doi: 10.3969/j.issn.0490-6756.2019.06.010

基于 DCBM 的马尔可夫谱聚类社区发现算法

任淑霞, 张书博, 吴 涛
(天津工业大学 计算机科学与软件学院, 天津 300387)

摘 要: 谱聚类划分算法是经典社区发现算法之一,由于目前构造的相似图承载的社区结构信息较少,导致聚类效果与理想效果具有较大差距,因此,提出了基于 DCBM 的马尔可夫谱聚类社区发现算法 MSCD. 首先,基于 DCBM 模型提出了以节点间连接概率为元素的概率矩阵,并建立了概率矩阵与相似矩阵之间的映射关系;其次,利用马尔可夫链重构了谱聚类的相似图;最后,使用重构的相似图对网络进行社区划分. 在人工合成网络和真实网络上与 SC,MRW-KNN 和 FluidC 三种典型算法进行了对比实验. 实验结果表明,MSCD 算法具有更加高效的聚类性能,能够揭示更加清晰的社区结构.

关键词: DCBM; 马尔可夫链; 概率矩阵; 谱聚类; 社区发现; 复杂网络
中图分类号: TP393.02 **文献标识码**: A **文章编号**: 0490-6756(2019)06-1049-08

Markov spectral clustering algorithm with DCBM for community detection

REN Shu-Xia, *ZHANG Shu-Bo*, *WU Tao*
(Department of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China)

Abstract: Spectral clustering algorithm is one of the classical community detection algorithms. Due to the current constructed similarity graphs carry less community structure information, the actual clustering effect has a big gap with the ideal clustering effect. Therefore, based on degree corrected stochastic block model and Markov chain, a novel spectral clustering approach for community detection, called MSCD, is proposed. Firstly, probability matrix composed of the connection probability between nodes is introduced based on DCBM, and the mapping relationship is established between probability matrix and similar matrix. Then, Markov chain is utilized to reconstruct the similar graph of spectral clustering. Finally, the reconstructed similar graph is used to partition the networks into clusters. Three typical algorithms of SC, MRW-KNN and FluidC are performed on synthetic networks and real networks. Comparative experiments show that the MSCD algorithm has more efficient clustering performance and can reveal a clearer community.

Keywords: DCBM; Markov chain; Probability matrix; Spectral clustering; Community detection; Complex network

1 引 言

现实世界中的众多复杂系统都可以抽象成复杂网络[1]而存在,比如万维网,社交网络以及生物网络,而复杂网络中的社区结构往往包含重要的后验信息,对认清节点间的关系、消息传递、相似推荐等具有重要的研究意义. 近年来,涌现出众多社区发现方法[2-3],经典算法有谱聚类算法[4],GN 算

收稿日期: 2018-10-01
基金项目: 国家自然科学基金(61403278)
作者简介: 任淑霞(1973-), 女, 山东梁山人, 副教授, 博士生, 研究领域为数据挖掘,大数据分析等. E-mail:t_rsx@126.com

法[5]和标签传播算法[6]等. 现代人们的生活已经离不开社会网络,为了深入理解社会网络的结构和功能,如何发现和分析社区结构显得至关重要.

在社区发现算法中,谱聚类算法是经典社区发现算法之一,谱聚类演化于图论,后因其优秀的性能被广泛应用于聚类中. 谱聚类社区发现算法相比传统聚类算法主要优点是处理稀疏数据的聚类非常有效,处理高维数据聚类的复杂度相对较低. 因此,许多学者开始研究谱聚类社区发现算法,并在经典谱聚类算法上进行拓展. Jin[7]提出基于特征向量比值的谱聚类社区发现算法,使用第一主导特征向量与其他各主导特征向量之间的比值进行聚类;Gulikers 等人[8]提出基于适当归一化邻接矩阵的谱聚类算法,并且可应用于度修正的随机块模型,且算法不依赖于参数作为输入.

然而,谱聚类存在不容忽视的缺点,主要表现在相似图的构造上,Luxburg 等人[9]对 ε-邻近法,K 邻近法以及全连接法三种主流的相似图构造方法进行了分析,并建议使用 K−NN 构建相似图,但这三种方法都是根据节点间的欧式距离构造相似图,节点间的直接相似性有时并不可靠. 如果相似矩阵能够更接近理想矩阵,谱聚类将会发挥更好的聚类性能. 因此,构造更接近理想矩阵的相似矩阵对于谱聚类社区划分的成功是至关重要的.

针对上述缺点,一些学者开始研究如何构造更理想的相似图. Yan 等人[10]针对传统谱聚类算法的高斯核参数 σ 敏感问题,设计了无高斯核函数参数和无最短路径相似性度量方法,降低了同一密度区域内数据对之间的相似度. Zang 等人[11]在密度敏感相似性测度的启发下,提出了一种改进的谱聚类方法,利用基于数据密度的相似性度量结合 DNA 遗传算法,以找到复杂数据的空间分布特征. Zhang 等人[12]提出一种基于随机游走模型处理高斯核相似矩阵的谱聚类算法,其利用高斯核函数计算转移概率,并由 m 步转移概率矩阵构建相似矩阵. 曹等人[13]提出一种基于 Markov 随机游走模型的稀疏相似图构造算法(MRW-KNN),其节点间的相似度同样采用高斯核函数计算,而将转移概率作为节点间是否具有相似度的阈值.

