【电子与信息科学 / Electronics and Information Science】

一种主动半监督大规模网络结构发现算法

柴变芳¹, 曹欣雨¹, 魏春丽¹, 王建岭²

1) 河北地质大学信息工程学院, 河北石家庄 050031; 2) 河北中医学院教务处, 河北石家庄 050200

摘要: 在线变分期望最大(online variational expectation maximization, onlineVEM)算法可快速发现大规模网络的聚类模式,但在网络结构复杂时算法的处理结果稳定性和准确性欠佳. 为更快更准地识别其聚类模式,提出一种主动半监督在线变分期望最大(active semi-supervised onlineVEM, AOnlineVEM)算法. 算法首先自动选择代表节点,确定类的个数,并基于代表节点初始化模型;然后迭代执行3个任务:运行在线算法onlineVEM、主动选节点及模型更新,直至算法达到准确率的设定阈值或收敛. 在不同结构的人工网络和真实网络上的实验结果表明, AOnlineVEM算法的准确性和效率均优于同类算法. AOnlineVEM算法利用主动选择的节点先验信息提高了网络聚类模式发现的稳定性及准确性,提高了在线算法的运行效率.

关键词: 计算机应用; 大规模网络; 半监督聚类; 主动学习; 在线变分期望最大算法; 成对约束

中图分类号: TP181

文献标志码: A

doi: 10.3724/SP.J.1249.2020.03243

An active semi-supervised structure exploring algorithm for large networks

CHAI Bianfang¹, CAO Xinyu¹, WEI Chunli¹, and WANG Jianling²

1) School of Information Engineering, Hebei GEO University, Shijiazhuang 050031, Hebei Province, P. R. China

2) Office of Educational Administration, Hebei University of Chinese Medicine, Shijiazhuang 050200, Hebei Province, P. R. China

Abstract: The algorithm of online variational expectation maximization (onlineVEM) can explore the clustering patterns of large networks fast. But the stability and accuracy of the algorithm are poor when the network structure is complex. In order to identify the clustering patterns faster and more accurately, an active semi-supervised online variational expectation maximization (AOnlineVEM) algorithm is proposed. Firstly, the algorithm selects the representative nodes automatically, determines the numbers of clusters, and initializes the model based on the representative nodes. Then, it iteratively executes three tasks: running the online algorithm onlineVEM, actively selecting nodes, and updating parameters until the algorithm reaches the preset threshold of accuracy or convergences. Experiments on artificial networks and real networks with different structures show that the accuracy and efficiency of AOnlineVEM algorithm are better than those of similar algorithms. The AOnlineVEM algorithm uses the priori information of actively selected nodes to improve the stability and accuracy of clustering pattern detection of networks and to improve the efficiency of online algorithm.

Key words: computer application; large networks; semi-supervised clustering; active learning; online variational expectation maximization (onlineVEM) algorithm; pairwise constraints

Received: 2019-04-15; Accepted: 2019-06-27

Foundation: National Natural Science Foundation of China (61503260); Natural Science Foundation of Hebei Province (F2019403070)

Corresponding author: Professor WANG Jianling. E-mail: wang_jl@126.com

Citation: CHAI Bianfang, CAO Xinyu, WEI Chunli, et al. An active semi-supervised structure exploring algorithm for large networks [J]. Journal of Shenzhen University Science and Engineering, 2020, 37(3): 243-250. (in Chinese)

当前社会正处于信息化浪潮的新阶段,大量数据时刻生成并可建模为大规模网络.快速、准确地识别这些网络数据潜在的聚类结构,是商家定位用户群体,制定精准营销策略的依据.网络中可能存在社区结构^[1-2]、星型结构或层次结构,甚至同时存在多种结构.关于这些网络的先验信息很少,成为人们发现网络潜在结构的难点.FORTUNATO^[2]采用社区发现算法识别网络中类内链接紧密和类间链接稀疏的结构,但当网络中可能不存在这种结构时,该算法不能发现真实的网络聚类模式.一些网络结构发现方法可识别网络的多类型潜在聚类模式,如基于随机块模型的算法^[3],但此类方法复杂度高,不能处理大规模网络.基于混合模型的网络结构发现算法——在线变分期望最大(online variational expectation maximization, onlineVEM)算法^[1],可发现大规模网络的多类型聚类结构,但其结果受初始参数影响,准确性和稳定性还有待提升.

