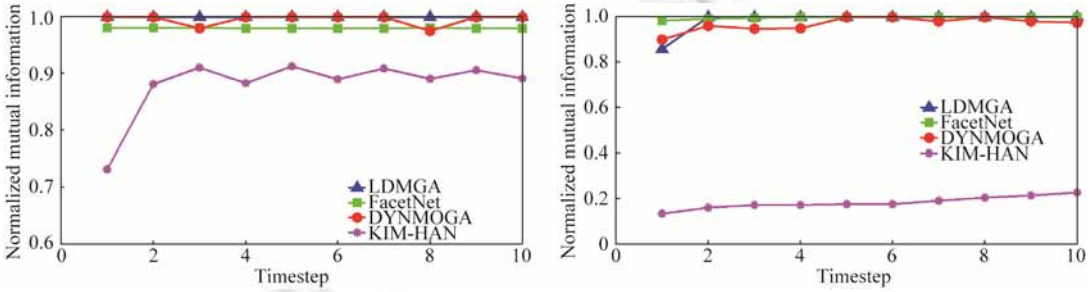
Fig.10 Birth and death

图 10 数据集 Birth and death

4.3 数据集#2

数据集#2 包含两种类型的数据集,一个是 SYN-FIX,另一个是 SYN-VAR.SYN-FIX 网络中有 128 个节点,被分成 4 个社团,每个社团中有 32 个节点.每个节点平均度数是 16,并且有 z 条边连接社团以外的节点.为了引入动态网络.随机地从每个社团选择 3 个节点,然后随机分配到其他 3 个社团中.SYN-VAR 网络中有 256 个节点,被分成 4 个社团,每个社团有 64 个节点.通过从每个社团中选择 8 个节点并且生成一个新的 32 个节点社团,生成 10 个连续的网络.这个过程持续 5 个时间点,然后节点数目恢复到最初的社团.所以,10 个时间点的网络社团数目为 4,5,6,7,8,8,7,6,5,4.社团中平均每个节点的度数被设置为社团大小的一半.此外,每个时间点,随机删除 16 个节点,同时增加 16 个新节点.本实验用的数据集 SYN-FIX 和 SYN-VAR 以及 KIM-HAN 算法的结果是由 DYNMOGA 算法作者所提供的.

从图 11(a)、图 11(b)的实验结果来看,当 z=3 时(如图 11(a)所示),LDMGA 算法的 NMI 值全部为 1,Error 值全部为 0,虽然 DYNMOGA 算法在时间点 3 和 8 上 NMI 值不为 1,但是都很接近 1.而 FacetNet 算法的 NMI 值也接近 1,但比 LDMGA 算法略差,KIM-HAN 的算法从时间点 2 以后其 NMI 值一直保持在 0.9 左右.所以,LDMGA 算法和 DYNMOGA 算法都能得到很接近真实社团的社团结构.但是,LDMGA 算法对于这种类型的网络出错的概率更小(如图 11(b)所示),相比于 DYNMOGA 算法,LDMGA 算法更适合应用在这种类型的网络上.当 z=5 时(如图 11(b)所示),由于 z 增大,导致网络的模糊度增大,社团结构开始减弱,LDMGA 算法、DYNMOGA 算法和 FacetNet 算法的结果依然非常相近.除了时间点 1,LDMGA 算法的结果和 FacetNet 算法的 NMI 值是相同的,LDMGA 算法能够准确发现社团的结构,而 KIM-HAN 算法的 NMI 值在 0.2 左右,该算法并不能很好地发现社团结构.

