文章编号: 1001-0920 (2014) 04-0713-06 DOI: 10.13195/j.kzyjc.2012.1730

网络重叠社区发现的谱聚类集成算法

黄发良[1], 黄名选[2], 元昌安[3], 姚志强[1]

(1. 福建师范大学 软件学院，福州 350007；2. 广西教育学院 科研处，南宁 530023；3. 科学计算与智能信息处理广西高校重点实验室，南宁 530023)

摘 要: 鉴于计算代价高昂的谱聚类无法满足海量网络社区发现的需求, 提出一种用于网络重叠社区发现的谱聚类集成算法 (SCEA). 首先, 利用高效的近似谱聚类 (KASP) 算法生成个体聚类集合; 然后, 引入个体聚类选择机制对个体聚类进行优选, 并对优选后的个体聚类建立簇相似图; 最后, 进行层次软聚类, 得到网络节点的软划分. 实验结果表明, 与代表性算法 (CPM, Link, COPRA, SSDE) 相比较, SCEA 能够挖掘出具有更高规范化互信息 (NMI) 的网络重叠社区结构, 且具有相对较好的鲁棒性.

关键词: 重叠社区发现；谱聚类；集成聚类

中图分类号: TP273 **文献标志码:** A

Spectral clustering ensemble algorithm for discovering overlapping communities in social networks

HUANG Fa-liang[1], *HUANG Ming-xuan*[2], *YUAN Chang-an*[3], *YAO Zhi-qiang*[1]

(1. Faculty of Software，Fujian Normal University，Fuzhou 350007，China；2. Scientific Research Office，Guangxi College of Education，Nanning 530023，China；3. Science Computing and Intelligent Information Processing of Guangxi Higher Education Key Laboratory，Nanning 530023，China. Correspondent：HUANG Fa-liang，E-mail: huangfliang@163.com)

Abstract: Considering that spectral clustering algorithms are unable to efficiently discover communities in massive networks for the high computation cost, a spectral clustering ensemble algorithm(SCEA) is proposed. Firstly, the effective KASP algorithm is used to produce clusters. Then the better clusters are chosen to construct a cluster ensemble, and the similarity graph for individual cluster is created. Finally, the resultant overlapping communities with hierarchical soft clustering are obtained. Experimental results show that, in contrast with some typical algorithms such as clique percolation method(CPM), Link, community overlap propagation algorithm(COPRA) and sampled spectral distance embedding(SSDE), the SCEA can discover the network communities with higher normal mutual information(NMI), and it exhibits satisfactory robustness.

Key words: overlapping community discovery；spectral clustering；ensemble clustering

0 引 言

网络社区结构可描述复杂系统内部构件之间的结构与功能的关联特征, 在各种不同复杂系统中发现社区结构知识具有重要的理论意义与应用价值. 近年来, 关于网络社区发现的研究备受关注, 涌现出大量算法[1]. 其中绝大多数方法都是硬划分算法, 即将网络节点集合划分为若干个互不相交的社区. 然而, 现实中网络社区之间往往是交错关联与相互重叠的, 例如, 同一个研究人员感兴趣的研究领域往往不只一个, 而是对相近或相关的若干个研究领域感兴趣, 因此科研合作网络的社区可能大都是相互重叠的. 显然, 硬划分的社区发现算法无法满足此需求. 为此, 人们陆续提出一系列方法来挖掘网络重叠社区, 如派系过滤算法 (CPM)[2], 社区重叠传播算法 (COPRA)[3], 基于最大完全图的层次凝聚聚类算法 (EAGLE)[4], 基于线图与粒子群优化技术的网络重叠社区发现算法

收稿日期: 2012-11-19；**修回日期:** 2013-06-12.

基金项目: 国家自然科学基金项目(61262028)；教育部人文社会科学研究青年基金项目(12YJCZH074)；福建省自然科学基金项目(2011J01339)；广西自然科学基金项目(2012GXNSFAA053235)；科学计算与智能信息处理广西高校重点实验室开放基金项目(GXSCIIP201212)；福建师范大学优秀青年骨干教师培养基金项目(fjsdjk2012082)；福建省教育厅科技项目 (JA13077).