近年来出现的半监督社区发现方法,利用先验信息提高了网络社区发现的准确率^[4-13].根据半监督社区发现算法使用先验的形式,将现有方法分为基于节点先验的方法^[4-6]和基于约束对先验的方法^[7-13].后者因先验信息容易获得而较为流行.根据约束对先验信息的获得方式,将现有算法大致分为2类:①随机选择约束对标记^[7-11],如YANG等^[9]提出基于非负矩阵分解(nonnegative matrix factorization, NMF)的超网络半监督社区发现算法(SuperNMF),随机选择约束对先验标记,利用约束先验改变网络的邻接矩阵,采用传统的社区发现方法获取网络节点聚类.②基于主动学习策略选择约束对进行人工标记,以较少的先验实现算法性能较大程度提升^[12-13].如YANG等^[13]提出迭代主动半监督社区发现算法ALISE(active link selection),主动选择最不确定的节点对进行人工标记,利用已标记链接重构网络拓扑结构.这些半监督社区发现算法主要发现网络中的社区结构,如果网络中不存在这种社区结构,或者存在其他类型结构,此类算法无效,且这些方法主要针对中小规模网络.

综合大规模在线网络结构发现方法、半监督社区发现方法和主动半监督聚类方法研究成果可知,通过主动选择信息量大的节点,利用约束对先验,对其进行类别标记,可用较少的代价获得高质量的先验,使算法不仅能更准确地识别网络的各种聚类模式,还能高效处理大规模网络.为此,本研究基于onlineVEM算法,提出一种用于大规模网络结构

发现的基于迭代框架的主动半监督在线变分期望最大(active semi-supervised onlineVEM, AOnlineVEM)算法.算法包含初始化阶段和在线算法迭代阶段.初始化阶段主动选择代表节点,并基于代表节点初始化模型参数.在线算法迭代阶段每次迭代执行3个任务:①运行onlineVEM算法;②不确定节点选择;③模型参数更新.在不同网络数据上与同类算法进行比对,结果表明,AOnlineVEM算法可利用少量先验信息较大程度提高网络结构发现的准确率.

1 大规模网络的主动半监督结构发现策略

onlineVEM算法是针对大规模网络结构发现设计的在线算法.算法先随机选择网络中部分节点构成子网络,利用期望最大(expectation maximization, EM)算法估计节点的类隶属度及混合模型参数,再迭代估计网络的剩余节点隶属度及模型参数.估计方法为:

迭代过程中选择1个节点 i ,它属于类 k 的概率 γ_{ik} 即为该节点的隶属度

$$\gamma_{ik} = \frac{\pi_k \prod_j \theta_{kj}^{A_{ij}}}{\sum_k \pi_k \prod_j \theta_{kj}^{A_{ij}}} \quad (1)$$

其中, $\theta_{kj}^{A_{ij}}$ 为类 k 的节点链接到节点 j 的概率, A_{ij} 为初始网络的邻接矩阵 A 的元素, $A_{ij}=0$ 或 1 ; π_k 为所有节点属于 k 的概率.

利用当前节点的隶属度更新模型参数,则第 $t+1$ 次迭代时所有节点属于类 k 的概率为

$$\pi_k^{t+1} = \frac{1}{n+1} N_k^{t+1} \quad (2)$$

其中, N_k^{t+1} 为第 $t+1$ 次迭代时节点属于类 k 的隶属度和, $N_k^{t+1} = N_k^t + \gamma_{(t+1),k}$.

