·人工智能及识别技术· 文章编号：1000-3428(2019)10-0227-07 文献标志码：A 中图分类号：TP391

基于聚类质量的半监督 INMF 动态社区检测算法

陈吉成，陈鸿昶，于洪涛

（国家数字交换系统工程技术研究中心，郑州 450002）

摘 要： 为实现复杂网络的快速分析，提出一种基于聚类质量的改进非负矩阵分解（INMF）算法，将其用于动态社区检测。从理论分析角度证明了演化谱聚类、INMF 和模块密度优化之间的等价性，并基于该等价性，在不增加时间复杂度的前提下，通过在 INMF 中加入先验信息给出一种半监督 INMF 算法。在人工构造和真实世界的动态网络上的实验结果表明，与 QCA、MIEN 算法相比，该算法的社区检测质量和社区检测效率更优。

关键词： 聚类质量；半监督；非负矩阵分解；动态社区检测；图模型

开放科学（资源服务）标志码（OSID）：

中文引用格式： 陈吉成，陈鸿昶，于洪涛. 基于聚类质量的半监督 INMF 动态社区检测算法[J]. 计算机工程，2019，45(10)：227-233.

英文引用格式： CHEN Jicheng, CHEN Hongchang, YU Hongtao. Semi-supervised INMF algorithm for dynamic community detection algorithm based on clustering quality[J]. Computer Engineering, 2019, 45(10): 227-233.

Semi-Supervised INMF Algorithm for Dynamic Community Detection Based on Clustering Quality

CHEN Jicheng, CHEN Hongchang, YU Hongtao

(China National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China)

【Abstract】 In order to realize the rapid analysis of complex networks, an Improved Non-Negative Matrix Factorization (INMF) algorithm based on Clustering Quality (CQ) is proposed, and applied to dynamic community detection. From the perspective of theoretical analysis, the equivalence between evolutionary spectral clustering, INMF and module density optimization is proved. Based on the equivalence, a semi-supervised INMF algorithm is given by adding a priori information to INMF without increasing the time complexity. Experimental results on artificial networks and real-world dynamic networks show that the algorithm has better community detection quality and community detection efficiency compared with QCA and MIEN algorithms.

【Key words】 Clustering Quality (CQ); semi-supervised; non-negative matrix factorization; dynamic community detection; graph model

DOI: 10.19678/j.issn.1000-3428.0052570

0 概述

通过将每个实体视为一个顶点网络，每条边看作一对顶点之间的相互作用，可实现社区网络图模型的构建。现实中存在各种各样的网络，如社交网络、互联网络和生物网络等。研究结果表明，许多网络具有社群（模块）结构，该结构被定义为子图，该子图具有更多的边连接同一组中的顶点，而相对较少的边连接不同组中的顶点。例如，蛋白质相互作用网络中的社区对应于生物过程至关重要的蛋白质复合物[1-2]。在科学合作网络中，社区对应于具有相同或相似研究方向的科学家群体[3]。

社区检测为分析社交网络系统的结构和功能提供了有效的分析方法，但设计有效且高效的社区检测算法仍具有挑战性，主要原因是没有统一的标准量化网络中的社区[4]。当前针对社区检测已提出许多方法，例如：文献[5]设计了采用模块度指标评价的动态社区发现算法，其本质上是一种贪婪社区检测算法，缺点是指标并未对模块密度问题进行综合分析，具有一定的片面性；文献[6]设计了采用

基金项目： 国家自然科学基金创新研究群体项目(61521003)。

作者简介： 陈吉成(1984—)，男，博士研究生，主研方向为社交网络挖掘、通信与信息系统；陈鸿昶、于洪涛，教授。

收稿日期： 2018-09-05 **修回日期：** 2018-10-23 **E-mail：** chenjichenglin@163.com

计算机工程 www.ecice06.com

Girvan 策略的动态社区检测算法,其基于层次聚类进行社区检测计算,但是层次聚类算法的计算复杂度过高,同时中心层次的构建存在先天缺陷;文献[7]研究设计了基于图模型的动态社区检测算法,图模型采用的是有向权重模型构建算法,可实现对信息流的有效压缩,获得更加优异的社区发现效果;文献[8]设计了基于信息传播方法的动态社区检测算法,算法对于社区检测过程的指标属性考虑不够充分,社区划分精度不高。以上算法对于复杂社区检测过程中的边缘密度具有一定要求,并且未对社区间的交互力问题进行有效分析。

为实现复杂网络的快速分析,较为理想的社区检测方式是对计算效率和精度进行同步考虑。因此,本文提出基于聚类质量(Clustering Quality,CQ)的改进非负矩阵分解(Improved Non-Negative Matrix Factorization,INMF)算法,并将其用于动态社区检测。

1 问题描述

1.1 符号定义

定义 1(动态网络) 令$\{1,2,\cdots,T\}$表示一组有限的时间步长。对于给定的变量,附加的下标 t 表示时间 t 上变量的值。动态网络 G 被定义为网络的一个序列 $G=\{G_1,G_2,\cdots,G_T\}$,其中 G_t 是时间 t 上的网络,具有顶点集 V_t 和边缘集 E_t。假设 G 中的所有网络都具有相同的顶点集,即 $G_t=(V_t,E_t)$ 等效为 $G_t=(V,E_t)$。

定义 2(三维邻接矩阵) 动态网络 G 可以用三维邻接矩阵表示 $\boldsymbol{A}=(a_{ijt})_{n\times n\times T}$,其中 n 是顶点的个数(即 $n=|V|$),如果网络 G_t 中第 i 个顶点与第 j 个顶点连接,则 $a_{ijt}=1$;否则 $a_{ijt}=0$。实际上,$\boldsymbol{A}=[A_1,A_2,\cdots,A_T]$是网络 G_t 的邻接矩阵,$\boldsymbol{A}_t=(a_{ijt})_{n\times n}$。如果 G_t 是无向网络,则邻接矩阵 $\boldsymbol{A}_t$ 是对称的。

