软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2753−2766 [doi: 10.13328/j.cnki.jos.004722]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于概率模型的大规模网络结构发现方法*

柴变芳[1,2], 贾彩燕[1], 于 剑[1]

[1](交通数据分析与挖掘北京市重点实验室(北京交通大学),北京 100044)

[2](石家庄经济学院 信息工程学院,河北 石家庄 050031)

通讯作者: 于剑, E-mail: jianyu@bjtu.edu.cn

摘 要: 随着万维网和在线社交网站的发展,规模大、结构复杂、动态性强的大规模网络应用而生.发现这些网络的潜在结构,是分析和理解网络数据的基本途径.概率模型以其灵活的建模和解释能力、坚实的理论框架成为各领域研究网络结构发现任务的有效工具,但该类方法存在计算瓶颈.近几年出现了一些基于概率模型的大规模网络结构发现方法,主要从网络表示、结构假设、参数求解这 3 个方面解决计算问题.按照模型参数求解策略将已有方法归为两类:随机变分推理(stochastic variational inference)方法和在线 EM(online expectation maximazation)方法,详细分析各方法的设计动机、原理和优缺点.定性和定量地对比、分析典型方法的特点和性能,并提出大规模网络结构发现模型的设计原则.最后,概括该领域研究的核心问题,展望未来发展趋势.

关键词: 大规模网络;结构发现;随机变分推理;在线 EM 算法;三角形模体

中图法分类号: TP181

中文引用格式: 柴变芳,贾彩燕,于剑.基于概率模型的大规模网络结构发现方法.软件学报,2014,25(12):2753−2766. http://www.jos.org.cn/1000-9825/4722.htm

英文引用格式: Chai BF, Jia CY, Yu J. Approaches of structure exploratory based on probabilistic models in massive networks. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2753−2766 (in Chinese). http://www.jos.org.cn/1000-9825/4722.htm

Approaches of Structure Exploratory Based on Probabilistic Models in Massive Networks

CHAI Bian-Fang[1,2], JIA Cai-Yan[1], YU Jian[1]

[1](Beijing Key Laboratory of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing 100044, China)

[2](Department of Information Engineering, Shijiazhuang University of Economics, Shijiazhuang 050031, China)

Corresponding author: YU Jian, E-mail: jianyu@bjtu.edu.cn

Abstract: The growth of the Internet and the emergence of online social websites bring up the development of massive networks which are large in scale, complex in structure, and dynamical in time. Exploring latent structure underlying a network is the fundamental solution to understand and analyze the network. Probabilistic models become effective tools in diverse areas of structure exploratory due to their flexibility in modeling, interpretability and the sound theoretical framework, however they incur computational bottlenecks. Recently, several approaches based on probabilistic models have been developed to explore structure in massive networks, which aim to solve the computational problems from three aspects: representations of a network, assumptions of the structure and methods of parameter estimation. This study classifies existing approaches as two categories by the methods of parameter estimation: approaches based on stochastic variational inference and online EM approaches, and analyzes in detail their designing incentives, principles, pros and cons. The properties and performance of classical models are compared and analyzed qualitatively and quantitatively, and as a result the principles are provided to develop approaches of structure detection in massive networks. Finally, the core problems of structure exploratory in massive networks are summarized based on probabilistic models and the development trend of this area is projected.

Key words: massive network; structure detection; stochastic varational inference; online EM algorithm; triangular motif

* 基金项目: 国家自然科学基金(61473030,61370129); 中央高校科研业务经费(2014YJS039); 河北省自然科学基金(F2013205192); 北京市科委项目(Z131110002813118); 北大方正集团有限公司数字出版技术国家重点实验室开放课题;

收稿时间: 2014-04-14; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

随着互联网的发展以及人类互动和沟通需求的扩展,社交网络开始影响人们的生活.社交网络传播速度和广度指数级增长,促使产生大规模网络数据.对这些数据的正确使用,可使分析者发现一些动向,更快地给出创造价值的重要洞察,带来社会化营销黄金时代,许多商家已与社交网站合作进行各种商业活动.如何挖掘庞大的网络数据的潜在模式,利用这些模式为人们提供良好的服务,进而获得更多的收益?网络结构发现方法为解决此问题提供了一些策略,如基于最优化的方法、基于启发式规则的方法、基于相似度的方法、基于概率模型的方法等.面对规模大、结构复杂、随时间变化速度快等特点的网络数据,基于概率模型的方法以其灵活的建模能力、可靠的解释性及坚实的概率图理论框架,成为处理大规模复杂网络数据的最佳选择.

