

核 k -means 聚类检测复杂网络社团算法

付立东

(西安科技大学计算机学院 西安 710054) (西安电子科技大学计算机学院 西安 710071)

摘要 为揭示复杂系统中的结构与功能之间的联系,复杂网络中的社团发现成为一项最基本的任务。最近,李等人提出了一种用来评估社团质量的函数,称之为模块密度函数(即 D 值),并利用一个核矩阵给出了模块密度目标函数与核 k -means 方法之间的等价性。基于这种等价性,通过过渡操作的核矩阵来优化模块密度函数并提出了一种新的核 k -means 算法。实验结果表明,这种算法在发现复杂网络社团上是有效的。

关键词 社团结构,模块密度,核 k -means 算法

中图法分类号 TP399 **文献标识码** A

Kernel k -means Clustering Algorithm for Detecting Communities in Complex Networks

FU Li-dong

(College of Computer, Xi'an University of Science and Technology, Xi'an 710054, China)

(School of Computer Science & Engineering, Xidian University, Xi'an 710071, China)

Abstract Discovery community structure is fundamental for uncovering the links between structure and function in complex networks. In this context, recently, Li et al. recently proposed modularity density objective function for community detecting called the D function and gave the equivalence between modularity density objective function and the kernel k -means by using a kernel matrix. In this paper, based on this equivalence, we used the kernel matrix to optimize the modularity density and developed a new kernel k -means algorithm. Experimental results indicate that the new algorithms are efficient at finding community structures in complex networks.

Keywords Community structures, Modularity density, Kernel k -means

复杂网络中,基于元素间相互作用模式来发现社团结构,越来越受到研究人员的关注^[1,2]。许多领域中,网络扮演着重要的角色,例如英特网、社会网络以及生物网络。所有复杂网络都可以模型化成图 $G=(V,E)$ 的形式,这里 V 表示一组顶点, E 表示顶点之间相互作用的一组边。在网络中,一个社团结构被定义成内部连接紧密、外部连接相对松散的顶点子集。这些社团结构暗含了复杂网络的功能,用来帮助研究人员理解复杂网络的增长机制。结果,复杂网络中社团刻画与检测成了当今最杰出的问题之一^[3,4]。

有许多算法用来发现复杂网络中的社团结构,例如划分技术^[5]、稀疏矩阵配方^[6]、代数连通性^[7]等。近来,许多算法悬挂在一种广泛使用的社团评估函数上,称之为 Q ^[8,9]。然而,最近的研究表明^[10],这个函数不能检测出小于一种内在尺寸的社团结构。这主要是因为一个网络的划分高敏感于该网络边的总数目。事实上,好的算法应当悬挂在好的评估函数质量上。

为了克服先前这种社团评估函数分辨极限,最近李等人提出了一种新的社团评估函数^[11],称之为模块密度(所谓 D 值)。这个函数 D 值在评估网络的划分,即评估社团结构时,不仅考虑了社团中边的多少,而且考虑了社团中节点的多少。

通过一种过渡操作核矩阵,他们同时证明了模块密度函数和目标核 k 函数的等价关系。本文利用这种等价性,提出了一种新的核 k -means 算法。实验结果表明,这种算法在发现复杂网络社团上是有效的。

1 核 k -means

我们介绍一种权重的核 k -means^[12,13]。给定一组 $v_i \in R^n$ 维的数据向量 $\{v_i\}_{i=1}^n$,权重核 k -means 的目的是发现 p 个不连接的划分 $\{V_c\}_{c=1}^p$,以便下列的目标函数最小化:

$$H(\{V_c\}_{c=1}^p) = \sum_{c=1}^p \sum_{v_i \in V_c} w_i ||\phi(v_i) - m_c||^2 \quad (1)$$