虽然已经提出很多基于优化谱聚类相似图的社区发现算法[10-13],但如何构建更健壮的相似图,提高社区划分质量一直是悬而未决的问题. 因此,本文提出一种基于马尔可夫的谱聚类社区发现方法,其主要思想是以度纠正随机块模型(DCBM,Degree-corrected stochastic block model)[14]推导为理论基础,利用马尔可夫链构建谱聚类的相似图,再使用谱聚类进行社区划分.

2 基于马尔可夫链的相似图构造

2.1 度纠正随机块模型 DCBM

随机块模型(BM)[15]是一个经典的网络生成模型,分两步生成网络:先给网络中的所有节点分配隶属社区,然后根据社区之间的概率决定两节点之间是否存在链接.

DCBM[14]是基于 BM 的改进模型,也是一个网络生成模型. 在 BM 中,同一社区的节点表现出相同的行为,但在现实网络中,同一个社区内有的节点比较活跃,有的节点比较懒散,所以,同社区的每个节点的表现行为是不同的. 为了弥补 BM 的缺点,DCBM 用一个参数 θ(简称异质性)来刻画每一个节点的个性,表示每个节点的不同活跃程度. 因此,DCBM 大致可描述为

$$P(A(i,j)=1)=\theta(i)B(c_i,c_j)\theta(j) \tag{1}$$

其中,矩阵 $\boldsymbol{A}$ 是无向网络的邻接矩阵,$A(i,j)$ 表示矩阵 $\boldsymbol{A}$ 的第 i 行第 j 列元素,$A(i,j)\in\{0,1\}$;$\theta(i)$ 表示节点 i 的异质性;$\boldsymbol{B}$ 是 $K\times K$ 的矩阵,刻画的是社区之间的节点连接概率,$\boldsymbol{K}$ 表示网络中的社区数量,c_i 表示节点 i 所属的社区号,$B(c_i,c_j)$ 表示社区 c_i 的节点与社区 c_j 的节点之间连接的概率,$B(c_i,c_j)=\dfrac{m_{c_ic_j}}{n_{c_i}n_{c_j}}$,$m_{c_ic_j}$ 表示社区 c_i 与社区 c_j 之间的连接边数量,n_{c_i} 表示社区 c_i 的节点数量.

由式(1)可知,节点之间相连接的概率不仅受社区间连接概率 $B(c_i,c_j)$ 的影响,还与节点 i,j 自身的异质性 $\theta(i),\theta(j)$ 有关. 虽然 DCBM 有大量的参数存在,但其在 BM 基础上,通过增加节点的异质性 θ,使模型更加符合现实网络,更能反映真实情况.

2.2 节点间连接概率及其概率矩阵

DCBM 模型在 MSCD 算法中起到至关重要的作用,本节使用 DCBM 模型推导得到节点间连接概率包含多层复杂的社区结构信息,提出了以节点间连接概率为元素的概率矩阵,并揭示了概率矩阵与相似矩阵之间的映射关系.

2.2.1 节点间连接概率包含多层复杂社区信息

假设网络 $N=(V,E)$ 存在 n 个顶点,K 个社区,每对节点之间边的数量服从泊松分布,B 是 $K*K$ 对称矩阵,用来描述社区之间的连接性,其中的元

素 b_{rs} 是社区 r 与社区 s 的连接概率，g_i 是节点 i 的社区号.

根据 DCBM 模型的定义可得，邻接矩阵元素 A_{ij} 的期望值为 $\theta_i\theta_j b_{g_ig_j}$，生成网络 N 的概率为

$$P(N|\theta,b,g)=\prod_{i<j}\frac{(\theta_i\theta_j b_{g_ig_j})A_{ij}}{A_{ij}!}\exp(-\theta_i\theta_j b_{g_ig_j})\times\prod_{i<j}\frac{(\frac{1}{2}\theta_i{}^2 b_{g_ig_j})A_{ij}/2}{(A_{ij}/2)!}\exp(-\frac{1}{2}\theta_i{}^2 b_{g_ig_j}) \quad (2)$$

其中，$P(N|\theta,b,g)$ 是通过节点间边的数量服从现实概率分布计算的，是指生成模拟现实网络 N 的概率，根据文献[14]的推导，概率 $P(N|\theta,b,g)$ 的非标准化的对数似然函数可简化为

$$L(N|g)=\sum_{rs}m_{rs}\log\frac{m_{rs}}{k_rk_s} \quad (3)$$

其中，m_{rs} 表示社区 r 和社区 s 之间连接边的数量，如果 r 等于 s，m_{rs} 等于社区 r 或社区 s 内连接边数的 2 倍，k_r 表示社区 r 内部节点的度之和.