现有的网络结构发现概率模型大多是基于简单随机块模型 SBM(stochastic block model)[1−13].SBM 模型是由社会科学家提出的一种可更好拟合实际网络的随机图模型,不仅能实现网络结构发现任务,还能识别体现网络中观结构的类间链接模式.目前出现了各种 SBM 模型的变型,如混合隶属度随机块模型 MMSB(mixed-membership stochastic block model)[2]、基于 PLSA(probabilistic latent semantic analysis)的限制型随机块模型[3]、通用随机块模型 GSB(general stochastic block)[4]、考虑度信息的随机块模型[5]、限制社区结构的泊松链接社区发现模型[6]、基于贝叶斯参数估计的限制随机块模型参数的 CSBM (constraint stochastic block model)[7]、节点隶属多社区的重叠随机块模型[8]、考虑节点度分布的 PPSB(popularity-productivity stochastic block)模型[9]、多模式社区发现块模型[10].这些模型主要从网络生成过程改进模型,生成更符合实际的网络.但这些模型的算法存在计算瓶颈问题,不能处理大规模网络数据,亟待提出新的方法解决此问题.

近来,一些大规模网络概率模型[14−22]从网络表示、结构假设和参数估计等多方面解决大规模网络结构发现问题.合理的网络表示单元为设计快速的网络结构发现算法奠定了基础,网络表示主要采用边和三角形模体两种表示方式.假设网络结构包括两种:社区结构(同类节点链接紧密、异类节点链接稀疏)和块结构(同类节点具有相同链接模式).模型参数求解方法将大数据挖掘算法迁移到网络数据上,现有两种求解大规模网络结构发现概率模型的主流技术:在线 EM 算法和随机变分推理算法,目前已有一些研究成果利用它们求解 SBM 模型 [19,20] 及 MMSB 扩展模型[14−18,21,22]参数.但已有的研究还不足以解决大规模网络结构发现遇到的问题,有必要总结现有技术和模型设计规律及存在问题,为设计更有效的模型提供参考.

依据模型参数估计方法将现有的大规模网络结构发现概率方法归为两类:

1) 在线 EM 方法,包括:Christophe Ambroise 研究小组成员于 2010 年基于 SBM 模型设计的在线 CEM (classification EM)算法[19]、在线 SAEM(stochastic approximation EM)和在线变分 EM[20].这些在线算法利用新观测数据对 SBM 模型参数进行更新,在算法准确性和速度上取得较好的折中;
2) 随机变分推理方法,包括:Blei 研究小组成员于 2012 年、2013 年基于 MMSB 模型[2]设计的仅考虑社区结构的模型[14−18]:a-MMSB(assortative MMSB)模型[14,15]、AMP(assortative MMSB with node popularities)模型[16]、HDPR(hierarchical dirichlet process relational)模型[18];Xing 研究小组成员于 2012 年、2013 年 NIPS[21,22] 会议提出的两个基于三角形模体表示的大规模社区发现模型:MMTM(mixed-membership triangular model)模型[21]和 PTM(parsimonious triangular model)模型[22].PTM 模型借鉴 MMSB 模型建模经验对三角形模体生成过程建模,利用三角形模体抽样策略、社区结构简化策略及随机变分技术提高算法运行效率.

大规模网络结构发现方法的研究还处于初始阶段,已有研究远远不能满足网络产生的大规模网络潜在模式发现的需求,还需要研究者借鉴已有的技术和模型,为大规模网络结构发现问题设计更好的模型和算法.目前,尚未看到有文献对现存基于概率模型的大规模网络结构发现技术进行总结和分析,本文对这些方法进行总结和分析,为该领域的研究者利用和改进这些技术提供帮助.

本文第 1 节定义大规模网络结构发现概率方法的相关概念及算法.第 2 节介绍基于在线 EM 算法的大规模网络结构发现方法.第 3 节介绍基于随机变分推理的大规模网络结构发现方法.第 4 节对相关技术和典型方法进行定性和定量的比对和分析,提出该类方法的设计原则.第 5 节总结该领域的核心研究课题,展望未来发展趋势.

