基于采样的大规模图聚类分析算法

张建朋[1,2],陈鸿昶[1],王 凯[1],祝凯捷[1,2],王亚文[1]
(1. 国家数字交换系统工程技术研究中心,河南郑州 450000;2. 荷兰埃因霍温理工大学计算机系,荷兰北布拉邦省 5600 MB)

摘 要: 针对当前聚类方法(例如经典的 GN 算法)计算复杂度过高、难以适用于大规模图的聚类问题,本文首先对大规模图的采样算法展开研究,提出了能够有效保持原始图聚类结构的图采样算法(Clustering-structure Representative Sampling,CRS),它能在采样图中产生高质量的聚类代表点,并根据相应的扩张准则进行采样扩张. 此采样算法能够很好地保持原始图的内在聚类结构. 其次,提出快速的整体样本聚类推断(Population Clustering Inference,PCI)算法,它利用采样子图的聚类标签对整体图的聚类结构进行推断. 实验结果表明本文算法对大规模图数据具有较高的聚类质量和处理效率,能够很好地完成大规模图的聚类任务.

关键词: 大规模图; 图采样; 图聚类; 整体推断; 聚类代表点; 扩张准则
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2019)08-1731-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.08.017

A Sampling-Based Graph Clustering Algorithm for Large-Scale Networks

ZHANG Jian-peng[1,2], CHEN Hong-chang[1], WANG Kai[1], ZHU Kai-jie[1,2], WANG Ya-wen[1]
(1. *National Digital Switching System Engineering & Technological R&D Center, Zhengzhou Henan* 450000, *China*;
2. *Dept of Computer Science, Technology University of Eindhoven, Eindhoven* 5600*MB, Netherland*)

Abstract: Since computational complexities of the existing methods such as classic GN algorithm are too costly to cluster large-scale graphs, this paper studies sampling algorithms of large-scale graphs, and proposes a clustering-structure representative sampling (CRS) which can effectively maintain the clustering structure of original graphs. It can produce high quality clustering-representative nodes in samples and expand according to the corresponding expansion criteria. Then, we propose a fast population clustering inference (PCI) method on the original graphs and deduce clustering assignments of the population using the clustering labels of the sampled subgraph. Experiment results show that in comparison with state-of-the-art methods, the proposed algorithm achieves better efficiency as well as clustering accuracy on large-scale graphs.

Key words: large-scale graphs; graph sampling; graph clustering; population inference; clustering representative nodes; expansion criteria

1 引言

目前,由于大规模的图数据在各个领域爆炸性增长,现有算法难以适用大规模图数据的聚类问题,这使得对大规模图整体进行建模和整体分析变得非常棘手和不切实际[1~6]. 在此背景下,图采样技术因其简单高效的特性为大规模图分析提供了切实可行的解决方案,其基本思想是在整体图结构中采集一个具有能够表征原始图潜在属性的采样子图,通过对采样子图进行结构分析,进而用来推断和评估整体图的固有结构[7~9]. 因而,众多的基于不同理论的采样算法被相继提出[10~16]. 这些采样算法大致分为三种主要类型[13]:基于边结构,基于顶点结构和基于拓扑结构的采样. 然而,现有方法针对图的内在聚类结构进行采样的研究尚未深入,现有大多数文献都忽视了图采样对固有的聚类结构变化的影响. Leskovec 等人[14]系统的分析了一系列图采样算法对图的各种基本属性进行了量化评估,实验结果表明随机游走(Random Walk Sampling, RWS)和森林火灾(Forest Fire Sampling,FFS)采样在保留图的基础属性上的表现优于其他采样算法. 然而,这些属性不足以保证采样子图能够有效地代表原始图固有的聚类结构. Maiya 等人[17]基于根据扩展子图的概念

收稿日期:2017-10-09;修回日期:2018-06-10;责任编辑:覃怀银
基金项目:国家自然科学基金群体项目(No. 61521003);国家重点研发计划项目(No. 2016YFB0800101)

提出了聚类结构保持的 XSN 采样算法,其基本思想是具有良好扩展属性的采样点往往更能够有效的代表聚类结构,在采样寻找满足扩展属性条件的邻居顶点并加入到采样图中,然而此方法对稀疏的大规模图的表现相对较差. 因此,本文需要研究设计能够保持聚类结构的采样算法和行之有效的聚类推断算法来解决上述问题.