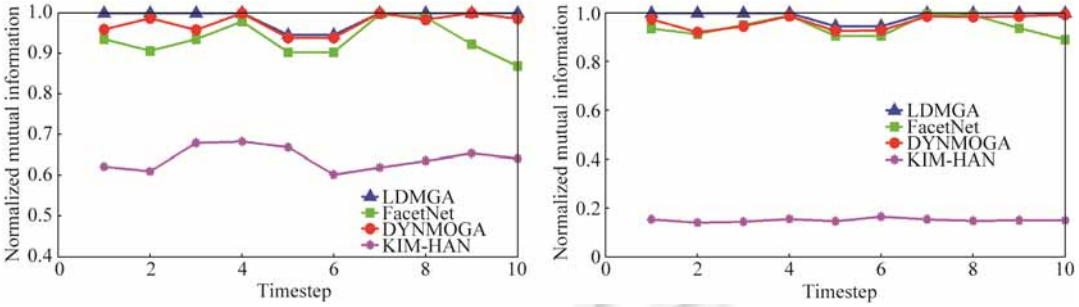


(a) z=3 (b) z=5

Fig.11 SYN-FIX NMI

图 11 SYN-FIX 标准化互信息

从图 12(a)、图 12(b)实验结果来看,在两个网络上,LDMGA 算法、DYNMOGA 算法和 FacetNet 算法有相同的趋势.LDMGA 算法除了在时间点 5、时间点 6 以外的其他时间点上 *NMI*=1,说明 LDMGA 算法能够准确发现社团结构.并且在时间点 5、时间点 6 的网络上,LDMGA 算法相比于另外 3 种算法的结果都要好,同时,LDMGA 算法的结果要明显高于 KIM-HAN 算法.当 z=5 时,社团结构模糊后,KIM-HAN 算法的 NMI 值均低于 0.2,KIM-HAN 算法发现社团结构的性能相较于 z=3 的时候下降得较为明显.



(a) z=3 (b) z=5

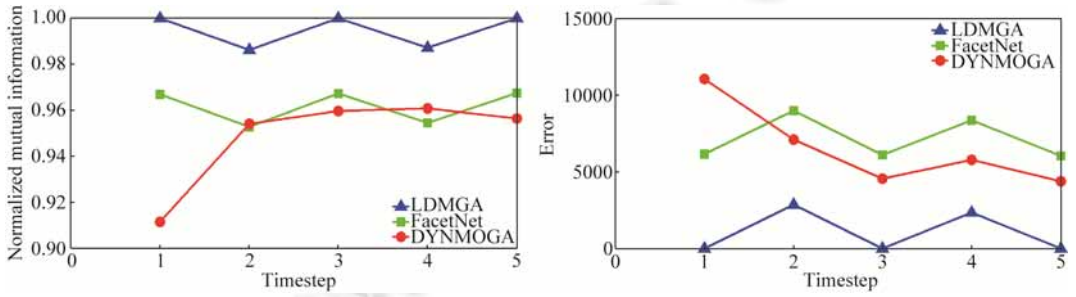
Fig.12 SYN-VAR NMI

图 12 SYN-VAR 标准化互信息

4.4 Power-Law 网络

Power-Law 网络采用 Lancichinetti 等人[30]提出的网络基准(LFR),它通过引入 power law degree distributions 的概念和不同的社团大小扩展了 Girvan 和 Newman 提出的标准[26].本文采用的 LER 网络数据由 DYNMOGA 算法[12]作者提供.该网络有 1 000 个节点,平均节点度数为 20,最大节点度数为 50,度分布的幂指数为–2,社团大小分布为–1,并且混合参数为 0.3.生成的网络只有一个节点的最大度数 50.70%的节点度数低于平均节点度数 20.为了引入动态网络,设置 5 个时间点,随机选择 10%的社团,重复进行分裂(时间点 2,时间点 4)和合并(时间点 3,时间点 5).

图 13(a)、图 13(b)展示了 LDMGA 算法、DYNMOGA 算法和 FacetNet 算法的 NMI 和 Error 结果.实验结果显示,从第 1 个时间点开始,LDMGA 算法就明显优于 DYNMOGA 算法和 FacetNet 算法的结果,LDMGA 算法和 FacetNet 算法的 NMI 结果趋势相同,但是 LDMGA 算法能够较好地发现社团结构,且 NMI 的值一直不低于 0.98.虽然 DYNMOGA 算法从第 2 个时间点开始,NMI 的结果稳定在 0.96,但是 LDMGA 算法的错误率(如图 13(b)所示)明显低于 DYNMOGA 算法和 FaceNet 算法,更加证明了 LDMGA 算法能够准确地发现节点所属的社团.