第 $t+1$ 次迭代时,类 k 节点链接到节点 j 的概率为

$$\theta_{kj}^{t+1} = \frac{H_{kj}^{t+1}}{G_k^{t+1}} \quad (3)$$

其中, H_{kj}^{t+1} 为第 $t+1$ 次迭代类 k 节点链接到 j 的概率和, $H_{kj}^{t+1} = H_{kj}^t + \gamma_{(t+1),k} A_{ij}$; G_k^{t+1} 为第 $t+1$ 次迭代类 k 中节点链接到所有节点的概率和, $G_k^{t+1} = G_k^t + \gamma_{(t+1),k} \deg_{t+1}$, $\deg_{t+1}$ 为第 $t+1$ 次节点的度.

onlineVEM算法每次迭代随机选择1个节点,

利用式(1)计算其隶属度,再利用式(2)和式(3)更新模型参数,迭代此过程直到算法准确率或迭代次数达到阈值(准确率阈值通常为算法两次迭代优化目标函数差值 $< 1 \times 10^{-6}$, 迭代次数阈值设为网络节点数的2倍)。若初始估计的模型参数不准确,会导致后期更新的模型参数是一个较差的局部最优解;每次随机选择1个节点则会使算法不能快速收敛。因此,可利用主动策略选择信息量大或难以识别的节点进行标记的特点,将这些信息作为监督信息,以提高算法的准确率。本研究采用3种策略对onlineVEM算法的准确性及效率进行改进。

策略1 主动选择代表节点进行标记。考虑到在线算法上次迭代参数对下次迭代的参数更新有较大影响,初始参数应保证有较高准确性,因此,需在初始化参数时尽量确定影响力大的节点的类型隶属度。为使代表节点对其同类节点具有较大影响,可通过主动策略选择代表节点标定初始化参数,并利用代表节点的精确标定最大程度影响剩余节点。基于 k -rank-D^[14-15] 选择最可能成为代表节点的候选集合,并用标定的约束对先验从候选集合中选择各类的代表节点,利用代表节点初始化模型参数。

策略2 主动选择不确定节点进行标记。当代表节点确定后,网络中仍存在一些难以确定类别的边界节点。通过节点类型隶属度信息主动选择不确定性高的节点,利用约束对先验确定其与哪个代表节点具有必联(must-link)关系,从而实现不确定节点类别的标记。

策略3 基于迭代框架的在线算法。迭代运行onlineVEM算法,每次基于上次算法迭代的结果,选择最不确定的节点进行标记,进而得到一个更优解,下次算法从该解开始寻找更优的解。该迭代框架能找到比运行1次onlineVEM算法更优的解。为提高算法运行效率,在迭代过程中,对于已标记类别信息的节点在后续算法中将不再估计其类型隶属度。

2 AOnlineVEM 算法

2.1 变量定义

定义 c 为类的个数; n 为网络节点数; H 为候选代表节点集合; N_k 为第 k 类节点集合, N_{k1} 为第 k 类的代表节点; N 为当前各子类组成的集合,即 $N = \{N_k\}_{k=1}^{\text{curk}}$, curk 为当前类别数; p_i 为节点 i 是代表节点的概率, p 为所有节点是代表节点的概率集合; Θ 为onlineVEM算法返回的模型参数。

2.2 AOnlineVEM 算法描述

算法包括模型初始化和提纯两个阶段。

2.2.1 模型初始化阶段

首先,基于 k -rank-D 获得网络候选代表节点集合;然后,依次主动选择各类的代表节点;最后,基于 Newman 混合模型^[16] 初始化模型参数。

1) 选择候选代表节点。 k -rank-D 算法假设:代表节点的中心度较大,周围节点的中心度较小且距离其他代表节点较远。节点 i 的中心度为 v_i ($i = 1, 2, \dots, n$), 初始中心度 $v_0 = 1/n$, 第 $t+1$ 次迭代时,所有节点的中心度为