根据文献[14]可知，通过迭代地最大化公式(3)，可得到网络社区结构的大致情况. 首先，将网络随机分成 K 个初始的社区节点集合；然后，重复地将一个顶点从一个组移动到另一个组，选择最能增加目标函数 $L(N|g)$ 的移动，每个顶点只能移动一次；当所有顶点都已移动完时，检查整个过程从开始到结束所经过的所有状态，选择具有最大函数值的状态，并将此状态作为下一迭代的起始点. 反复执行上述迭代过程，当完成这样的迭代而函数值没有任何的增加时，算法结束.

然而，由于多次迭代造成较大的计算量以及参数 θ 的不确定性，上述算法的划分效果不尽人意. 但是，根据式(3)的参数 m_{rs}、k_r 和上述算法可知，节点间连接概率蕴含着社区间连边数，社区内节点的度以及社区结构等多层复杂社区信息. 这里所提的节点间连接的概率是式(2)中 $P(N|\theta,b,g)$ 的因数，指的是节点间存在连接边以及边的数量的概率.

2.2.2　概率矩阵及其与相似矩阵的映射关系　节点间连接概率拥有多层复杂的社区信息，基于节点间连接概率构建的矩阵同样蕴含着多元复杂的网络社区信息. 因此，提出以节点间连接概率为元素的概率矩阵.

定义 1　概率矩阵. 给定网络 N，节点数量为 n，以网络节点间连接概率为元素组成的矩阵称为概率矩阵，记作 P，可定义为 $P=(P_{ij})$，其中，P_{ij} 表示节点 i 与节点 j 之间连接的概率，$i,j\in\{0,1,\cdots,n-1\}$.

P_{ij} 又可看作节点 i 与节点 j 之间连边的权重值 w_{ij}，而相似矩阵 $\boldsymbol{S}$ 表示网络中节点的相似度，S_{ij} 也可看作连边的权重值 w_{ij}，综上所述，概率矩阵 P 可近似为相似矩阵 $\boldsymbol{S}$。

$$s_{ij}=w_{ij}=p_{ij},\boldsymbol{S}=\boldsymbol{W}=\boldsymbol{P} \quad (4)$$

其中，W 表示由节点间相似度构建的权重矩阵.

2.3　利用马尔可夫链计算概率矩阵 P

本节基于马尔可夫链计算概率矩阵 $\boldsymbol{P}$，首先，通过转移步长确定节点间是否存在连接，其次，建立转移概率与节点间的连接概率之间的关系，最后，利用转移概率矩阵计算概率矩阵 $\boldsymbol{P}$.

2.3.1　马尔可夫链　离散时间马尔可夫链由具有马尔可夫性质的序列随机变量 $X_1,X_2,X_3,\cdots$ 构成，描述的是一种状态序列. 假设在网络 N 上随机漫步，$\boldsymbol{A}$ 是 N 的邻接矩阵，V 是 N 的节点集，变量 $X=\{X_t=i\}$ 表示行走 t 步之后，变量 X 的节点位置；$P\{X_t=i\}$ 表示在 t 步之后，到达节点 i 的概率. 对于 $i_t\in V$，$P\{X_t=i_t\mid X_0=i_0,X_1=i_1,\cdots,X_{t-1}=i_{t-1}\}=P\{X_t=i_t\mid X_{t-1}=i_{t-1}\}$，也就是说，移动到下一状态的概率仅由当前状态决定，具有序列依赖性.

2.3.2　转移概率及转移概率矩阵　转移概率是指从一个随机状态出发，经过数次的转移，出现 1,2,…,m 状态的概率，设 X_t 为离散时间马尔可夫链，对任意 $k\geqslant 0$，$l\geqslant 1$，$i,j\in V$，$pr_{ij}(k,k+l)=P\{X_{k+l}=j\mid X_k=i\}$，该式表示马尔可夫链于状态 k 时位置在 i，移动 l 步后到达 j 的转移概率，简称 l 步转移概率，Pr_l 表示与之相应的 l 步转移概率矩阵. 那么，一步转移概率为 $pr_{ij}(k,k+1)$(简写为 pr_{ij})，与之对应的一步转移概率矩阵为 Pr，显然，

$$\boldsymbol{Pr}_l=\boldsymbol{Pr}^l,\ \boldsymbol{Pr}=\boldsymbol{D_A}^{-1}\boldsymbol{A} \quad (5)$$

其中，$\boldsymbol{Pr}^1$ 表示一步转移概率矩阵 $\boldsymbol{Pr}$ 的 l 次幂，$\boldsymbol{Pr}$ 自乘 l 次，$\boldsymbol{D}_A=\mathrm{diag}(d_{A_0},d_{A_1},\ldots,d_{A_{n-1}})$，$d_{A_i}=\sum_{j=0}^{n-1}A_{ij}$.