作者简介: 黄发良(1975−), 男, 副教授, 博士, 从事知识发现与智能计算的研究；黄名选(1966−), 男, 教授, 从事知识发现与数据挖掘等研究.

(LGPSO)[5]以及基于Link的算法[6].

实用性很强的谱聚类[7]引起了重叠社区发现研究人员的广泛关注. 文献[8]在NG模块度函数的基础上, 定义了重叠社区模块度度量函数, 通过对谱嵌入特征空间中的行向量进行模糊k均值聚类来实现重叠社区结构的发现; 文献[9]提出了利用基于随机漫步的扩展策略来发现重叠社区的方法, 其中扩展的种子节点集是网络节点谱聚类的聚类结果; 文献[10]提出了SSDE算法, 在谱嵌入的基础上运用高斯混合模型(GMM)算法对网络节点实行软划分, 从而达到发现重叠社区的目的. 尽管优势颇多的谱聚类为网络重叠社区发现的研究提供了一个强有力的方法, 但也存在一些缺陷, 如计算量较大、构造相似性矩阵时对尺度参数敏感等. 值得指出的是, 文献[11]提出了一种谱聚类集成算法, 但与本文不同的是, 该算法是应用于遥感图像分割的.

本文提出一种谱聚类集成算法(SCEA), 并将其用于网络重叠社区发现, 该算法主要分为产生个体聚类、选择个体聚类和集成个体聚类3个基本步骤. 第1步的主要任务是选取快速的近似谱聚类(KASP)算法[12], 产生聚类集成所需的聚类个体集合, 这既可有效缓解谱聚类计算代价高的问题, 又能避免精确选择谱聚类的尺度参数; 第2步是对第1步中产生的个体聚类进行选取; 第3步是将选取的各个体聚类进行综合, 得到最终的集成结果. 最后, 通过大量的实验表明, SCEA具有较好的有效性和稳定性.

1 研究背景

1.1 谱聚类算法

谱聚类算法是建立在代数图论基础之上的, 其基本思想为: 首先, 采用某种相似度策略建立起原始数据集隐含的数据对象相似关系图; 然后, 利用谱映射对此相似图进行子空间学习形成低维表示空间; 最后, 在低维空间中采用k-means等算法进行后聚类学习.

1.2 聚类集成

聚类集成是集成学习思想在无监督学习领域中的应用, 其核心就是将同一数据集的不同聚类合并成一个比单个聚类更能刻画数据真实分布的结果聚类. 研究表明, 与单一聚类算法相比, 聚类集成具有更好的鲁棒性、适用性、稳定性、并行性及扩展性[13]. 聚类集成主要涉及到3个关键问题. 一是如何高效地产生准确性高且多样性丰富的个体聚类集合, 当前主要的方法有: 1) 采用具有不同参数的同种聚类算法生成个体聚类; 2) 采用不同的聚类算法来生成个体聚类; 3) 利用由bootstrap或特征选择技术产生的数据样本随机子空间来生成个体聚类. 二是根据什么标准从个体聚类集合中选取个体聚类进行聚合操作. 三是如何设计一致性函数来指导聚类合成.

2 谱聚类集成算法

2.1 个体聚类的产生

本文选取KASP算法来生成个体聚类, 这主要出于两个方面的考虑: 1) 复杂网络的海量性(网络规模达到百万、千万甚至更大的数量级)使诸如SM[14]之类的具有高昂计算代价(常为$O(n^3)$)的传统谱算法无法有效完成网络节点的聚类; 2) 虽然该策略也可以很大程度地降低谱聚类的计算成本, 但利用Nystrom降秩策略[15]并不能显性给出数据约简量与聚类准确率之间的直接关联. KASP算法的基本思路是: 首先, 利用时间复杂度相对较低的聚类算法k-means对网络节点进行预聚类; 然后, 对k-means的结果簇质心进行谱聚类; 最后, 根据k-means聚类产生的网络节点簇标记与谱聚类簇质心的社区标签形成最终的网络社区结构. 这样会大大提升算法的时间效率, 其提升的程度大小取决于算法中的压缩比参数$\beta = n/k$.

