一种基于交叉熵的社区发现算法

于 海 赵玉丽 崔 坤 朱志良
(东北大学软件学院 沈阳 110819)

摘 要 作为复杂网络中的一个极其重要的研究领域,社区结构的搜寻和发现研究具有重要的应用价值.该文将信号处理领域的交叉熵概念引入到网络社区结构的发现算法中,提出了一种基于交叉熵的社区发现算法.算法利用Modularity值作为判别依据,使用交叉熵方法中的重要抽样方法提高收敛速度,从而在提高社区发现算法运算效率的同时,提高算法的精确性.针对计算机生成网络的社区划分结果表明,该算法所得*MNI*值和划分正确节点所占比例高于Girvan-Newman算法.在真实网络上的仿真结果表明,该社区划分算法的Modularity值高于Girvan-Newman算法,且不低于极值优化算法,进一步验证了该文提出算法的社区划分准确性优于已有的Girvan-Newman算法和极值优化算法.

关键词 复杂网络;社区发现;交叉熵
中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2015.01574

Community Detection Algorithm Based on Cross-Entropy Method

YU Hai ZHAO Yu-Li CUI Kun ZHU Zhi-Liang
(*Software College*, *Northeastern University*, *Shenyang* 110819)

Abstract Community detection algorithm is a very significant research topic in the complex network theory, which can be applied in communities' structures search and discovery applications. In this paper, the concept of Cross-Entropy in the field of signal processing is introduced and a community detection algorithm based on Cross-Entropy is proposed. The algorithm defines modularity as the quality function, which uses importance sampling in Cross-Entropy to speed up the convergence, thus the efficiency and accuracy of communities' detection can be improved simultaneously. Comparing with the Girvan-Newman algorithm over networks the computer generated, the proposed algorithm achieves higher NMI and the proportion of correctly division nodes. Moreover, the simulation results over real-world networks further reveal that the proposed algorithm accomplishes higher value of Modularity than Girvan-Newman algorithm, and no less than External Optimization algorithm. It is further verified that the proposed algorithm is more accurate than Girvan-Newman and External Optimization ones.

Keywords complex networks; community detection; cross-entropy

1 引 言

近年来,复杂网络引起了研究人员的广泛关注[1-4].现实中很多系统都可以被看作复杂网络,如Internet、万维网、社会关系网络、食物网络和科技文献引用网络等[5-10].研究人员发现,这些网络中节点的分布并不是随机的,而是呈现出一些有规律的结

收稿日期:2013-07-08;最终修改稿收到日期:2015-01-05.本课题得到国家自然科学基金(61374178,61402092)、辽宁省自然科学基金(201202076)、中央高校基本科研业务费专项资金(N130417004,N130317001)资助.于 海,男,1971年生,博士,副教授,主要研究方向为混沌、复杂网络理论与应用.E-mail:yuhai@mail.neu.edu.cn.赵玉丽,女,1985年生,博士,主要研究方向为复杂网络、信道编码.崔 坤,男,1986年生,硕士,主要研究方向为复杂网络、计算机算法.朱志良,男,1962年生,博士,教授,主要研究领域为混沌、复杂网络、软件工程.

构特性，如“小世界”性和“无标度”性等[10].

复杂网络的一个重要特性是社区结构，即网络可以由若干社区构成，社区内的节点连边较多，而社区之间存在的连边较少[11-13]. 社区内的节点有很多共同的特性，如在社会网络中，社区表示了真实的社会团体；万维网中的社区则表示了一系列相关主题的网站集合；软件系统中的社区表示了实现某一类功能的组件[14-17]. 因此，发现并分析这些社区能够帮助研究人员更好地理解和研究复杂网络，揭示复杂网络的内在机制，改善网络行为，具有重要的应用价值.

研究人员提出了很多算法来解决社区发现问题[18-20]. 2004 年，Girvan 和 Newman[21] 提出了一种根据最大边介数来划分社区的方法，被称为 Girvan-Newman 算法，他们首次提出了一个定量描述——Modularity 来评价社区结构的合理性. 这种方法已经在很多真实的复杂网络中得到了应用，例如：动物社会网络、新陈代谢网络、电子邮件网络等.

时间复杂度和准确性是复杂网络社区发现算法面临的主要问题. 利用最优的 Modularity 搜索社区的计算复杂度是一个 NP 完全问题. 为降低计算的复杂度，基于组合优化算法的社区发现方法被相继提出. 2005 年，Medus 等人[22] 提出了基于模拟退火算法的社区发现算法，Duch 和 Arenas[23] 使用极值优化算法实现社区划分；2007 年，Xu 等人[24] 将 Mixed integer 算法应用到社区发现中，提高了社区发现的准确度. 尽管这些结果从不同角度研究了社区发现问题，丰富了网络社区结构的意义，但这些算法仍然存在着不同的缺点，如 Girvan-Newman 算法由于算法的局部性和不可逆性，划分精度不高；而极值优化算法虽然大大提高了社区划分的 Modularity 值，但是在处理划分有歧义的节点时，准确率依然不高. 因此，针对不同的网络结构，如何在降低计算复杂度的前提下，提高社区发现算法的准确性是社区发现研究需要解决的重要问题之一.

为解决这一问题，本文提出了一种利用组合优化算法中的交叉熵(Cross Entropy)方法来划分社区的算法，实验结果表明在不增加算法复杂度的前提下，本文所提方法在准确性上优于上述算法.