本文首先对图的采样算法展开研究,提出了能够有效保持原始图聚类结构的图采样算法(CRS),进而利用所得到的采样子图,提出了快速的整体样本聚类推断算法(PCI). 在真实网络数据集上的实验表明,本文算法对大规模图数据具有较高的聚类质量和处理效率,很好的完成了大规模图的聚类任务.

2 基于采样的大规模图数据聚类算法

2.1 相关概念及定义

正式地,我们考虑 $G=(V,E)$,其中 $V=\{v_1,v_2,\cdots,v_N\}$ 表示图的顶点集合;$E=\{e_1,e_2,\cdots,e_M\}$ 表示图的边的集合,其中 $N=|V|$ 表示顶点总数,$M=|E|$ 表示边的总数. 相关定义如下:

定义 1 采样子图 $G_s=(V_s,E_s)$:定义 G_s 为采样子图;$E_s\subseteq$E 为采样边;$V_s\subseteq V$ 为采样顶点;

定义 2 采样率 p:$p=|V_s|/|V|$ 为图的采样率;

定义 3 聚类结构 π:π 表示将顶点集 V 划分为有限非空子集 $\pi=\{C_1,\cdots,C_k\}$ 的集合,其中每个 $C_i(i\in[1,k])$ 是一个聚类簇,使得 $\cup C_i=\pi$. 注意这里的聚类簇之间允许彼此独立或重叠.

定义 4 聚类代表点列表 L:$L=\{l_1,l_2,\cdots,l_k\}$ 是聚类代表点的列表,其中 $l_i(i\in[1,k])$ 表示第 i_{th} 个聚类簇的聚类代表点.

定义 5 邻居集合 $N(S)$:$N(S)$ 表示与顶点集 S 的相连接的邻居集合,即:$N(S)=\{w\in V-V_s|\exists v\in S$ s. t. $(v,w)\in E\}$. 特别地,当 $S=\{u\}$,$N(S)$ 可以写为 $N(u)$,$|N(u)|$ 表示顶点 u 的度数.

定义 6 重要邻居集合 $H(u)$:$H(u)$ 表示具有与顶点 u 相等或更高度数的邻居集合. 即:$H(u)=\{v|v\in N(u)$ s. t. $|N(v)|\geq|N(u)|\}$.

定义 7 新邻居集合 EN(v,S):EN(v,S) 表示顶点 v 一旦被采样到集合 S 中所能提供的新的邻居集合,即:EN$(v,S)=N(v)-N(S)\cup S$.

定义 8 内部顶点集合 IN(S):IN(S) 表示顶点集 S 的邻居顶点中满足使|EN$(v,S)|=0$ 的集合,即:IN$(S)=\{v|v\in N(S)$ s. t. |EN$(v,S)|=0\}$.

定义 9 激活顶点 Unsampled_Node:Unsampled_Node 表示顶点还未被采样到采样子图中.

定义 10 非激活顶点 Sampled_Node:Sampled_Node 表示顶点已经被采样.

根据上述定义,我们首先详细介绍所提的聚类结构保持的采样算法(CRS);其次,根据采样子图,提出整体聚类的推断算法(PCI);最后,给出基于采样的大规模图数据聚类分析算法的整体描述.

2.2 聚类结构保持的采样算法

由于现有研究对聚类结构采样的研究尚未成熟[18,19],我们提出了一种新的聚类结构保持的采样算法(CRS)算法,它能有效产生聚类代表点,并根据相应的扩张准则进行采样扩张. 此算法能够很好地保持原始图的内在聚类结构. 算法主要由两个阶段:初始化聚类代表点和采样扩张阶段.