2.2 个体聚类的选取

聚类集成学习的传统框架隐含着两个缺陷: 1) 缺乏个体聚类选择的操作, 直接将所有个体聚类都作为成员聚类, 采用平等对待的方式进行集成; 2) 简单地认为个体聚类的多样性与聚类结果准确率成正比. 然而, 研究结果表明[16-17], 与直接集成所有个体聚类相比较, 集成个体聚类集合中的某个优选子集能获得质量更高的聚类结果, 并且聚类准确率与个体聚类多样性并非是简单的正相关关系, 需要由数据样本的分布特性来决定其二者的相关性.

考虑到复杂网络数据分布的多样性使得寻求个体聚类集合多样性最大化的选择策略难以满足其社区发现任务的要求, 在文献[16-17]的基础上, 本文提出了一种用于网络社区划分方案的选择算法CCChooser. 该算法的基本思想是: 从个体谱聚类生成的社区划分候选方案集合中, 选出具有最大规范互信息平均值ANMI的社区划分方案作为当前社区划分方案; 然后, 通过一个社区迭代评分的过程持续修正当前社区划分方案; 最后, 以修正后的当前社区划分方案为基准选出待聚合的社区划分方案集合. 该算法基本流程可描述如下:

Step 1: 令$Y = \{Y_1, Y_2, \cdots, Y_\Lambda\}$为个体谱聚类生成的社区划分候选方案集合, 计算$Y$中的每个社区划分候选方案$Y_i = \{C_1, C_2, \cdots, C_k\}$的代表性

$$\mathrm{ANMI}(Y_i) = \frac{1}{\Lambda}\sum_{j=1}^{\Lambda}\mathrm{NMI}(Y_i, Y_j),$$

选择具有最大ANMI的社区划分方案作为当前社区划分方案C^*, 其中$\mathrm{NMI}(Y_i, Y_j)$表示社区划分候选方案Y_i与Y_j的规范化互信息.

Step 2: 将所有社区划分候选方案中的社区组织成社区集合, 并给出其中社区随机初始评分V.

Step 3: 对社区集合中的社区i进行阈值为α的随机概率选择, 若没有选中, 则从社区集合中选取具有最大评分V的社区, 并以此社区对当前社区划分方案进行high-voting投票, 形成新的社区划分方案C^{new}, 并重新计算此方案的ANMI, 记为$\mathrm{ANMI}_{\mathrm{new}}$, 计算选中当前社区的回报$R = \mathrm{ANMI}_{\mathrm{new}} - \mathrm{ANMI}$, 更新此方案的$\mathrm{ANMI} = \mathrm{ANMI}_{\mathrm{new}}$, 并且更新该社区的评分$V_i^{\mathrm{new}} = \alpha R + (1-\alpha)V_i^{\mathrm{old}}$. 迭代以上过程, 直到评分收敛为止.

Step 4: 计算所有社区划分候选方案与C^{new}的NMI, 并按NMI降序排列社区划分候选方案, 从中选择前$\lceil \Lambda/2 \rceil$个方案构成待聚合的社区划分方案集合ToBeFused并返回.

假定网络的节点数为n, 则算法CCChooser的时间复杂度可估算如下: 对于Step 1, 其操作是计算Λ个社区划分候选方案的ANMI, 由于一对社区划分候选方案NMI所需的时间为$O(n^2)$, Step 1的时间复杂度为$O(\Lambda^2 n^2)$; 对于Step 2, 令平均每个社区划分方案包含s个社区, 则其时间复杂度为$O(\Lambda s)$; 对于Step 3, 令评分收敛平均所需迭代次数为T, 则该迭代评分过程的复杂度为$O(T\Lambda n^2)$; 对于Step 4, 计算所有社区划分候选方案与C^{new}的NMI所需的时间为$O(\Lambda n^2)$, 对Λ个社区划分候选方案的NMI进行排序所需的时间为$O(\Lambda \log \Lambda)$, 从而可得Step 4的时间复杂度为$O(\Lambda n^2)$. 综上所述, 算法CCChooser的时间复杂度为$O(\max(T\Lambda n^2, \Lambda^2 n^2))$.