2　基于交叉熵的社区发现算法

2.1　社区结构的定量描述

目前，有关复杂网络中的社区结构还没有一个被广泛认可的定义. 在网络社区结构研究中通常会选用 Newman 和 Girvan 在文献[21]中定义的 Modularity 来定量描述网络中的社区结构.

Modularity 是网络社区结构的一个定量描述，指网络中连接社区内部节点的边所占的比例与相同规模和社区结构的随机网络中连接社区内部节点的边所占比例相减得到的差值. 文献[21]中给出了随机网络的构造方法：保持每个节点的社区性质，节点间的边根据节点的度随机连接.

对于一个已经划分社区的网络 G，定义一个 $n\times n$ 的对称矩阵 $\boldsymbol{A}$ 表示网络中节点的邻接关系，当且仅当节点 i 和节点 j 存在连边时 $A_{ij}=1$，否则 $A_{ij}=0$. 若网络中边的数量为 m，c_i 为节点 i 所在的社区，则描述社区划分的 Modularity 函数 Q 有如下表示形式[14,25]：

$$Q=\frac{1}{2m}\sum_{ij}\left[\boldsymbol{A}_{ij}-\frac{k_ik_j}{2m}\right]\delta_{c_i,c_j} \tag{1}$$

其中：k_i 表示节点 i 的度；$k_ik_j/2m$ 表示随机网络中节点 i 和节点 j 之间的连边数的期望. δ_{ij} 是 Kronecker 函数，当 $c_i=c_j$ 时值为 1，否则为 0.

Q 值的大小反映了网络社区内部连接的强弱. 通常，Q 值在 0～1 之间，Q 值越接近 1，社区结构越明显，一般以 $Q=0.3$ 作为网络具有明显社区结构的下界. 然而，Q 值作为社区结构定性定义的一个度量方式，表明了社区结构的内聚性，并不能作为社区划分准确度的评价标准. 目前，在评价社区划分准确度方面主要借助于诸如划分正确节点所占比例和规范化交互信息(Normalized Mutual Information，NMI)等统计信息[26-27].

2.2　交叉熵和交叉熵方法

1951 年，Kullback 和 Leibler 在文献[28]中首次提出了交叉熵的概念. 作为一种概率分布与其近似分布接近程度的一种度量，交叉熵广泛地应用于数字信号处理领域.

定义 1.　交叉熵. 若存在基于随机变量 $X=\{X_1,X_2,\cdots,X_n\}$ 的两个概率分布函数 $P(\cdot)$ 和 $Q(\cdot)$，则分布函数 $Q(\cdot)$ 对于 $P(\cdot)$ 的交叉熵为

$$CE(P,Q)=\sum_{[X_1,X_2,\cdots,X_n]}P(\cdot)\log\frac{P(\cdot)}{Q(\cdot)} \tag{2}$$

1999 年，Rubinstein[29] 在解决组合优化问题中首次引入了交叉熵方法. 交叉熵方法最初用于估计复杂随机网络中稀有事件发生的概率，但是经实验发现它对于优化问题也非常适用，是一种寻找估计方差的自适应算法.

组合优化通常是指在有限的数学结构上,寻找一个满足给定约束条件并使其目标函数值达到最大或最小的解.对于一个如下式所示的组合优化问题:

$$\gamma^* = \max_{x\in\chi} S(x) \tag{3}$$

其实质为求解 $S(x)\geq\gamma$ 的概率估计:

$$\ell = P(S(x)\geq\gamma) \tag{4}$$

若χ上存在一个 $\gamma\in R$ 的示性函数集合$\{I_{\{S(x)\geq\gamma\}}, x\in\chi,\gamma\in R\}$,其具体表示为

$$I_{\{S(x)\geq\gamma\}} = \begin{cases} 1, & S\geq\gamma,\ S\geq 0 \\ 0, & S<\gamma,\ S\geq 0 \end{cases} \tag{5}$$

则估计 ℓ 可以通过重要抽样方法实现.

设 $f(x,p)$和 $f(x,q)$是χ上的等价概率分布,则 ℓ 的估计$\hat{\ell}$ 为

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} I_{\{S(x_i)\geq\gamma\}} \frac{f(x,p)}{f(x,q)} \tag{6}$$

这样就可以将"确定性"的优化问题转变成一个相关的"随机"优化问题,并可以使用交叉熵方法来解决.使用交叉熵方法解决优化问题可以分为两个基本步骤,首先选择一个与初始概率分布函数等价的概率分布函数生成随机样本,从中选取结果比较好的样本通过一定机制更新概率分布函数[25,28].通过不断地更新概率分布函数,最终会产生"更好"的分布函数,解决系统的优化问题.

交叉熵方法是一种衡量两种概率分布相似程度的快速算法,其收敛性优于遗传算法、Pareto 优化算法等优化排序算法[30].特别是采用基于交叉熵的重要抽样方法可以大幅度地减少测试样本,提高效率.因此考虑将交叉熵方法引入到社区发现中,以提高划分的准确性.

2.3 基于交叉熵方法的社区发现算法分析

设 G 是一个由 n 个节点组成网络,用节点序列$\{y_1,y_2,\cdots,y_r\}$来表示 G 上的一个社区,设向量 $\boldsymbol{y}=\{y_1,y_2,\cdots,y_r\}$,其中 $y_i\in\{1,2,\cdots,n\}$,且对于 $i=1,2,\cdots,r\leq n$ 都满足 $y_1\neq y_2\neq\cdots\neq y_r$.用$\chi$表示所有可能存在的向量 $\boldsymbol{y}$,即χ为网络 G 上所有可能的 $\boldsymbol{y}$ 组成的向量空间.

