

基于谱聚类的复杂网络社团发现算法

蔡晓妍 戴冠中 杨黎斌

(西北工业大学自动化学院 西安 710072)

摘要 复杂网络社团发现的研究对于控制疾病传播、网络病毒的传播等具有重大意义。针对已有社团发现算法时间复杂度过高,不适用于结构未知的大型网络等问题,结合谱聚类在识别未知分布数据集聚类方面的优势,以及模块度函数能够在大型网络中搜寻出最佳社团数目的能力,提出了基于谱聚类的社团发现算法——SCCF算法。实验结果表明,与已有的社团发现算法相比,SCCF算法效率更高,并且能够在网络节点数上万的大型网络中得到高质量的社团结构。

关键词 复杂网络,社团结构,谱聚类,模块度

Community-finding Algorithm in Complex Networks Based on Spectral Clustering

CAI Xiao-yan DAI Guan-zhong YANG Li-bin

(College of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract Research on community finding is very helpful to control virus spreading in networks. Most of the proposed community-finding algorithms are not suitable for very large networks because of their time-complexity. Combined with the advantage of solving the clustering of unknown distributed data set of the spectral clustering, and the ability of modularity function in finding good community number in large networks, a community-finding algorithm based on spectral clustering was proposed. Experimental results indicate that the new algorithm is efficient and effective at finding good community structure in large networks.

Keywords Complex networks, Community structure, Spectral clustering, Modularity

1 引言

随着对网络性质的物理意义和数学特性的深入研究,人们发现许多实际网络都存在一个共同性质——社团结构,即整个网络由若干个社团构成,每个社团内部节点之间的连接非常紧密,而各个社团之间的连接相对来说比较稀疏^[1]。研究复杂网络的社团结构对于控制疾病传播、网络病毒的传播等具有重大意义。

研究者们已经提出了两类检测网络中社团的方法——计算机科学中的图分割^[2]和社会学中的分层聚类^[3]。图分割方法大多采用二分法,缺点在于难以把网络分解为合理个数的社团;而分层聚类方法对外围节点的分类经常出错,不适合结构未知的复杂网络。

为了解决上述问题,本文结合谱聚类算法在识别未知分布数据集聚类方面的优势,以及模块度函数在衡量社团结构方面的技术基础,提出了一种基于谱聚类的社团发现算法——SCCF算法,并对其进行了时间复杂度分析。实验结果表明,与已有的社团发现算法相比,SCCF算法效率更高,并且能够在节点数上万且结构未知的大型网络中得到高质量的社团结构。

2 谱聚类算法

谱聚类算法的思想来源于谱图划分理论^[4],它将聚类问题看成是一个无向图的多路划分问题。假设一无向加权图 $G(V, E, W)$, 其表现形式为一对称矩阵 $W = [w_{ij}]_{n \times n}$, 其中 n 表示图 G 所含的节点数, w_{ij} 表示连接顶点 i 与 j 的权值, D 为对角阵 $D_{ii} = \sum_j w_{ij}$ 。定义一个图划分判据, 如 Shi 和 Malik^[5] 提出的 2-way 目标函数 $Ncut$:

$$Ncut(A, B) = \frac{cut(A, B)}{vol(A)} + \frac{cut(A, B)}{vol(B)} \quad (1)$$

$$vol(A) = \sum_{i \in A} \sum_j w_{ij} \quad (2)$$

$$cut(A, B) = \sum_{i \in A, j \in B} w_{ij} \quad (3)$$

$cut(A, B)$ 是子图 A, B 间的边, 又称“边切集”。通过最小化 $Ncut$ 函数可使同一类内的点具有较高的相似性, 而不同类之间的点则相似性较低。

考虑约束条件 $x^T W e = x^T D e = 0$, 则

$$\min Ncut(A, B) = \min \frac{x^T (D - W) x}{x^T D x} \quad (4)$$

将 x 松弛到连续域 $[-1, 1]$, 求解 $\min Ncut(A, B)$ 的问题就可以转化为下式:

到稿日期:2008-10-15 返修日期:2008-12-30 本文受国家 863 计划项目(2005AA147030)资助。

蔡晓妍 博士生, 主要研究方向为智能信息处理、网络与信息安全, E-mail: caixiaoyan1982@hotmail.com; 戴冠中 教授, 博士生导师, 主要研究方向为自动控制、信息安全; 杨黎斌 博士生, 研究方向为网络与信息安全、嵌入式系统。

$$\arg \min_{x^T D_1=0} \frac{x^T (D-W)x}{x^T D x} \quad (5)$$

根据 Rayleigh 商原理,式(5)的优化问题等同于求解下列等式的第二小特征值问题:

$$(D-W)x=\lambda D x \quad (6)$$

可以认为谱聚类方法是对图划分判据的逼近,也可以利用类似于 PCA (principle component analysis)子空间方法中的嵌入思想来解释谱聚类方法^[6]。代表性算法有 Meila 等人提出的 MS 算法^[7],该算法同时使用矩阵的多个特征向量构造一个简化的数据空间,使得该空间中数据的分布结构更加明显。