定义 3(度矩阵) 假设所有的网络是无向和未加权的。网络 G_t 中的第 i 个顶点的度数定义为连接到它的边数,即 $d_{it}=\sum_j a_{ijt}$。动态网络 G 的度矩阵定义为 $\boldsymbol{D}=[D_1,D_2,\cdots,D_T]$,其中 $\boldsymbol{D}_t$ 是具有 G_t 度序列的对角矩阵,即 $\boldsymbol{D}_t=\mathrm{diag}(d_{1t},d_{2t},\cdots,d_{nt})$。

网络中的社区检测是将顶点划分成不连续的组,其中同一簇中的边缘数量大于组中的边缘数量。具体而言,对于给定网络 $G_t=(V,E_t)$,社区结构对应于硬划分$\{C_{1t},C_{2t},\cdots,C_{kt}\}$(表示为$\{C_{it}\}_{i=1}^k$),如果 $i\neq j$ 且 $V=\sum_i C_{it}$,则使得 $C_{it}\cap C_{jt}=\varnothing$。$C_{it}$是时间 t 的第 i 个本地社区(集群)。$\{C_{it}\}_{i=1}^k$可利用 $n\times k$ 归一化指示符矩阵 $\boldsymbol{Z}_t=(z_{ijt})_{n\times k}$进行表示,其中如果在时间 t 上,第 i 个顶点属于第 j 个局部簇,则 $z_{ijt}=1$,否则 $z_{ijt}=0$。为简化,利用局部簇的大小参数对矩阵 $\boldsymbol{Z}_t$ 列进行归一化处理,即 $\tilde{z}_{\cdot jt}=z_{\cdot jt}/\sqrt{(C_{jt})}$,其中 $\tilde{z}_{\cdot jt}$是 $\boldsymbol{Z}_t$ 的第 j 列,$\tilde{\boldsymbol{Z}}_t=[\tilde{z}_{\cdot 1t},\tilde{z}_{\cdot 2t},\cdots,\tilde{z}_{\cdot kt}]$,$\tilde{\boldsymbol{Z}}_t$ 是列正交的,容易验证 $\tilde{\boldsymbol{Z}}_t'\tilde{\boldsymbol{Z}}_t=\boldsymbol{I}_k$,其中,$\boldsymbol{I}_k$ 是单位矩阵,$\tilde{\boldsymbol{Z}}_t'$是矩阵 $\tilde{\boldsymbol{Z}}_t$ 的转置。

1.2 社区定量评价函数

给定网络 $G_t=(V_t,E_t)$以及硬划分$\{C_{it}\}_{i=1}^k$,定义 $L(C_{it},C_{jt})=\sum_{i\in C_{it},j\in C_{jt}}a_{ijt}$和 $\bar{C}_{it}=V\backslash C_{it}$。模块化指标 Q 可定义为[9]:

$$Q(\{C_{it}\}_{i=1}^k)=\sum_{i=1}^k\frac{L(C_{it},C_{it})}{L(V,V)}-\left(\frac{L(C_{it},V)}{L(V,V)}\right)^2 \tag{1}$$

模块密度指标 Q_{D} 可定义为:

$$Q_{\mathrm{D}}(\{C_{it}\}_{i=1}^k)=\sum_{i=1}^k\frac{L(C_{it},C_{it})-L(C_{it},\bar{C}_{it})}{|C_{it}|} \tag{2}$$

负相关性指标可定义为:

$$NA=\mathrm{Tr}(\boldsymbol{A}_t)-\sum_{i=1}^k\frac{L(C_{it},C_{it})}{|C_{it}|} \tag{3}$$

其中,$\mathrm{Tr}(\boldsymbol{A}_t)$是矩阵 $\boldsymbol{A}_t$ 的迹,即$\sum_{i=1}^n a_{iit}$。文献[10]提出的模块化指标是目前常用的一种衡量网络中社区稳定度的方法;模块密度指标主要评价社区中社区间的距离程度,对网络社区的稀疏性进行表征;负相关性指标评价的主要是社区之间的相关性。

2 基于 CQ 的聚类算法等价性描述

2.1 基于 CQ 的聚类算法分析

基于聚类时间(Clustering Time,CT)的定义,可利用聚类质量指标对聚类算法的等价性进行定义。利用前一时间步长网络以及当前网络聚类结果的一致性进行定义,图 1 表示了在 $t-1$ 和t 时刻网络间的节点连接及节点间的权重(暂不考虑虚线部分)。由于聚类只考虑将同类的节点归类,因此图中虚线清晰地将 $t-1$ 和 t 时刻的聚类结果进行表示,而没有考虑节点间的权重。

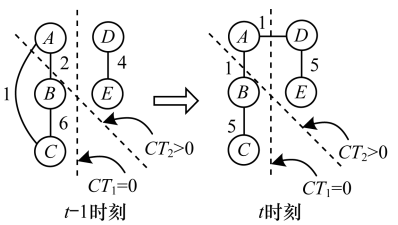


图 1 基于 CQ 的聚类过程

在该算法中,前一时间步的局部社区用于指导当前时间步的社区发现,并证明进化非负矩阵分解、演化谱聚类和模块密度在该过程中的等价性。虽然它们看似无关,但在轨迹优化方面是等价的,其关键问题

www.ecice06.com

是如何将目标函数转化为轨迹优化。在 CQ 算法框架中,CT 表示当前划分对历史数据的聚类程度。网络 G_t 的划分 $Z_t^{[1]}$ 和 $Z_t^{[2]}$ 是等价的,即 $CS_1 = CS_2$。