2.3　个体聚类的集成

本节根据选取的各个体聚类所包含簇的关系建立簇相似图, 然后通过层次软聚类来实现聚类集成.

簇相似图的节点与加权边分别对应于个体聚类中的簇与各簇之间的相似关系, 假定从个体聚类集合中选取的λ个个体聚类分别为$M_1, M_2, \cdots, M_\lambda$, 个体聚类$M_i$含有$K_i$个簇, 其对应的簇隶属矩阵为

$$M_i = (M_{i1}, M_{i2}, \cdots, M_{iK_i}) = \begin{bmatrix} d_{11} & \cdots & d_{1K_i} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nK_i} \end{bmatrix}.$$

其中: M_{ik}表示第i个个体聚类中的第k个簇; $d_{jk}(j = 1, 2, \cdots, n,\ k = 1, 2, \cdots, K_j)$的取值为: 若第$j$个数据点属于第$k$个簇, 则$d_{jk} = 1$, 反之$d_{jk} = 0$. 节点间的Jaccard相似度为

$$\mathrm{sim}(M_{ik}, M_{rs}) = \frac{P_{11}}{P_{00} + P_{10} + P_{01}}.$$

其中: P_{11}表示属于簇M_{ik}且属于M_{rs}的数据计数, P_{00}表示既不属于簇M_{ik}也不属于M_{rs}的数据计数, P_{10}表示属于簇M_{ik}但不属于M_{rs}的数据计数, P_{01}表示不属于簇M_{ik}但属于M_{rs}的数据计数. 最终的簇相似图用矩阵L表示, 其元素$L(i, j)$为K个簇中的簇i与簇j的相似度, 其中$K = K_1 + K_2 + \cdots + K_\lambda$为待集成的个体聚类所含有的簇总数.

由于在簇相似图的层次聚类中存在两类簇: 一类是聚类对象, 即个体聚类中的簇; 另一类是在该算法执行过程中生成的处于不同层次的簇. 为了区分二者, 本文将前者称为原始簇, 将后者称为生成簇.

借鉴min-max图划分策略[18]可以给出生成簇相似度的计算方法: 对于生成簇GM_1与GM_2, 其相似度为

$$\mathrm{sim}(\mathrm{GM}_1, \mathrm{GM}_2) = \frac{s(\mathrm{GM}_1, \mathrm{GM}_2)}{s(\mathrm{GM}_1, \mathrm{GM}_1) * s(\mathrm{GM}_2, \mathrm{GM}_2)},$$

其中

$$s(\mathrm{GM}_1, \mathrm{GM}_2) = \sum_{M_i \in \mathrm{GM}_1} \sum_{M_j \in \mathrm{GM}_2} L(i, j).$$

关于聚类树截断点的最优选取问题, 常见方法是由用户指定相似度阈值或最终簇数, 而对原始数据分布缺乏先验知识的用户很难给出合理的阈值或簇数. 又考虑到本文要解决的问题是发现重叠社区, 基于此, 借鉴文献[4]中的重叠社区结构质量测度指标Q_{ov}值, 本文提出了一种用于最优化聚类树截断点的测度指标—— 截断适应度

$$\mathrm{CF}(T) = \frac{1}{2m} \sum_{c \in P_T} \sum_{i,j} \frac{1}{O_i O_j} \Big(A_{ij} - \frac{k_i k_j}{2m} \Big) \delta_{ic} \delta_{jc}. \quad (1)$$

其中: P_T为聚类结果树的第T层所对应的网络重叠社区; m为网络节点数; k_i为网络中节点i的度; A为网络邻接矩阵, A_{ij}表示节点i与节点j的连接权重; 节点i所属社区的数目为O_i; 若节点i属于社区c, 则函数δ_{ic}的取值为1, 反之取值为0. 由截断适应度的定义可知, 最优截断点为

$$\mathrm{cut} = \arg\max_{\mathrm{level}=1}^{H}(\mathrm{CF}(T)),$$

其中level为聚类结果树的树高.