谱聚类算法是一种配对聚类方法,算法仅与数据点的数目有关,而与维数无关,因而可以避免由于特征向量的过高维数所造成的奇异性问题。谱聚类通过特征分解,可以获得聚类判据在放松了的连续域中的全局最优解。与其它聚类算法相比,谱聚类具有识别未知分布数据集聚类方面的能力,非常适合于许多实际问题,而且执行起来比较容易。

3 基于谱聚类的复杂网络社团发现算法

3.1 模块度函数

由于社团发现算法大都在给定社团数目的情况下进行社团的探测,而在社团数目未知的情况下就无法评价得出的社团结构是否有效^[8]。为了解决这个问题,Newman 和 Girvan 定义了一个用以评价网络分解满意度的指标,称为模块度 (Modularity)^[9]。模块度的定义为:假定网络已经被分裂成 g 个社团,定义一个 $g \times g$ 维的矩阵 $b = [b_{ij}]_{n \times n}$,其中元素 b_{ij} 表示社团 i 和社团 j 之间的边数占总边数的比例, $a_i = \sum_j b_{ij}$ 表示与第 i 个社团中的节点相连的边在所有边中所占的比例。在此基础上,用下式来定义模块性的衡量标准:

$$Q = \sum_i (b_{ii} - a_i^2) \quad (7)$$

式(7)的物理意义是,网络中连接两个同种类型节点的边(即社团内部的边)的比例减去在同样的社团结构下任意连接节点的边的比例的期望值。如果社团内部边的比例不大于任意连接时的期望值,则有 $Q=0$ 。 Q 的上限为 1,而 Q 越接近于这个值,就说明社团结构越明显。实际网络中,该值通常位于 0.3~0.7 之间。

3.2 基于谱聚类的复杂网络社团发现算法(SCCF 算法)

由于谱聚类相比传统聚类算法,具有聚类准确度高、收敛速度快并且能够识别未知分布数据集聚类方面等优势,而模块度函数能够在大型网络中搜寻出最佳社团数目,结合两者优点,提出了一种基于谱聚类的社团发现算法(SCCF 算法)。算法描述如表 1 所列。

表 1 基于谱聚类的社团发现算法

算法:基于谱聚类的社团发现算法(SCCF 算法)	
输入:复杂网络 $G(V, E, W)$, $k_{\max}$ 和 $k_{\min}$ ($k_{\max}$ 一般设为 $n/3$, $k_{\min}$ 一般设为 2)	
输出:聚类结果集 $\{V_1, V_2, \dots, V_K\}$	
STEP I 初始化 P 为包含所有节点的一个整簇,且令 $Q=0$;	
STEP II 对于 P 中各簇 V_c	
Step1	构造权重矩阵 $W \in \mathbb{R}^{n \times n}$ 及对角阵 $D \in \mathbb{R}^{n \times n}$;
Step2	算概率转移矩阵 $M=D^{-1}W$;
Step3	用隐式循环 Lanczos 方法计算矩阵 M 的前 $k_{\max}$ 个最大特征值所对应的特征向量 $u_1 u_2 \dots u_{k_{\max}}$, 构造矩阵 $U_K = [u_1 u_2 \dots u_{k_{\max}}] \in \mathbb{R}^{n \times K}$, 并将 U_K 的行向量转变为单位向量;
Step4	for ($k=k_{\min}$; $k \leq k_{\max}$; $k++$)

- i) 将 U_K 的每一行看作是 $\mathbb{R}^K$ 空间中的一个点,对其使用 k -means 算法获取 k 个新的子簇, $V_{c_1}, V_{c_2}, \dots, V_{c_k}$;
 - ii) 令 $P'=P$, 并将 P' 中的 V_c 替换为 $V_{c_1}, V_{c_2}, \dots, V_{c_k}$;
 - iii) 计算 $Q(P')$;
- Step5 选取能够使 $Q(P')$ 值达到最大的 k^* , 即 $k^* = \arg \max_k Q(P')$;
- Step6 若 $Q(P_k^*) > Q(P)$, 将 P 中相应的 V_c 替换为 $V_{c_1}, V_{c_2}, \dots, V_{c_k^*}$, 并令 $Q(P) = Q(P_k^*)$; 否则,保持 P 中的簇结构不变;
- Step7 若 P 中某个簇分裂成了若干个子簇,或者 P 中还含有其它簇,则返回 STEP II, 再对这些簇进行计算。