2.3.3　计算概率矩阵 P　通过马尔可夫链的多步转移概率计算概率矩阵 P，利用转移步长来判定节点间是否存在连接，即若节点 i 在规定的转移步长内无法到达节点 j 则视为节点 i 与节点 j 不存在连接；反之，则视为节点 i 与节点 j 存在连接. 将多步转移概率代入概率矩阵的定义中，P_{ij} 可重写为

$$P_{ij}=\begin{cases}0, \text{ 节点 } i \text{ 与节点 } j \text{ 不存在连接}\\ pr_{ij}(k,k+t), \text{ 节点 } i \text{ 与节点 } j \text{ 存在连接}\end{cases} \tag{6}$$

其中,k 表示节点 i 的初始状态;t 表示规定的转移步长,称为马尔可夫链的时间尺度. 根据式(5)和式(6),概率矩阵 P 可用多步转移概率矩阵表示

$$\boldsymbol{P}=\boldsymbol{P}\boldsymbol{r}_t=\boldsymbol{P}\boldsymbol{r}^t \tag{7}$$

其中,Pr 为一步转移概率矩阵.

然而,不同步长的转移概率矩阵所反映的点与点之间的相似关系有所不同,因此,问题的关键在于多少步长的转移概率矩阵能够真实地反映节点之间的相似度,如何确定马尔可夫链的时间尺度. 近年来一些文献已经对马尔可夫的时间尺度问题进行了相关分析,如 Delvenne 等人[16]利用不断优化动态马尔可夫链的聚类的自协方差来确定时间尺度;Wang 等人[17]所用的时间尺度是通过大量实验确定的,且不大于划分网络的直径;Herrmann 等人[18]重复使用 MCL 扩张和融合两种机制直到算法收敛为止,从而计算出随机步数 $t=30$ 次.

在计算概率矩阵 P 中时间尺度 t 是至关重要的,当 t 较大时,会出现过拟合的现象,即起始点经过 t 步移动后到达任何点的概率都相差无几,甚至社区外的点比社区内的点的相似度高,在划分社区时,两点之间可能无边存在且距离很远,却划分在同一社区内,无法真实地反映社区结构. 所以,为了避免上述情况的发生,本文根据文献[16-18]的参数选择情况,将 t 置为 5,并代入式(7)得到概率矩阵 P,即

$$\boldsymbol{P}=\boldsymbol{P}\boldsymbol{r}_5=\boldsymbol{P}\boldsymbol{r}^5 \tag{8}$$

MSCD 算法将 t 置为 5 的基本思路与经典谱聚类构造相似图的基本思想是相似的,即舍弃全连接法,偏向 K 近邻法,其目的是防止过拟合,当社区划分数量过大时,全连接法会将无连边的两个节点划分到一个社区,而 K 近邻法会降低出现这种情况的频率,但需要设置合适的近邻点数. 因此,MSCD 算法选择将 t 置为 5,这样既可以防止过拟合,又能够选择合适的点作为近邻点.

2.4 重构谱聚类的相似图

谱聚类中的相似矩阵是对称矩阵,这有利于构建拉普拉斯矩阵 L(Graph Laplacians),而上述的概率矩阵 P 显然不是一个对称矩阵,但 P 具有其特殊性,也能够构建拉普拉斯矩阵.

拉普拉斯矩阵 L 的性质主要有以下三个:

(1) L 是对称矩阵,$L=D-W$. 其中,$D=\mathrm{diag}(d_0,d_1,\ldots,d_{n-1})$,$d_i=\sum_{j=0}^{n-1}W_{ij}$,$W=S$.

(2) 对于任意向量 f,都满足

$$f^TLf=\frac{1}{2}\sum_{i,j=1}^{n}w_{ij}(f_i-f_j)^2$$

(3) L 是半正定的,且特征值大于等于 0.

在谱聚类中,拉普拉斯矩阵 L 的性质 2 至关重要,性质 1 是前提条件,保证 L 可逆,可求特征值及其特征向量,性质 2 用来逼近谱聚类中的目标函数,性质 3 分解特征,按特征值大小获取相应特征.

MSCD 算法将相似矩阵换成概率矩阵,即 $W=P$,$d_i=\sum_{j=0}^{n-1}W_{ij}=\sum_{j=0}^{n-1}P_{ij}=1$,$D=I$,则 $L_p=D-W=I-P$,其中,I 为单位矩阵;P 是一个对角线元素为 0;其他位置的元素对称存在,只是数值不同的矩阵. 显然,$I-P$ 与 P 的区别是对角线元素为 1,其他元素位置不变,数值取反,因此,$I-P$ 是可逆的,即 L_p 可逆,符合拉普拉斯矩阵的性质 1.

对于任意向量 f,L_P 都满足下面等式,

$$\begin{aligned}&f^TL_pf=f^TDf-f^TWf=\\&f^TIf-f^TPf=\sum_{i=1}^{n}i_if_i^2-\sum_{i,j=1}^{n}p_{ij}f_if_j=\\&\frac{1}{2}\Big(\sum_{i=1}^{n}i_if_i^2-2\sum_{i,j=1}^{n}p_{ij}f_if_j+\sum_{j=1}^{n}i_jf_j^2\Big)=\\&\frac{1}{2}\sum_{i,j=1}^{n}p_{ij}(f_i-f_j)^2=\frac{1}{2}\sum_{i,j=1}^{n}w_{ij}(f_i-f_j)^2\end{aligned} \tag{9}$$

综上可得,L_P 也符合性质 2 和性质 3,因此,概率矩阵 P 可以构造拉普拉斯矩阵,能够作为谱聚类的相似矩阵来构建相似图.