2.2.1 初始化聚类代表点

聚类代表点的初始化在整个过程中起着很重要的作用,现有的全局启发式方法是简单地选择具有 k 个最高度数的顶点作为各自聚类簇的聚类代表点. 然而,该方法并不适用于度分布呈幂律的真实网络图. 这是因为小聚类簇中的聚类代表点的度数很可能要比大聚类簇的普通顶点度数要小,从而导致小聚类簇中无法产生聚类代表点. 为此我们提出局部聚类代表点选取方法来解决这个问题. 本方法的基本思想是聚类代表点应该满足以下三个准则:

(1)聚类代表点要具有尽可能大的全局度数.

(2)任意两个聚类代表点应该拥有少于指定数量的共同邻居(共同邻居约束 ε).

(3)聚类代表点至多有 $k-1$ 个比其度数高的邻居顶点(k 为聚类数目). 这些邻居顶点有可能是来自其他相对较大的聚类簇的聚类代表点(重要邻居约束).

首先,我们按照顶点度数对顶点集 V 进行降序排列,然后从排序队列中选择拥有最高度数的顶点作为第一个聚类代表点,然后跳到下一个顶点,检查它是否满足约束(即重要邻域和公共邻居约束). 如果满足,我们将顶点添加到聚类代表点列表 V_s 中,直到找到 k 个聚类代表点. 本方法不仅允许两个聚类代表点之间相互连接,而且能够更准确地选择局部密度较小的聚类簇的聚类代表点.

2.2.2 内部顶点优先的采样扩张

在实际网络图中,聚类簇的大小的分布是呈现幂率分布的特点,现有方法无法很好的对其进行有效的采样扩张. 为了克服这个缺点,我们提出了内部顶点优先的采样扩张方法以适应不同大小的聚类簇的扩张需求. 其基本思想是当从各个聚类代表点选择邻居顶点时,通过最小化|EN$(v,C_i)|(i\in[1,k])$ 来添加与聚类簇紧密连接的内部顶点. |EN$(v,C_i)|$ 表示通过加入邻居顶点 v 到聚类簇 C_i 中所能带来的新邻居. 较小的|EN$(v,C_i)|$ 值有助于将潜在的内部顶点加入到聚类簇中,

从而更好的保留聚类簇结构信息. 由于 IN(C_i)中的那些顶点(即满足|EN(v,C_i)| =0 的顶点)是聚类簇 C_i 的内部顶点,这些顶点不为聚类簇提供任何新的邻居.

算法 1　内部顶点优先的采样扩张

输入:原始图 G;聚类簇数目 k;采样子图大小 n;聚类代表点集合 $L=\{l_1,l_2,\cdots,l_k\}$
输出:,采样图聚类标签 label_{V_s};采样子图 G_s
1:$\Omega\leftarrow\varnothing$, Unsampled_Node$\leftarrow V-L$
2:###采样扩张
3:For each $l_i\in L$
4:　$C_i\leftarrow C_i\cup\{l_i\}$
5:While $|V_s|<=n$
6:　For each $C_i\in\Omega$
7:　　If IN(C_i)$\neq\varnothing$
8:　　Internal_nodes$\leftarrow$IN(C_i)
9:　　$C_i\leftarrow C_i\cup$Interanl_nodes
10:　　标记 Interanl_nodes 为 Sampled_Node;
11:　　Else
12:　　$v\leftarrow$Unsampled_nodes s. t. $\min_{v\in N(C_i)}|\text{EN}(v,C_i)|$
13:　　$C_i\leftarrow C_i\cup\{v\}$
14:　　标记 v 为 Sampled_Node
15:###获取聚类标签
16:　$C_i\leftarrow C_i\cup$Interanl_nodes
17:　$\text{label}_{V_s}\leftarrow\{C_1,C_2,\cdots,C_k\}$#将聚类簇转化为聚类标签
18:###满足指定采样数量
19:　If $|V_s|>n$
20:　　迭代删除 V_s 中最小度数的顶点直到满足指定采样数
21:For all $e=\{u,v\}\in E$ s. t. $\{u,v\}\in V_s$
22:　　$E_s\leftarrow E_s\cup\{e\}$
返回:G_s