式中, $m_c = \frac{\sum_{v_i \in V_c} w_i \phi(v_i)}{\sum_{v_i \in V_c} w_i}$, w_i 是向量 v_i 的非负权重,并且 ϕ 是一种映射 U 中的向量到一个高维空间的函数,故不需要知道 ϕ 的确切表达。显而易见,如果所有的权重等于 1 并且 ϕ 是一种恒等函数,式(1)可退化为标准的 k -means 方法。特别地,仅当所有的权重等于 1 时,这个权重的核 k -means 涵盖了标准的核 k -means 定义。如果扩展权重的核 k -means 函数中的距离为 $||\phi(v_i) - m_c||^2$,可得到:

到稿日期:2009-10-26 返修日期:2010-01-19 本文受国家自然科学基金重点项目(60933009),教育部高校博士点基金资助项目(2008 07010013),国家自然科学基金项目(60970065)资助。

付立东(1973—),男,博士生,讲师,主要研究领域为复杂网络中的社团发现、中心度量等,E-mail:fulidong2005@163.com。

$$\|\phi(v_i) - m_c\|^2 = \phi(v_i) \cdot \phi(v_i) - \frac{2 \sum_{v_j \in V_c} w_j \phi(v_i) \cdot \phi(v_j)}{\sum_{v_j \in V_c} w_j} + \frac{\sum_{v_j, v_l \in V_c} w_j w_l \phi(v_j) \phi(v_l)}{(\sum_{v_j \in V_c} w_j)^2} \quad (2)$$

注意到式(2)包含的数据点仅是以内积的形式出现。因此,我们能使用给定的核矩阵 K 去计算映射空间中节点之间的距离,这里 $K_{ij} = \phi(v_i) \cdot \phi(v_j)$ 。我们假设输入的是内积形式的核矩阵 K ;已经显示任何正半定定义的矩阵 K 被看成一个核矩阵。应用这个核矩阵 K ,距离计算式(2) $d(v_i, m_c) = \|\phi(v_i) - m_c\|^2$ 可以写成

$$K_{ii} - \frac{2 \sum_{v_j \in V_c} w_j K_{ij}}{\sum_{v_j \in V_c} w_j} + \frac{\sum_{v_j, v_l \in V_c} w_j w_l K_{jl}}{(\sum_{v_j \in V_c} w_j)^2} \quad (3)$$

结果,我们可以设计一种相似于 k -means 算法,以单调递减这个权重的核 k -means 目标函数。

2 模块密度与社团结构

让 $G=(V, E, A)$ 表示一个无向网络,其中 V 是一个包含 n 个顶点的集合, E 是一个包含 m 条边的集合, A 是一个 $n \times n$ 维的对称邻接矩阵。 A_{ij} 代表着顶点 i 和顶点 j 之间的边的权重。若要划分网络 G 的顶点到 p 个非重叠社团结构,那么模块密度函数被定义成:

$$D(\{V_c\}_{c=1}^p) = \sum_{c=1}^p \frac{\text{links}(V_c, V_c) - \text{links}(V_c, \overline{V_c})}{|V_c|} \quad (4)$$

式中, $\text{links}(V_c, V_c)$ 等于 $\sum_{i,j \in V_c} A_{ij}$, 度量了社团内部边的权重和, $\overline{V_c}$ 是 V_c 的补集。进一步,让 V_c 的度等于社团 V_c 内的顶点到所有其他顶点的边数,即 $\text{degree}(V_c) = \text{links}(V_c, V)$, 则有 $\text{links}(V_c, \overline{V_c}) = \text{degree}(V_c) - \text{links}(V_c, V_c)$ 。结果, $\text{links}(V_c, \overline{V_c})$ 度量了不同社团之间边的权重。 $|V_c|$ 是社团 V_c 内顶点的数目,度量了一个社团的大小。

因此,模块密度 D 值等于所有社团平均内部度减去所有社团平均外部度。较大的 D 值意味着较强壮的社团结构。如果一个划分给出了所有社团平均内部度不大于平均外部度,那么模块密度 D 值小于等于零。