$$v^{t+1} = \left[(1 - \beta) P + I \frac{\beta}{n} \right] \quad (4)$$

其中,重启概率 $\beta = 0.15$; P 为 $n \times n$ 维转移概率矩阵,其元素 P_{ij} 为节点 i 与节点 j 之间的转移概率,

$$P_{ij} = \frac{A_{ij}}{\sum_j A_{ij}}; I \text{ 为单位矩阵。}$$

节点 i 的离散度 δ_i 指节点 i 距离下一个最近代表节点 j 的最小距离,即 $\delta_i = \min_{j: v_j > v_i} d_{ij}$, d_{ij} 为节点 i 和 j 的距离。节点 i 是代表节点的概率为

$$p_i = \frac{v_i \delta_i}{\max_{j=1}^n v_j \max_{j=1}^n \delta_j} \quad (5)$$

将所有节点的 p_i ($i = 1, 2, \dots, n$) 值按降序排列,选择前几个元素对应的节点作为候选节点集合 H 的元素。不同规模网络选择的节点数不同,但不超过 $\sqrt{n}$ 。如果网络规模较大,考虑节点度的信息,对一些度值大的节点采样并计算其为代表节点的概率。

2) 主动确定代表节点集合。对 H 中的节点按照成为代表节点的可能性进行排序,选择代表性最大的节点作为初始类节点,并记为 N_1 的元素。然后,依次选择下一个 H 中未查询的节点 i , 确定其类别。如果节点 i 与已有 N_k 存在必联关系,则添加到 N_k ; 如果 i 与 $\{N_k\}_{k=1}^{\text{curk}}$ 中节点都不存在必联关系,则再构造一个类集合,并将节点 i 加入到新集合中,作为新的类代表节点,直到类的个数达到需要或 H 中所有元素处理完,将此时的代表节点集合 N 作为初始类集合。

3) 初始化模型参数。基于 Newman 混合模型利用 EM 算法估计 N 集合中代表节点隶属各类的概率;随机初始化 N 集合外所有剩余节点的类型隶属度;最后,计算初始模型参数。

2.2.2 模型提纯阶段

通常基于模型初始化阶段得到的参数计算网络

聚类的结果并不准确,需多次运行在线参数估计算法,选择不易识别类隶属度的节点进行标记,逐步提升模型性能,迭代该过程,直到达到阈值.每次迭代需进行 3 个操作:运行在线算法 onlineVEM、主动选择节点及模型参数更新.首先,运行 onlineVEM 算法更新未标记节点的隶属度及模型参数.然后,采用式(6)估计节点 i 的熵,选择熵最大的前几个不确定节点(本研究选择 3)作为待标记节点集合.最后,查询每个待标记节点 i 与 N 中代表节点是否存在必联关系,若存在,则其与代表节点属于同类,将节点 i 加入代表节点所在的 N_k 并更新节点隶属度和模型参数.迭代运行上述过程直至达到准确率和先验限制的阈值.

$$\text{entropy}(i) = - \sum_{k=1}^K \gamma_{ik} \lg \gamma_{ik} \quad (6)$$

ASonlineVEM 算法伪代码和源代码请扫描论文末页右下角二维码见图 S1 和图 S2.第 1 步的中心度和节点间距离可通过离线并行算法计算,不是影响算法运行时间的主要因素.第 1 步的运行时间主要为计算节点离散度和代表性所用的时间,时间复杂度为 $O(n)$.第 2 步的运行时间是对少量节点的计算,但这与对大规模网络节点和边相比占用时间很少.第 3 步初始化模型参数的复杂度为 $O(n)$.第 4 步迭代的主要运行时间是在计算节点的熵和运行 onlineVEM 算法上,其复杂度分布为 $O(nc)$ 和 $O(mc)$, m 为网络边数.虽然第 4 步需多次迭代,但 onlineVEM 算法总是以上次迭代得到的参数作为初始参数,因此,算法每次迭代都会很快收敛.第 4 步的时间复杂度为 $O(tmc)$, t 为迭代次数,网络节点越少则 t 越小.对于大规模网络,即 m 较大时,可通过每次迭代对邻接边进行采样来提高效率.因此,ASonlineVEM 算法可用来处理大规模网络,总的时间复杂度为 $O(tmc)$.