1 相关概念和算法

大规模网络结构发现概率方法是基于概率模型的方法,下面描述该类模型相关概念及求解算法[23−32].

定义 1(三角形模体). 网络潜在结构发现概率模型常以边为建模对象,研究表明:将网络表示为更大的网络模体单元更利于社区发现任务[21,22],其是网络中频繁出现的 k-子图,三角形模体是由 3 个顶点形成的子图,有 4 类三角形模体:(a) 完全三角形;(b) 2 边三角形,该类三角形以 3 个顶点为中心;(c) 1 边三角形;(d) 空三角形.完全三角形和 2 边三角形对局部社区结构贡献大,当节点的最大度为 D 时,这两类三角形的数量被证明不超过 ND^2;当节点度很大时,对三角模体进行 δ采样,保证每个节点只采样 δ个链接,生成三角形模体的数量被证明不超过 $N\delta^2$[21].

定义 2(潜在结构发现概率图模型). 潜在结构发现概率图模型可用如图 1 所示的贝叶斯网络抽象表示,网络结构发现概率模型属于该类模型,其对网络所有构成单元建模,如边、三角形模体.建模变量包括:所有单元的局部变量 $z=\{z_n\}_{i=1}^N$、全局变量β、观测单元变量 $x=\{x_n\}_{i=1}^N$、超参数α.

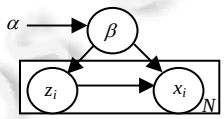

Fig.1 A graphical model

图 1 概率图模型

根据概率图模型将所有变量的联合分布分解为如下形式:

$$p(x,z,\beta\mid\alpha)=p(\beta\mid\alpha)\prod_{i=1}^{N}p(x_i,z_i\mid\beta) \tag{1}$$

定义 3(指数簇分布). 概率图模型的变量分布通常假设为指数簇分布,假设图 1 全局变量β给定下局部完全数据分布、β先验分布、β后验分布服从指数簇分布,定义如下:

$$p(x_i,z_i|\beta)=h(x_i,z_i)\exp\{\beta^T t(x_i,z_i)-a_l(\beta)\} \tag{2}$$

$$p(\beta)=h(\beta)\exp\{\alpha^T t(\beta)-a_g(\alpha)\} \tag{3}$$

$$p(\beta|x,z,\alpha)=h(\beta)\exp\{\eta_g(x,z,\alpha)^T t(\beta)-a_g(\eta_g(x,z,\alpha))\} \tag{4}$$

$h(\cdot)$和 $a(\cdot)$是基度量和对数正则项标量函数,$\eta(\cdot)$和 $t(\cdot)$是自然参数和充分统计量.

后验分布的计算公式如下:

$$p(\beta\mid x,z,\alpha)\propto p(\beta,x,z\mid\alpha)=p(\beta\mid\alpha)\prod_{n=1}^{N}p(x_i,z_i\mid\beta,\alpha) \tag{5}$$

则根据公式(2)~公式(5)及变量分布的共轭关系,可得如下等式:

$$\eta_g(x,z,\alpha)=\left(\alpha_1+\sum_{i=1}^{N}t(x_i,z_i),\alpha_2+N\right) \tag{6}$$

$$t(\beta)=(\beta,-a_l(\beta)),\alpha=(\alpha_1,\alpha_2) \tag{7}$$

定义 4(网络结构发现概率模型). 网络结构发现概率模型不对网络结构做任何假设,认为具有相同链接模式的节点属于同类,相比发现紧密子图的社区发现概率模型,可对更多类型的网络结构建模,如社区结构、星型结构、多分结构、层次结构等.SBM 模型及其变型[2,4,5,9,10]属于网络结构发现模型.

网络结构发现概率方法建模及求解过程:首先,根据网络表示设定网络生成过程假设、变量集合、变量分布;然后,根据网络生成过程将变量可视化为概率图模型;最后,根据图模型得到模型观测变量、局部变量和全局变量的联合分布,进而得到局部变量、全局变量后验分布 $p(z,\beta|x)$.

