DOI:10.11918/j.issn.0367-6234.201704101

密度峰值聚类的自适应社区发现算法

金志刚, 徐珮轩

(天津大学 电气自动化与信息工程学院, 天津 300072)

摘 要: 为减少社区发现算法中参数的选择对社区划分的影响,同时使算法能够自适应地进行社区划分,本文提出一种基于核密度估计的密度峰值聚类的社区发现算法 KDED.首先,定义一种基于信任度的距离度量,将社交网络中的用户关系量化为距离矩阵, 使用矩阵元素的大小度量用户关系的紧密程度;然后对距离矩阵进行核密度估计,统计各个节点在网络中的影响大小,结合热扩散模型改进计算流程,使其自适应不同规模的数据集以提高计算精度;结合密度峰值聚类原理和社区属性确定社区中心节点后,可根据节点间的距离得到社区内部层次结构和社区外部的自然结构;最后将剩余节点按距离分配到相应的社区当中以完成社区划分.仿真结果表明:通过可视化软件可观察到,通过 KDED 算法得到的社区划分结果具有清晰的自然结构和内部层次结构;随着社区规模的提升以及划分难度增加, KDED 算法具有出色的稳定性;在真实数据集以及 LFR 基准网络上均得到较为接近真实划分结果的社区划分,自适应性良好,验证算法的可行性与有效性.

关键词: 社区发现;密度峰值聚类;信任度;核密度估计;自适应

中图分类号: TP393 **文献标志码**: A **文章编号**: 0367-6234(2018)05-0044-08

An adaptive community detection algorithm of density peak clustering

JIN Zhigang, XU Peixuan

(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: In order to reduce the influence of the selection of parameters in community detection algorithm on the results of community partitioning and detect adaptively, a community detection algorithm called KDED based on kernel density estimation for density peak clustering is proposed. Firstly, a distance measure based on trust is defined, and the user relationship in the social network is quantified as a distance matrix. Then the kernel density is estimated by the distance matrix to calculate the influence of each node on the network. The thermal diffusion model improves the computational flow so that it adapts to different sizes of data sets to improve computational accuracy. The clustering center is determined by density peak clustering principle and community property, and the node is allocated to the corresponding community which can obtain the hierarchical structure within the community and the natural structure among the community. The simulation results show that it can be observed by the visualization software that the community division result obtained by the KDED algorithm has a clear natural structure and internal hierarchical structure. The KDED algorithm has the best stability with the increase in the size of the community and the difficulty of detection compared with the classical algorithm. And it gets a community partition closer to the real partition in the real data set and the LFR benchmark network, which verifies the feasibility and effectiveness of the algorithm.

Keywords: community detection; density peak clustering; trust degree; kernel density estimation; adaptive

在社交网络中,社区(community)被用来表示具有相似性、共同爱好、或联系的共同体,它具有“内部节点之间连接紧密,与外部连接稀疏”的特点.社区成员之间因社交关系不均匀,使社区表现出社区结构特性[1]. 通过社区发现(community detection)算法分析复杂网络的拓扑结构和层次结构,理解其形成过程,有助于揭示复杂系统的内部规律,为信息传播控制[2]、维护社会安全[3]、推荐系统建设等领域提供有力支撑,已成为 social network analysis 领域的一个研究热点.

社区发现算法源于 Girvan 和 Newman 提出的 GN 算法[1],但该算法需要预知社区的个数来确定算法终止条件,并且时间复杂度较高. 标签传播法(label propagation algorithm, LPA)[4]是 Raghavan 提出的一种基于图的半监督学习的社区发现算法,虽复杂度低,但其每次迭代结果不稳定,准确度不高,后来张鑫等[5]提出稳定标签法作为改进. 而后很多社区发现算法也相继被提出,如:基于随机游走的算法[6]、基于信息编码的算法[7]以及基于网络拓扑结

收稿日期: 2017-04-20.
基金项目: 国家自然科学基金(61571318)
作者简介: 徐珮轩(1992—),男,硕士研究生;
金志刚(1972—),男,教授,博士生导师.
通信作者: 金志刚,zgjin@ tju.edu.cn.

构的算法[8]等. 这些算法都在一定程度上提高社区发现的精度,然而绝大部分社区发现算法都对参数敏感,一旦参数选择不当就会对结果产生很大影响. 除此之外,划分结果的真实性逐渐被更多的学者所关注,基于模块度指标优化的社区发现算法,对于较大规模社区具有倾向性,不善于识别小规模社区,具有分辨率限制[9]. 因此,社区发现算法在复杂度、准确度、自适应等方面都有着提升的空间.

在一定意义上,社区发现就是某种类型的聚类,一般被认为是广义的图聚类. 二者区别就在于:常见的聚类一般是假设任意对象都是互连的,只是距离(或者相似度)不同,数据集可表示为一张完全连通图,如:划分、层次、密度聚类等,而社区发现过程中可能存在孤立节点;此外,聚类一般是将数据集划分为 k 个不相交的子集,而社区发现包含非重叠社区与重叠社区,允许子集重叠. 但这两者的过程极为相似,因此出现很多将聚类算法改成社区发现的例子,如 DBSCAN 改进为 SCAN,层次聚类等.