本方法的算法伪代码在算法 1 中给出. 初始地,我们将聚类代表点 l_i 添加到第 i 个聚类簇 $C_i(i\in[1,k])$中. 随后,本算法包括两个基本步骤.

步骤 1:每次迭代将符合条件的内部顶点添加到每个聚类簇 C_i 中. 具体地,如果聚类簇 C_i 在其邻居中有多个内部顶点 IN(C_i),则将它们都添加到采样顶点中并标记其所属聚类标签. 否则,寻找具有最小度数的邻居顶点并将其加入到采样集中.

步骤 2:由于各个聚类簇每次迭代所加入的内部顶点数目不同,那么步骤 1 结束后,顶点数目可能比指定的采样数目要多. 因此迭代地找出具有最小度值的采样顶点进行删除,直到满足目标顶点数 n 为止.

内部顶点优先扩展策略的主要优点是能够很好的保持聚类代表点及其相应的追随顶点,在有效采样的同时也给出了采样点所属聚类标签,为后续整体样本聚类推断奠定了基础.

2.3　整体聚类推断算法

基于采样子图,本文提出了一种新的整体聚类推理方法(PCI)来标记未被采样的顶点. 它主要由两个步骤构成:(1)初始化非采样顶点的聚类标签和(2)整体聚类标签传播.

具体而言,在粗筛选阶段,首先,为了充分利用采样点已有的标签信息,本算法根据各个未标记顶点中已有聚类标签邻居的比例,按照降序排列对未标记的顶点进行排序. 其次,对于排序队列 U 中的每个未标记顶点,我们将未标记顶点的聚类标签设置为邻居中最大频率出现的聚类标签. 然后,将它从未标记的集合 U 中删除,并将此顶点添加到已标记的顶点集 V_s 中,直到 V 中没有更多的顶点可以被标记为止. 对于未与采样子图连接的孤立顶点,它们被指定到其自身的单顶点聚类簇中.

在微调整阶段,本文利用具有线性复杂度的标签传播 LP(Label Propagation)聚类算法[6]来进一步优化整体的聚类结构. 本文的微调整方法是将粗筛选阶段的聚类簇作为初始聚类簇,聚类标签为初始聚类标签,并设定采样子图的聚类标签在迭代过程保持不变. 我们在此基础上执行标签传播算法,从而进一步优化未标记顶点的聚类归属. 因此,本文所提 PCI 算法计算复杂度接近线性,并对标签传播算法的初始化过程进行改造,能够有效和准确地将聚类标签分配到整体的聚类中.

2.4　聚类算法整体描述

综上所述,基于采样的大规模图数据聚类(Sample-Based Scalable Clustering,SBSC)算法的整体描述在算法 2 给出. 其主要分为两个步骤对大规模图数据进行聚类分析:首先,根据所提的聚类结构保持的 CRS 采样算法对原始图进行有效采样和形成初始聚类标签. 其次,利用本文所提的整体聚类推断 PCI 算法对整体样本进行聚类推断. 整体算法描述如下:

算法 2　基于采样的图聚类算法

输入:原始图 G;聚类簇数目 k;采样子图大小 n;共同邻居约束 ε
输出:采样子图 G_s,原始图聚类划分 π
1:####CRS 采样阶段
2:执行初始化聚类代表点算法,获取初始聚类代表点集合 L
3:执行内部顶点优先的采样扩张算法,获取采样子图 G_s
4:#### PCI 整体聚类推断阶段
5:执行初始化非采样顶点的聚类标签
6:执行标签传播聚类算法,获取整体聚类标签
返回:原始图聚类划分 π

3　大规模网络数据集实验分析

为了验证提出的 SBSC 算法的聚类效果,本文在大规模网络数据集上进行算法对比,对比算法包括:模块

度优化的快速算法[5](Blondel),标签传播算法[6](LP)、Bigclam[20]算法以及 High-order[21] 聚类算法. 本文算法参数设置:采样率:$p=0.2$;共同邻居约束:$\varepsilon=3$,其余对比算法则使用其默认参数.