4 时间复杂度分析

从 3.2 节的算法描述中可以看出,SCCF 算法的时间复杂度主要集中在 STEP II。在 Step3 中,用隐式循环 Lanczos 方法计算得出 $k_{\max}$ 个特征向量所需的时间复杂度为 $O(mk_{\max}h + nk_{\max}^2h + k_{\max}^3h)$, 其中 m, n 分别表示复杂网络 $G(V, E, W)$ 所含边的个数和节点数, h 是 Lanczos 方法收敛时所需迭代的次数^[10]。由于 $k_{\max} \leq n$, 因此, Step3 的时间复杂度可被简化为 $O(mk_{\max}h + nk_{\max}^2h)$ 。在 Step4 中,需要运行 $k_{\max}$ 次 k -means 算法,运行一次 k -means 算法的时间复杂度为 $O(nk_{\max}l)$, l 为 k -means 算法收敛时所需迭代的次数。综上,运行一次 STEP II 的时间复杂度为 $O(mk_{\max}h + nk_{\max}^2h + nk_{\max}^2l)$ 。

SCCF 算法的时间复杂度依赖于循环调用 STEP II 的次数。最坏情况下,循环次数等于最终产生的簇个数 K ($K \leq n$)。一般情况下,循环次数为 $\log_{k_{\max}} K$ 。因此,计算特征向量所需的时间复杂度为 $O((mk_{\max}h + nk_{\max}^2h) \log_{k_{\max}} K)$, 运行 k -means 算法所需的时间复杂度为 $O(nk_{\max}^2l \log_{k_{\max}} K)$, 运行整个 SCCF 算法的时间复杂度为 $O((mk_{\max}h + nk_{\max}^2h + nk_{\max}^2l) \log_{k_{\max}} K)$ 。当复杂网络为稀疏网络时,即 $m \sim n$, SCCF 算法的时间复杂度为 $O(n(k_{\max}h + k_{\max}^2h + k_{\max}^2l) \log_{k_{\max}} K)$, 可以看出这与网络的节点数 n 成近似线性关系。

5 实验结果

为了测试算法的实际性能,选用 Zachary 网络^[11]、2000 年美国大学美式足球赛(Football)网络^[12]、爵士乐队(Jazz)合作网络^[13]和 2008 年 4 月人民网强国论坛的“深入讨论”板块数据(共 95247 条 BBS 数据),对经典的社团发现算法 GN^[14], NM^[15], MS 算法^[7]以及我们提出的 SCCF 算法进行测试比较。主要测试项为算法的执行时间和衡量网络分解满意度的指标—— Q 值。测试环境为 Pentium 2.8GHz, 512MB 内存, Windows XP。编译环境为 Matlab 6.5.1。测试结果如表 2 与表 3 所列。

表 2 几种真实网络的 Q 值比较

Network	n	m	k^*	Q			
				GN	NM	MS	SCCF
Zachary	34	78	3	0.392	0.397	0.417	0.420
Football	115	613	11	0.486	0.491	0.612	0.617
Jazz	198	5484	4	0.391	0.395	0.428	0.443
人民网	12603	95247	52	—	—	—	0.739

表 3 几种算法的执行时间比较 T/s

Network	GN	NM	MS	SCCF
Karate	0.2	0.1	0.3	0.3
Football	0.3	0.3	0.1	0.6
Jazz	0.5	0.3	0.5	0.7
人民网	—	—	—	62.7

(下转第 95 页)

参考文献

- [1] Zhao B Y, Kubiawicz J D, Joseph A D. Tapestry: An Infrastructure for Fault-tolerant Wide-area Location[D]. University of California at Berkeley, 2001
- [2] Stoica I, et al. Chord: a scalable peer-to-peer lookup protocol for Internet applications [J]. Networking, IEEE/ACM Transactions, 2003, 11(1): 17-32
- [3] Rowstron A I T, Druschel P. Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems [C]// Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms Heidelberg. Springer-Verlag, 2001
- [4] Krishnamurthy S, et al. A Statistical Theory of Chord under Churn[C]// Proceedings of The 4th Annual International Workshop on Peer-To-Peer Systems (IPTPS 05). Ithaca, NY, USA, 2005
- [5] Mo Z, Yafei D, Xiaoming L. A measurement study of the structured overlay network in P2P file-sharing systems[M]. Hindawi Publishing Corp, 2007
- [6] Silvia Bianchi S S, Felber P, Kropf P. Adaptive Load Balancing for DHT Lookups[C]// Proceedings of the 15th International Conference on Computer Communications and Networks (ICCCN'06). Arlington, VA, October 2006