设 S 为 $\boldsymbol{y}$ 上的一个实值评价函数

$$S(\boldsymbol{y}) = S(\{y_1,y_2,\cdots,y_r\}) \tag{7}$$

则可以在概率密度为 $f(\cdot;\boldsymbol{v})$的条件下,研究 $S(\boldsymbol{y})$比一个给定数 γ 大的概率问题.那么社区划分问题可以使用交叉熵方法按照组合优化问题来求解,即将划分社区看成如下所示的优化问题:

$$S(\boldsymbol{y}^*) = \gamma^* = \max_{x\in\chi} S(\boldsymbol{y}) \tag{8}$$

其中:S 为 Modularity 函数;γ^* 表示最优解,可通过产生的社区向量 $\boldsymbol{y}$ 计算.

为了得到网络中的一个社区,选择 $\boldsymbol{x}$ 表示一个可能存在的社区划分,即 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_r\}$.特别地,若节点 i 在第一个社区中,则 $x_i=1$,否则 $x_i=0$,即 $\boldsymbol{x}$ 可以表示为一组由 0、1 构成的向量.

这样,划分社区的问题就转变为求 $S(\boldsymbol{x})$在向量空间χ的最大值问题.在χ上定义示性函数$\{I_{\{S(x)\geq\gamma\}}\}$表示 $S(\boldsymbol{x})$值不小于阈值 γ 的解集.设$\{f(\cdot;\boldsymbol{v}),\boldsymbol{v}\in V\}$为$\chi$上的概率密度函数,$\boldsymbol{v}$ 表示概率密度函数的参数向量,则对于一组特定参数向量 $\boldsymbol{u}\in V$,式(8)的相关估计为

$$\ell(\gamma)=\boldsymbol{P}_{\boldsymbol{u}}(S(\boldsymbol{X})\geq\gamma)=\sum_{\boldsymbol{x}} I_{\{S(\boldsymbol{x})\geq\gamma\}} f(\boldsymbol{x};\boldsymbol{u})=\boldsymbol{E}_{\boldsymbol{u}} I_{\{S(\boldsymbol{X})\geq\gamma\}} \tag{9}$$

其中:$\boldsymbol{X}$ 为概率密度函数 $f(\cdot;\boldsymbol{u})$生成的随机样本;$\boldsymbol{P}_{\boldsymbol{u}}$是基于 X 的概率度量;$\boldsymbol{E}_{\boldsymbol{u}}$是期望算子.这样,当 γ 取接近于最优解的 Modularity 值时,$S(\boldsymbol{X})\geq\gamma$ 是一个小概率事件.使用 Monte Carlo 方法求解式(9)需要足够大的样本数量[30],因此可以利用式(6)得到参数估计:

$$\hat{\ell}(\gamma) = \frac{1}{N}\sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i)\geq\gamma\}} \frac{f(\boldsymbol{X}_i;\boldsymbol{u})}{g(\boldsymbol{X}_i)} \tag{10}$$

其中,$g(\boldsymbol{x})$为重要抽样密度.通过优化参数 $\boldsymbol{v}$,使得 $f(\boldsymbol{x},v)$与最优重要抽样密度 $g^*(\boldsymbol{x},\boldsymbol{v})$之间交叉熵(即 Kullback-Leibler 距离)最小,从而得到社区的最优划分.利用 Kullback-Leibler 距离公式可以推导出最优参数 $\boldsymbol{v}^*$ 的解为[30]

$$\boldsymbol{v}^* = \arg\max_{\boldsymbol{v}} \boldsymbol{E}_{\boldsymbol{u}} I_{\{S(\boldsymbol{X})\geq\gamma\}} \ln f(\boldsymbol{X};\boldsymbol{v}) \tag{11}$$

利用迭代求解,则式(11)可采用以下表述形式[30]:

$$\boldsymbol{v}_{t,j} = \frac{\boldsymbol{E}_{\boldsymbol{v}_{t-1}} I_{\{S(\boldsymbol{X})\geq\gamma_t\}} \boldsymbol{X}_j}{\boldsymbol{E}_{\boldsymbol{v}_{t-1}} I_{\{S(\boldsymbol{X})\geq\gamma_t\}}} \tag{12}$$

其中:t 表示迭代次数;j 表示参数向量中第 j 个参数.在算法实现时为了方便选取 γ_t的值,引入分位数 ρ 表示 $S(\boldsymbol{x})$不小于阈值 γ_t所占的比例,那么 γ_t的估计值 $\hat{\gamma}_t$为

$$\hat{\gamma}_t = S_{\lceil(1-\rho)N\rceil} \tag{13}$$

这里,$S_1,S_2,\cdots,S_N$是 $S(\boldsymbol{X}_1),S(\boldsymbol{X}_2),\cdots,S(\boldsymbol{X}_N)$按递增方式排序的序列,$\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\in\chi$ 为产生的社区样本,N 是样本数量.因此$\{S(\boldsymbol{x})\geq\hat{\gamma}_t\}$可以表示为$\{S_{\lceil(1-\rho)N\rceil},S_{\lceil(1-\rho)N\rceil+1},\cdots,S_N\}$.

由此可以得到满足式(8)的一组$\{x_1,x_2,\cdots,x_r\}$,则$\{x_1,x_2,\cdots,x_r\}$表示的节点为所求的 Q 值最大的一组样本,即为网络 G 最好的一种社区划分方式.