当考虑全局变量β的先验分布 $p(\beta|\alpha)$时,常用贝叶斯估计算法(如变分贝叶斯推理)估计模型变量的分布;否则,用最大似然估计算法(如 EM 算法)估计模型变量的某个局部最优值.

EM 算法易实现、鲁棒性强,是最大似然参数估计的常用算法,已有研究者设计了针对独立同分布数据的在线 EM 算法[27−32].EM 算法通过最大化全局参数的期望估计局部和全局变量最优解,多次迭代估计局部变量和

全局变量实现变量的更新,*E* 步根据上次迭代的参数计算所有观测对象的局部变量,*M* 步根据局部变量更新全局变量.在线 EM 算法是 CappéO 提出的在线 EM 算法版本,相比以前在线算法更简单,其在随机 *E* 步估计新观测对象的局部变量,*M* 步基于新观测的局部变量和旧全局变量更新当前全局变量[31,32].

变分贝叶斯推理用分布簇 $q(z,\beta)$近似 $p(z,\beta|x)$,最小化两个分布 KL 散度估计最优的 $q^*(z,\beta)$[23,24]:

$$\begin{aligned} KL(q(z,\beta)\|p(z,\beta|x)) &= E_q[\log q(z,\beta)] - E_q[\log p(z,\beta|x)] \\ &= E_q[\log q(z,\beta)] - E_q[\log p(x,z,\beta)] + \log p(x) \\ &= -L(q) + const \end{aligned} \tag{8}$$

最小化 KL 散度相当于最大化证据的下界 $L(q)$,$L(q)$称观测对数似然下界 ELBO(evidence lower bound).最简单的变分分布簇 $q(z,\beta)$是平均场分布簇:

$$q(z,\beta) = q(\beta|\lambda)\prod_{i=1}^{N}\prod_{k=1}^{K} q(z_{ik}|\phi_{ik}) \tag{9}$$

固定全局变分参数,将 ELBO 重写为局部变分参数$\{\phi_{ik}\}$的函数,对目标函数求导得ϕ_{ik}迭代函数.固定局部变分参数,将目标函数 ELBO 重写为全局变分参数λ的函数,对目标函数求导得λ更新式子.迭代更新局部变分参数$\{\phi_{ik}\}$和全局变分参数λ,直到 ELBO 收敛.

变分贝叶斯推理每次迭代需利用上次迭代获得的变分参数更新所有局部变分参数.局部变分参数估计从一个随机全局变分参数值开始,估计所有数据点的局部变分参数,而初值不能反映实际参数变量分布,降低了估计准确率;且每次迭代都需计算所有数据的局部变分参数,很耗时[17].

针对变分贝叶斯推理的缺点,2013 年,Hoffman、Blei 等人提出了随机变分推理[17],使概率图模型的变分推理可应用到大规模数据上.该方法每次迭代首先抽样部分数据估计其局部变分参数,然后利用这些局部变分参数估计中间全局变分参数,基于中间全局变分参数和上次迭代全局变分参数加权平均得到当前的全局变分参数.该方法相当于随机梯度算法,采用自然梯度代替传统梯度,研究证明,基于自然梯度的最大似然估计收敛速度高于传统梯度算法[25].随机变分推理每次随机抽样一个或一批数据点,为描述简单,以一个数据点说明随机变分推理过程.通过抽样数据的 ELBO $L_l(\lambda)$近似整个数据集合上的 ELBO $L(\lambda)$:

$$L(\lambda) = E_q[\log p(\beta)] - E_q[\log q(\beta)] + \sum_{i=1}^{N}\max_{\phi_i}(E_q[\log p(x_i,z_i|\beta)] - E_q[\log q(z_i)]) \tag{10}$$

$$L_l(\lambda) = E_q[\log p(\beta)] - E_q[\log q(\beta)] + N\max_{\phi_l}(E_q[\log p(x_l,z_l|\beta)] - E_q[\log q(z_l)]) \tag{11}$$

$L_l(\lambda)$的期望等于 $L(\lambda)$,因此,关于全局变分参数的函数 $L_l(\lambda)$的自然梯度是 $L(\lambda)$的噪音的无偏估计.