2.2　演化谱聚类

在 CQ 框架下,使用负相关性的演化谱聚类的总成本可定义为:

$$Cost_{NA} = \alpha CS_{NA} + (1-\alpha) CT_{NA} = \alpha NA_t \mid_{Z_t} + (1-\alpha) NA_{t-1} \mid_{Z_t} \tag{4}$$

其中,$\mid_{Z_t}$表示基于社区划分 $\boldsymbol{Z}_t$ 的谱聚类社区优化评价,时间成本 $CT_{NA} = NA_{t-1} \mid_{Z_t}$表示对不属于网络 G_{t-1}的 $\boldsymbol{Z}_t$ 的惩罚。将式(4)转化为轨迹优化问题:

$$Cost_{NA} = \mathrm{Tr}[\alpha \boldsymbol{A}_t + (1-\alpha)\boldsymbol{A}_{t-1}] - \mathrm{Tr}[\tilde{\boldsymbol{Z}}_t'(\alpha \boldsymbol{A}_t + (1-\alpha)\boldsymbol{A}_{t-1})\tilde{\boldsymbol{Z}}_t] \tag{5}$$

其中,第一项 $\mathrm{Tr}[\alpha \boldsymbol{A}_t + (1-\alpha)\boldsymbol{A}_{t-1}]$为常数,因为它独立于局部分区 $\tilde{\boldsymbol{Z}}_t$。因此,最小化 $Cost_{NA}$等同于最大化最后一项,即:

$$\min Cost_{NA} \propto \max \mathrm{Tr}[\tilde{\boldsymbol{Z}}_t'(\alpha \boldsymbol{A}_t + (1-\alpha)\boldsymbol{A}_{t-1})\tilde{\boldsymbol{Z}}_t] \tag{6}$$

2.3　模块密度优化

通过优化模块密度 Q_{D} 可优化聚类的总成本,定义成本函数:

$$Cost_{Q_{\mathrm{D}}} = \alpha CS_{Q_{\mathrm{D}}} + (1-\alpha) CT_{Q_{\mathrm{D}}} = \alpha (Q_{\mathrm{D}} \mid_{Z_t} + (1-\alpha)(Q_{\mathrm{D}} \mid_{Z_t} \tag{7}$$

为导出 $Cost_{Q_{\mathrm{D}}}$的轨迹优化,需先导出$(Q_{\mathrm{D}})_t \mid_{Z_t}$的轨迹优化问题:

$$(Q_{\mathrm{D}})_t \mid_{Z_t} = \sum_{i=1}^{k} \frac{L(C_{it}, C_{it}) - L(C_{it}, \bar{C}_{it})}{\mid C_{it} \mid} = \mathrm{Tr}[\tilde{\boldsymbol{Z}}_t'(2\boldsymbol{A}_t - \boldsymbol{D}_t)\tilde{\boldsymbol{Z}}_t] \tag{8}$$

将式(8)代入式(7),可使 $Cost_{Q_{\mathrm{D}}}$的整体成本优化问题转换为轨迹优化问题:

$$Cost_{Q_{\mathrm{D}}} = \mathrm{Tr}[\tilde{\boldsymbol{Z}}_t'(\alpha(2\boldsymbol{A}_t - \boldsymbol{D}_t) + (1-\alpha)(2\boldsymbol{A}_{t-1} - \boldsymbol{D}_{t-1}))\tilde{\boldsymbol{Z}}_t] \tag{9}$$

对比式(6)和式(9),除了使用的矩阵不同外,它们具有相同的优化形式,表明模块密度优化 Q_{D} 与演化谱聚类是等价的,即:

$$\max Cost_{Q_{\mathrm{D}}} \propto \min Cost_{NA} \tag{10}$$

该结果从理论上解释了为何拓扑结构可用于动态社区检测。

2.4　非负矩阵分解

矩阵分解是进行图中挖掘的有效模式,旨在通过将目标矩阵近似为两个低秩矩阵的乘积来学习原始数据的表示。具体地,对于给定 $n \times m$ 矩阵 $\boldsymbol{F}$,非负矩阵分解将 $\boldsymbol{F}$ 分解为 2 个非负矩阵 $\boldsymbol{R}_{n\times s}$和 $\boldsymbol{S}_{s\times m}$,使得[11]:

$$\boldsymbol{F} \approx \boldsymbol{RS}, \text{ s.t. } \boldsymbol{R} \geqslant 0, \boldsymbol{S} \geqslant 0 \tag{11}$$

在通常情况下,s 比 $\min\{n, m\}$ 小得多;对称非负矩阵分解(Symmetric Non-Negative Matrix Factorization, SNMF)是非负矩阵分解的一个扩展,它将矩阵 $\boldsymbol{F}$ 近似分解成 $n \times s$ 矩阵 $\boldsymbol{R}$,使得:

$$\boldsymbol{F} \approx \boldsymbol{RR}', \text{ s.t. } \boldsymbol{R} \geqslant 0 \tag{12}$$

给定网络 $\boldsymbol{G}_t$ 的邻接矩阵,利用 SNMF 可以得到局部簇 $\tilde{\boldsymbol{Z}}_t$:

$$\boldsymbol{A}_t \approx \tilde{\boldsymbol{Z}}_t \tilde{\boldsymbol{Z}}_t' \tag{13}$$

通过最小化平方误差来求解上述方程,即:

$$SNMF_t = \min_{\tilde{\boldsymbol{Z}}_t \geqslant 0,\ \tilde{\boldsymbol{Z}}_t'\tilde{\boldsymbol{Z}}_t = \boldsymbol{I}_k} \| \boldsymbol{A}_t - \tilde{\boldsymbol{Z}}_t \tilde{\boldsymbol{Z}}_t' \|^2 \tag{14}$$