Rodriguez 等在 14 年提出一种基于密度峰值的聚类算法[10]. 该方法只依赖于节点之间的距离,根据节点分布的局部密度选择聚类中心,降低参数选择对算法影响的影响,使得该方法能够获得稳定的聚类结果. 黄岚[11]等通过相似度定义网络中各节点之间的距离,将密度峰值聚类算法应用于社区发现中;Yan[12]等引入箱线图选取核心节点,将密度峰值聚类算法应用于重叠社区发现;Wang[13]等采用启发式算法划分社区的自然结构. 但是上述几种算法虽然成功将密度峰值聚类算法的原理应用到社区发现当中,但却又引入一些新的额外参数. 并且参数的选取并没有非常合适的方法,需要大量实验数据来获得,通常只对某一种或几种数据集的结果较好,无法自适应地对社区进行划分,算法在性能存在缺陷.

因此,本文提出一种基于核密度估计的密度峰值聚类的社区发现算法(Community detection algorithm of density peak clustering based on kernel density estimation, KDED),首先引入一种基于信任度的距离度量,将社交网络中的用户关系量化为距离矩阵;然后通过基于热扩散理论的核密度估计方法对距离度量进一步处理,利用密度峰值聚类确定聚类中心,并将节点分配到相应的社区当中. 选取 Fastgreed 算法、Informap 算法、Walktrap 算法、LPA 算法在真实数据集进行比较,与 Fastgreed 算法、Informap 算法、Walktrap 算法、LPA 算法、VDDPC 算法[12]、LCCD 算法[13]在 LFR 人工基准网络上进行比较,结果充分验证 KDED 算法的可行性和有效性.

1　密度峰值聚类算法

Rodriguez 等所提出的基于密度峰值的聚类算法[10]通过“距离”来寻找聚类中心. 聚类中心基于两个假设:1)聚类中心的密度大于环绕它的邻居节点;2)聚类中心到比其局部密度高的节点的距离较远.

设数据集中的数据点总数为 N, 对于数据集中的任意点 i, 使用局部密度 ρ_i 和到更高密度节点的距离 δ_i 进行描述. 其中局部密度 ρ_i 的定义为

$$\rho_i = \sum_j \chi(d_{ij} - d_c). \tag{1}$$

式中:设 $x = d_{ij} - d_c$,若 $x < 0$,则 $\chi(x) = 1$,否则 $\chi(x) = 0$, d_{ij} 为数据点 i 到 j 的距离, d_c 为截断距离. i 点的局部密度 ρ_i 等于到其距离小于截断距离 d_c 点的个数. 因此 d_c 的选取将直接影响 ρ_i 的取值,进而决定聚类结果的准确性.

节点 i 到比其局部密度更高节点的距离 δ_i 的定义为

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij}). \tag{2}$$

式中:若 i 点为数据集中局部密度最大的点,则 $\delta_i = \max_j(d_{ij})$,选取 i 点到相邻节点距离的最大值作为其 δ_i;否则,选取局部密度大于 i 点的数据点到 i 点的距离中的最小值,作为 i 点的 δ_i.

由上述定义可知,作为聚类中心的数据点一定会比一般节点具有更高的 ρ_i 和 δ_i,可根据此原理选取社区中心,再将一般节点分配到其最近的聚类中心,完成整个社区的划分.

2　峰值密度聚类的自适应社区发现算法

2.1　基于用户信任度的距离度量

由于社交网络图中的点没有坐标值,仅通过边来表征两点是否存在一定联系,所以不能用传统的 Euclidean 距离和 Manhattan 距离来计算相异度;Jaccard 距离可以突出节点间的局部关系,但无法度量节点之间的全局距离;而最短距离作为一种常用距离度量,在社区划分当中不考虑社区内部样本的方差,只考虑均值,因此分类精度不高. 因此本文首先提出一种将节点关系转化为距离度量的方法. 根据 Ziegler C N[14]提出的“在社交网络中,人们之间的信任和偏好相似度存在着一种正相关的联系”的观点,提出假设:1)具有相同爱好、兴趣的群体内部成员,彼此会获得更多信任;2)成员偏向和受信任的成员进行联系,因此会具有更短的相对距离;3)最受信任的几个成员普遍也存在联系,满足“富人俱乐部”效应. 基于 3 点假设,结合 Jaccard 相似度,提出基于“信任度”的距离度量,如式(3)~(6)所示:

$$D(i,j)=\frac{1}{\alpha(i,j)\times\beta(i,j)}, \tag{3}$$

$$\alpha(i,j)=\frac{|C(i,j)|+1}{|N_i\cup N_j|}, \tag{4}$$

$$\beta(i,j)=\begin{cases}\dfrac{|E(C(i,j))|}{(|C(i,j)|)(|C(i,j)|-1)/2}+1 & |C(i,j)|>2,\\ |C(i,j)|\leqslant 2\end{cases} \tag{5}$$