2.4　算法描述及复杂度分析

综上所述, SCEA的基本流程可描述如下:

Step 1: 调用具有不同参数设置的KASP算法产生Λ个社区划分方案.

Step 2: 调用个体聚类选择算法CCChooser构造待聚合的社区划分方案集合ToBeFused.

Step 3: 构造存在于ToBeFused中社区的相似矩

阵 L.

Step 4: 进行层次软聚类, 生成网络社区划分方案谱系树.

Step 5: 计算网络社区划分方案谱系树的最优截断点, 并将此截断点对应的社区结构作为结果输出.

假定原网络的节点数为 n, 则算法的复杂度可以估计如下: 对于 Step 1, 由于 Λ 个社区划分方案的产生方式可以是并行的也可以是串行的, 考虑到并行计算时间复杂度分析的复杂性, 在此只讨论串行方式, 故 Step 1 的时间复杂度为 $O(\Lambda\cdot\max(k^3,knt))$. 其中: $k=n/\beta$, t 为 KASP 算法中的 k-means 迭代次数. 由第 2.2 节的分析可知, Step 2 的时间复杂度为 $O(\max(T\Lambda n^2,\Lambda^2n^2))$, 其中 T 为评分收敛平均所需迭代次数. Step 3 的时间复杂度可按以下方法估算: 由于集合 ToBeFused 的大小为 $\lceil\Lambda/2\rceil$, 而平均每个社区划分方案包含 s 个社区, 故该步的时间复杂度为 $O((\lceil\Lambda/2\rceil\times s)^2)=O(\Lambda^2s^2)$. 对于 Step 4, 由第 2.3 节的分析可知, 该层次聚类中的生成簇相似度计算方式是类似于 average-linkage 的, 故其时间复杂度为 $O((\lceil\Lambda/2\rceil\times s)^3)=O(\Lambda^3s^3)$, 引入数据结构优先级队列可将此步的时间复杂度降低到

$$O\left(\left(\left\lceil\frac{\Lambda}{2}\right\rceil\times s\right)^2\times\log\left(\left\lceil\frac{\Lambda}{2}\right\rceil\times s\right)\right)=O(\Lambda^2s^2\log(\Lambda s)).$$

对于 Step 5, 其复杂度可按以下方法估算: 若令该谱系树的第 level 层包含 r 个社区, 社区的平均大小为 $[n\alpha/r]$, α 为社区重叠因子 (平均每个节点隶属的社区数在复杂网络中常表现为一个大于 1 的较小常量), 则社区内关联度的计算复杂度为

$$O(r\times(n\alpha/r)^2)=O(n^2\alpha^2/r),$$

社区间关联度的计算复杂度为

$$O(r\times((n\alpha/r)\times(n\alpha/r)\times(r-1)))\approx O(n^2\alpha^2).$$

故 Step 5 的复杂度为 $O(n^2\alpha^2H)$. 综上所述, SCEA 的时间复杂度为

$$O(\max(\Lambda k^3,\Lambda knt,T\Lambda n^2,\Lambda^2n^2,\Lambda^2s^2\log(\Lambda s),n^2\alpha^2H)).$$

3 实 例

本节以红楼梦中重要人物的基本关系网络对算法加以说明, 利用 SCEA 可发现如图 1 所示的社区结构. 从图 1 可以看出: com (史府)$\bigcap$com (荣国府)=[史侯], com (宁国府)$\bigcap$com (荣国府)=[尤二姐, 贾源, 贾演], com (王府)$\bigcap$com (荣国府)=[刘姥姥, 王熙凤, 王夫人], com (薛府)$\bigcap$com(荣国府)=[邢岫烟, 薛宝钗], com (薛府)$\bigcap$com(王府)=[薛姨妈]. 从这些交集节点可以看出, 红楼梦中的 4 大家族主要是通过婚姻关系来形成的社会团体: “尤二姐”嫁给“贾琏”, “王熙凤”嫁给“贾琏”, 以及“王夫人”嫁给“贾政”等. 除此之外, 还有一些有趣的共享节点“刘姥姥” (非婚姻关系的连接点), “贾演”与“贾源” (兄弟关系). SCEA 不仅能发现隐含于红楼梦网络中的最重要关系, 即婚姻纽带关系, 而且可发现难以预料的非婚姻关系连接点.