其中,$\| \boldsymbol{B} \|$ 是矩阵 $\boldsymbol{B}$ 的 Frobeneus 范数。将等式(14)改写为轨迹优化,即:

$$SNMF_t = \max_{\tilde{\boldsymbol{Z}}_t \geqslant 0,\ \tilde{\boldsymbol{Z}}_t'\tilde{\boldsymbol{Z}}_t = \boldsymbol{I}_k} \mathrm{Tr}(\tilde{\boldsymbol{Z}}_t' \boldsymbol{A}_t \tilde{\boldsymbol{Z}}_t) \tag{15}$$

利用松弛正交性 $\tilde{\boldsymbol{Z}}_t'\tilde{\boldsymbol{Z}}_t = \boldsymbol{I}_k$,式(15)可改写为下列形式:

$$SNMF_t \propto \max_{\tilde{\boldsymbol{Z}}_t \geqslant 0} \mathrm{Tr}(\tilde{\boldsymbol{Z}}_t' \boldsymbol{A}_t \tilde{\boldsymbol{Z}}_t) \tag{16}$$

结合式(4)可定义非负矩阵分解的整体成本函数形式为:

$$Cost_{\mathrm{INMF}} = \alpha CS_{\mathrm{SNMF}} + (1-\alpha) CT_{\mathrm{SNMF}} = \alpha SNMF_t \mid_{Z_t} + (1-\alpha) SNMF_{t-1} \mid_{Z_t} \tag{17}$$

基于式(16),将式(7)重新表述为:

$$Cost_{\mathrm{INMF}} \propto \alpha \mathrm{Tr}(\tilde{\boldsymbol{Z}}_t' \boldsymbol{A}_t \tilde{\boldsymbol{Z}}_t) + (1-\alpha)\mathrm{Tr}(\tilde{\boldsymbol{Z}}_t' \boldsymbol{A}_{t-1} \tilde{\boldsymbol{Z}}_t) = \mathrm{Tr}[\tilde{\boldsymbol{Z}}_t'(\alpha \boldsymbol{A}_t + (1-\alpha)\boldsymbol{A}_{t-1})\tilde{\boldsymbol{Z}}_t] \tag{18}$$

对比式(6)、式(9)及式(18)可得出,SNMF、演化谱聚类和模块密度优化是等价的,即:

$$\begin{cases} \min Cost_{\mathrm{INMF}} \propto \min Cost_{NA} \\ \min Cost_{\mathrm{INMF}} \propto \max Cost_{Q_{\mathrm{D}}} \end{cases} \tag{19}$$

由此证明 SNMF 和演化谱聚类之间的等价性,表明 SNMF 算法也可以用于动态社区检测。

3　半监督 INMF 算法

3.1　半监督局部社区聚类发现

为基于时间平滑框架在时刻 t 发现局部簇,通过线性函数同时考虑 G_t 和 G_{t-1},定义该线性函数为[12]:

$$\boldsymbol{A}_t^* = \alpha \boldsymbol{A}_t + (1-a)(\boldsymbol{A}_t - \boldsymbol{A}_{t-1}) \tag{20}$$

其中,参数 α 控制 CS 和 CT 的相关重要性。基本的假设是:如果一组顶点的连通性在 G_t 和 G_{t-1}中都很强,那么它们很可能在时刻 t 是本地集群。通常选取参数 $\alpha = 0.8$。许多研究成果表明[5-6],在聚类算法中加入先验信息可显著提高算法精度,如 Pagerank 算法、半监督 NMF 算法和半监督谱聚类算法。本文提出结合非负矩阵算法和谱聚类算法

www.ecice06.com 计算机工程

的半监督改进非负矩阵分解算法。采用谱聚类的原因为:1)谱聚类通常可获得良好的聚类性能;2)演化谱聚类和非负矩阵分解算法之间的等价性为它们组合奠定了坚实的理论基础。

构建先验信息的策略包括两个步骤:1)获得高质量的顶点群;2)部分信息的构建。首先使用由指示符矩阵 $\boldsymbol{Z}_t^{\mathrm{sc}}$ 表示谱聚类算法来获得时刻 t 的局部聚类,其中如果第 i 个顶点属于第 j 个局部社区,则 $Z_{ijt}^{\mathrm{sc}}=1$;否则 $Z_{ijt}^{\mathrm{sc}}=0$。为确保先验信息的质量,对于每个局部社区,需移除社区内的顶点,直到社区的诱导子网络的密度达到阈值 β。基本假设是强连通顶点很可能被分组为一个社区。由于在随机网络中普遍存在小的密集子图,因此保留了至少 5 个顶点的子图作为先验信息,细化的局部社区由 $\boldsymbol{Z}_t^{\mathrm{rc}}$ 表示。将先验信息结合到 $\boldsymbol{A}_t^*$ 中:

$$\hat{\boldsymbol{A}}_t^*=\boldsymbol{A}_t^*+\gamma\boldsymbol{Z}_t^{\mathrm{rc}}(\boldsymbol{Z}_t^{\mathrm{rc}})' \tag{21}$$

其中,γ 是控制先验信息的相关权重参数。为获得时间 t 上的局部簇,本文所提算法通过最小化平方误差对矩阵 $\hat{\boldsymbol{A}}_t^*$ 进行分解,即:

$$\boldsymbol{J}_t=\|\boldsymbol{A}_t^*-\boldsymbol{B}_t\boldsymbol{F}_t\|^2,\text{s.t. }\boldsymbol{B}_t\geqslant 0,\boldsymbol{F}_t\geqslant 0 \tag{22}$$

为求解上述方程,将 J_t 的导数设为:

$$\frac{\partial\boldsymbol{J}_t}{\partial\boldsymbol{B}_t}=0,\frac{\partial\boldsymbol{J}_t}{\partial\boldsymbol{F}_t}=0 \tag{23}$$