$$C(i,j)=N_i\cap N_j. \tag{6}$$

式中：$D(i,j)$ 代表基于信任度的距离度量，使用描述点关系的因子 $\alpha(i,j)$ 和描述边关系的因子 $\beta(i,j)$ 刻画点 i 和点 j 之间的紧密程度，距离越近代表着节点间的关系越紧密. 因子 $\alpha(i,j)$ 基于 Jaccard 相似系数，代表节点 i 与节点 j 的共同邻居节点在其邻居节点并集中的比例，$|C(i,j)|$ 为节点 i 与节点 j 的公共邻居个数，N_i 和 N_j 分别为与节点 i 与 j 邻居节点的数量. 因子 $\beta(i,j)$ 描述点 i 和点 j 公共邻居节点相连边之间的紧密程度，$|E(C(i,j))|$ 为节点 i 与节点 j 公共邻居中的连边的数量，当节点 i 与节点 j 的公共邻居点数大于 2 时，认为边的连接强度较大，计算两点相连的边数与公共节点间可能连接的最大边数之比；否则，直接置 1，这样会使得联系疏远的节点，在距离度量中得到体现，符合“信任”假设. 此外，为避免因子等于 0 时对距离度量产生影响，均在因子的定义中做加 1 处理.

这种定义方法将成员社交关系的邻接矩阵转化为距离度量，充分考虑节点之间的局部拓扑联系，可以使潜在社区内部的距离变得紧密，而不同社区的节点距离较远，为划分社区的层次结构奠定基础.

2.2 局部密度 ρ_i 的等效

峰值密度聚类算法中的 ρ_i、δ_i 与截断距离 d_c 有着很大关系，但只能通过大量实验选取出一个比较合理的值，这样对社区发现的结果产生较大影响，易出现误差.

非参数密度估计是数据统计分析的重要工具，它用于评估偏度，多模态，总结贝叶斯后验，判别分析和约束[15-16]，它对于数据集的建模更加灵活，并且不受规范偏差的影响. 核密度估计（kernel density estimation，KDE）是较为常用的非密度估计方法之一，引入核方法，可将低维特征空间的样本通过核函数映射到高维特征空间，增加对特征的优化，使其特征在高维特征空间线性可分，可突出样本点之间根据距离远近体现出的关系紧密程度，更易聚类.

借鉴 Rashid 等提出的将热扩散模型（heat diffusion）应用到聚类算法的思想[17]，在得到距离矩阵的基础上，选取 Gaussian 核函数进行核密度估计，统计每个点对于其他节点影响的大小，代替局部密度 ρ，可以合理规避 d_c 选取的不合理对实验结果的影响. Gaussian 核密度为

$$k(d,d_j;h)=\frac{1}{\sqrt{2\pi h}}\mathrm{e}^{-\frac{d_j^2}{2h}}. \tag{7}$$

式中：k 代表核函数，h 为带宽，d_j 为距离矩阵中的分量，代表当前节点到点 j 的距离. 在根据上式对距离矩阵进行处理后，会得到一个关于 k 的矩阵，纵列意义为节点自身的影响在其他节中得到的反馈，与峰值密度函数中局部密度的意义一致.

但 Gaussian 核密度估计具有一些局限性，例如，敏感参数带宽 h 难以选择，积分时的边界偏差以及计算过程中欠平滑或过平滑的问题[16]. 因此，使用热扩散方法解决此问题.

热扩散模型中，将核密度估计视为求解扩散过程中偏微分方程的唯一解，使用时间参数 t 代替带宽 h. 通过热扩散对 KDE 求解的思想源自维纳过程（weiner process），它是一种连续时间随机过程，其下一时间状态可直接由之前的时间状态算出，与 Gaussian 核密度估计结合的形式如下所示：

1）准备概率通过 d 纬数据点 $\{d_1,d_2,d_3,\cdots,d_n\}$ 平均分配.

2）使用 Gaussian 核估计估计点 i 到点 j 的转移概率，

$$P_{\text{transition}(d,d_j;t)}=\frac{1}{n}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi t}}\exp\Big[-\frac{(d-d_j)^2}{2t}\Big],$$

此种情况下，KDE 被转化为时长为 t 过程中的概率分布函数，但函数形式类似于 Gaussian 核函数.

$$\hat{f}_t(d;t)=\frac{1}{2}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi t}}\exp\Big[-\frac{(d-d_j)^2}{2t}\Big]. \tag{8}$$

式(8)代表迭代过程，因此要满足扩散过程的偏微分方程（partial differential equation，PDE）条件，即

$$\frac{\partial}{\partial t}\hat{f}(d;t)=\frac{1}{2}\frac{\partial^2}{\partial d^2}\hat{f}(d;t),\quad d\in D,t>0$$

式中，$D\equiv\mathbb{R}$ 并且初始条件 $\hat{f}(d;0)=\Delta(d)$，其中：

$$\Delta d=\frac{1}{n}\sum\nolimits_{j=1}^{n}\delta(d-d_j),$$

为 D 的经验密度，$\delta(d-d_j)$ 是狄拉克函数. 为消除 Gaussian 函数在有限节点时带来的边界误差需进行边界修正，同时满足 Neumann 边界条件，$\hat{f}(d;t)$ 在域的上下界关于 t 的偏导数应该为 0.