3.1 实验数据

本文对两类网络图数据进行实验验证:一类是人工合成网络图,本文采用 LFR 生成网络[22];另一类是真实网络大数据集,本文采用斯坦福 SNAP 的网络大数据进行实验分析. 下面分别对其进行简要介绍.

3.1.1 人工合成图

本文首先采用最常见的 LFR 网络图作为人工合成网络来进行实验分析. 具体参数设置如表 1 所示.

表 1 LFR 人工合成图参数

网络	顶点数 N	平均度数 D	最大度数 D_{max}	最小簇 C_{min}	最大簇 C_{max}	混合参数 μ
LFR_1000	1000	20	100	50	200	0.1 ~ 0.6
LFR_10000	10000	40	1000	50	200	0.1 ~ 0.6

表 1 中根据生成网络图的顶点数目 N 不同(分别为 1000、10000),将人工合成网络图分为两类,分别为 LFR_1000 和 LFR_10000. 混合参数 μ 是网络图中所有顶点的外部度数与总度数的比例,μ 越大表示聚类结构越不明显,本文将混合参数 μ 取值在区间[0.1,0.6]中,间隔为 0.1. 因此,每一类网络图应包含有 6 组不同 μ 值的网络图.

3.1.2 真实网络图

真实网络图实验考虑了多个现实世界的 SNAP 大规模网络图. 本文选取每个网络的前 5000 个质量较高的聚类簇进行实验,并进行预处理将聚类簇密度过低、簇大小小于 3 的聚类簇以及重复聚类簇删除,因为此类聚类簇会导致聚类算法的聚类质量显著下降. 经过预处理的网络图的统计信息如表 2 所示.

表 2 真实大规模网络的基本信息

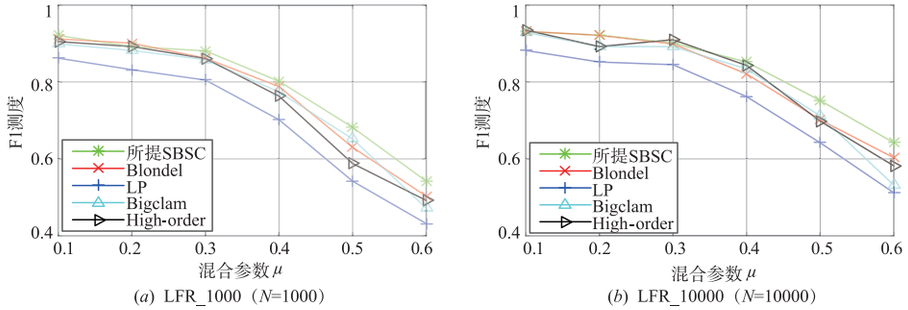
网络	顶点数	边数	聚类簇数	最大聚类簇	最小聚类簇
Amazon	8279	23,023	1138	27	3
DBLP	26,867	89,659	3721	38	6
LiveJournal	43,985	886,545	3528	407	3
Orkut	297,579	7,809,034	3664	787	3
YouTube	12,084	29,664	3579	31	2

3.2 聚类评价指标

本文实验采用了 Precision(查准率),Recall(查全率),NMI 指标来评价聚类算法与真实的聚类标签的差异程度. 同时,实验采用模块度 Q 来进行聚类结果性能的评价. 最后,对各个算法的运行时间进行比较. 相应指标简要描述如下:

3.3 人工网络图上的聚类结果和分析

图 1 表明了不同的聚类算法在图顶点为 1000 和 10000 时,F1 测度值随着混合参数 μ 增大的变化曲线. 5 种聚类算法在 LFR 人工合成数据集上的实验结果表明,当网络图数据的规模越大,各个算法的聚类质量均有所下降. 另外,在相同的图数据的规模下,随着混合参数 μ 值的增加,合成图的聚类结构越来越不明显,各个算法的聚类质量均明显降低. 由图 1 可以看到,本文所提算法在聚类结构较为明显的合成图的聚类质量与 Blondel 算法效果相当,但是在聚类结构相对复杂(μ 值的增加)的合成图中,本文算法的效果优于其它对比算法,进一步验证了本文方法的有效性. 由此可见,所提的 SBSC 算法能够适用于图中顶点数目较大的情况,且当聚类结构更为复杂时,仍能保持较好的聚类效果.



(a) LFR_1000 (N=1000) (b) LFR_10000 (N=10000)

图 1 合成网络图上的实验结果

3.4 真实网络图上的聚类结果和分析

在五个真实的网络数据集上分别运行五种聚类算法,每个算法分别在对应的网络数据集上运行 5 次并计算度量指标的均值和标准差,实验结果如表 3 所示.

表 3　五种算法在大规模真实网络中的实验结果

网络集	算法	Precision	Recall	NMI	模块度 Q	时间(秒)
Amazon	SBSC	0.986 ± 0.008	**0.932 ± 0.012**	**0.827 ± 0.019**	**0.952 ± 0.001**	72.5
	Blondel	0.913 ± 0.021	0.930 ± 0.018	0.794 ± 0.020	0.923 ± 0.001	302.4
	LP	0.780 ± 0.025	0.662 ± 0.029	0.642 ± 0.035	0.791 ± 0.007	**63.2**
	Bigclam	**0.990 ± 0.005**	0.930 ± 0.009	0.826 ± 0.010	0.943 ± 0.005	462.4
	High-order	0.923 ± 0.038	0.860 ± 0.034	0.734 ± 0.026	0.931 ± 0.009	202.8
DBLP	SBSC	0.853 ± 0.023	**0.921 ± 0.021**	**0.803 ± 0.027**	0.532 ± 0.003	173.2
	Blondel	0.832 ± 0.025	0.902 ± 0.024	0.795 ± 0.029	0.502 ± 0.009	673.2
	LP	0.452 ± 0.032	0.687 ± 0.029	0.569 ± 0.027	0.399 ± 0.015	**132.2**
	Bigclam	0.853 ± 0.027	0.914 ± 0.021	0.794 ± 0.027	**0.550 ± 0.011**	982.2
	High-order	**0.855 ± 0.043**	0.920 ± 0.051	0.802 ± 0.032	0.547 ± 0.009	264.2
LiveJournal	SBSC	0.932 ± 0.024	**0.953 ± 0.023**	**0.632 ± 0.025**	**0.901 ± 0.005**	827.6
	Blondel	0.902 ± 0.029	0.923 ± 0.030	0.594 ± 0.034	0.883 ± 0.010	2532.9
	LP	0.872 ± 0.034	0.884 ± 0.032	0.504 ± 0.029	0.717 ± 0.017	**623.4**
	Bigclam	**0.953 ± 0.023**	0.932 ± 0.027	0.624 ± 0.030	0.893 ± 0.014	2232.2
	High-order	0.892 ± 0.027	0.911 ± 0.043	0.554 ± 0.061	0.842 ± 0.040	1332.9
Orkut	SBSC	0.823 ± 0.032	**0.743 ± 0.034**	**0.492 ± 0.029**	**0.694 ± 0.006**	**4223.9**
	Blondel	0.824 ± 0.038	0.702 ± 0.040	0.490 ± 0.032	0.667 ± 0.017	8529.3
	LP	0.809 ± 0.052	0.683 ± 0.061	0.432 ± 0.067	0.346 ± 0.023	4782.4
	Bigclam	0.812 ± 0.029	0.740 ± 0.027	0.453 ± 0.023	0.652 ± 0.009	14529.2
	High-order	**0.826 ± 0.062**	0.733 ± 0.054	0.487 ± 0.039	0.659 ± 0.015	6782.4
YouTube	SBSC	0.382 ± 0.042	0.428 ± 0.039	0.432 ± 0.045	0.536 ± 0.022	**73.3**
	Blondel	0.402 ± 0.045	0.354 ± 0.043	0.357 ± 0.032	0.552 ± 0.018	482.2
	LP	0.357 ± 0.076	0.243 ± 0.083	0.293 ± 0.073	0.336 ± 0.031	89.3
	Bigclam	0.408 ± 0.040	**0.452 ± 0.039**	0.452 ± 0.042	0.531 ± 0.025	520.3
	High-order	**0.410 ± 0.042**	0.450 ± 0.051	**0.453 ± 0.056**	**0.532 ± 0.028**	127.2