- 抽样数据的局部变分参数更新:计算抽样数据的局部变分参数ϕ_{ik}.
- 中间全局变分参数更新:根据抽样数据计算全局变分参数噪音梯度 $\hat{\nabla}_\lambda L_i = \alpha + N(E_{\phi_i(\lambda)}[t(x_i,z_i)],1) - \lambda$,则第 *m* 次迭代的中间全局变分参数 $\hat{\lambda}_m$ 计算如下:

$$\hat{\lambda}_m = \alpha + N(E_{\phi_i(\lambda)}[t(x_i,z_i)],1) \tag{12}$$

- 全局变分参数更新:基于随机近似(robbins-monro)方法[24]得到第 *m* 次迭代全局变分参数的估计:

$$\lambda^{(m)} = \lambda^{(m-1)} + \rho_m(\hat{\lambda}_m - \lambda^{(m-1)}) = (1-\rho_m)\lambda^{(m-1)} + \rho_m\hat{\lambda}_m \tag{13}$$

其中,ρ_m是全局变分参数的更新步长,$\rho_m=(m+\tau)^{-\kappa}$,$\kappa\in(0.5,1]$.

执行上述 3 步迭代,直到 ELBO 收敛.

2 基于在线 EM 算法的大规模网络结构发现方法

网络数据的节点间存在关系,利用在线 EM 算法求解大规模关系数据的结构发现问题时,最大化完全数据对数似然的期望涉及局部变量在观测网络下分布 $p(z|X)$的计算,而该值不可分解,需要采取近似策略求解.Zanghi 等人[19,20]基于 MixNet 模型设计了 $p(z|X)$的几种计算方法.MixNet[12]是 SBM 模型一种特例,下面也将 MixNet 模型称作 SBM 模型.

SBM 模型假设网络的每个节点隶属一个社区,网络中的链接生成过程分两步:

1) 将网络节点 *i* 以概率α_r指派到第 *r* 个社区;

2) 每对节点(i,j)产生链接 X_{ij} 的概率服从 $Bernoulli(B)$分布,其中,B 表示类间链接概率矩阵,B_{ql} 表示类 q 和类 l 间的链接概率.

在线 CEM 算法、在线 SAEM 算法和在线变分 EM 算法在估计模型参数过程中,对 $p(z|X)$进行了 3 种不同的假设,下面分析这 3 种典型在线 EM 算法的设计动机、思想、原理及优缺点.

2.1 在线CEM算法

相比 Erdös-Rényi 随机图模型,SBM 可以更好地拟合实际网络,但其参数估计算法限制其只能处理 200 个节点的网络[11].Daudin 等人提出变分方法估计 SBM 模型参数,可处理成千上万个节点的网络[12].为处理更大规模的网络,Zanghi 等人[19]假设每个节点类标 z 为 0/1,则 $\log p(z|X)$的计算简化为每个节点类概率的乘积,即 $\log p(z\mid X)=\sum_{i,q} z_{iq}\log\alpha_q$,进而提出一个增量式 CEM 大规模网络结构发现算法.在线 CEM 算法 E 步为新增节点指派类标,M 步根据新指派和上次迭代参数更新当前参数,迭代 E 步和 M 步,直到收敛.

SBM 模型 CEM 算法的完全数据对数似然函数如下:

$$L(X,z;\Phi)=\log p(X,z;B,\alpha)==\sum_{i,q} z_{iq}\log\alpha_q+\sum_{i>j,q,l} z_{iq}z_{jl}\log(B_{ql}^{X_{ij}}(1-B_{ql})^{1-X_{ij}}) \tag{14}$$

令 $m-1$ 个节点的网络邻接矩阵为 X^{m-1},类指派矩阵为 $z^{m-1}(\Phi)$(Φ为参数集合),一个新节点 x_m 加入时,m 个节点的网络的完全数据的对数似然表示如下:

$$L_C^m(X^m,z^m(\Phi^{m-1});\Phi)=L_C^{m-1}(X^{m-1},z^{m-1};\Phi)+\sum_q z_{mq}\left\{\log\alpha_q+\sum_{j\neq m,l} z_{jl}\log(B_{ql}^{X_{mj}}(1-B_{ql})^{1-X_{mj}})\right\} \tag{15}$$

由根据公式(15)可知:新增节点 x_m 的类指派只与新节点链接及参数有关,最大化公式(15)第 2 项得新节点类标 q^*;在 M 步基于新节点类标和第 $m-1$ 次的参数更新当前参数.当所有节点都加入后,如果硬件允许还可搜索能增加对数似然的节点,为其指派新的类标,再更新参数,直到迭代次数达到阈值或完全数据似然不再增加.该增量式的在线 CEM 算法可证明是参数的随机梯度算法[28].