(a) 标准化互信息 (b) 错误率

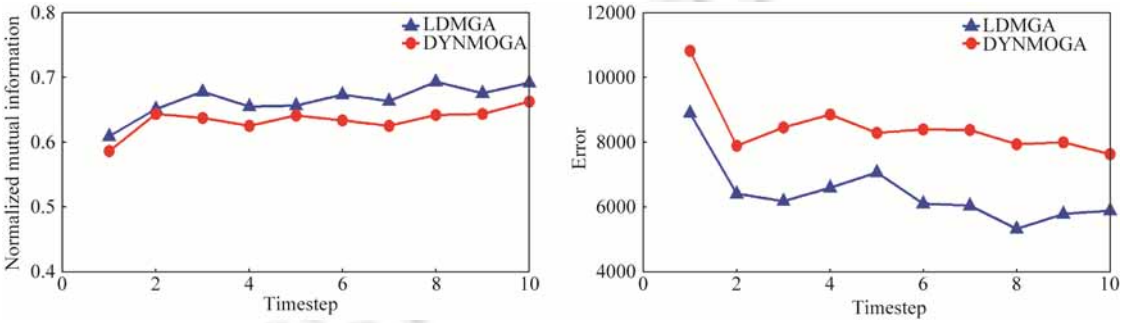
Fig.13 Power-Law

图 13 Power-Law 网络

4.5 真实数据集

Cell Phone Calls这个数据集来自于VAST 2008 mini challenge 3[31].Cell Phone Calls数据集中包含了虚拟的Paraiso运动成员间的手机通话记录,2006年6月中10天的记录.这些记录组成网络,在网络中节点表示每个手机,边表示手机之间有通话记录.该网络是节点数量400,10个时间点的有权网络.每条边表示当天的两个成员的通话次数的总时间.因为真实社团结构未知,本文仿照Lin等人[27]提出的相同的方法.首先考虑整体的网络并且计算社团结构,即只考虑一个目标函数Q.在本文中,真实的社团划分是用FacetNet算法实现的,LDMGA算法依据FacetNet算法划分的社团结果来评价结果质量.真实网络社团平均模块度为0.3,并且平均社团数目为25.本实验采用的数据集Cell Phone Calls由DYNMOGA算法作者所提供.

图14(a)、图14(b)所示为LDMGA和DYNMOGA算法在Cell Phone Calls数据集上NMI和Error的结果.LDMGA算法在NMI和Error上都明显优于DYNMOGA算法.LDMGA算法的NMI值在0.67左右,而DYNMOGA算法的NMI值在0.64左右,且从图14(b)可以明显看出,LDMGA算法的Error明显小于DYNMOGA算法.由此可见,LDMGA算法能够在真实网络中平衡社团质量和时间平滑性,发现良好的社团结构.