2.3 ASonlineVEM 算法收敛性证明

ASonlineVEM 算法是基于 onlineVEM 的迭代算法,每次迭代首先基于当前隶属度和模型参数运行 onlineVEM 算法,再选择 3 个节点,利用先验更新其隶属度值. CHAI 等^[1]证明了 onlineVEM 算法收敛,并设第 t 次迭代算法收敛到一个局部最优值 Q' ,此时,模型参数为 Θ' ,节点隶属度为 γ' .然后,利用先验而非式(1)修正选择节点的隶属度,使节点更快地收敛到正确的隶属度,隶属度更新为 γ' ,基于式(2)和式(3)将模型参数更新为 Θ' .第 $t+1$ 次迭代将以第 t 次的参数作为初始值,其对应似然收

敛到更优,即 $Q^{t+1} \geq Q'$.因此,经过一定迭代次数可使算法收敛,后续实验中通过算法运行结果也可验证其收敛性.

3 实验及结果分析

选取不同结构的人工网络和真实网络数据集对 ASonlineVEM 算法进行测试,并与半监督社区发现 SuperNMF 和 ALISE 算法比较.其中,主动节点标记使用的约束对先验比例 R_p 为网络节点对的百分比;网络节点对为 $n(n-1)/2$, n 为网络节点数.

使用如式(7)的标准化互信息^[17](normalized mutual information, NMI)来衡量算法准确性.它可较好地评估基准标签与计算得到的标签之间的相互关系和吻合程度,常用来评价网络结构发现的质量. NMI 值越接近 1,表示算法聚类准确度越高.

$$\text{NMI}(X, Y) = \frac{2I(X; Y)}{H(X) + H(Y)} \quad (7)$$

其中, X 是计算结果; Y 是真实结果; $I(X; Y) =$

$$\sum_x \sum_y p(x, y) \lg \frac{p(x, y)}{p(x)p(y)}; H(X) = - \sum_x [p(x) \lg p(x)]; H(Y) = - \sum_y p(y) \lg p(y); p(x) \text{ 和 } p(y) \text{ 分别为 } x \text{ 和 } y \text{ 的概率分布; } p(x, y) \text{ 为 } x \text{ 和 } y \text{ 的联合概率分布.}$$

算法使用 Matlab 2014 软件在 Intel® Core™ i5-6200U CPU, 8 Gbyte 内存的 Windows 7 (64 bit) 计算机上运行.考虑到算法每次运行都存在差异性,取 5 次结果的均值作为实验结果.

基准网络 Girvan-Newman (GN) 网络中,顶点与社团外顶点连边数的均值 Z_{out} 决定社区结构的清晰程度, Z_{out} 越大,网络结构越复杂.图 1 分别为 $Z_{\text{out}} = 8$ 和 $Z_{\text{out}} = 9$ 时 3 种算法的 NMI 结果对比.可见,ASonlineVEM 算法准确率明显高于其他网络,尤其当 $Z_{\text{out}} = 9$ 时, $R_p = 1.6\%$, ASonlineVEM 算法准确率为 0.92, ALISE 算法仅 0.33.随着 Z_{out} 增加,对比算法准确率明显下降,ASonlineVEM 准确率较高,说明后者能准确发现更复杂的网络结构.