得到以下更新规则:

$$\boldsymbol{B}_t=\boldsymbol{B}_t\frac{\hat{\boldsymbol{A}}_t^*\boldsymbol{F}_t'}{\boldsymbol{B}_t\boldsymbol{F}_t\boldsymbol{F}_t'} \tag{24}$$

$$\boldsymbol{F}_t=\boldsymbol{F}_t\frac{\boldsymbol{B}_t'\hat{\boldsymbol{A}}_t^*}{\boldsymbol{B}_t'\boldsymbol{B}_t\boldsymbol{F}_t} \tag{25}$$

基于矩阵 $\boldsymbol{F}_t$ 构造 G_t 局部簇的指示矩阵 $\boldsymbol{Z}_t$,即如果 $j^*=\underset{j}{\operatorname{argmax}}F_{ij}$,则可得 $Z_{ij^*t}=1$;否则 $Z_{ij^*t}=0$。该过程如图 2 所示。

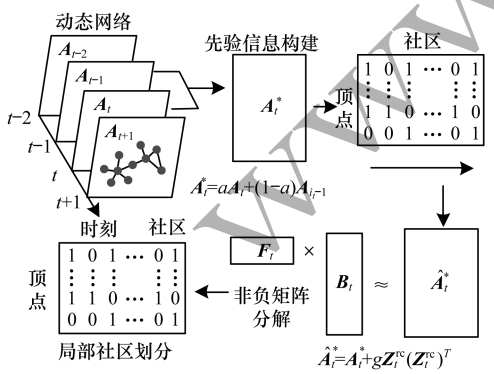


图 2 基于先验信息的时间平滑框架局部聚类发现过程

3.2 动态社区发现

给定时间步长 $t-1$ 和 t 的局部簇,通过局部簇映射来发现动态社团。令 C_t 为时间 t 时的局部簇集合,C_{it} 是 C_t 的第 i 个局部集群。构造 C_t 和 C_{t-1} 之间的二分图,其中每个局部聚类被视为顶点,$C_{i,t-1}$ 和 C_{jt} 之间的链接权重被定义为重叠顶点的数目,即:

$$l_{ijt}=|C_{i,t-1}\cap C_{jt}| \tag{26}$$

潜在的假设是:若更多的顶点重叠,则更可能多的局部集群将被映射。类似地,构造从 C_1 到 C_T 的 T 划分图如图 3 所示。

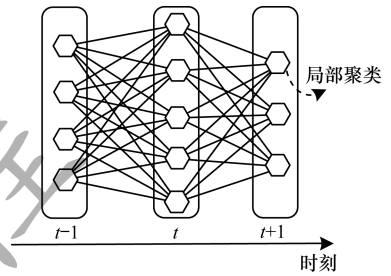


图 3 基于局部簇映射的动态社区划分

利用互信息定义 T 图的拓扑相似度[13-14]:

$$MI(C_{i,t-1},C_{jt})=\frac{l_{ijt}}{\sum_j l_{ijt}}\lg\frac{l_{ijt}\sum_{i,j}l_{ijt}}{\sum_i l_{ijt}\sum_j l_{ijt}} \tag{27}$$

假设局部簇之间的互信息越高,两个局部簇就越有可能相关。贪婪搜索程序用于从 C_{t-1} 到 C_t 的局部映射集群:对于每一次迭代,选择具有最大 MI 值的局部簇对;一旦 $C_{i,t-1}$ 和 C_{jt} 之间的映射完成,C_{t-1} 内的其他局部簇就不能映射到 C_{jt};此过程继续,直到没有可用的映射。动态社区是这些映射的局部集群。半监督非负矩阵分解算法具体步骤如下:

算法 1 半监督非负矩阵分解算法

输入 动态网络 G、局部聚类数量 $\{k_t\}_{t=1}^T$、局部簇密度阈值 β、先验信息权重 γ

输出 局部集群与演化社区 $\{\boldsymbol{Z}_t\}_{t=1}^T$

//局部社区发现

1. 对于每个时间步长 t,根据式(26)构造 A_t^*;

2. 利用谱聚类算法获取局部信息 Z_t^{rc};

3. 根据式(27)构造划分矩阵 A_t^*;

4. 根据式(28)、式(29),利用 NMF 获得局部聚类发现 $\{Z_t\}_{t=1}^T$;

//改进社区发现

5. 利用局部聚类 $\{Z_t\}_{t=1}^T$,构造 T 划分图;

6. 利用贪心搜索算法发现改进社区;

7. 返回查询结果

3.3 参数选择

本文算法包含 3 个参数:k_t,β 和 γ,其中,k_t 是网络 G_t 中的局部簇数量,β 是先验信息的密度阈值,γ 是控制先验信息的重要性占比。

在聚类中确定簇数目的方法有很多,例如:在谱聚类算法中,使用两个连续特征值之间的间隙来选择适当的 k_t 参数,当网络规模较小时,容易获得较大的间隙,

www.ecice06.com 计算机工程

当网络规模较大时,由于难以定义阈值,因此难以找到间隙;目标函数引导策略也是常用方法,其中 k_t 是与给定目标函数(例如模块化 Q)的最大(最小)值相对应的值。本文在矩阵谱理论的基础上选择 k_t 取值。给定网络 G_t,邻接矩阵 $\boldsymbol{A}_t$ 的谱是特征值 $\lambda_{1t},\lambda_{2t},\cdots,\lambda_{nt}$ 以及相应的特征向量 $x_{1t},x_{2t},\cdots,x_{nt}$ 的集合。