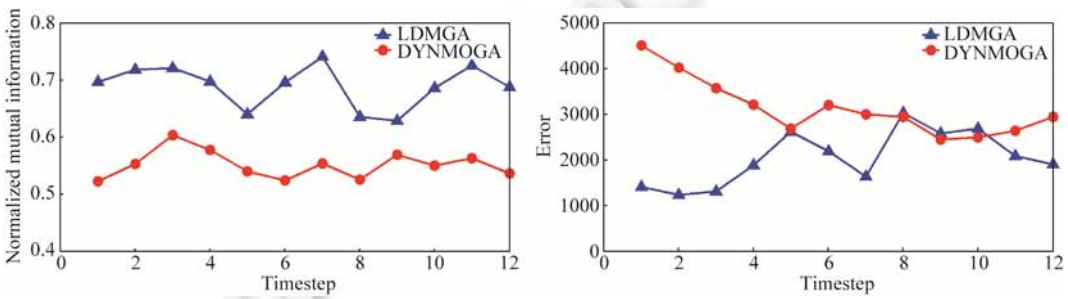
(a) 标准化互信息 (b) 错误率

Fig.14 Cell Phone Calls

图14 Cell Phone Calls数据集

Enron mail数据集[32]:第2个数据集是U.S.公司1992年~2002年潜在的异常邮件.原始的数据集包含517 431封邮件,来自151个用户,在3 500个文件夹中.该数据集在DYNMOGA作者数据清理后,数据规模被减小到50 000. 2001年的数据按每个月被分为一个子集,共分为12个子集,采用和Cell Phone Calls1同样的实验方法,用FacetNet算法获得真实社团结构.整体网络平均社团数目为11.本实验采用的Enron mail数据集由DYNMOGA算法作者所提供.

图15(a)、图15(b)所示为LDMGA和DYNMOGA算法在Enron mail数据集上NMI和Error的结果.LDMGA算法的NMI值在0.7左右,而DYNMOGA算法的NMI值在0.55左右.明显地,LDMGA算法比DYNMOGA算法可更准确地发现社团结构.由此可见,在Enron网络上,LDMGA算法依然能够发挥优势,在快照质量和历史开销之间寻得平衡.



(a) 标准化互信息 (b) 错误率

Fig.15 Enron mail

图15 Enron mail数据集

4.6 可扩展性分析

用遗传算法解决优化问题最大的阻碍就是较长的计算时间.而且,进化算法中一个主要的问题是适应度函数的重复计算.当群体数量很大时,这个问题会变得很严重,尤其是针对多目标优化方法.从之前的时间复杂度分析中可以看出,在本文的算法中适应度计算是非常有效率的,对于较大的网络也是非常有效的.

为了证明本文算法的可扩展性强,采用 Girvan 等人[26]提出的标准生成数据集.该网络分为 4 个社团,节点平均度数 $avgDegree$=16,z=5,每个时间点上有 10%的节点被选择进入其他社团,且节点数目 n 的变化范围是{128,256,512,1024,2048,4096,8192,16384},相应的边 m 的变化范围是{1938,4018,8184,16158,33026,65256,131388,285782},群体数量 p 的变化范围是{50,100,200},迭代次数 g 的变化范围是{50,100}.表 2、表 3 显示的结果是一个时间点上,针对不同的 p,g 组合,LDMGA 算法和 DYNMOGA 算法运行的时间.从表 2、表 3 中可以看出,尤其是当节点数量成倍增长时,LDMGA 算法的运行时间明显低于 DYNMOGA 算法,DYNMOGA 算法本身运行一次的时间就是 LDMGA 算法运行时间的 6 倍以上.LDMGA 算法的时间复杂度为 $O(gp\log p\times(n+m))$,在种群数量和迭代次数确定的情况下,LDMGA 算法运行时间呈线性增长,所以该算法的可扩展性较强.

Table 2 Running time g=50

表 2 算法运行时间 g=50

LDMGA/DYNMOGA	节点 2^7	节点 2^8	节点 2^9	节点 2^{10}	节点 2^{11}	节点 2^{12}	节点 2^{13}	节点 2^{14}
p=50	3.394s/17.8s	5.700s/33s	10.656s/100s	18.69s/250s	42.42s/500s	102.21s/1 500s	253.58s/3 800s	790.15s/12 000s
p=100	6.099s/31.7s	11.483s/62.78s	21.666s/200s	43.579s/500s	105.272s/1 000s	277.092s/3 100s	547.22s/7 900s	1 485.39s/24 000s
p=200	13.684s/69.25s	24.573s/132s	44.454s/400s	88.529s/900s	208.453s/2 200s	594.152s/6 000s	1 175.83s/15 500s	2 971.10s/48 000s