综上，考虑到 Neumann 边界条件和域 $[0,1]$ 的概率密度，该 PDE 的解析解可以通过 θ 核的形式求解，以此实现 Gaussian 核的转化.

$$\hat{f}(d;t)=\frac{1}{n}\sum_{j=1}^{n}\theta(d,d_j;t),\ d\in[0,1],$$

其中 θ 核函数为

$$\theta(d,d_j;t)=\sum_{i=-\infty}^{\infty}\emptyset(d,2k+d_i;t)+\emptyset(d,2k-d_i;t),$$

进一步变换得到

$$\hat{f}(d;t)=\frac{1}{n}\sum_{j=1}^{n}\sum_{k=-\infty}^{\infty}e^{-k^2\pi^2t/2}\cos(k\pi d)\cos(k\pi d_j),$$

最终，上式可近似为

$$\hat{f}(d;t)\approx\sum_{k=0}^{n-1}a_k e^{-k^2\pi^2t/2}\cos(k\pi d). \qquad (9)$$

其中 n 为足够大的正整数，a_k 为

$$a_k=\begin{cases}1, & k=0;\\ \dfrac{1}{n}\sum_{i=1}^{n}\cos(k\pi d_i), & k=1,2,\cdots,n-1.\end{cases}$$

式(9)是 KDE 的完全自适应和替代形式，并考虑最佳带宽选择和边界校正.此外，式(9)可以使用快速傅立叶变换求解，复杂度为 $O(n\log_2 n)$ [18]. 对于小带宽，式(9)的表现与 Gaussian 核相似；对于大的带宽其表现接近 Uniform 核，这样可以为 KDED 算法提供更好的性能并且更加接近与真实的密度，改善 Gaussian 核函数在边界处存在的误差，Krishanaswamy S 等说明通过热扩散理论快速评估 KDE 的优越性[19].

2.3　到高密度点的距离 δ_i 的计算

依照第一章介绍的密度峰值聚类的原理，δ 在社区发现过程中保证社区之间相对稀疏的特性，其计算基于局部密度 ρ 以及距离矩阵 D. 首先将节点的 ρ 值进行排列；对于节点 i，统计局部密度 ρ 值大于 ρ_i 的点到 i 的距离 $\{d\}$；δ_i 是到该点的距离的最小值，即 $\delta_i=\min\{d\}$，若节点 i 为局部密度最大的点，则取 $\delta_i=\max_j(d_{ij})$. 值得注意的是，依照定义，对于局部密度或全局密度最大的点，显然会是一个社区中心，其 δ 值会普遍大于到它最近邻居的距离，若以 ρ 和 δ 为坐标系画出决策图，则此点会出现在决策图的右上角；而对于其它的社区中心，其 δ 值也会大于一般节点，结合筛选条件，即可发现所有的社区中心.

2.4　最优带宽以及核心节点选取

本文使用改进的 Sheather-Jones(ISJ)算法计算最优带宽. 它的优势是可以使用快速余弦变换来估计带宽，而无需考虑分布上的正态性假设. Botev 等[15]提出非线性方程的独特解，可自适应地找出 KDE 的最优带宽 t，

$$t=\xi\gamma^{[l]}(t). \qquad (10)$$

ISJ 算法的详细原理本文不再叙述，最佳带宽 t 可以缩放内核函数来估计更精确的密度. 本文使用 $T_{\text{boundary}}=sqrt(t)/3.3$ 修复集群的边界点.

在依次计算每个节点对应的 ρ_i、δ_i 的后，选取 $\delta\geqslant E(\delta_i)+\sigma(\delta_i)$ 作为社区划分中的核心节点，每一个核心节点代表一个社区. 然后将一般节点分配到距其最近的核心节点所在的社区. $E(\delta_i)$ 和 $\sigma(\delta_i)$ 分别代表均值和方差. 到此，完成社区发现的所有过程.

算法 1. 峰值密度聚类的自适应社区发现算法

输入：数据集 G

输出：社区划分

1)初始化邻接关系

2)根据式(3)~(6)计算距离矩阵

3)根据式(10)计算 t

4)根据式(9)的 KDE 替代式计算 ρ

5)利用密度峰值聚类原理计算 δ

6)选择核心节点

7)将一般节点划分到距其距离最近的核心节点处构成社区

8)整理社区划分结果

3　实验与分析

3.1　实验数据集

为验证 KDED 算法的可行性和有效性，本文选取 4 个真实数据集和 2 种规模的人工合成数据集进行实验仿真.

3.1.1　真实数据集

表 1 列出 4 个真实数据集，且这些数据集具有已知的社区结构，分别为 Karate 数据集、Dolphins 数据集、Football 数据集和 Polbooks 数据集.

Karate 网络反应某大学空手道俱乐部之间的成员关系，节点连接的边代表成员的关系紧密. 由于校长与俱乐部主管间存在矛盾，该网络分裂为两个社区.

Dolphins 网络反应 Doubtful Sound 海峡中海豚的群居情况. 一个节点代表一只海豚，节点相连的边代表海豚之间接触频繁. 经过长期观察，该网络被分为两个社区.

Football 网络数据集中每一个节点代表一只球队，节点之间的连边代表球队之间曾经进行过比赛. 由于球队属于不同的州，每个州内的比赛场次多，因此该网络被分为 12 个社区.

Polbooks 网络反映亚马逊政治书籍的分类. 网络数据集中每一个节点代表一本在亚马逊销售的政治类书籍，节点之间的连边代表两本书被同时购买的频率高. 由于政治类书籍代表的政治立场不同，所以该网络被划分为 3 个社区，分别代表“自由”、“保守”和“中立”.