Lancichinetti-Fortunato-Radicchi (LFR) 基准网络结构比 GN 网络的更复杂.设网络节点数分别为 1 000,最小和最大社区节点数分别为 20 和 100,节点度分布参数为 2,社区大小分布参数为 1,混合参数 μ 决定了发现网络结构的难易程度,设 μ 为 0.7 和 0.8.图 2 为在 LFR 基准网络上 3 种算法的

图1 三种算法在 GN 网络上的 NMI 对比结果
Fig. 1 NMI comparison results of three algorithms on GN networks

图2 三种算法在 LFR 网络的 NMI 对比结果
Fig. 2 NMI comparison results of three algorithms on LFR networks

NMI 结果. 由图 2 可见, ASonlineVEM 算法的的准确率明显高于对比的半监督社区发现算法. 综上可

见, 在具有社区结构的基准网络上, ASonlineVEM 算法的社区发现准确率要高于对比算法, 尤其在网络结构不清晰时更能体现算法的优势.

人工网络 采用 ASonlineVEM、SuperNMF 和 ALISE 算法对人工生成具有社区结构 (network1) 和二分结构 (network2 和 network3) 的网络进行实验, 参数设置如表 1. 图 3 为 3 种算法在 3 个网络上的性能实验结果. 由图 3 可见, 在社区结构网络中, 3 种算法效果相当; 而在二分结构网络中, 相同先验比例下 ASonlineVEM 算法准确率高于对比算法. 如在 network2 中, $R_p = 0.12\%$ 时, ASonlineVEM 算法的准确率约是对比算法的 2 倍. 在人工网络上, ASonlineVEM 的准确率高于对比算法.

表 1 人工网络数据集

Table 1 Synthetic network dataset				
网络	节点数	边数	类别数	类间概率矩阵
network1	2 000	539 888	3	$[0.6, 0.1, 0.1;$
				$0.1, 0.6, 0.1;$
				$0.1, 0.1, 0.6]$
network2	2 000	806 718	3	$[0.5, 0.1, 0.5;$
				$0.1, 0.5, 0.5;$
				$0.5, 0.5, 0.5]$
network3	5 000	3 851 286	5	$[0.5, 0.1, 0.1, 0.1, 0.5;$
				$0.1, 0.5, 0.1, 0.1, 0.5;$
				$0.1, 0.1, 0.5, 0.1, 0.5;$
				$0.1, 0.1, 0.1, 0.5, 0.5;$
				$0.5, 0.5, 0.5, 0.5, 0.5]$

真实网络 本研究使用的真实网络数据集包括经典网络 Dolphins、Football、Adjnoun 和 Friendship^[18], 以及 Facebook 网络 Baylor、USC、Maryland 和 NYU^[19]. 真实网络数据集的具体参数请扫描论文末页右下角二维码见表 S1. 图 4 为分别采用 ASonlineVEM、SuperNMF 和 ALISE 算法在小规模经典网络数据集上测得的 NMI 结果. 由图 4 可见, 不同先验比例下 ASonlineVEM 算法性能均优于其他算法.

ASonlineVEM 和 ALISE 算法都是基于迭代框架的主动半监督社区发现算法, 图 5 为 2 种算法在真实大规模复杂网络 Baylor 和 USC 数据集上的 NMI 性能对比. 由图 5 可见, 在大规模复杂结构网络上, ASonlineVEM 算法的准确性都优于 ALISE 算法.

图3 三种算法在人工网络上的 NMI 对比结果
Fig. 3 NMI comparison results of three algorithms on synthetic networks

对于具有社区结构的网络, ASONlineVEM 算法初始通过主动策略选择代表节点进行标定, 保证初始参数有较高的准确性, 再利用代表节点的精确标定最大程度影响剩余节点. 迭代过程主动选择不确定性且信息量大的节点进行标记, 提高了算法发现网络结构的准确率. SuperNMF 算法随机选择先验导致无法充分利用高质量先验信息. ALISE 算法可发现社区网络的聚类结构, 但非社区网络 network2 和 network3 的聚类结构属于二分结构, SuperNMF 和 ALISE 算法主要处理的是社区结构发现问题, 无法发现多类型网络聚类结构.