在线 CEM 算法在 E 步的计算复杂度为 $O(NK)$,M 步复杂度为 $O(NK^2)$,该算法通常在整个节点上迭代 cN 次,其中,c 为常数,则整个复杂度为 $O(cN^2K^2)$.虽然整个复杂度与 CEM 算法一样,但是 CEM 算法每次迭代需要在所有节点上计算参数,在线 CEM 算法只需要计算一个节点的相关参数,相当于遍历整个数据少量次.

在线 CEM 在每次迭代过程中为节点指派类标,简化了 $p(z|X)$的计算,但该方法估计参数有偏.2010 年,Zanghi 等人[20]基于随机近似 EM 算法设计了在线 SAEM(stochastic approximation EM),基于变分方法设计了在线变分 EM 算法.

2.2 在线SAEM算法

Zanghi 等人针对 SBM 模型设计了大规模网络结构发现在线 SAEM 算法.SAEM 算法采用随机近似理论估计完全数据对数似然,在线 SAEM 算法将 SAEM 算法扩展为在线版本,用新加入节点信息更新当前参数. SAEM 算法属于 EM 框架,其在 E 步利用随机近似理论近似完全数据对数似然,在 M 步最大化近似的完全数据对数似然求解模型参数.SAEM 算法 E 步求解的完全数据对数似然 $\hat{L}_{m+1}(\beta\mid\beta^m)$ 不存在闭解,其利用第 m 次完全数据似然的近似值 $\tilde{L}_m(\beta\mid\beta^{m-1})$ 与 $m-1$ 次迭代的完全数据对数似然的差值作为梯度来近似 $\hat{L}_{m+1}(\beta\mid\beta^m)$:

$$\hat{L}_{m+1}(\beta\mid\beta^m)=\hat{L}_m(\beta\mid\beta^{m-1})+\rho_k(\tilde{L}_m(\beta\mid\beta^{m-1})-\hat{L}_m(\beta\mid\beta^{m-1})) \tag{16}$$

其中,m 是迭代标号;ρ_k 是迭代步长;$\tilde{L}_m(\beta\mid\beta^{m-1})$ 通过蒙特卡洛模拟得到,为 t 个完全数据对数似然的平均值;当前完全数据对数似然需要根据 $p(z|X)$的吉布斯抽样估计缺失类标.

在线 SAEM 算法与 SAEM 算法存在两处不同:根据 $p(z|X)$进行缺失数据采样不同,SAEM 算法需用吉布斯采样估计所有网络节点的类标,在线 SAEM 算法只估计新加入节点的类标;计算 $\hat{L}_{m+1}(\beta\mid\beta^m)$ 中连续两次完全数据对数似然差值时,SAEM 算法涉及所有节点类标,在线 SAEM 算法仅与新加入节点的类标相关.在线 SAEM 算法的第 m 次的目标函数可写为第 $m-1$ 次目标函数与新加入节点类标的函数,通过最大化目标函数序列化地更

新模型参数β.每次迭代通过如下 3 步实现:

1) 根据 $m-1$ 个节点的指派 z^{m-1} 和 m 个节点的网络 X^m,基于完全数据似然充分统计量的可加性计算第 m 个新节点 x_m 的指派为 q 的概率 $p(z_{m,q}=1|X^m,z^{m-1})$,根据该值模拟抽样 t 次 x_m 的类指派;
2) 按照公式(16),根据 t 个采样均值估计 $\hat{L}_{m+1}(\beta|\beta^m)$,其只与上次迭代的参数和新节点指派和链接信息有关;
3) 最大化 $\hat{L}_{m+1}(\beta|\beta^m)$ 更新充分统计量,进而更新参数β.

在线 SAEM 算法用两次迭代的似然差值作为目标函数的梯度,用两次似然值差值代替随机梯度算法梯度,该算法可证明是模型参数β的随机梯度算法.该算法每次迭代的计算复杂度为 $O(NK^2)$,整个复杂度为 $O(N^2K^2)$.与在线 CEM 算法的复杂度类似,但实验证明,准确率有所提高.