3 核 k -means 算法

为了发现社团结构的有效检测方法,李等人^[11]介绍了模块密度函数。通过一种过渡操作核矩阵,他们同时给出了模块密度函数和目标核 k 函数的等价关系。

给出一组数据向量 $U = \{u_i\}_{i=1}^N$, $u_i \in R^n$ 以及一个 $N \times N$ 的核矩阵:

$$K = \sigma I + 2A - B \quad (5)$$

式中, I 是单位矩阵; B 是一个对角的度矩阵,其元素 $B_{ii} = \sum_{j=1}^N A_{ij}$; σ 是一个被选择为足够大的实数,以便这个核矩阵 K 始终处于正定。为寻找 $\{U_c\}_{c=1}^p$ 个不重叠社团,函数 D 和核 k 函数 H 被表达成^[11]:

$$H = (N - p)\sigma - D \quad (6)$$

那么可以通过式(2)中的核矩阵来使用核 k -means,以优化模块密度函数 D ,从而在复杂网络中发现社团结构。这里仅仅考虑无权重网络中的社团检测问题,因此设置每个顶点的权

重为 1。那么该算法可以详细描述为:

输入:核矩阵 K , p :聚类数目, $t_{\max}$:迭代操作的最大次数, $\{V_c^{(0)}\}_{c=1}^p$:操作的初始聚类

输出: $\{V_c\}_{c=1}^p$:最终的图划分

1)初始化来自 $\{V_c^{(0)}\}_{c=1}^p$ 的 p 个聚类

2)计算核矩阵 $K = \sigma I + 2A - B$

3)设置 $t=0$

4)对每一个顶点 x_i 及每一个聚类 c , 计算

$$d(x_i, m_c) = K_{ii} - \frac{2 \sum_{v_j \in V_c} K_{ij}}{|V_c|} + \frac{\sum_{v_j, v_l \in V_c} K_{jl}}{|V_c|^2}$$

5)发现 $c^*(x_i) = \arg\min_c d(x_i, m_c)$

6)计算这个更新的聚类

$$V_c^{(t+1)} = \{x_i : c^*(x_i) = c\}$$

7)如果没有收敛或 $t_{\max} > t$, 设置 $t=t+1$, 并回到第 3)步; 否则, 停止并输出最终的聚类 $\{V_c^{(t+1)}\}_{c=1}^p$ 。

4 算法检验

本节在一个众所周知的预定义计算机产生的网络社团结构中^[2], 检验了我们提出的核 k -means 算法, 并与基于 Q 函数的谱分算法做了对比^[9]。这个人工网络包含 128 个顶点, 这些顶点被划分成相同大小的 4 个社团结构, 每个社团结构包含 32 个顶点。边使用了两个概率被随机放置, 以使得每一个顶点的平均度等于 16, 即 $z_{in} + z_{out} = 16$, 这里 z_{in} 是每一个顶点连接的、平均落在相同社团中的边数目, z_{out} 是每一个顶点连接的、平均落在不同社团间的边数目。图 1 展示了该随机网络中 128 个顶点在 $z_{in} = 12$, $z_{out} = 4$ 时被划分成 4 个社团结构时的大概情况。这个随机网络被广泛应用于检验社团发现算法^[2,11]。

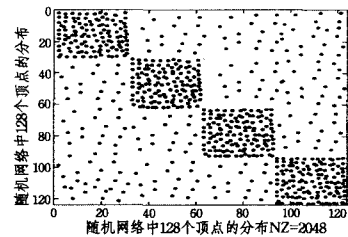


图 1 计算机产生的网络中 128 个顶点在 $z_{in} = 12$, $z_{out} = 4$ 时被划分成 4 个社团结构时的大概情况

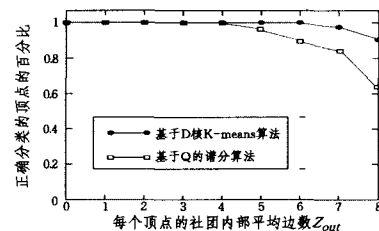


图 2 两种方法在预知社团结构的计算机产生图上对顶点的分类结果(刻画了随着 z_{out} 的变化两种方法能正确分类出顶点的百分比)