2.4　算法实现

通常，网络的社区会多于两个，因此在划分社区时使用二分迭代法. 即先将整个网络划分成两个社区，再分别将每一个社区继续划分为两个更小的社区，不断迭代操作直到 Modularity 不再增加为止. 算法的具体描述如下：

(1) 初始化概率密度函数参数，这里使用向量 $\boldsymbol{v}=\{x_1,x_2,\cdots,x_n\}$ 表示节点分布不同社区的概率. 其中，x_i 为节点 i 在第 1 个社区中的概率，$1-x_i$ 为节点 i 在第 2 个社区中的概率. 定义初始化概率向量 $\boldsymbol{v}_0=\{1,1/2,1/2,\cdots,1/2\}$，并设置迭代次数 $t=1$；

(2) 使用概率向量 $\boldsymbol{v}$ 生成随机样本 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N$，采用式(1)计算样本所对应社区的 Modularity 值；

(3) 通过式(13)计算 $\hat{\gamma}_t$，并取 $\lceil(1-\rho)N\rceil$ 个重要抽样，其中 $(1-\rho)$ 是分位点，即从原有的 N 个采样中选择使得 Modularity 值较大的 $\lceil(1-\rho)N\rceil$ 个重要样本，通过式(12)更新概率向量 $\boldsymbol{v}_t$；

(4) 如果 t 满足

$$\hat{\gamma}_t=\hat{\gamma}_{t-1}=\cdots=\hat{\gamma}_{t-d}$$

则结束，否则 $t=t+1$，返回第 2 步. 其中 d 是控制最优值逼近的迭代次数，即停止位.

在上述算法执行完毕后，可以得到一个近似最优的解，即找到一个 Modularity 近似最大的社区划分. 本文提出的交叉熵算法使用二分迭代法，理论上当真实社区个数为偶数个时，划分的社区结构相对准确，但在划分个数为奇数时，存在潜在的问题. 因此，本文在二分迭代法的基础上增加了调优操作，即如果某些节点明显不属于某个社区，那么将这些节点从一个社区移动到另一个社区中，如果 Modularity 值增加，则接受此移动操作，否则不接受，提高了社区划分的准确性，解决了社区个数为奇数时存在的问题.

3　实验结果与分析

为了检验本文提出的基于交叉熵的社区发现算法的有效性，将此算法应用到计算机生成网络和 Zachary 俱乐部网络、Dolphins 社会关系网络等实际网络上进行社区划分，并将实验结果同 Girvan 和 Newman 在文献[21]中提出的 Girvan-Newman 算法以及文献[23]中的极值优化算法进行比较.

在基于交叉熵的社区发现算法中，参数 ρ 和 d 的选择会影响算法的执行效率. 经过大量实验得出，$\rho=0.4$ 和 $d=7$ 时算法执行效果较好.

3.1　计算机生成网络

本节采用划分正确节点所占的比例和规范化交互信息对比不同算法找到的社区结构与标准社区结构的相似程度，进而评价社区划分的准确度. 划分正确节点所占的比例和 NMI 越大，则社区划分结果越准确.

如图 1 所示，实验中选择的计算机生成网络由 $n=128$ 个节点构成，分为 4 个社区，每个社区 32 个节点. 每个节点有平均 Z_{in} 个边连接社区内的节点，有平均 Z_{out} 个边连接其他社区的节点，并且满足 $Z_{\text{in}}+Z_{\text{out}}=16$.

本文使用上述网络生成原则创建了 20 个随机网络，并分别使用基于交叉熵的社区发现算法、Girvan-Newman 算法和极值优化算法对这 20 个计算机生成网络进行社区划分，计算相应的划分正确节点所占的比例，其平均值如图 1 所示.

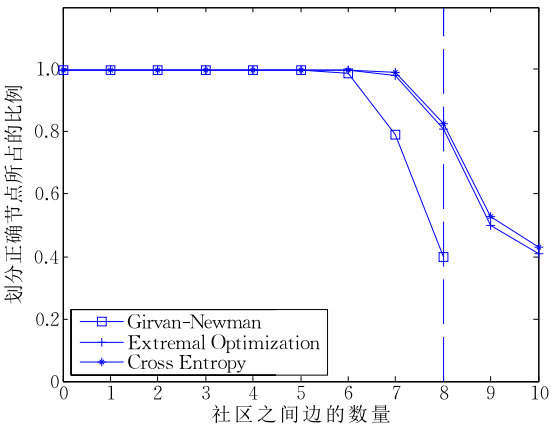

图 1　计算机生成网络社区划分正确节点所占比例($n=128$)

在图 1 中，当 $Z_{\text{out}}<5$ 时，3 种方法都可以准确地划分社区. 随着社区间连边的增加，不同的算法划分正确节点所占比例开始出现差别. 从图 1 中可以看出本文提出的交叉熵算法和极值优化算法都明显优于Girvan-Newman 算法，Girvan-Newman 算法在 $Z_{\text{out}}=8$ 时的划分正确节点所占比例只有约 40%，而交叉熵算法和极值优化算法的划分正确节点所占比例都超过 80%. 在 $Z_{\text{out}}=9$ 时，极值优化算法的划分正确节点所占比例为 50%，而交叉熵算法的划分正确节点所占比例为 53%；在 $Z_{\text{out}}=10$ 时，极值优化算法的划分正确节点所占比例为 41%，而交叉熵算法为 43%.