分析不同网络的准确率结果发现, 随着先验比例的增加, ASONlineVEM 算法较其他同类算法准确率更高、收敛速度更快、稳定性更佳, 表明该算法

图4 三种算法在经典真实网络的 NMI 对比结果
Fig. 4 NMI comparison results of three algorithms on classic real networks

效果均优于其他算法.

ASONlineVEM 算法不仅能准确识别网络的聚类模式, 还能高效处理不同规模网络. 对于规模较小

(a) Baylor数据集

(b) USC数据集

图5 AOnlineVEM 和 ALISE 算法在 Facebook 网络上的 NMI 对比结果

Fig. 5 NMI comparison results of AOnlineVEM and ALISE algorithms on Facebook networks

的基准网络, 如 $Z_{out} = 9$ 的 GN 网络, 当 $R_p = 2\%$ 时, AOnlineVEM、SuperNMF 和 ALISE 算法运行时间 t 分别为 0.258 7、0.611 6 和 0.552 0 s. 在网络节点较少的真实网络 Football 上, 当 $R_p = 14\%$ 时, AOnlineVEM、SuperNMF 和 ALISE 算法的运行时间分别为 0.185 4、0.364 0 和 4.296 4 s. 可见, AOnlineVEM 算法运行效率最高, 其他小规模网络上也有相同运行结果.

针对节点较多或边数较多的大规模复杂网络, 为简单起见, 给出部分大规模网络上 AOnlineVEM 和 ALISE 算法的运行时间对比结果. network2 和 network3 人工网络的时间对比结果如图 6. 图 7 为两种算法在 Facebook 数据集上数据传输一次所需要的时间对比结果. 总的来说, 在不同规模网络上, AOnlineVEM 算法明显比 ALISE 算法高效.

AOnlineVEM 算法在迭代过程中, 对已标记类别信息的节点将不在后续算法中估计其类隶属度, 只计算剩余节点的类隶属度, 大幅节省了算法耗时. ALISE 算法每次迭代选择网络中最不确定的链接进行人工标记, 而大规模网络的边数往往远大于

节点数, 这使 ALISE 算法每次迭代计算边不确定性的复杂度较大, 导致算法运行时间长.

(a) network2

(b) network3

图6 AOnlineVEM 和 ALISE 算法在人工网络上的运行时间对比

Fig. 6 Time comparison of AOnlineVEM and ALISE algorithms on synthetic networks

图7 AOnlineVEM 和 ALISE 算法在 Facebook 网络上运行时间对比

Fig. 7 Time comparison of AOnlineVEM and ALISE algorithms on Facebook networks

结 语

提出了主动半监督大规模网络结构发现算法 AOnlineVEM, 基于网络数据可快速且准确地识别潜在的多类型聚类结构, 利用主动学习方法选择构造高质量先验提高了算法准确率, 利用在线技术提

高了算法运行效率,有利于在推荐系统和舆情分析中用户群体划分等领域的实际应用。但是,该算法目前主要针对的是静态大规模网络,如何实现动态大规模网络的结构发现算法以及在大数据平台下的大规模网络结构发现算法仍需深入研究。

基金项目: 国家自然科学基金资助项目 (61503260); 河北省自然科学基金资助项目 (F2019403070)

作者简介: 柴变芳 (1979—), 河北地质大学副教授、博士。研究方向: 复杂网络分析和机器学习等。

E-mail: chaibianfang@163.com

引文: 柴变芳, 曹欣雨, 魏春丽, 等. 一种主动半监督大规模网络结构发现算法[J]. 深圳大学学报理工版, 2020, 37(3): 243-250.