表 1 真实数据集

Tab.1 Real data set

数据集	节点数	边数
Karate	34	78
Dolphins	62	158
Football	115	612
Polbooks	105	441

3.1.2 人工网络数据集

Lancichinetti A 提出著名的 LFR 人工基准网络构造模型[20],已成为测试社区发现算法性能的一种主要方法. LFR 基准网络有两大优势:(1)它模拟节点度数和社区大小的无标度特征;(2)已知社区划分结构.

本文中一共采用两组 LFR 基准网络,具体参数如表 2 所示. 其中表中 N 为网络的节点数, K 为平均度数, $K_{\max}$ 为最大度数, t_1 和 t_2 分别为度分布和社区大小分布的幂指数,本文统一使用 $t_1 = 2$ 和 $t_2 = 1$, u 为模糊参数, $C_{\min}$ 和 $C_{\max}$ 为社区规模的最大值和最小值. 通过 u 调节网络结构的模糊程度来验证算法性能. 具体参数如表 2 所示.

表 2 LFR 基准网络参数

Tab.2 Parameters of LFR benchmark networks

网络	N	K	$K_{\max}$	Cmin	Cmax	t_1	t_2	u
N_1	1 000	20	50	10	50	2	1	0.1 ~ 0.9
N_2	5 000	20	50	30	80	2	1	0.1 ~ 0.9

3.2 评价指标

3.2.1 模块度 Q

模块度 Q 是由 Girvan 和 Newman 提出的一种被广泛应用于社区划分结果的评价指标,它反应社区结构的稳定性,其定义为

$$Q = \sum_{ij} \left[\frac{A_{ij}}{2m} - \frac{k_i * k_j}{(2m)(2m)}\right]\delta(c_i, c_j). \quad (10)$$

Q 表示网络中社区内部节点之间的连边所占的比例和另一个随机网络中社区内部节点连边所占比例的期望相减得到的差值.

式中:$\boldsymbol{A}$ 是网络 G 对应的邻接矩阵, k_i 和 k_j 对应点 i 和 j 的度数, m 是总连接边数. 实际连边数越是高于随机期望,说明节点越有集中在某些社区内的趋势,即网络的模块化结构越明显. 其取值范围为 - 1 到 1. 数值越接近 1 表示社区划分的结果越好.

3.2.2 标准化互信息

标准化互信息(normalized mutual information, NMI)是由 Danon 等提出的一种衡量已知社区结构网络的评价标准,它反应社区划分结果与实际划分结果的相似性,其定义为

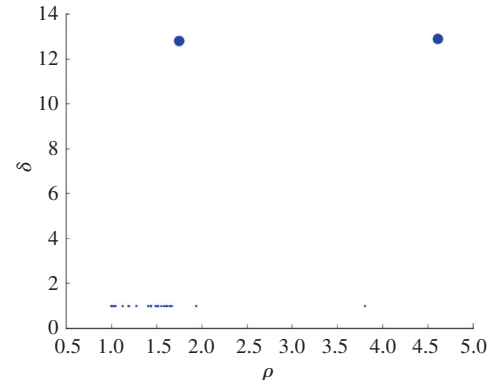
$$N_{\mathrm{NMI}}(R \mid P) = 1 - [H(R \mid P) + H(P \mid R)]/2. \quad (11)$$

式中: R 为真实划分结果, P 为预测划分结果, $H(R \mid P)$ 为 R 关于 P 的归一化条件熵. N_{NMI} 的取值范围为 0 到 1,其值越高表明所划分的社区结构与已知的社区结构越接近. 当其取值为 1 时,所划分的社区结构与已知社区结构相同.

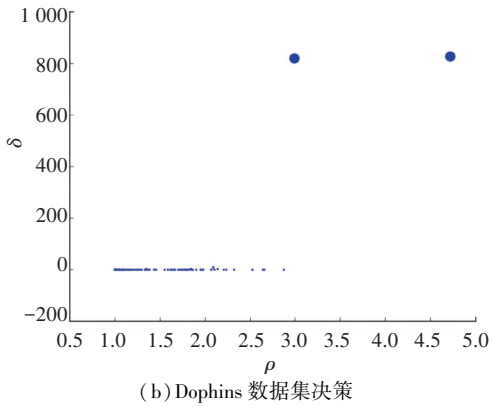
3.3 实验结果与分析

本文将所提出的 KDED 算法与 Fastgreed 算法、Informap 算法、Walktrap 算法、LPA 算法在真实数据集上进行对比,与 Fastgreed 算法、Informap 算法、Walktrap 算法、LPA 算法、VDDPC 算法、LCCD 算法在人工基准网络上进行比较. 图 1 和图 2 为 KDED 算法在真实数据集上的到的决策图和社区划分结果,表 3 列出 5 种算法在真实数据集上得到的仿真结果,图 3 描绘 7 种算法在人工数据集上的仿真结果. 图 1(a) ~ (d)为在 Karate 数据集、Dolphins 数据集、Football 数据集和 Polbooks 数据集上得到的决策图. 可以看出,核心节点的局部密度 ρ 以及到更高密度节点的距离 δ 明显大于一般节点,这会使后续核心节点的筛选过程更为直观、准确;此外,在选取核心节点与以及划分社区结构之后,结合距离矩阵可以得到十分清晰的社区内部的层次结构. 图 2(a) ~ (d)为 KDED 算法在 4 个真实数据集上得到的社区划分结果. 这样首先从划分结果的角度上粗略地说明本文提出的基于信任度的距离度量以及使用核密度估计规避截断距离 d_c 的方法是合理有效的.