图 2 为 $n=160$ 个节点的计算机生成网络中社区划分的划分正确节点所占的比例，其中社区数量为 4，每个社区包含 40 个节点，$Z_{\text{in}}+Z_{\text{out}}=20$.

以上述两种计算机生成网络为基础，应用式(14)

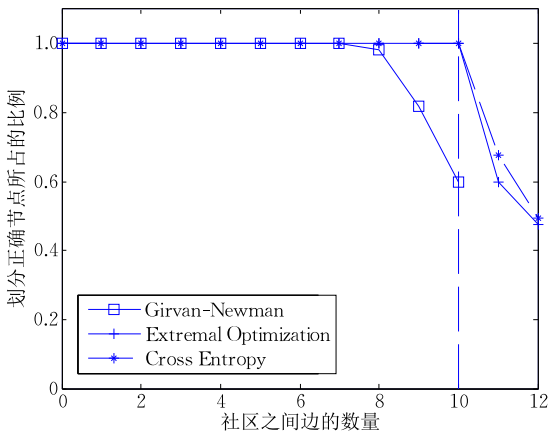

图 2 计算机生成网络社区划分正确节点所占比例($n=160$)

计算 20 个随机网络可以达到的最大规范化交互信息[26-27],进一步比较 3 种社区划分算法的准确度,结果如表 1 所示.

$$NMI=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}M_{ij}\log(M_{ij}n/M_{i.}M_{.j})}{\sum_{i=1}^{C_A}M_{i.}\log(M_{i.}/n)+\sum_{j=1}^{C_B}M_{.j}\log(M_{.j}/n)} \tag{14}$$

其中 $\boldsymbol{M}$ 为含混矩阵;元素 M_{ij} 表示第 i 个真实社区与第 j 个划分的社区都存在的节点数量;n 为网络规模;C_A 是真实的社区结构数量;C_B 是划分的社区结构数量;矩阵 $\boldsymbol{M}_{ij}$ 中第 i 行的和是 $M_{i.}$,第 j 列的和是 $M_{.j}$.

表 1 计算机生成网络社区划分的 *NMI* 值

规模	$n=128$				$n=160$			
Z_{out}	5	6	7	8	7	8	9	10
NMI_{GN}	1.0	0.97	0.82	0.55	1.0	1.0	0.84	0.46
NMI_{EO}	1.0	1.00	0.95	0.63	1.0	1.0	1.00	1.00
NMI_{CE}	1.0	1.00	0.96	0.62	1.0	1.0	1.00	1.00

从表 1 可知,在平均 Z_{out} 较小时,社区结构明显,3 种社区划分算法的 *NMI* 值均为 1.0. 随着 Z_{out} 增大,网络结构更趋向于随机网络,社区结构不明显,*NMI* 值较小. 基于交叉熵的社区发现算法的 *NMI* 值明显高于 Girvan-Newman 算法的 *NMI* 值,且与基于极值优化算法的 *NMI* 值近似相同. 综合分析 3 种社区划分算法中划分正确节点所占的比例和 *NMI* 值可知,针对计算机生成网络的社区划分,本文提出的交叉熵算法与 Girvan-Newman 算法相比,在准确度方面得到了明显的改善,与极值优化算法相比,也有一定的提高.

3.2 实际网络

测试用的实际网络包括:Zachary 俱乐部网络[31]、Dolphins 社会关系网络[32]、Les Miserables 作家关系网络[33]、爵士乐演奏家网络[34]、C. elegans 新陈代谢网络[35]等,实验结果如表 2 所示.

表 2 中,网络规模指的是网络的节点数,Q_{GN}、Q_{EO} 和 Q_{CE} 分别为 Girvan-Newman 算法、极值优化算法和交叉熵算法分析 5 种实际网络的 Modularity 值,C_{GN}、C_{EO} 和 C_{CE} 为 3 种算法划分得到的社区的数量.

表 2 3 种算法划分实际网络的最大 Modularity 值

网络	规模	Q_{GN}	Q_{EO}	Q_{CE}	C_{GN}	C_{EO}	C_{CE}
Zachary	34	0.4013	0.4188	0.4198	5	4	4
Dolphins	62	0.5194	0.5264	0.5285	5	4	5
Les Miserables	77	0.5381	0.5563	0.5600	11	6	6
Jazz	198	0.4379	0.4452	0.4452	4	5	4
C. elegans	453	0.4001	0.4342	0.4418	10	12	12

从表 2 可以看出,对于 5 个实际网络的 Modularity 值,交叉熵算法均大于 Girvan-Newman 算法,而对 Zachary 网络、Dolphins 网络、Les Miserables 网络和 C. elegans 网络的社区划分中,交叉熵算法也优于极值优化算法,仅在 Jazz 网络划分中,两种算法得到的 Modularity 值相同. 也就是说交叉熵算法得到的社区内部的连接性强,更能够体现社区的本质. 此外,通过将文献[31-35]中的真实网络结构与 3 种算法的划分结果相比较,可以发现交叉熵算法明显好于 Girvan-Newman 算法. Girvan-Newman 算法得到的社区划分中有一些社区只有很少的节点,如 Girvan-Newman 算法用于 Les Miserables 网络的社区划分时会得到包含 1 个节点和 2 个节点组成的社区,这样导致算法划分结果较差,与真实的社区结构相差较多.

分析 Girvan-Newman 算法和极值优化算法可知,两者的时间复杂度分别为 $O(n^3)$ 和 $O(n\ln n)$. 与上述两种算法相比,基于交叉熵的社区发现算法采用重要采样技术,收敛速度很快. 因此,基于交叉熵的社区发现算法能够准确划分复杂网络中社区结构的同时,使得寻找社区结构的效率也有所提高.