结合式(4)和式(8)可知,网络 N 的相似矩阵为

$$S=W=P=Pr^5,pr_{ij}=\frac{a_{ij}}{\sum_{m=0}^{n-1}a_{im}} \tag{10}$$

其中,i, $j\in\{0,\ n-1\}$,Pr 是一步转移概率矩阵,Pr^5 是五步转移概率矩阵,Pr 的 5 次幂,pr_{ij} 是一步转移概率,a_{ij} 是无向无权网络 N 的邻接矩阵 A 的元素. 式(10)表示转移概率 p_{ij} 近似为节点的相似度 s_{ij},并利用 s_{ij} 使无向无权网络变换成无向权重网络 N_W,边 e_{ij} 的权值为 w_{ij},

$$s_{ij}=w_{ij}=p_{ij}=pr_{ij}(k,k+5) \tag{11}$$

特别注意,w_{ij} 是无向权重网络 N_W 的邻接矩阵 W 的元素;$pr_{ij}(k,\ k+5)$ 是五步转移概率.

3　基于马尔可夫的谱聚类社区发现算法

3.1　MSCD 算法思想

根据谱图理论，聚类分析可以有许多不同的目标函数，主流谱聚类切图函数是 RatioCut 和 NCut[19]. NCut 切图和 RatioCut 切图类似，但是两者的分母不尽相同，

$$\text{RatioCut}(A_1,\dots,A_k)=\frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i,\bar{A}_i)}{|A_i|} \tag{12}$$

$$\text{NCut}(A_1,\dots,A_k)=\frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i,\bar{A}_i)}{vol(A_i)} \tag{13}$$

其中，$A_1,A_2,\cdots,A_k$ 是无向图 $N(V,E)$ 的 k 个子图点的集合，满足 $A_i\cap A_j=\varnothing$，$A_1\cup A_2\cup\dots\cup A_k=V$，$\bar{A}_i=V-A_i$；对于任意两个子图点的集合 $A,B\subset V$，满足 $W(A,B)=\sum_{i\in A,j\in B}w_{ij}$，$w_{ij}$ 为节点 i 与节点 j 连边的权重；$\text{vol}(A_i)=\sum_{j\in A_i}d_j$ 表示 A_i 子集中所有节点度之和；$|A_i|$=子图 A_i 中节点的个数.

一般来说，NCut 切图要优于 RatioCut 切图，因为子图中个数多并不代表权重高，而谱聚类就是通过最大化子图内连边的权重和，最小化子图间连边的权重和，切掉相应子图间的连边，得到图中的社区. 因此，MSCD 算法是基于 NCut 切图的社区划分算法，也就是说，优化目标为

$$\underbrace{\text{argmin}}_{F} tr(F^T D-1/2L_P D-1/2F)\ \text{s.t.}\ F^TF=I \tag{14}$$

通过求解 $D^{-1/2}L_PD^{-1/2}$ 的前 k 个最小的特征向量，并对它们进行标准化，得到特征矩阵 F. 由于计算中使用维度规约而缺失少量信息，导致得到的优化后的特征向量 f 对应的矩阵 F 不能够完全指示各样本的归属，所以，最后需要对 $n\times k$ 维度的矩阵 F 进行一次传统的聚类（如 K-Means）才能将网络 N 准确地划分为 k 个社区.

3.2　MSCD 算法步骤

算法 1　MSCD

输入： 网络 $N=(V,E)$ 以及其邻接矩阵 A，社区数量 k，其中，V 表示点集，$v_i\in V,i=1,\cdots,n$.

输出： 网络 N 的 k 个社区划分，即 $C_1,C_2,\cdots,C_k$.

Begin

1）根据式(8)，计算概率矩阵 P.

2）根据式(10)，计算权重邻接矩阵 W.

3）根据小节 2.4，计算未规范化的拉普拉斯矩阵 L_P，即 $L_P=D-W$.

4）构建规范化的拉普拉斯矩阵 L_n，即 $L_n=D-1/2L_PD-1/2=D-1/2(D-W)D-1/2=I-D-1W$，其中，$I$ 是 $n\times n$ 的单位矩阵.

5）令 $L_nf=\lambda f$，计算 L_n 最小的 k 个特征值以及各特征值对应的特征向量 $f_1,\cdots,f_k$.

6）将 $f_1,\cdots,f_k$ 按行方式排列，组成 $n\times k$ 的特征矩阵 F.

7）将 F 中的每一行作为 k 维的样本，共 n 个样本，令 $y_i\in F^n$ 表示矩阵 F 的第 i 行，$i=1,\cdots,k$.

8）利用 K-Means 方法对数据集 $\{y_i\}$ 进行聚类，聚类维数为 k，当 $y_i\in C_j$，$v_i\in C_j$，以此获得社区划分 $C_1,C_2,\cdots,C_k$.

End.