对这个实验的计算结果概括在图 2 中。图 2 展示了我们的基于模块密度 D 的核 k -means 方法及基于模块 Q 谱分方法^[9]随着 z_{out} 的变化能正确检测出顶点的百分比。从图 2 可看出, 当 $z_{out} \leq 5$ 时, 两种方法都能正确检测出多于 95% 的顶

(下转第 256 页)

满足最优性原理是使用动态规划法的前提条件,因此,许多文献用最优化原理的证明过程来分析最优解的求解过程,本质上说,文献[19]是这样思考 Hamilton 回路的,文献[20]也是这样讨论 TSP 的。本文的研究结果表明,最优性原理本身并没有提供最优的解结构,具体而言,对于 TSP 问题,虽然一条长度为 n 的最短路径一定是一长度为 1 的路径和一长度为 $n-1$ 的最短路径组合,然而,这种组合未必是求解长度为 n 的路径的最优过程,结合肖尔茨猜想的证明,本文探讨了最优解组成的复杂性,为进一步研究提供了理论基础。从工程应用而言,本文提供的算法具有如下特点:

1. 计算不依赖于初始解与经验值。
2. 没有欧氏空间的平面约束条件,有利于推广到其他相关非平面的应用领域。
3. 作为一种精确算法,其解可以作为近似算法的评价依据。
4. 在降低时间复杂度的同时,为以后优化留下了很大的研究空间——数据压缩与矩阵并行运算的相关研究将是该算法有潜力的发展方向。

参考文献

- [1] 文中华,姜云飞. 用分层关联方法求有向图中所有 Hamilton 回路的算法[J]. 计算机研究与发展, 2005, 42(10): 1809-1814
- [2] 王彦棋. 逐点循环递归法求哈密顿回路[J]. 哈尔滨工业大学学报, 2004, 36(1)
- [3] 邹鹏,周智,等. 求解旅行商问题的循环局部搜索算法的运行时间和性能分布分析[J]. 计算机学报, 2006, 29(1)
- [4] 刘新,刘任任,侯经. 求解旅行商问题的整体优先算法[J]. 计算机应用, 2007, 27(5)
- [5] 贺一,刘光远. 禁忌搜索算法求解旅行商问题研究[J]. 西南师范

大学学报:自然科学版, 2002, 27(3)

- [6] 耿素云,屈婉玲. 离散数学(修订版)[M]. 北京:高等教育出版社, 2004
- [7] 高尚,韩斌,吴小俊,等. 求解旅行商问题的混合粒子群优化算法[J]. 控制与决策, 2004, 19(11)
- [8] 柴世红. 两段式遗传算法求解 CTSP[J]. 大众科技, 2008(4)
- [9] 顾大权,侯太平,左莉,等. 用数据搅动算法求解 TSP 问题[J]. 计算机应用, 2004(24)
- [10] 杨忠,鲍明,张阿舟. 求解中国旅行商问题的新结果[J]. 数据采集与处理, 1993, 8(3)
- [11] 胡作玄. 本世纪未解的数学难题(下)[J]. 百科知识, 1999(5)
- [12] 靳蕃,范俊波,谭永东. 神经网络与神经计算机[M]. 成都:西南交通大学出版社, 1991
- [13] 杨启宇. 关于无穷大基数的矛盾[J]. 成都大学学报:自然科学版, 2004, 23(4)
- [14] 温邦彦. 自然数和偶数的个数一样多吗? [J]. 重庆工学院学报:自然科学版, 2008, 22(11)
- [15] 欧阳耿. 无定义的实无穷概念是第三次数学危机的真正根源[J]. 宜春学院学报, 2005, 27(4)
- [16] 杜国平. 潜无穷、实无穷探析[J]. 自然辩证法通讯, 2009, 31(3)
- [17] 朱梧根. 数学与无穷观的逻辑基础[M]. 大连:大连理工大学出版社, 2008
- [18] 柯华庆. 直觉主义数学哲学的两个阶段[J]. 学术研究, 2005(2)
- [19] 文中华,陈志红. 一个求简单图中所有 Hamilton 回路的算法[J]. 湘潭大学:自然科学学报, 2005, 27(4)
- [20] 徐伯庆,宣国荣,柴佩琪. 中国旅行商问题的二叉树描述及其求解[J]. 模式识别与人工智能, 2000, 13(2)
- [21] 王红梅. 算法设计与分析[M]. 北京:清华大学出版社, 2006
- [22] 郭文强,秦志光,冯昊. 一种基于归零矩阵的 TSP 求解算法[J]. 计算机科学, 2009, 36(8): 254-257