下面具体比较 3 种算法划分 Zachary 网络和 Dolphins 社会关系网络的划分结果.

(1) Zachary 俱乐部网络

Zachary 社会关系网包含了 34 个节点和 78 条边,是复杂网络与社区分析研究领域中经常选用的一个小型测试网络. 该网络是 20 世纪 70 年代初,Wayne Zachary 研究一所美国大学的空手道俱乐部的社会关系时建立的. 网络中的节点表示一个俱乐部成员,节点间的连接表示两个成员经常一起出现

在俱乐部活动之外的其他场合，即在俱乐部活动之外，他们有相对密切的联系和来往. Wayne Zachary 利用俱乐部成员的内部关系和外部关系构建了社会网络. 但是，在他学习期间俱乐部管理员和老师之间因为俱乐部费用问题将俱乐部一分为二，成为了两个小的俱乐部[31].

使用 Girvan-Newman 算法、极值优化算法和本文提出的交叉熵算法分析 Zachary 网络的实验结果分别如图 3～图 5 所示. 图中在相同直线分割区域内的节点属于同一社区.

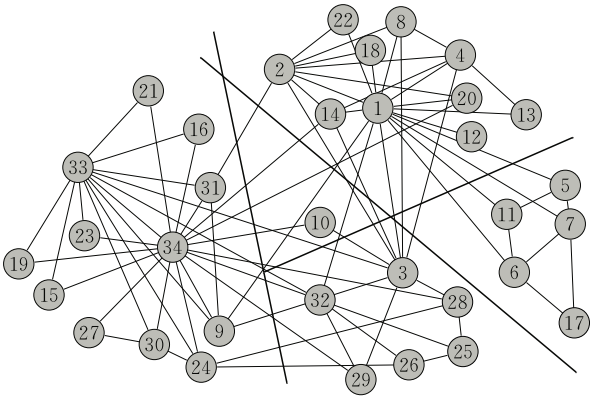

图 3　Girvan-Newman 算法的 Zachary 网络社区划分结果

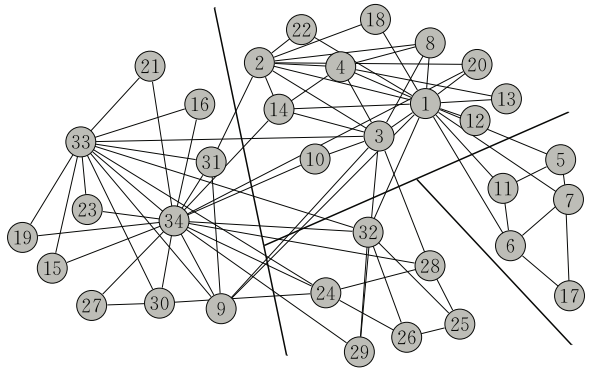

图 4　极值优化算法的 Zachary 网络社区划分结果

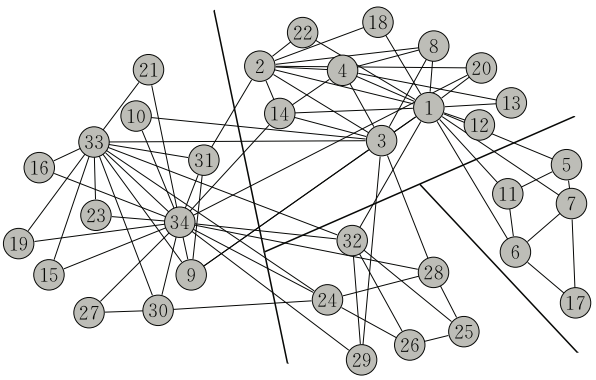

图 5　交叉熵算法的 Zachary 网络社区划分结果

如图 3 所示，Girvan-Newman 算法将 Zachary 网络划分为 5 个社区，Q 值为 0.4013. 在真实的 Zachary 网络社区中，节点 10 与节点 34、30、24 等组成一个社区，节点 1、2、12 等组成了另外一个社区. 其中，节点 10 与两个社区的连边一样多，Girvan-Newman 算法错误地将其划分为一个单独的社区，导致 Modularity 值较低. 极值优化算法和交叉熵算法都将网络划分为 4 个社区，虽然与网络真实的社区结构相比，极值优化算法和交叉熵算法得到了更多的社区，但是得到 Modularity 值也要更大.

交叉熵算法划分该网络的 Q 值为 0.4198(图 5)，略大于极值优化算法的 0.4188(图 4). 特别的是，极值优化算法将节点 10 划分到了中心节点 1 所在的社区，而交叉熵算法则避免了这个错误，因此比极值优化算法得到了更好的社区划分和更大的 Modularity 值.

(2) Dolphins 社会关系网络

Dolphins 社会关系网络也是社区划分中经常选用的一个真实网络，是 Lusseau 等人通过对海豚之间接触频率进行 7 年的观察和研究构建的. Dolphins 社会关系网络是由生活在新西兰 Doubtful Sound 峡湾的 62 只宽吻海豚构成，网络包含 62 个节点和 159 条边[32]. 其中节点代表海豚，连边代表海豚间的接触. 研究发现，Doubtful Sound 峡湾的海豚属于不同家族，较大的海豚家族包含 42 个成员，而较小的家族仅包含 20 只海豚.