不失一般性,假设 $\lambda_{1t}\geqslant\lambda_{2t}\geqslant\cdots\geqslant\lambda_{nt}$。矩阵可以在特征值和特征向量的基础上重建,即:

$$\boldsymbol{A}_t=\sum_{i=1}^{n}\lambda_{it}x_{it}x'_{it} \tag{28}$$

选择 k_t 作为最小值 k,可得:

$$k_t=\arg\min_k\sqrt{\|\sum_{i=1}^{k}\lambda_{it}x_{it}x'_{it}\|/\|\boldsymbol{A}_t\|}>\delta \tag{29}$$

其中,δ 是控制近似值的参数。一方面,随着特征向量个数的增加,$\boldsymbol{A}_t$ 和局部解之间的误差(即 $\|\boldsymbol{A}_t-\sum_{i=1}^{k}\lambda_{it}x_{it}x'_{it}\|^2$)减小。另一方面,随着特征值个数的增加,特征是冗余的。因此,对 δ 来说,用相当少的特征值来实现合理、近似的值,可使结果趋向于良好的方向。设置 δ、γ 和 β 作为调谐参数,并根据经验进行选取。

3.4　计算复杂度分析

对本文算法的空间复杂度进行分析,在给定动态网络 G 的情况下,三维邻接矩阵 $\boldsymbol{A}_{n\times n\times T}$ 需要的空间复杂度为 $O(n^2T)$。在给定邻接矩阵的情况下,指标矩阵的空间复杂度为 $O(nk_t)$。分解矩阵 $\boldsymbol{A}$ 所需的总计算复杂度为:

$$O(n^2T)O(n(k_1+k_2+\cdots+k_n))=O(nT\max\{k_1,k_2,\cdots,k_n\}) \tag{30}$$

因为 $k_i\leqslant n(1\leqslant i\leqslant T)$,所以式(30)的计算复杂度为 $O(n^2T)$。在社区映射过程中,T 划分图的空间复杂度是 $O(T(\max_i,k_i)^2)$。因此,总的空间复杂度为 $O(n^2T)$,从而证明本文算法在空间复杂度上的计算是有效的。

然后,对时间复杂度进行分析。对于每个时间 t,本文算法包含 3 个主要组成部分:采用谱聚类算法的先验信息构造、局部聚类矩阵分解以及模块映射过程。谱聚类的时间复杂度为 $O(n^3)$。在时间 t 上,局部簇的非负矩阵分解的时间复杂度为 $O(rn^2k_t)$,其中 r 是迭代次数。局部聚类映射过程的时间复杂度为 $O(k_{t-1}k_t)$。因为 $k_i<n$,所以每个时间步长的时间复杂度为:

$$O(n^3+rn^2k_t+k_{t-1}k_t)=O(n^3+rn^2k_t) \tag{31}$$

本文算法的整体时间复杂度为 $O(T(n^3+rn^2\max\{k_t\}_{t=1}^{T}))$,分析结果显示,本文算法的时间复杂度与 INMF 相当。

4　实验结果与分析

4.1　实验环境设置

在实验中,选取的实验模型分为两组:1)具有静态属性的社区发现模型;2)具有动态属性的社区发现模型。实验平台的硬件情况:CPU-i5-6200k,内存大小为 16 GB ddr4-2400,系统为 Win7 旗舰版。

实验模型的具体选取情况如下:1)具有静态属性的社区发现模型,主要有微博网络(Polblogs)、海豚网络(Dolphins)、新陈代谢网络(Cel)、爵士乐网络(Jazz)、邮箱网络(Email)、足球社网络(Football)、空手道网络(Karate);2)具有动态属性的社区发现模型,主要有 arXive-print 网络和 Enron Email 网络。

选取的对比实验算法为 CNM 算法、GN 算法、LPA 算法、Spin-glass 算法、Similarity 算法、NEC 算法、QCA 算法(以上算法相关定义见文献[15])及 MIEN 算法[16]。实验包含 3 个主要过程:静态社区发现模型有效性验证,静态社区发现模型可视化验证,动态社区发现模型性能验证。

4.2　静态社区发现模型有效性验证

为验证不同算法的社区划分性能,在此选取模块度以及社区划分数量作为实验对比指标,选取的对比实验算法在上述具有静态属性的社区发现模型上的性能测试结果如表 1 所示。

表 1　具有静态属性的社区发现模型上的性能测试结果

实验对象	节点数/边数	模块度/社区数						
		GN 算法	CNM 算法	LPA 算法	NEC 算法	Similarity 算法	Spin-glass 算法	本文算法
Karate	33/77	0.37/2	0.39/4	0.38/4	0.37/5	0.42/5	0.41/4	0.42/4
Dolphins	62/158	0.51/3	0.51/4	0.49/5	0.51/6	0.52/5	0.52/5	0.53/4
Jazz	197/2 741	0.43/3	0.40/35	0.35/3	0.28/6	0.43/5	0.43/5	0.43/4
Football	114/612	0.54/5	0.62/16	0.45/15	0.58/7	0.62/6	0.62/10	0.64/11
Polblogs	1 221/16 713	0.42/9	0.45/207	0.27/66	0.43/8	0.36/7	0.42/16	0.45/2
Cel	452/2 024	0.38/8	0.47/38	0.32/27	0.33/6	0.37/6	0.43/10	0.46/7
Email	1 132/5 450	0.48/15	0.58/63	0.42/48	0.01/6	0.48/6	0.55/12	0.57/6

www.ecice06.com 计算机工程

在表 1 实验对比数据中,除了本文算法是通过本文实验结果获得外,其他对比算法均来自于文献[15]。根据表 1 实验结果可知,Spin-glass 算法和本文考虑聚类质量的半监督 INMF 动态社区检测算法在所选取的静态社区发现模型上,均可获得较为理想的社区发现结果。从模块度指标的评估结果看,Spin-glass 算法的指标分布区间是 0.41 ~ 0.62,本文算法的社区发现结果的指标分布区间是 0.42 ~ 0.64。CNM 算法、GN 算法、LPA 算法、Similarity 算法、NEC 算法的指标分布区间均低于 Spin-glass 算法和本文算法。上述实验结果验证了本文算法在社区发现模块度指标上的性能优势。