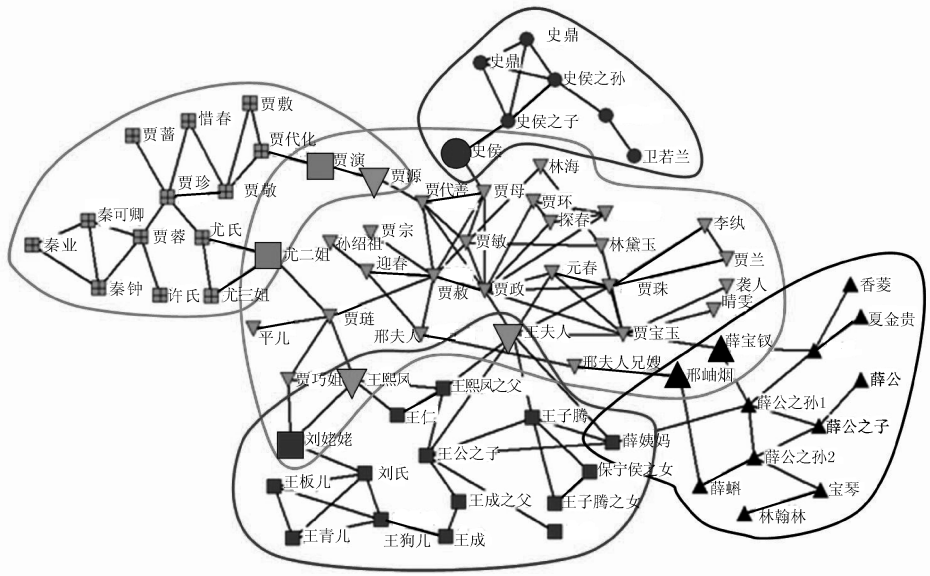


图 1 红楼梦人物关系网络隐含的重叠社区结构

4 实验与讨论

本文利用 3 个真实网络与 3 个人工网络对 SCEA 的性能进行比较分析, 3 个真实网络分别是 Karate 网络, Dolphins 网络与 HLM 网络, 而 3 个人工网络 (SynNet1, SynNet2 与 SynNet3) 是借助网络数据生成器[19]产生的.

4.1 算法有效性分析

为了评价SCEA的聚类有效性, 本文引入NMI[20]评价标准, 通过比较网络真实社区结构与 SCEA 发现的网络社区结构的结构相似性来验证算法的有效性.

将 SCEA 与其他 4 种代表性算法在 6 个数据集上进行实验比较, 结果如表 1 所示. 由表 1 可以看出, 不论是真实网络还是人工网络, 利用 SCEA 所得的社区结构与真实社区相一致的程度远远高于其他算法. 需要说明的是, 表 1 中的算法参数 k (CPM 算法), t(Link 算法) 与 v(COPRA 算法) 分别表示完全图阶数、边相似度阈值与网络节点最多可允许隶属的社区数.

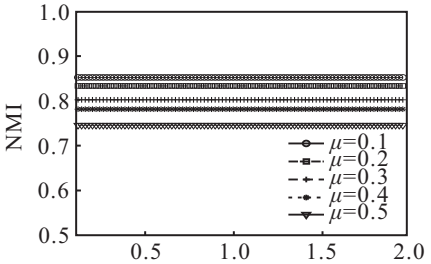
表 1　SCEA 的聚类有效性比较

Network	CPM($k=3$)	Link($t=0.15$)	COPRA($v=2$)	SSDE	SCEA
karate	0.335	0.409	0.347	0.346	0.903
dolphins	0.461	0.209	0.649	0.602	0.824
HLM	0.211	0.331	0.638	0.43	0.861
SynNet1	0.348	0.338	0.684	0.257	0.863
SynNet2	0.288	0.382	0.678	0.284	0.884
SynNet3	0.275	0.329	0.693	0.312	0.892