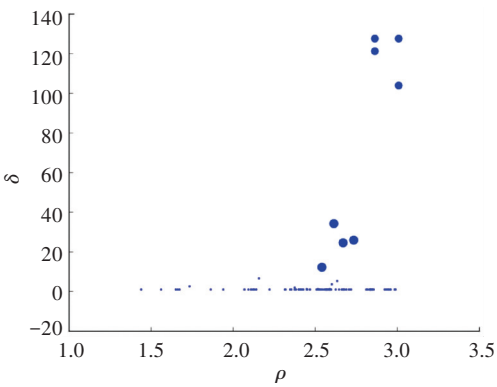
接下来评估划分结果的性能. 表 3 为本文所提出的 KDED 算法与其他 4 个算法在 Karate、Dophins、Polbooks、Football4 个真实网络数据所得到的结果. 可以观察到,KDED 算法在 Q 值上与几种算法相近,没有十分优秀的表现,在 karate 网络中取得最好的 Q 值,在其他的网络中与各种经典算法效果相近;但正如前文所讲,在模块度 Q 长期作为衡量社区发现算法性能好坏的地位受到部分学者质疑之后, NMI 逐渐成为社区发现领域比较被看重的指标,在已知社区划分的真实网络数据集中,通过 NMI 可以更为直观的观察出算法的准确性,更具有参考价值. LPA 具有令人满意的时间效率,但由于标签传播的不确定性,其性能远远不能令人满意. 通过数据可以观察到,相比较于其它算法, KDED 算法具有很高的 NMI 值,虽然在 Football 数据集上的 NMI 略低于 LPA 算法,但二者相差不大. 这意味着本文提出的 KDED 算法在不同规模的真实网络下可以得到较为符合实际社区划分的结果,初步验证其可行性与准确性.