2.3 在线变分EM算法

在线 SAEM 算法通过 collapsed 吉布斯抽样计算新节点的类指派,进而可计算 $p(z|X)$和完全数据对数似然的期望.完全数据对数似然的期望是多个抽样上的均值,对于不同数据,抽样多少还需要进一步设置参数,计算相对复杂.在线变分 EM 算法相对简单,且 Daudin 等人设计的 SBM 模型是基于变分 EM 算法,实现该算法的在线版本很有意义.变分 EM 算法利用平均场变分原则,假设 $p(z|X)$的近似变分分布 R 满足 $\log R(z)=\sum_i\sum_q z_{iq}\log\tau_{iq}$,$\tau_{iq}$ 为节点 i 为 q 类的概率.变分算法最小化变分分布 R 与真实分布 $p(z|X)$的 KL 散度,相当于最大化完全数据似然期望与变分分布 R 的熵的和.完全数据似然和熵具有可加性,利用该可加性实现在线变分 EM 算法:加入第 m 节点后的完全数据似然目标函数为 $L(X^m,R(z^m)|\beta^m)+\sum_{i=1}^m\lambda_i\left(\sum_{q=1}^K\tau_{iq}-1\right)$,固定 $i<m$ 的$\{\tau_{i,q}\}$,求$\{\tau_{m,q}\}_q$;固定$\{\tau_{m,q}\}_q$,求解模型参数,参数迭代等式可写为第 $m-1$ 次迭代的充分统计量和$\{\tau_{m,q}\}_q$ 的函数.迭代计算所有节点的$\{\tau_{m,q}\}_q$,并更新模型参数.

在线变分 EM 算法每次迭代更新$\{\tau_{m,q}\}_q$ 和参数的复杂度与在线 CEM、在线 SAEM 相同,总复杂度也为 $O(N^2K^2)$.3 种在线算法的迭代次数可与节点数一致,也可为其整数倍.虽然复杂度与源 EM 算法一样,但在线版本的算法遍历节点数少,即,N^2K^2 的实际次数要比非在线算法少,因此耗时较少.实验结果表明:3 类在线算法的运行时间近似,在线变分 EM 算法估计参数的效果最好.

3 基于随机变分推理的大规模网络结构发现方法

MMSB 模型[2]是 2008 年 Airodi 等人提出的处理生物网络数据的一个模型,该模型结合了随机块模型和混合隶属度模型的优点:既可对各种类型的网络结构灵活建模,又可对节点隶属多个社区的特性建模.MMSB 模型的参数推理采用传统的变分推理方法,算法复杂度是 $O(N^2K^2)$(N 表示节点个数,K 表示类个数),不能处理类个数多、规模大的网络.2012 年、2013 年,Blei 小组的研究成员基于 MMSB 模型提出 3 个典型模型:a-MMSB[14,15]、AMP 模型[16]、HDPR(hierarchical dirichlet process relational)模型[18],假设网络为类内链接紧密的 assorative 社区结构,并采用随机变分推理技术求解模型参数,使模型求解算法可处理大规模网络社区结构发现问题.Xing 及其小组成员 Ho,Yin 提出的 MMTM 模型和 PTM 模型用三角形模体表示网络,MMTM 模型采用吉布斯采样估计网络结构,PTM 模型基于随机变分推理算法和并行技术拟合与节点数成比例的抽样三角形模型[21,22].下面详细描述这些模型的设计动机、原理、优缺点及随机变分推理的应用方式.

3.1 a-MMSB模型

MMSB 模型假设任意两个类间都存在链接,类间链接概率参数维数为 K^2,变分贝叶斯参数估计算法复杂度为 $O(N^2K^2)$,不能处理社区个数多的大规模网络.a-MMSB 模型[14]假设网络中存在社区结构,社区内节点链接概率较大,社区间节点链接概率较小,仅对节点的同配性(assortativity)建模,社区链接概率参数为 K 维,采用随机变分贝叶斯推理算法估计模型参数,使该算法可处理大规模网络.a-MMSB 模型假设网络生成过程如下:

1) 任意社区 k 链接强度$\beta_k\sim Beta(\eta)$;

2) 节点 $i\in\mathcal{N}$的隶属度向量$\theta_i\sim Dir(\alpha)$;

3) 按照如下步骤确定每对节点$\langle i,j\rangle$是否产生链接:

3.1) 为节点 i 指派社区 $z_{i\to j}\sim Mult(\theta_i)$;

3.2) 为节点 j 指派社区 $z_{i\leftarrow j}\sim Mult(\theta_j)$;

3.3) 根据 $Bernoulli(\lambda)$分布确定 i 和 j 是否产生链接,如果 i 和 j 属于一个社区 k,$\lambda=\beta_k$;否则,$\lambda=\varepsilon$.