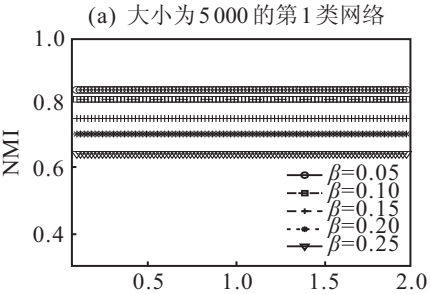
4.2　算法鲁棒性分析

由于尺度参数在传统谱聚类中具有非常重要的作用, 不同的尺度参数值会导致大相径庭的聚类结果. 由于从现实数据中很难获得有关尺度参数选取的先验知识, 这极大地限制了谱聚类的实际应用, 那么作为谱聚类集成算法的 SCEA 是否面临着同样的问题? 为了探究此问题, 本文采用网络数据生成器产生两类具有不同拓扑特征的人工网络: 第 1 类是混合系数 μ 逐步递增改变的网络 (μ = 社区间边数/网络总边数); 第 2 类是重叠节点比率 β 逐步递增改变的网络 (β = 重叠节点数/网络节点总数). 由图 2 可知, 在各个混合系数 (重叠节点比率) 取值不同的网络中, SCEA 的结果社区 NMI 值并没有随着尺度参数逐步递增而发生某种线性或非线性的改变, 而是一个常量值. 尽管较小的混合系数 (重叠节点比率) 会导致较大的 NMI 值, 但各个混合系数对应的 NMI 都处于 0.65 ~ 0.85 之间这样一个较好区间内. 由此可见, 就尺度参数而言, SCEA 在挖掘网络重叠社区上具有很强的鲁棒性. 另外还需指出的是, 混合系数与重叠节点比率对 SCEA 的影响可根据其二者的定义作出这样的解释: 混合系数越大, 也就是社区间边数越大, 意味着网络模块性越低; 而重叠节点比率越大, 社区间重叠程度越高, 意味着网络社区之间分界越模糊; 不论是低模块性还是高模糊性都会增加社区挖掘问题的难度, 从而会降低挖掘算法结果社区的有效性.

SCEA 的时间效率与压缩比参数有着紧密的联系, 那么压缩比与 SCEA 的有效性又有着怎样的关系? 与尺度参数的分析类似, 本文利用网络生成器生成 5 个第 1 类网络与 5 个第 2 类网络, 网络节点数都为 10 000. 由图 3 可以看出, 在各个混合系数 (重叠节点比率) 取值不同的网络中, SCEA 的结果社区 NMI 值并没有随着压缩比的增加而显著降低, 而是出现微小幅度的波动, 处于 0.7 ~ 0.88 这样一个区间中. 由此可见, 就压缩比而言, SCEA 在挖掘网络重叠社区上具有很强的鲁棒性. 特别的, 压缩比的变化会在一定范围内消除混合系数 (重叠节点比率) 的影响. 例如, 在图 3(a) 中, 当压缩比为 4 时, 混合系数 0.4 与 0.5 对应的 NMI 是一样的; 当压缩比为 8 时, 混合系数 0.2 对应的 NMI 成为 5 种取值情况中的最好情形. 类似的现象也存在于图 3(b) 中, 当压缩比为 4 时, 重叠节点比率 0.1 对应的 NMI 成为 5 种取值情况中的最好情形; 当压缩比为 8 时, 重叠节点比率 0.2 与 0.25 对应的 NMI 值相等.