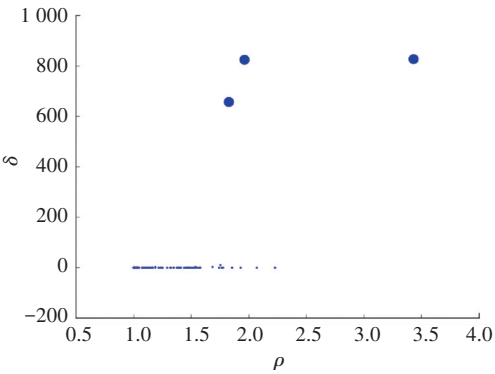
(a) Karate 数据集决策



(b) Dophins 数据集决策



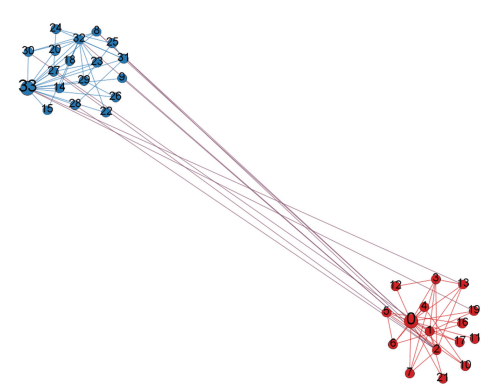
(c) Football 数据集决策



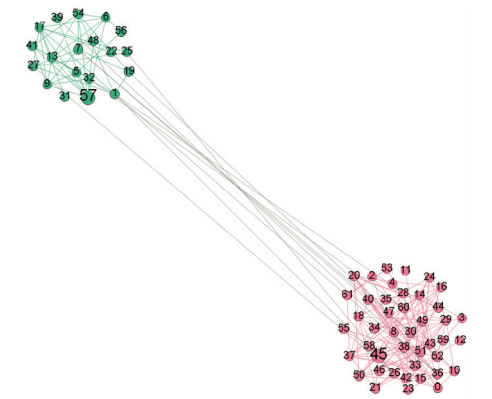
(d) Polbooks 数据集决策

图 1　KDED 算法下的决策

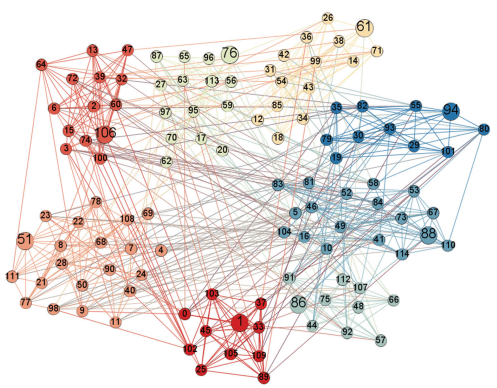
Fig.1　Decision diagramof KDED algorithm



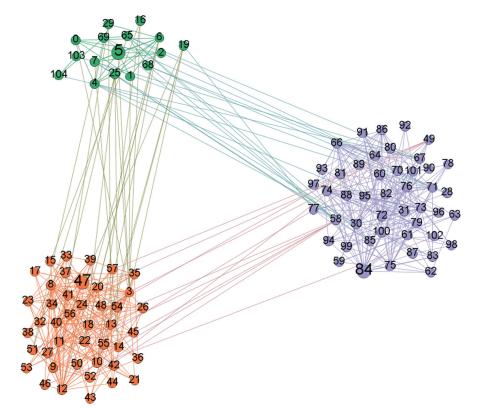
(a) Karate 社区划分



(b) Dophins 社区划分



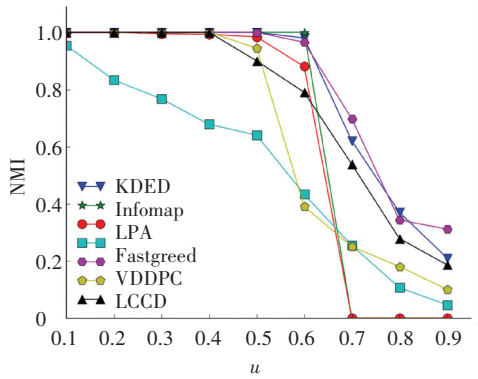
(c) Football 社区划分



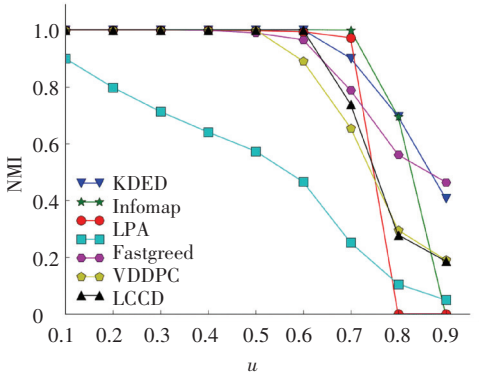
(d) Polbooks 社区划分

图 2　KDED 算法下的社区划分

Fig.2　Community division under KDED algorithm



(a)1 000 节点



(b)5 000 节点

图 3 LFR 人工基准网络下的仿真结果

Fig.3 Simulation results of LFR benchmark networks

表 3 真实网络仿真结果

Tab.3 Real network simulation results

网络	本文 KDED		Fastgreed		Informap		Walktrap		LPA	
	Q	N_{NMI}	Q	N_{NMI}	Q	N_{NMI}	Q	N_{NMI}	Q	N_{NMI}
Karate	0.402	1	0.381	0.692	0.371	0.699	0.353	0.504	0.360	0.836
Dolphins	0.447	0.778	0.482	0.604	0.519	0.481	0.487	0.418	0.417	0.735
Football	0.585	0.741	0.550	0.698	0.601	0.924	0.603	0.877	0.583	0.850
Polbooks	0.506	0.551	0.502	0.531	0.523	0.493	0.507	0.543	0.505	0.546

图 3 为 KDED 算法与其他几种算法在 LFR 基准网络中的实验结果,节点的规模为 1000 节点(a)和 5000 节点(b),刻画 NMI 随着模糊度 u 的增加的变化趋势.从曲线趋势走向来看,随着模糊参数 u 的增加,社区之间的界限逐渐模糊,划分难度增加,曲线的下降代表着 NMI 值逐渐降低,也意味着算法性能的下降. 在 1000 节点规模的网络中,$u < 0.6$ 时,KDED 算法的 $N_{\mathrm{NMI}} = 1$,这意味着可以得到符合真实社区划分的结果,与其他算法相比,当 $u > 0.5$ 后,LPA 算法和 Informap 算法的 NMI 值迅速下降至 0,通过后台观察社区划分结果发现其划分的社区个数为 1 ,这说明算法因社区间模糊度过高,已无法继续划分有效社区. 从总体上看, Informap 算法在 $u \leqslant 0.6$ 时取得最好的整体效果,但在 $u > 0.6$ 后算法性能急剧下降; Fastgreed 算法在整个实验过程中表现较差,VDDPC 与 LCCD 算法在 $u > 0.4$ 后算法性能下降较快;KDED 算法与 Walktrap 算法整体性能最好,当 $u < 0.6$ 时,两种算法均能划分出真实的社区结构,即 $N_{\mathrm{NMI}} = 1$,随后其性能下降. 值得注意的是,在节点规模与社区规模增加时,如图 3(b),曲线变化的趋势是不同的,KDED 与 Walktrap 在 $u \geqslant 0.7$ 时性能上有明显的下降, Walkstrap 性能下降的速度快于 KDE;但是 Informap 与 LPA 算法在大规模社区的检测上具有更好的性能, $0.6 \leqslant u \leqslant 0.7$ 时二者的性能优于 KDED,但是 KDED 的曲线没有发生急剧下降至 0 的情况,总体性能较为平稳.

相比于一般社区发现算法只能发现社区外部的自然结构的情况,KDED 可以通过各个节点在社区中到中心节点距离来发现社区内部的层次机构. 在算法原理的叙述中可以充分地体现,可以为需要解社区内部结构的学科带来便利,如:舆论检测、推荐系统. 基于以上分析, KDED 算法具有比较突出的性能和最好的稳定性,还可发现社区的层次结构,可以认为它的表现是非常具有竞争力的.