4　实验分析

MSCD 算法分别在计算机合成网络和现实世界网络上进行实验，并与 SC[20]，MRW－KNN[13] 和异步流体社区算法（FluidC）[21] 相比较. 实验选择 SC 和 MRW-KNN 这两种算法的原因在上文中已经提过，选择 FluidC 作为比较算法的原因有三个，(1) FluidC 是 2017 年提出的较为杰出的算法；(2) 此算法已经在 networkx 上实现，说明其在社区发现算法中具有一定的重要性；(3) 它的设计思想完全与另外三种算法不同. 因此，实验不仅有同一类型算法的比较，还有与不同类型算法的比较，这样可以使实验更加完整，更加具有说服力.

本文实验在合成网络和现实网络上的实验指标分别为规范互信息 NMI[22] 和模块度 Q[23] 两大主流评估指标. 实验的硬件环境是 CPU 为 2.5 GHZ 的 Intel(R) Core(TM)2 Quad，内存为 8 G；软件环境是 Windows 10 + Python + PyCharm Community Edition 2018.

4.1　计算机合成网络

4.1.1　LFR 基准合成网络　本文使用 LFR 基准[24] 随机生成合成网络，并通过测量不同算法对合成网络社区划分的 NMI 来评估各算法的性能. LFR 基准比 GN 基准[23] 拥有更加符合现实世界网络的属性，比如复杂网络无尺度的程度大小和簇范围的分布，因此，LFR 基准生成网络更能够检测算

法社区划分的能力.

在实验中,本文利用不同网络尺度 N 的基准图和不同的社区间连接边指数 mu 进行算法性能评估,mu 又称混合参数[24],是节点与其他社区存在连接边的平均比例,除此之外,LFR 基准图还需要设置一些其他的超参数(见表 1)来保持实验用图的一致性.

表 1 LFR 基准生成网络的超参数

Tab. 1 Hyperparameters of LFR benchmark

参数	说明	取值
*tau*1	网络中度分布的幂律指数	4
*tau*2	网络中社区尺度的幂律指数	2
average_degree	网络中节点的平均度数	10
min_community	每个社区至少拥有的节点数	30
seed	随机数生成器的种子	10

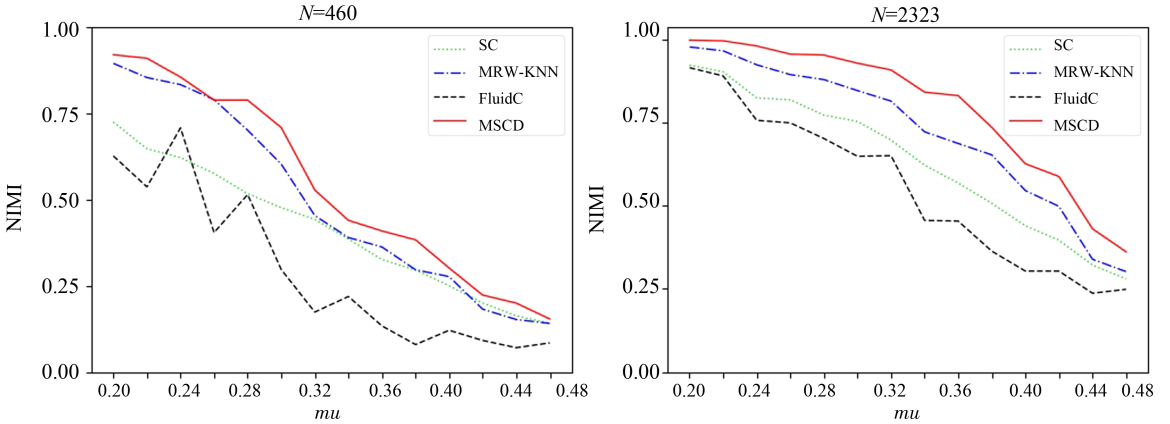

图 1 N=460 和 N=2 323

Fig. 1 N=460 and N=2 323

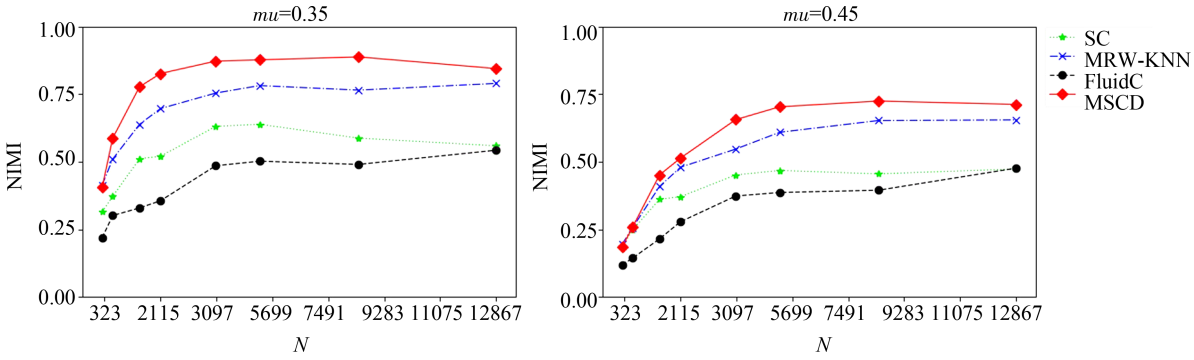

图 2 mu=0.35 和 mu=0.45

Fig. 2 mu=0.35 and mu=0.45

4.1.2 实验结果分析 本文按照规定的超参数标准对四种算法进行了多组对比试验,具体分析如下.