4.3 静态社区发现模型可视化验证

在可视化实验验证阶段,因为受到文章篇幅的限制,仅选取静态属性的社区发现模型中的 Karate、Dolphins 和 Jazz 3 种社区发现模型进行可视化实验验证。同时,选取这 3 种社区发现模型的另一个原因是其模型中的节点规模数量相对较少,如果节点规模数量过大,则不利于直观显示实验结果。可视化实验验证结果如图 4 ~ 图 6 所示。

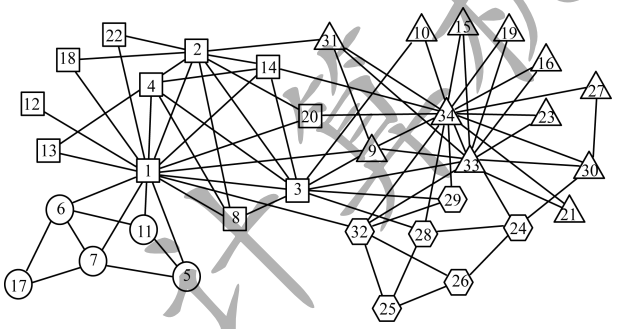


图 4 Karate 模型社区划分结果

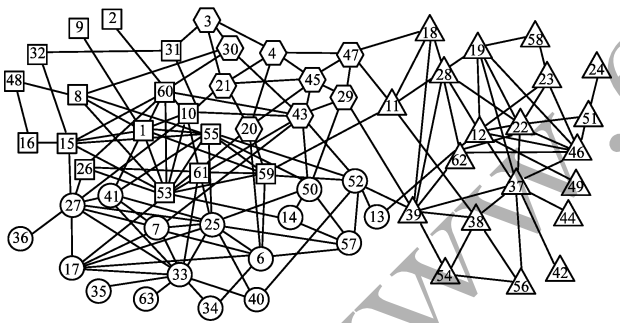


图 5 Dolphins 模型社区划分结果

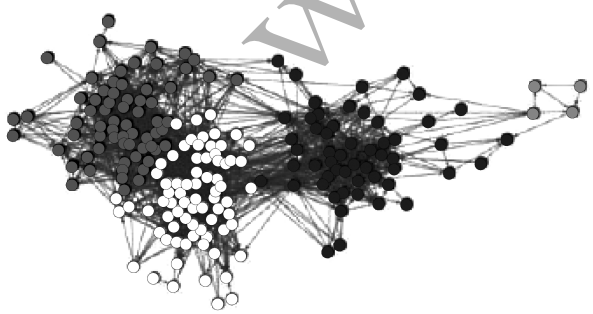


图 6 Jazz 模型社区划分结果

在上述实验模型中,Karate 模型是针对俱乐部内部成员之间的社会关系构建的社区发现模型,成员模型的数量是 34,也就是其节点是 34 个。如图 4所示的 Karate 模型中,节点 1 是俱乐部中的教练角色,其余节点是俱乐部中的经理角色,相互之间存在冲突,且具有不可调和性,进而影响俱乐部内部成员之间的关系,存在派系问题。在本文算法的可视化社区检测结果中,共可得到 Karate 模型的 2 个派系 4 个社区检测结果,该划分结果同原始模型构建过程中的真实情况完全吻合。Dolphins 模型主要是为研究海豚群体而建立的社区分析模型,该模型的原始对象由2 个不同的海豚家族构成,模型的构建节点数量是 62。图 6 为本文算法对 Dolphins 模型的社区划分结果,可视化实验结果显示,本文算法可将海豚群体划分为 4 个社区检测结果,该划分结果同原始模型构建过程中的真实情况完全吻合。Jazz 模型对象主要是为研究爵士乐队社会关系构建的社区发现模型,本文算法的划分结果如图 5所示,因为其中所涉及的节点数量过多,这里并未对其节点的顺序号进行一一标注,根据图 6的 Jazz 模型的社区发现结果可知,本文算法可将 Jazz 模型划分成 4 个社区,该划分结果同原始模型构建过程中的真实情况完全吻合。上述实验结果显示,本文算法在选取的 Karate、Dolphins 和 Jazz 这 3 种静态社区检测模型上可获得非常精确的社区划分结果。

4.4 动态社区发现模型性能验证

本文算法主要考虑的是针对社区划分质量的动态社区发现问题,为验证本文算法的有效性,在此选取 QCA 算法[15]、MIEN 算法[16]进行算法性能测试。动态社区划分实验对象选取 Enron Email 网络。实验过程中的硬件配置同上,测试指标使用的是社区检测算法获得的社区数量、模块度指标以及计算时间 3 组评价结果。实验对比数据如图 7 所示。

Enron Email 模型主要是针对 Enron 公司中高管之间的邮件关系进行设计,其考虑的高管节点数量是 151。选取网络模型中一半的边建立实验模型,此外构建 21 组具有静态属性的模型结构,其中边缘的增加速度是 1 000。根据实验对比结果,本文算法相对于 QCA 动态社区发现算法和 MIEN 动态社区发现算法,在社区数量、模块度指标以及计算时间 3 组评价指标上具有更加优异的性能表现。