Table 3 Running time g=100

表 3 算法运行时间 g=100

LDMGA/DYNMOGA	节点 2^7	节点 2^8	节点 2^9	节点 2^{10}	节点 2^{11}	节点 2^{12}	节点 2^{13}	节点 2^{14}
p=50	4.816s/39.477s	5.8756s/66.625s	8.973s/160.289s	36.644s/500s	82.237s/1 000s	223.869s/3 600s	481.79s/6 800s	1 201.82s/18 000s
p=100	10.673s/67.868s	20.646s/126.659s	36.199s/318.197s	73.273s/800s	165.51s/2 000s	419.365s/6 000s	931.60s/11 000s	2 274.07s/38 000s
p=200	22.929s/146.37s	39.152s/277.280s	73.621s/596.435 9s	146.577s/1 900s	330.605s/5 000s	814.49s/13 000s	2 034.16s/23 000s	3 931.91s/71 000s

5 结 论

本文提出了一种基于标签的多目标优化算法,该算法在动态网络上能够发现较好的社团结构,同时满足时间平滑性的要求.多目标优化的思想能够在每个时间点提供一种良好的平衡,既能够根据当前时间点的网络发现良好的社团结构,又能够使连续时间点上的网络结构差异性较小.在多目标优化算法中加入标签算法,能够有效地提高算法的精度,即聚类的效果.此外,基于标签的变异算法很好地增强了社团结构,同时缩短了运行的时间.实验结果表明,本文提出的算法在仿真数据和真实数据上都优于目前优秀的算法.最值得一提的是,LDMGA 算法在运行时间上明显优于 DYNMOGA 算法,更适用于大规模的数据挖掘.

References:

[1] Leskovec J, Krevl A. Large Network Dataset Collection. SNAP Datasets: Stanford, 2014.

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

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

[4] Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group formation in large social networks: Membership, growth, and evolution. In: Mierswa I, Wurst M, Klinkenberg R, eds. Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Philadelphia: ACM, 2006. 44–54. [doi: 10.1145/1150402.1150412]

[5] Tantipathananandh C, Berger-Wolf T, Kempe D. A framework for community identification in dynamic social networks. In: Tantipathananandh C, Berger-Wolf T, Kempe D, eds. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. San Jose: ACM, 2007. 717–726. [doi: 10.1145/1281192.1281269]

[6] Shan B, Jiang SX, Zhang S, Gao H, Li JZ. IC: Incremental algorithm for community identification in dynamic social networks. Ruan Jian Xue Bao/Journal of software, 2009,20:184−192 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/09022.htm

[7] Chakrabarti D, Kumarand R, Tomkins A. Evolutionary clustering. In: Mierswa I, Wurst M, Klinkenberg R, eds. Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Philadelphia: ACM, 2006. 554–560. [doi: 10.1145/1150402.1150467]

[8] Yang B, Liu DY. Force-Based incremental algorithm for mining community structure in dynamic network. Journal of Computer Science and Technology, 2006,21(3):393–440. [doi: 10.1007/s11390-006-0393-1]

[9] Sun JM, Papadimitriou S, Yu PS, Faloutsos C. GraphScope: Parameter-Free mining of large time-evolving graphs. In: Tantipathananandh C, Berger-Wolf T, Kempe D, eds. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Jose: ACM, 2007. 687–696. [doi: 10.1145/1281192.1281266]

[10] Lin YR, Chi Y, Zhu SH, Sundaram H, Tseng BL. FacetNet: A framework for analyzing communities and their evolutions in dynamic networks. In: Lin Y, Chi Y, Zhu S, eds. Proc. of the 17th Int'l Conf. on World Wide Web. Beijing: ACM, 2008. 685–694. [doi: 10.1145/1367497.1367590]

[11] Kim MS, Han JW. A particle-and-density based evolutionary clustering method for dynamic networks. Proc. of the VLDB Endowment, 2009,2(1):622–633. [doi: 10.14778/1687627.1687698]

[12] Folino F, Pizzuti C. An evolutionary multi-objective approach for community discovery in dynamic networks. IEEE Trans. on Knowledge and Data Engineering, 2014,99(8):1. [doi: 10.1109/TKDE.2013.131]

[13] Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans. on Evolutionary Computation, 2002,6(2):182–197. [doi: 10.1109/4235.996017]

[14] 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).