使用交叉熵算法、Girvan-Newman 算法和极值优化算法分析 Dolphins 社会网络的实验结果分别如图 6～图 8 所示. 图中相同直线分割区域内节点属于同一社区.

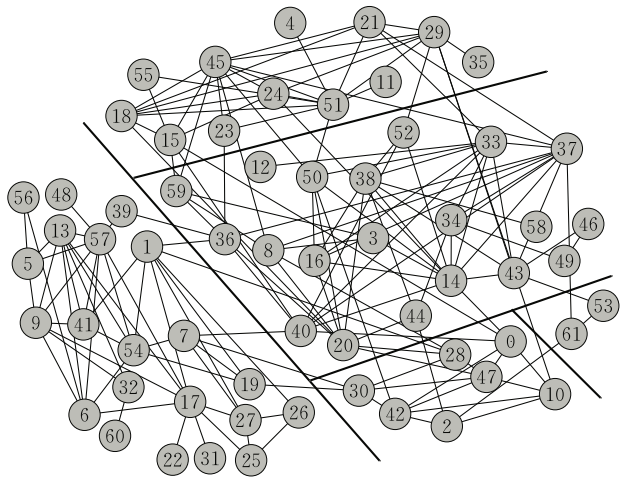

图 6　Girvan-Newman 算法的 Dolphins 网络社区划分结果

如图 6～图 8 所示，3 种算法都将 Dolphins 社会关系网络划分成了较小粒度的网络，社区数量都要比实际网络的多. 在真实的 Dolphins 社会关系网络中，节点 53、61 与节点 58、43、34、33 等同属于

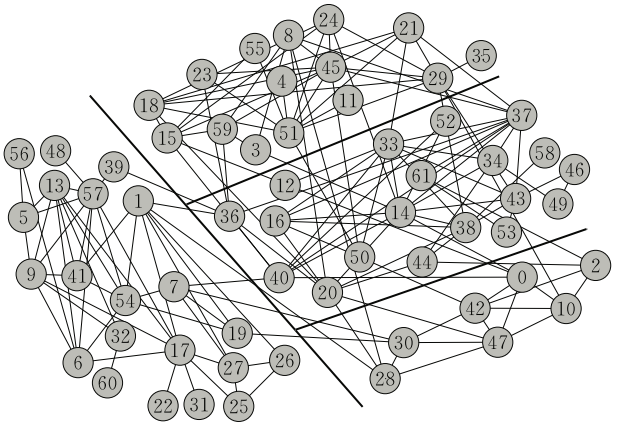

图 7 极值优化算法的 Dolphins 网络社区划分结果

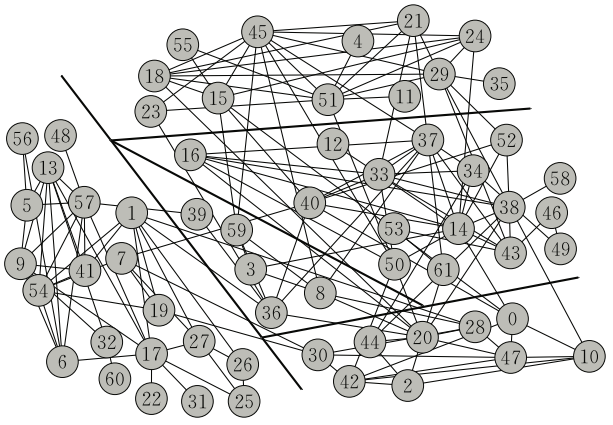

图 8 交叉熵算法的 Dolphins 网络社区划分结果

1 个社区. Girvan-Newman 算法将网络划分为 5 个社区，Q 值为 0.5194，且该算法将节点 53 和 61 单独划分成了 1 个社区. 而极值优化算法和交叉熵算法都修正了 Girvan-Newman 算法中划分错误的 53 和 61 两个节点，得到了更好的划分结果. 极值优化算法将网络划分为 4 个社区，Q 值为 0.5264. 交叉熵算法将 Girvan-Newman 算法和极值优化算法社区划分有歧义的 5 个节点 3、8、36、39、59 划分成 1 个社区，提高了 Modularity 值，为 0.5285.

4 结 论

社区结构是复杂网络的一个极其重要的特性，在复杂网络中搜寻和发现社区结构具有重要的应用价值，在生物学、计算机科学和社会学等多个研究领域均具有重要意义.

近年来，针对不同类型的大规模复杂网络，人们提出了很多发现社区结构的算法. 本文提出了一种基于交叉熵的社区发现算法，算法利用对社区 Modularity 值的组合优化进行社区划分. 该算法不仅无须事先指定社区个数等算法参数，而且由于采用重要抽样方法更新概率密度分布函数，因而具有收敛速度快的优点.

通过对计算机生成网络以及 Zachary 俱乐部网络、Dolphins 社会关系网等 5 个实际网络进行划分的实验结果显示，基于交叉熵的社区发现算法在社区划分准确率上明显好于 Girvan-Newman 算法和极值优化算法，且可以得到 3 个算法中最大 Modularity 值. 此外，该算法在一定程度上可以消除社区结构及社区间具有不确定性的重叠节点现象，进一步提高了社区划分的准确率.

参 考 文 献

[1] Strogatz S H. Exploring complex networks. Nature, 2001, 410(6825): 268-276

[2] Kim H J. Analysis of a complex network of physics concepts. Modern Physics Letters B, 2012, 26(28): 1250186(9pp.)