在表 3 中可以看出,在五个大规模真实网络数据中,通过对查准率,查全率,NMI 指标和模块度进行分析,本文算法在有效性方面与当前聚类效果最优的 Blondel 和 Bigclam 算法聚类效果相当,甚至更优. 这是因为本文算法的采样算法能够较好的保留原始图的聚类结构,并利用采样子图的聚类结构来推断未被采样的顶点的聚类标签,并经过精细调整标签,使其聚类质量明显提高. 由于 High-order 聚类是基于高阶连接模式进行聚类,这些高阶连接模式包含了出现在图中的所有交互作用,该算法能够识别出富含某个特定高阶模体的密集区域. 但是这种高阶连接模式需要预先指定,并需要根据不同真实图结构进行人工调整从而发现聚类簇. 本文实验中的 High-order 算法中的高阶模体(Motif)采用默认设置的模体 M_{edge}进行实验仿真,实验结果表明此算法对具有不同聚类结构的真实图的聚类效果差异较大. 标签传播(LP)算法的聚类效果受限于初始顶点的迭代顺序,因此 LP 算法具有很大的波动性,聚类结果取决于初始顶点的迭代顺序,因此聚类效果相对较差.

此外,我们对算法运算时间进行了比较分析,本文算法优于 Blondel 和 Bigclam 算法,与线性复杂度的 LP 算法执行效率相当. 因为本文算法的算法复杂度为线性对数时间 $O(N\log N)$. 综上分析,本文算法不仅能够具有较高的聚类质量,而且具有较低的时间复杂度,能够有效应对大规模数据的聚类问题.

4　总结

为了有效应对大规模图的聚类问题,本文首先提出了能够有效保持原始图聚类结构的图采样算法,它能够很好地采样原始图的内在聚类结构. 其次,根据所

得到的采样子图,提出了快速的整体样本聚类推断算法,它能有效利用采样子图的聚类标签对整体图的聚类结构进行推断. 在多种网络数据集上的实验结果表明本文基于采样的聚类算法对大规模图数据具有较高的聚类质量和处理效率,能够很好地完成大规模图的聚类任务.

参考文献

[1] Strogatz S H. Exploring complex networks[J]. Nature, 2001,410(6825):268 – 276.

[2] Boccaletti S,Ivanchenko M,Latora V,et al. Detecting complex network modularity by dynamical clustering[J]. Physical Review E,2007,75(4):45102.

[3] Newman M E,Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E,2004,69(2):26113.

[4] Newman M E. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6):66133.

[5] Blondel V D, Guillaume J, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, no. 10:P10008.

[6] Raghavan U N,Albert R,Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E,2007,76(3):36106.

[7] Jeong H,Tombor B,Albert R,et al. The large-scale organization of metabolic networks[J]. Nature, 2000, 407(6804):651 – 654.