[15] Li S, Chen Y, Du H, Feldman MW. A genetic algorithm with local search strategy for improved detection of community structure. Complexity, 2010,15(4):53–60. [doi: 10.1002/cplx.20300]

[16] Pizzuti C. A multi-objective genetic algorithm for community detection in networks. In: Charot F, Hannig F, Teich J, eds. Proc. of the 21st IEEE Int'l Conf. on Tools with Artificial Intelligence. IEEE, 2009. 379–386. [doi: 10.1109/ICTAI.2009.58]

[17] Shi C, Yan Z, Wang Y, Cai Y, Wu B. A genetic algorithm for detecting communities in large-scale complex networks. Advances in Complex Systems, 2010,13(1):3–17. [doi: 10.1142/S0219525910002463]

[18] Jin D, He D, Liu D, Baquero C. Genetic algorithm with local search for community mining in complex networks. In: Grégoire É, ed. Proc. of the 22nd IEEE Int'l Conf. on Tools with Artificial Intelligence. Arras: IEEE, 2010. 105–112. [doi: 10.1109/ICTAI.2010.23]

[19] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation. Technical Report, CMU-CALD-02-107, Pittsburgh: Carnegie Mellon University, 2002.

[20] Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007,76(3):1–12. [doi: 10.1103/PhysRevE.76.036106]

[21] Tasgin M, Herdagdelen A, Bingol H. Community detection in complex networks using genetic algorithms. Eprint Arxiv, 2007. arXiv:0711.0491

[22] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010,12(10): 2011–2024. [doi: 10.1088/1367-2630/12/10/103018]

[23] Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005,2005(9):9008. [doi: 10.1088/1742-5468/2005/09/P09008]

[24] Liu CT, Hu BG. Mutual information based on Renyi's entropy feature selection. In: Proc. of the 2009 IEEE Int'l Conf. on Intelligent Computing and Intelligent Systems (ICIS 2009). Shanghai: IEEE, 2009,1:816–820. [doi: 10.1109/ICICISYS.2009.5358033]

[25] Smit SK, Eiben AE. Parameter tuning of evolutionary algorithms: Generalist vs. specialist. Applications of Evolutionary Computation, 2010,6024:542–551. [doi: 10.1007/978-3-642-12239-2_56]

[26] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences of the United States of America, 2002,99(12):7821–7826. [doi: 10.1073/pnas.122653799]

[27] Lin YR, Zhu S, Sundaram H, Tseng BL. Analyzing communities and their evolutions in dynamic social networks. ACM Trans. on Knowledge Discovery from Data (TKDD), 2009,3(2):8. [doi: 10.1145/1514888.1514891]

[28] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Trans. on Knowledge Discovery from Data (TKDD), 2009,3(4):16. [doi: 10.1145/1631162.1631164]

[29] Greene D, Doyle D, Cunningham P. Tracking the evolution of communities in dynamic social networks. In: Memon N, ed. Proc. of the 2010 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM). Odense: IEEE, 2010. 176–183. [doi: 10.1109/ASONAM.2010.17]

[30] 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]

[31] Cell Phone Calls. http://www.cs.umd.edu/hcil/VASTchallenge08/

[32] Enron Email Dataset. http://www.cs.cmu.edu/~enron/

附中文参考文献:

[6] 单波,姜守旭,张硕,高宏,李建中.IC:动态社会关系网络社区结构的增量识别算法.软件学报,2009,20:184–192. http://www.jos.org.cn/1000-9825/09022.htm

[14] 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.

牛新征(1978－),男,河南唐河人,博士,副教授,CCF 高级会员,主要研究领域为移动计算,数据挖掘.

司伟钰(1992－),女,硕士生,主要研究领域为数据挖掘,机器学习.

佘堃(1967－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络计算,人工智能.