[3] Tam W M, Lau F C M, Tse C K. Complex-network modeling of a call network. IEEE Transactions on Circuits and Systems I, 2009, 56(2): 416-429

[4] Newman M E J. The structure and function of complex networks. SIAM Review, 2003, 45(2): 167-256

[5] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology. Computer Communications Review, 1999, 29(4): 251-262

[6] Albert R, Jeong H, Barabasi A L. Internet-diameter of the World-Wide Web. Nature, 1999, 401(6749): 130-131

[7] Barabasi A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509-512

[8] Albert R, Barabasi A L. Topology of evolving networks: Local events and universality. Physical Review Letters, 2000, 85(24): 5234-5237

[9] Barabasi A L, Albert R, Jeong H. Scale-free characteristics of random networks: The topology of the world-wide web. Physica A, 2000, 281: 69-77

[10] Chen Guan-Rong, Wang Xiao-Fan, Li Xiang. Introduction to Complex Networks: Models, Structures and Dynamics. Beijing: China Higher Education Press, 2012

[11] Estrada E. Community detection based on network communicability. Chaos, 2011, 21: 016103(7pp.)

[12] Newman M. Detecting community structure in networks. European Physical Journal B, 2004, 38(2): 321-330

[13] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006, 74(3): 036104

[14] Fortunato S. Community detection in graphs. Physics Reports, 2010, 486(3-5): 75-174

[15] Guimera R, Amaral L A N. Functional cartography of complex metabolic networks. Nature, 2005, 433(7028): 895-900

[16] Palla G, Der'enyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814-818

[17] Flake G W, Lawrence S, Giles C L, Coetzee F M. Self-organization and identification of web communities. Computer, 2002, 35(3): 66

[18] Newman M E J. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582

[19] Fortunato S, Barthelemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(1): 36-41

[20] Arenas A, Duch J, Fernandez A, Gomez S. Size reduction of complex networks preserving modularity. New Journal of Physics, 2007, 9: 176

[21] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113

[22] Medus A, Acuna G, Dorso C O. Detection of community structures in networks via global optimization. Physica A, 2005, 358(2-4): 593-604

[23] Duch J, Arenas A. Community detection in complex networks using extremal optimization. Physical Review E, 2005, 72(2): 027104

[24] Xu G, Tsoka S, Papageorgiou L G. Finding community structures in complex networks using mixed integer optimisation. European Physical Journal B, 2007, 60(2): 231-239

[25] Morarescu I C, Girard A. Opinion dynamics with decaying confidence: Application to community detection in graphs. IEEE Transactions on Automatic Control, 2011, 56(8): 1862-1873

[26] Danon L, Diaz-Guilera A, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(09), P09008

[27] Huang Jian-Bin, Sun He-Li, Song Qin-Bao, et al. Revealing density-based clustering structure from the core-connected tree of a network. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(8): 1876 -1889

[28] Kullback S, Leibler R A. On information and sufficiency. The Annals of Mathematical Statistics, 1951, 22(1): 79-86

[29] Rubinstein R Y. The simulated entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability, 1999, 2: 127-190

[30] De Boer P T, Kroese D P, Mannor S, Rubinstein R Y. A tutorial on the cross-entropy method. Annals of Operations Research, 2005, 134(1): 19-67

[31] Zachary W. Information-flow model for conflict and fission in small-groups. Journal of Anthropological Research, 1997, 33: 452

[32] Lusseau D, Schneider K, Boisseau O J, Haase P, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations—can geographic isolation explain this unique trait? Behavioral Ecology and Sociobiology, 2003, 54: 396-405

[33] Knuth D E. The Stanford GraphBase: A Platform for Combinatorial Computing. Boston: Addison-Wesley, 1993

[34] Gleiser P M, Danon L. Community structure in jazz. Advances in Complex Systems, 2003, 6(4): 565-573

[35] Jeong H, Tombor B, Albert R, et al. The large-scale organization of metabolic networks. Nature, 2000, 407(6804): 651-654

YU Hai, born in 1971, Ph. D., associate professor. His research interests include chaos, complex network theories and applications.

ZHAO Yu-Li, born in 1985, Ph. D. Her research interests include complex network, channel coding.

CUI Kun, born in 1986, M. S. His research interests include complex network, computer algorithm.

ZHU Zhi-Liang, born in 1962, Ph. D., professor. His research interests include chaos, complex network, software engineering.

Background

The problem of community structure detection in complex networks has been intensively investigated in recent years. It has been proven that nodes in the same community also have some similar properties. Consequently, community detection, as a major topic in Complex Networks, has important reference and application value. Generally, whether a community detection algorithm is designed well or not is evaluated by modularity. In recent years, some community detection schemes based on combinatorial optimization algorithm and modularity have been published, but they are either irreversibility or lack of accuracy.

This paper proposes new community detection algorithm based on Cross-Entropy from signal processing. The algorithm defines modularity as the quality function. It uses importance sampling in Cross-Entropy to speed up the convergence, thus the efficiency and accuracy of community detection can be improved simultaneously. Moreover, variables such as number of communities do not need to be defined in advance. Some theoretical analyses illustrate it is an effective algorithm to detect communities. Besides, simulation results also show that this algorithm is more accurate than Girvan-Newman and External Optimization ones.

This research was supported by the National Natural Science Foundation of China (Grant Nos. 61374178 and 61402092), the Liaoning Provincial Natural Science Foundation of China (Grant No. 201202076), the Fundamental Research Funds for the Central Universities (Grant Nos. N130417004 and N130317001).