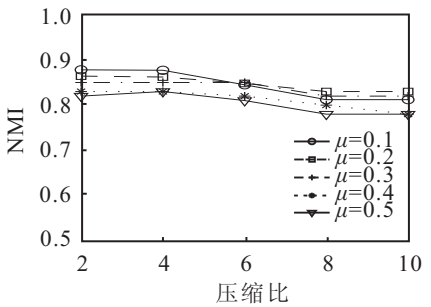


(a) 大小为 5 000 的第 1 类网络

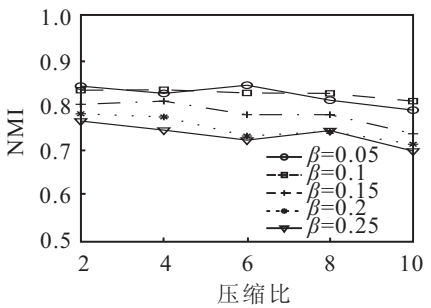


(b) 大小为 5 000 的第 2 类网络

图 2　尺度参数对 SCEA 有效性的影响



(a) 大小为 10 000 的第 1 类网络



(b) 大小为 10 000 的第 2 类网络

图 3　压缩比对 SCEA 有效性的影响

5 结 论

社区模式挖掘是复杂网络研究中的一个重要课题, 其本质是一个网络节点的聚类分析问题, 常用的基于个体谱聚类的社区发现算法都存在着时间复杂度高、构造相似性矩阵时对尺度参数敏感等不足之处, 导致不能有效地发现网络中隐含的真实社区结构. 本文从集成学习的角度出发, 提出了谱聚类集成算法, 并引入高效的谱聚类算法KASP以构造用于网络社区发现的个体聚类器, 借鉴层次软聚类的思想将各个个体聚类器的结果社区结构进行集成, 最后形成网络重叠社区结构. 采用真实网络与人工网络测试本文方法的有效性与鲁棒性, 实验结果显示, 与当前最具代表性的算法(CPM, Link, COPRA 和 SSDE)相比, SCEA 能更加有效地揭示网络的重叠社区结构, 且具有相对较好的鲁棒性.

参考文献(References)

[1] 黄发良. 信息网络的社区发现及其应用研究[J]. 复杂系统与复杂性科学, 2010, 7(1): 64-74.
(Huang F L. Studies on community detection and its application in information network[J]. Complex Systems and Complexity Science, 2010, 7(1): 64-74.)

[2] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[3] Gregory S. Finding overlapping communities in networks by label propagation[J]. New J of Physics, 2010, 12(10): 103018 .

[4] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and Its Applications, 2008, 388(8): 1706-1712

[5] 黄发良, 肖南峰. 基于线图与PSO的网络重叠社区发现[J]. 自动化学报, 2011, 37(9): 1140-1144
(Huang F L, Xiao N F. Discovering overlapping communities based on line graph and PSO[J]. Acta Automatica Sinica, 2011, 37(9): 1140-1144.)

[6] Ahn Y-Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature, 2010, 466(7307): 761-764.

[7] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[J]. Advances in Neural Information Processing Systems, 2002, 2: 849-856.

[8] Zhang S, Wang R S, Zhang X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering[J]. Physica A, 2007, 374(1): 483-490.

[9] Wei F, Qian W, Wang C, et al. Detecting overlapping community structures in networks[J]. World Wide Web, 2009, 12(2): 235-261.

[10] Magdon-Ismail M, Purnell J. Fast overlapping clustering of networks using sampled spectral distance embedding and GMMs[R]. New York: Department of Computer Science Rensselaer Polytechnic Institute, 2010.

[11] Zhang X, Jiao L, Liu F, et al. Spectral clustering ensemble applied to SAR image segmentation[J]. IEEE Trans on Geoscience and Remote Sensing, 2008, 46(7): 2126-2136.

[12] Yan D, Huang L, Jordan M I. Fast approximate spectral clustering[C]. Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. Paris, 2009: 907-916.

[13] Topch Y A, Jain A K. Clustering ensembles: Models of consensus and weak partitions[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881.

[14] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[15] Fowlkes C, Belongie S, Chung F, et al. Spectral grouping using the nystrom method[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214-225.

[16] Tumer K, Agogino A K. Ensemble clustering with voting active clusters[J]. Pattern Recognition Letters, 2008, 29(14): 1947-1953

[17] Azimi J, Fern X. Adaptive cluster ensemble selection[C]. Proc of the 21st Int Jont Conf on Artifical Intelligence. Pasadena, 2009: 992-997.

[18] Ding C, He X. Cluster merging and splitting in hierarchical clustering algorithms[C]. Proc of IEEE Int Conf on Data Mining. Maebashi, 2002: 139-146.

[19] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008, 78(4): 046110.

[20] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. J of Statistical Mechanics: Theory and Experiment, 2009, (3): 1-22.

(责任编辑: 滕 蓉)