- [7] Datta A, Schmidt R, Aberer K. Query-load balancing in structured overlays[C]// Proceedings of the Seventh IEEE International Symposium on Cluster Computing and the Grid (CC-GRID'07). Riode Janeiro, Brazil, 2007
- [8] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval [M]. Addison Wesley, 1999
- [9] Saroiu S G, Krishna, Steven G. A Measurement Study of Peer-to-Peer File Sharing Systems[C]// Proceedings of Multimedia Computing and Networking 2002 (MMCN '02). 2002
- [10] Rhea S, et al. Handling Churn in a DHT[D]. EECS Department, University of California, Berkeley, 2003
- [11] Moni N, Udi W. Novel architectures for P2P applications: the continuous-discrete approach[C]// Proceedings of the fifteenth annual ACM symposium on Parallel algorithms and architectures. ACM, San Diego, California, USA, 2003
- [12] Valerie K, Jared S. Choosing a random peer[C]// Proceedings of the twenty-third annual ACM symposium on Principles of distributed computing. ACM, St. John's, Newfoundland, Canada, 2004
- [13] Xiaoming W, Dmitri L. Load-balancing performance of consistent hashing: asymptotic analysis of random node join [M]. IEEE Press, 2007
- [14] Valerie K, et al. Choosing a Random Peer in Chord[M]. Springer-Verlag New York, 2007

(上接第 50 页)

从表 2 的测试数据中可以看出, SCCF 算法在 4 种复杂网络中均能返回具有最高 Q 值的社团结构, 这说明用 SCCF 算法划分得到的社团结构是最令人满意的。而 GN, NM 和 MS 算法由于超额的系统资源占用, 在由 BBS 数据构成的大型虚拟社会网络中甚至没能完成实验。

从表 3 可以看出, GN, NM 和 WS 算法在仅含有几百个节点的小型网络中能够有效地进行社团结构探测, 但在节点数上万的虚拟社会网络中探测时, 这 3 种算法均失效。而 SCCF 算法却可以在很短的时间内处理具有上万个节点的大型网络。

结束语 复杂网络的社团发现已经成为当今一个非常具有挑战性和前景性的研究领域。本文用基于谱聚类算法的思想来优化网络社团结构的划分。与已有的社团发现算法相比, 本算法时间复杂度较低, 且能够在结构未知的大型网络中得到高质量的社团结构。

参考文献

- [1] Strogatz S H. Exploring complex networks [J]. Nature, 2001, 410: 268-276
- [2] Garey M R, Johnson D S. Computers and Intractability: A Guide to the Theory of NP-Completeness [M]. San Francisco: W. H Freeman Publishers, 1979
- [3] Scott J. Social Network Analysis: A Handbook [M]. 2nd ed. London: Sage Publications, 2002
- [4] Fiedler M. Algebraic connectivity of graphs[J]. Czech, Math J,

- 1973, 23: 298-305
- [5] Shi J, Malik J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905
- [6] Gu M, Zha H, Ding C, et al. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering [R]. CSE-01-007. Penn State University, 2001
- [7] Meila M, Shi J. Learning segmentation by random walks[C]// NIPS. 2000: 873-879
- [8] Muff S, Rao F, Cflisch A. Validation of network clustrizations [J]. arXiv: cond-mat, 2005: 0503252
- [9] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69(2): 026113
- [10] White S, Smyth P. A spectral clustering approach to finding communities in graph[J]. SIAM Data Mining, 2005
- [11] Zachary W. An information flow model of conflict and fission in small groups[J]. J. Anthropol. Res., 1993, 33: 452-473
- [12] Girvan M, Newman M. Community structure in social and biological networks[J]. Proc Natl Acad Sci USA, 2002, 99: 7821-7826
- [13] Gleiser P, Danon L. Community structure in jazz. Advances in Complex Systems, 2003, 6: 565
- [14] Newman M. Modularity and community structure in networks [J]. Proc Natl Acad Sci USA, 2006, 103: 8577-8582
- [15] Donetti L, Munoz M A. Detecting Network Communities: a new systematic and efficient algorithm[J]. cond -mat, 2004: 0404652