综合上述实验结果可知,本文算法相对于选取的对比实验算法在社区检测质量和社区检测效率上性能更优,验证了本文算法在动态社区发现上的性能优势。

计算机工程 www.ecice06.com

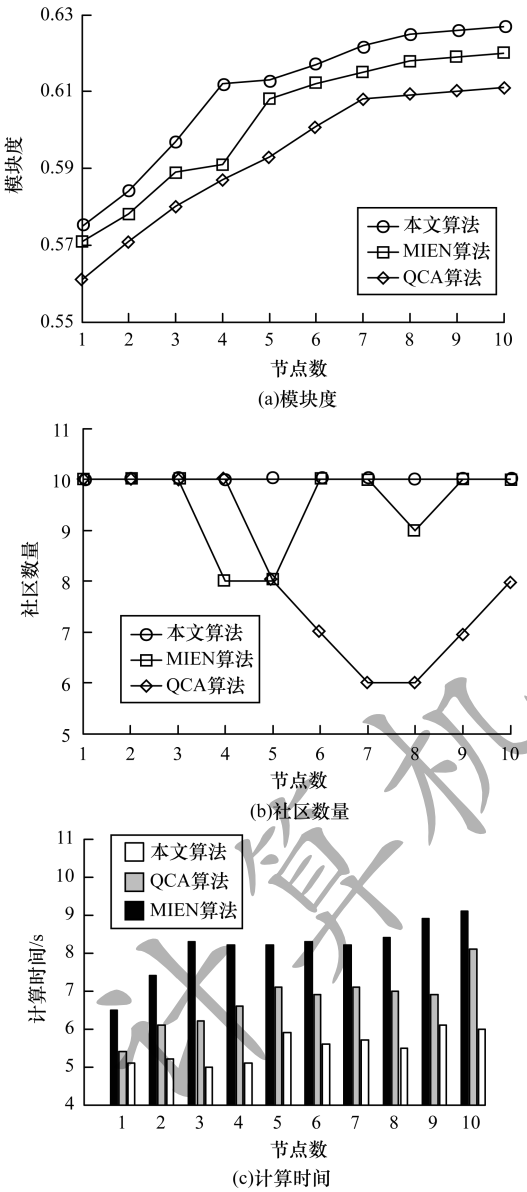


图 7　Enron Email 模型社区划分结果

5　结束语

本文为推导改进聚类算法之间的理论关系,证明了模块密度优化、演化谱聚类和 INMF 之间的等价性,从而扩展演化谱聚类和局部聚类之间的等价性。结合 INMF 和谱聚类,提出半监督 INMF 算法,并将其引入到 INMF 目标函数中。在不增加时间复杂度的情况下,通过使用人工构造和真实世界的动态网络,证明了该算法的性能优势。但由于半监督 INMF 算法前一时间步社区中的错误会被传递到当前时间步的社区检测中,从而影响算法准确性,因此下一步将对此进行研究及改进。

参考文献

[1] 刘海洋,王志海,黄丹,等. 基于评分矩阵局部低秩假设的成列协同排名算法[J]. 软件学报,2015,26(11):2981-2993.

[2] LIU Tongliang, TAO Dacheng. On the performance of manhattan nonnegative matrix factorization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016,27(9):1851-1863.

[3] 邹丹,窦勇,郭松. 基于 GPU 的稀疏矩阵 Cholesky 分解[J]. 计算机学报,2014,37(7):1445-1454.

[4] LI Jun, JOSÉ M, PLAZA A, et al. Robust collaborative nonnegative matrix factorization for hyperspectral unmixing[J]. IEEE Transactions on Geoscience and Remote Sensing,2016,54(10):6076-6090.

[5] WANG S S, CHERN A, TSAO Y. Wavelet speech enhancement based on nonnegative matrix factorization[J]. IEEE Signal Processing Letters,2016,23(8):1101-1105.

[6] KITAMURA D, ONO N, SAWADA H, et al. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing,2016,24(9):1626-1641.

[7] ZHANG Xianchao, ZONG Linlin, LIU Xinyue, et al. Constrained clustering with nonnegative matrix factorization[J]. IEEE Transactions on Neural Networks and Learning Systems,2016,27(7):1514-1526.

[8] SHI Lei, ZHANG Lefei, ZHAO Lingli, et al. Adaptive laplacian eigenmap-based dimension reduction for ocean target discrimination[J]. IEEE Geoscience and Remote Sensing Letters,2016,13(7):902-906.

[9] VENKATARAMAN A, YANG D Y J, PELPHREY K A, et al. Bayesian community detection in the space of group-level functional differences[J]. IEEE Transactions on Medical Imaging,2016,35(8):1866-1862.

[10] NEWMAN M. Detecting community structure innetworks[J]. European Physical Journal B,2004,38(2):321-330.

[11] SAMMARCO M, CAMPISTA M E M, AMORIM M D. Scalable wireless traffic capture through community detection and trace similarity[J]. IEEE Transactions on Mobile Computing,2016,15(7):1757-1769.

[12] 罗会兰,万成涛,孔繁胜. 基于 KL 散度及多尺度融合的显著性区域检测算法[J]. 电子与信息学报,2016,38(7):1594-1601.

[13] 李艳雄,吴水,贺前华. 基于特征均值距离的短语音段说话人聚类算法[J]. 电子与信息学报,2012,34(6):1404-1407.

[14] KIM Y D, CHOI S. Variational Bayesian view of weighted trace norm regularization for matrix factorization[J]. IEEE Signal Processing Letters,2013,20(3):261-264.

[15] SAMMARCO M, CAMPISTA M E M, AMORIM M D. Scalable wireless traffic capture through community detection and trace similarity[J]. IEEE Transactions on Mobile Computing,2016,15(7):1757-1769.

[16] KARIMI-MAJD A M, FATHIAN M, AMIRI B. A hybrid artificial immune network for detecting communities in complex networks[J]. Computing, 2015,97(2):483-507.

编辑　陆燕菲

www.ecice06.com 计算机工程