[8] 崔颖安,李雪,王志晓,张德运. 社会化媒体大数据多阶段整群抽样方法[J]. 软件学报,2014,(4):781 – 796.
Cui Ying-An, Li Xue, Wang Zhi-Xiao, Zhang De-Yun. Sampling online social media big data based multi stage cluster method[J]. Journal of Software,2014,(4):781 – 796. (in Chinese)

[9] 张新猛,蒋盛益. 基于核心图增量聚类的复杂网络划分算法[J]. 自动化学报,2013,39(7):1117 – 1125.
Zhang Xin-meng, Jiang Sheng-yi. Complex network community detection based on core graph incremental clustering[J]. Acta Automatica Sinica,2013,39(7):1117 – 1125. (in Chinese)

[10] Hübler C,Kriegel H,Borgwardt K,et al. Metropolis algorithms for representative subgraph sampling[A]. Eighth IEEE International Conference on Data Mining[C],Pisa, Italy:IEEE,2008.

[11] Hastings W K. Monte Carlo sampling methods using Markov chains and their applications[J]. Biometrika, 1970,57(1),97 – 109.

[12] Salehi M,Rabiee H R,Rajabi A. Sampling from complex networks with high community structures[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2012, 22(2):023126.

[13] Ahmed N K, Neville J, Kompella R. Network sampling: from static to streaming graphs[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2014, 8(2):7.

[14] Leskovec J, Faloutsos C. Sampling from large graphs[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Philadelphia,USA:ACM,2006.

[15] Lovász L. Random walks on graphs[J]. Combinatorics, PaulErdos is Eighty,1993,2 (1 – 46):4.

[16] Fortunato S,Hric D. Community detection in networks: a user guide[J]. Physics Reports,2016,659:1 – 44.

[17] Maiya A S,Berger-Wolf T Y. Sampling community structure[A]. Proceedings of the 19th International Conference on World Wide Web[C]. NC,USA:ACM,2010.

[18] Mall R,Langone R,Suykens J A. FURS: fast and unique representative subset selection retaining large-scale community structure[J]. Social Network Analysis and Mining,2013,3(4):1075 – 1095.

[19] Khorasgani R R,Chen J,Zaiane O R. Top leaders community detection approach in information networks[A]. 4th SNA-KDD Workshop on Social Network Mining and Analysis[C]. Washington DC,USA:ACM,2010.

[20] Yang J,Leskovec J. Overlapping community detection at scale:a nonnegative matrix factorization approach[A]. Proceedings of the 6th ACM International Conference on Web Search and Data Mining[C]. Rome, Italy: ACM,2013.

[21] Benson,A. R. ,Gleich,D. F. ,& Leskovec,J. Higher-order organization of complex networks[J]. Science,2016,353(6295),163 – 166.

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

作者简介

张建朋 男,1988 年 3 月出生于河北省廊坊市. 现为国家数字交换系统工程技术研究中心助理研究员,埃因霍温理工大学博士研究生. 主要研究方向为大数据分析.

E-mail:j. zhang. 4@ tue. nl

陈鸿昶　男,1964 年 4 月出生于河南省密县. 国家数字交换系统工程技术研究中心教授. 主要研究方向为电信网信息关防.
E-mail:chenhongchang@ ndsc. com. cn

王　凯　男,1980 年 01 月出生于河南许昌. 国家数字交换系统工程技术研究中心副研究员. 主要研究方向为电信网信息关防.
E-mail:wangkai@ ndsc. com. cn

祝凯捷　男,1991 年 01 月出生于河南省郑州市. 现为埃因霍温理工大学博士研究生. 主要研究方向为图数据库基础.
E-mail:kaijie. ndsc@ gmail. com

王亚文　男,1990 年 8 月出生于河南省郑州市. 现为国家数字交换系统工程技术研究中心博士研究生. 主要研究方向为计算机视觉.
E-mail:15738321455@163. com