(1) 图 1 是以 mu 为自变量,以 NMI 为因变量,对四种算法进行实验的结果. 图 1 的网络节点数分别为 460 和 2 323;社区数量为 9 和 47. 根据图 1 的数据可得,当 $mu<0.45$ 时,MSCD 的划分质量明显优于其他算法;当 $mu>0.45$ 时,四种社区发现算法的划分质量趋于相同,且 NMI 较小没有参考价值.

(2) 图 2 是以 N 为自变量,以 NMI 为因变量的测试结果. 在实验中,mu 取值分别为 0.35,0.45,N 的取值是离散无规律的值(N=323,591,1 451,2 123,3 890,5 298,8 459,12 867)而不是连续值或者相同间隔的值,使用离散值的原因有两个:(1) LFR 基准网络生成过程中会出现自环的情况,而 MSCD 不适用于拥有自环的网络;(2) FluidC 只能对连通图社区划分,而 LFR 基准网络可能生成非连通图. 由图 3 的实验数据可得,当 $N<5\,000$ 时,随着 N 的增大,MSCD 的划分效果越来越好,提升越来越快,并逐渐拉大了与其他算法之间的距离;当 $N>5\,000$ 时,四种算法的划分质量趋于稳定,MSCD 的划分效果明显优于其他算法.

4.2 现实世界网络

4.2.1 现实网络 由于真实网络可能具有与合成网络不同的拓扑属性,本组实验决定采用 8 个真实世界网络(详见表 2)来进一步评估四种算法的性能. 但是,符合实验标准的真实数据相当少且网络

规模比较小，因为真实数据必须要满足 4 种算法的适用范围，比如网络是无向无权图，网络中不允许出现自环，网络必须是连通图等等，所以，在实验过程中，对部分真实网络统一进行了标准化处理（例如，去掉孤立点，抹除自环的边等），以适应算法的适用范围，最重要的是，标准化处理可以增大实验网络的数量和规模.

表 2　真实网络

Tab. 2　Real-world networks

序号	网络名称	顶点数量	边的数量
1	American football	115	613
2	US Air lines	332	2 126
3	C. elegans metabolic network	453	2 025
4	Graph Drawing Contests	638	1 020
5	Small & Griffith and Descendants	1 024	4 916
6	E-mail network URV	1 133	5 451
7	US Power Grid	4 941	6 594
8	PGP network	10 680	24 316

4.2.2　实验结果分析　本文在 8 个真实网络上对四种算法进行了实验，并将模块度 Q 作为四种算法的评估指标，因 8 个真实网络的节点数量差距悬殊，网络的最大社区数量相差甚远，实验分为(a)和(b)两组，两组的真实网络分别对应表 2 中的前四个网络和后四个网络，实验结果具体分析如下：

图 3 中的实验数据 Q 是根据不同社区数量进行多次迭代求得的均值，每组的迭代次数为 11 次，社区数量的增加步长为 1，(a)组的社区数量 8～18，(b)组的社区数量 95～105.

由图 3 可知，MSCD 的 Q 值几乎都大于 0.3，说明 MSCD 对真实网络的划分效果良好. 虽然 MSCD 对于网络 2 和网络 8 的划分效果比 FluidC 稍差一些，但是 MSCD 对其他 6 个真实网络的划分效果明显优于其他三种算法. 综上所述，MSCD 对大部分真实网络的划分质量优于其他三种算法.

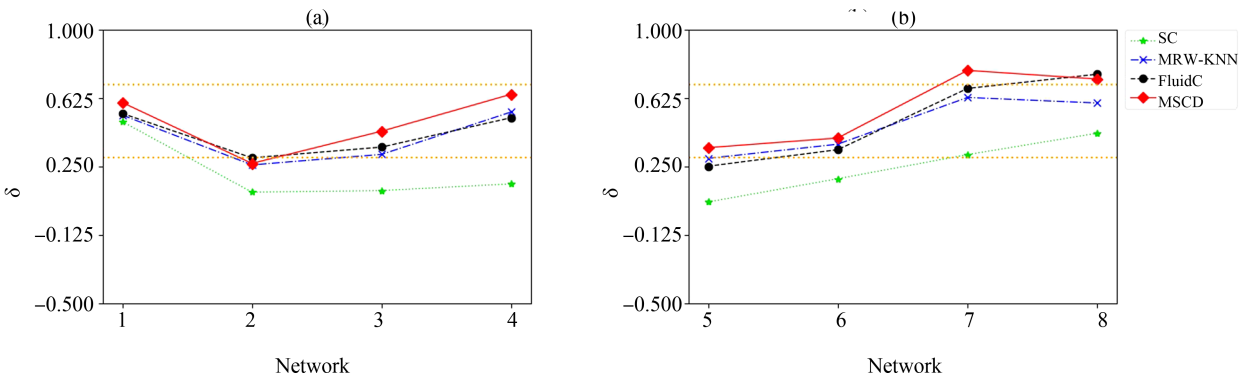

图 3　真实网络的实验（两条参考黄线的值分别是 0.3 和 0.7）

Fig. 3　Experiments on real-world networks (the two yellow reference lines are 0.3 and 0.7 respectively)