参考文献 / References:

- [1] CHAI Bianfang, JIA Caiyan, YU Jian. An online expectation maximization algorithm for exploring general structure in massive networks [J]. *Physica A: Statistical Mechanics and Its Applications*, 2015, 438: 454-468.
- [2] FORTUNATO S. Community detection in graphs [J]. *Physics Reports*, 2010, 486 (3/4/5): 75-174.
- [3] 赵学华, 杨博, 陈贺昌. 一种高效的随机块模型学习算法[J]. *软件学报*, 2016, 27(9): 2248-2264.
ZHAO Xuehua, YANG Bo, CHEN Hechang. An efficient stochastic block model learning algorithm [J]. *Journal of Software*, 2016, 27(9): 2248-2264. (in Chinese)
- [4] YA Fangli, CAI Yanjia, JIAN Qiangli, et al. Enhanced semi-supervised community detection with active node and link selection [J]. *Physica A: Statistical Mechanics and Its Applications*, 2018, 510: 219-232.
- [5] EATON E, MANSBACH R. A spin-glass model for semi-supervised community detection [C]// *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Atlanta, USA: AAAI Press, 2012: 900-906.
- [6] MA Xiaoke, GAO Lin, YONG Xuerong, et al. Semi-supervised clustering algorithm for community structure detection in complex networks [J]. *Physica A: Statistical Mechanics and Its Applications*, 2010, 389: 187-197.
- [7] ZHANG Zhongyuan. Community structure detection in complex networks with partial background information [J]. *Europhysics Letters*, 2013, 101(4): 48005.
- [8] ZHANG Zhongyuan, SUN Kaidi, WANG Siqi. Enhanced community structure detection in complex networks with partial background information [J]. *Scientific Reports*, 2013, 3(11): 3241.
- [9] YANG Liang, JIN Di, HE Dongxiao, et al. Improving the efficiency and effectiveness of community detection via prior-induced equivalent super-network [J]. *Scientific Reports*, 2017, 7(1): 634.
- [10] YANG Liang, CAO Xiaochun, JIN Di, et al. A unified semi-supervised community detection framework using latent space graph regularization [J]. *IEEE Transactions on Cybernetics*, 2015, 45(11): 2585-2598.
- [11] YANG Liang, GE Meng, JIN Di, et al. Exploring the roles of cannot-link constraint in community detection via multi-variance mixed Gaussian generative model [J]. *PLoS ONE*, 2017, 12(7): e0178029.
- [12] CHENG Jianjun, LENG Mingwei, LI Longjie, et al. Active semi-supervised community detection based on must-link and cannot-link constraints [J]. *PLoS ONE*, 2014, 9(10): e110088.
- [13] YANG Liang, JIN Di, WANG Xiao, et al. Active link selection for efficient semi-supervised community detection [J]. *Scientific Report*, 2015, 5(3): 9039.
- [14] JIA Caiyan, LI Yafang, CARSON M B, et al. Node attribute-enhanced community detection in complex networks [J]. *Scientific Reports*, 2017, 7(1): 2626.
- [15] LI Yafang, JIA Caiyan, YU Jian. A parameter-free community detection method based on centrality and dispersion of nodes in complex networks [J]. *Physica A: Statistical Mechanics and its Applications*, 2015, 438: 321-334.
- [16] NEWMAN M E J, LEICHT E A. Mixture models and exploratory analysis in networks [J] *Proceedings of the National Academy of Sciences*, 2007, 104(23): 9564-9569.
- [17] LIU Xin, CHENG Huimin, ZHANG Zhongyuan. Evaluation of community detection methods [J]. *IEEE Transactions on Knowledge and Data Engineering*. (2019-04-17). <http://ieeexplore.ieee.org/stamp.jsp?tp=&number=8693534>.
- [18] NEWMAN M. Network data [DB/OL]. (2013-04-19). <http://www-personal.umich.edu/~mejn/netdata/>.
- [19] TRAUD A L, MUCHA P J, PORTER M A. Social structure of Facebook networks [J]. *Physica A: Statistical Mechanics and its Applications*, 2012, 391(16): 4165-4180.

【中文责编: 英子; 英文责编: 木柯】

补充材料