4 总 结

1)本文提出的 KDED 算法,具有 2 个特点: 1. 提出一种基于信任度的新的距离度量,该距离度量促进同社区内节点的相互靠拢,使得其更易被聚类算法所应用;2. 通过核密度估计规避仅能通过主观决策或的参数选择,使算法自适应地根据网络特性进行有效的社区划分,划分结果更加准确、有效.

2)通过在真实数据集以及 LFR 人工网络下的实验仿真,对于已有社区划分的数据集,KDED 算法的 NMI 值普遍高于现有算法,证明 KDED 算法所划分的社区结构,更加接近真实划分结果,验证算法的可行性与有效性.

3)KDED 不仅有效发现自然社区结构,而且可以展示社区内部的层次结构,将来可以对 KDED 进行一些有意义的延伸,如:结合一些改进,还可以用

于发现重叠的社区.

参考文献

[1] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826. DOI: 10.1073/pnas.122653799.

[2] VINCENT D B. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 2008(10): 155-168. DOI: 10.1088/1742-5468/2008/10/P10008.

[3] JOHNSON N F, ZHENG M, VOROBYEVA Y, et al. New online ecology of adversarial aggregates: ISIS and beyond[J]. Science, 2016, 352(6292): 1459. DOI: 10.1126/science.aaf0675.

[4] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76(2): 036106. DOI: 10.1103/PhysRevE.76.036106.

[5] 张鑫, 刘秉权, 王晓龙. 稳定标签传播的社区发现方法[J]. 哈尔滨工业大学学报, 2016, 48(11): 47-52. DOI: 10.11918/j.issn.0367-6234.2016.11.008.

ZHANG Xin, LIU Bingquan, WANG Xiaolong. Community detection method based on stable label propagation[J]. Journal of Harbin Institute of Technology, 2016, 48(11): 47-52. DOI: 10.11918/j.issn.0367-6234.2016.11.008.

[6] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences, 2007, 105(4): 1118-1123. DOI: 10.1073/pnas.0706851105.

[7] ZHOU Haijun. Network landscape from a Brownian particle´s perspective.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2003, 67(4Pt1): 041908. DOI: 10.1103/PhysRevE.67.041908.

[8] LIU Wei. Detecting Communities Based on Network Topology[J]. Scientific Reports, 2014, 4(4): 5739. DOI: 10.1038/srep05739.

[9] AGARWAL G, KEMPE D. Modularity-maximizing graph communities via mathematical programming[J]. The European Physical Journal B, 2008, 66(3): 409-418. DOI: 10.1140/epjb/e2008-00425-1.

[10] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496. DOI: 10.1126/science.1242072.

[11] 黄岚, 李玉, 王贵参, 等. 基于点距离和密度峰值聚类的社区发现方法[J]. 吉林大学学报(工), 2016, 46(6): 2042-2051. DOI: 10.13229/j.cnki.jdxbgxb201606038.

HUANG Lan, LI Yu, WANG Guisheng, et al. Community Detection Method Based on Point Distance and Density Peak Clustering[J]. Journal of Jilin University Engineering Edition, 2016, 46(6): 2042-2051. DOI: 10.13229/j.cnki.jdxbgxb201606038.

[12] HUANG Lan, WANG Guisheng, WANG Yan, et al[J]. International Journal of Modern Physics B, 2016, 30(24): 1650167. DOI: 10.1142/S0217979216501678.

[13] HENNIG C, HAUSDORF B. Design Of Dissimilarity Measures: A New Dissimilarity Between Species Distribution Areas[J]. Data Science & Classification, 2006: 29-37. DOI: 10.1007/3-540-34416-0_4.

[14] ZIEGLER C N, LAUSEN G. Analyzing Correlation between Trust and User Similarity in Online Communities[C]// International Conference on Trust Management. Springer Berlin Heidelberg, 2004: 251-265. DOI: 10.1007/978-3-540-24747-0_19.

[15] BOTEV Z I, GROTOWSKI J F, KROESE D P. Kernel density estimation via diffusion[J]. Annals of Statistics, 2010, 38(5): 2916-2957. DOI: 10.1214/10-AOS799.

[16] LEHMANN E L. Model Specification: The Views of Fisher and Neyman, and Later Developments[M]// Selected Works of E. L. Lehmann. 2012: 955-963. DOI: 10.1007/978-1-4614-1412-4_78.

[17] RASHID M. Clustering by fast search and find of density peaks via heat diffusion[J]. Neurocomputing, 2016, 208(6191): 210-217. DOI: 10.1016/j.neucom.2016.01.102.

[18] XU Xiaoyuan, YAN Zheng, XU Shaolun. Estimating wind speed probability distribution by diffusion-based kernel density method[J]. Electric Power Systems Research, 2015, 121: 28-37. DOI: 10.1016/j.epsr.2014.11.0.

[19] KRISHNASWAMY S, SPITZER M H, MINGUENEAU M, et al. Systems biology. Conditional density-based analysis of T cell signaling in single-cell data[J]. Science, 2014, 346(6213): 1250689. DOI: 10.1126/science.1250689.

[20] LANCICHINETTI A, FORTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80(1Pt2): 016118. DOI: 10.1103/PhysRevE.80.016118.

（编辑　苗秀芝）