(上接第 213 页)

点数。然而,当 $5 < z_{out} \leq 8$ 时,暗示着这 4 个社团的顶点越来越难以检测,基于模块 Q 谱分算法的曲线急速下滑,意味着成功检测已知社团顶点的偏离较大。特别地,当通过谱分方法或半指导方法初始化我们的核 k -means 算法时,这个核方法仍然能正确分类出 82% 以上的顶点,比基于 Q 谱分算法准确性高。换句话说,随着 z_{out} 的变化,我们的基于核 k -means 算法比基于 Q 谱分算法更能正确检测出这 4 个已知的社团结构。

结束语 本文根据最近提出的用来评估与发现社团的模块密度 D 函数与核 k -means 之间的等价性,提出了一种新的社团检测核 k -means 算法。从计算机产生的人工网络的实验结果中可以看出,我们提出的核 k -means 方法比基于 Q 算法有较好的准确性。

由于核 k -means 方法对聚类开始的划分比较灵活,因此可以通过谱分方法、半指导或指导方法来初始化核 k -means 方法,以尽可能地提升社团检查结果。

参考文献

- [1] Fortunato S, Latora V, Marchiori M. Method to find community structure based on information centrality[J]. Phys Rev E, 2004, 70(13): 056105
- [2] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc Natl Acad Sci USA, 2002, 99(12): 7821

- [3] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex network in nature and society[J]. Nature, 2005, 435: 814-818
- [4] Newman M E J. The structure and function of complex networks[J]. Proc Natl Acad Sci USA, 2006, 103: 8577-82
- [5] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell Syst Tech J, 1970, 49: 292-307
- [6] Pothén A, Simon H D, Uou K P. Partitioning sparse matrices with eigenvectors of graphs[J]. Siam J Matrix Anal Appl, 1990, 11(3): 430-452
- [7] Fiedler M. Algebraic connectivity of graphs[J]. Czech math J, 1973, 23: 298-305
- [8] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Phys Rev E, 2006, 74(3): 036104
- [9] Newman M E J. Modularity and community structure in networks[J]. Proc Natl Acad Sci USA, 2006, 103: 8577-8582
- [10] Fortunato S, Barthélemy M. Resolution limit in community detection[J]. Proc Natl Acad Sci USA, 2007, 104(1): 36-41
- [11] Li Z P, Zhang S H, Wang R S, et al. Quantitative function for community detection[J]. Phys Rev E, 2008, 77(3): 036109
- [12] Dhillon I S, Guan Y, Kulis B. Kernel K-means, Spectral Clustering and Normalized Cuts[C]//Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, 2004: 551-556
- [13] Dhillon I S, Guan Y, Kulis B. A unified view of kernel K-means, spectral clustering and graph cuts[R]. University of Texas at Austin, 2004