5　结　论

社区发现是复杂网络领域中的一个重要研究方向，而谱聚类是一种经典的社区发现算法，且具有坚实的理论基础，但是，由于目前谱聚类的相似图包含较少的社区信息，谱聚类的划分效果并不理想. 因此，本文提出一种改进的谱聚类社区发现算法 MSCD，该算法是以 DCBM 模型推导作为理论基础，利用马尔可夫链重构了谱聚类的相似图，并通过重构的相似图实现社区划分. 通过大量的实验表明，MSCD 能够准确地划分 LFR 基准网络和现实世界网络，在划分质量上优于大部分社区发现算法，尤其是对大型复杂网络的社区检测.

MSCD 是一种静态社区发现算法，且只能划分无权图和非重叠网络，所以，未来的工作主要从以下几个方面进行：(1)考虑如何确定最优社区划分数量 k，因为 MSCD 的 k 是人工设置的参数；(2)考虑能否将其扩展到动态网络；(3)考虑如何划分有权图和重叠网络.

参考文献：

[1]　范大鹏，张凤斌. 一种基于并行免疫网络的大数据分类算法[J]. 江苏大学学报：自然科学版，2018，39：581.

[2]　冯成强，左万利，王英. 基于相似度投票的社区划分改进算法[J]. 吉林大学学报：理学版，2018，56：601.

[3]　Žalik K R. Community detection in networks using new update rules for label propagation [J]. Computing，2017，99:1.

[4]　Newman M E. Spectral methods for community de-

tection and graph partitioning. [J]. Phys Rev E, 2013, 88: 042822.

[5] Girvan M, Newman M E J. Community structure in social and biological networks [J]. PNAS, 2002, 99: 7821.

[6] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Phys Rev E, 2007, 76: 036106.

[7] Jin J. Fast community detection by SCORE [J]. Ann Stat, 2012, 43: 672.

[8] Gulikers L, Lelarge M, Massoulié L. A spectral method for community detection in moderately-sparse degree-corrected stochastic block models [J]. Mathematics, 2017. doi: 10.1017/apr.2017.18.

[9] Luxburg U. A tutorial on spectral clustering [J]. Stat Comput, 2007, 17: 395.

[10] Yan J, Cheng D, Zong M, *et al*. Improved spectral clustering algorithm based on similarity measure [M]// Advanced Data Mining and Applications. USA: Springer International Publishing, 2014.

[11] Zang W, Jiang Z, Ren L. Improved spectral clustering based on density combining dna genetic algorithm [J]. Int J Pattern Recogn, 2016, 31: 1750010.

[12] Zhang X, You Q. An improved spectral clustering algorithm based on random walk [J]. Front Comput Sci, 2011, 5: 268.

[13] 曹江中, 陈佩, 戴青云, 等. 基于 Markov 随机游走的谱聚类相似图构造方法[J]. 南京大学学报:自然科学, 2015, 51: 772.

[14] Karrer B, Newman M E J. Stochastic blockmodels and community structure in networks [J]. Phys Rev E, 2011, 83: 016107.

[15] Bickel P J, Chen A. A nonparametric view of network models and Newman-Girvan and other modularities [J]. PNAS, 2009, 106: 21068.

[16] Delvenne J C, Newman M. Stability of graph communities across time scales [J]. PNAS, 2010, 107: 12755.

[17] Wang W, Liu D, Liu X, *et al*. Fuzzy overlapping community detection based on local random walk and multidimensional scaling [J]. Phys A, 2013, 392: 6578.

[18] Herrmann S, Ochoa G, Rothlauf F. Communities of local optima as funnels in fitness landscapes [C]//Genetic and Evolutionary Computation Conference. [S.l.]: ACM, 2016.

[19] Shi J, Malik J. Normalized cuts and image segmentation [J]. IEEE Trans Pattern Anal Mach Intell, 2000, 22: 888.

[20] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm [J]. Proc Nips, 2001, 14: 849.

[21] Parés F, Gasulla D G, Vilalta A, *et al*. Fluid communities: a competitive, scalable and diverse community detection algorithm [C]// Complex Networks & Their Applications VI. Cham: Springer, 2017, 689: 229.

[22] Bai L, Liang J, Du H, *et al*. A novel community detection algorithm based on simplification of complex networks [J]. Knowl-Based Syst, 2018, 143: 58.

[23] Newman M E, Girvan M. Finding and evaluating community structure in networks. [J]. Phys Rev E, 2004, 69: 026113.

[24] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms [J]. Phys Rev E, 2008, 78: 046110.

引用本文格式:

中 文: 任淑霞, 张书博, 吴涛. 基于 DCBM 的马尔可夫谱聚类社区发现算法[J]. 四川大学学报: 自然科学版, 2019, 56: 1049.

英 文: Ren S X, Zhang S B, Wu T. Markov spectral clustering algorithm with DCBM for community detection [J]. J Sichuan Univ: Nat Sci Ed, 2019, 56: 1049.