通过估计参数后验分布 $p(z,\beta,\theta|Y,\alpha,\eta)$得到网络结构,用 mean-field 变分簇 $q(z,\beta,\theta)$近似后验分布:

$$q(z,\beta,\theta)=\prod_k q(\beta_k\mid\lambda_k)\prod_i q(\theta_i\mid\gamma_i)\prod_{i,j} q(z_{i\to j}\mid\phi_i)q(z_{i\leftarrow j}\mid\phi_j) \tag{17}$$

其中,$q(\beta_k)=Beta(\beta_k;\lambda_k)$,$q(z_{i\to j}=k|\phi_{i\to j})=\phi_{i\to j,k}$,$q(z_{i\leftarrow j}=k|\phi_{i\leftarrow j})=\phi_{i\leftarrow j,k}$,$q(\theta_i)=Dir(\theta_i;\gamma_i)$.

通过最大化观测网络对数似然 $\log p(X)$下界 ELBO 求解最优的变分分布,ELBO 中与变量β,θ变分参数λ和γ相关项为全局项(global terms),与 z 变分参数$\phi_{i\to j}$和$\phi_{i\leftarrow j}$相关的项为局部项(local terms).随机变分推理不需要估计所有局部变分参数,仅估计 N^2 个节点对上的抽样集合的相关局部变分参数.根据这些局部变分参数估计全局变分参数更新的噪音自然梯度,进而基于随机梯度算法更新全局变分参数.随机变分分参数估计算法如下:

1) 初始化全局变分参数γ和λ;

2) 从节点对集合中抽样边集合 S;

3) 局部变分参数估计:计算 S 中每对节点(i,j)的最优局部变分参数$\phi_{i\to j}$和$\phi_{i\leftarrow j}$;

4) 全局变分参数估计:根据局部变分参数更新γ和λ;

重复步骤 2)~步骤 4),直到 ELBO 收敛或达到最大迭代步数.

第 2)步可采用不同的抽样方法,但是必须保证抽样的子集服从某种特殊分布,保证从抽样估计的噪音梯度必须是真正梯度的无偏估计.抽样的策略有很多,如随机节点对抽样、随机节点抽样、链接抽样.其中,链接抽样是从网络的链接集合抽样,实验证明:该抽样方法能以最高的准确率恢复真实的社区结构,可处理百万级节点的网络.如果将抽样子集看作每次新产生的观测,随机最优化过程可解释为在线随机最优化算法.将在线产生节点看作随机抽样子集,在线方法就是随机梯度算法.因此,两种算法都可看作随机梯度算法.

a-MMSB 模型建模阶段简化网络的链接模式,在参数学习阶段采用随机变分推理技术,每次迭代计算局部变分参数复杂度为 $O(SK)$,S 是每次迭代抽样的节点对数量;估计全局变分参数复杂度是 $O(NK)$.该模型参数的随机变分推理算法总的计算复杂度相当于 $O(cMK)$(c 为常数).因此,a-MMSB 模型的随机变分推理算法可用来处理大规模网络社区发现问题,但是其只能发现链接紧密的社区结构;MMSB 模型复杂度较高,但可估计的网络结构类型更多.a-MMSB 模型还可计算节点的桥接度(bridgeness),即节点连接多个社区的程度.

3.2 AMP模型

网络社区结构发现模型通常对两个节点属性建模:一个是节点的流行性(popularity),基于该属性生成的网络符合实际网络的属性,如度服从幂率分布;另一个是节点的同质性(homophily)或相似性(similarity),基于该属性生成的网络具有社区结构.为更好地解释网络数据,网络概率模型需要对节点的流行性和同质性建模.2012 年,文献[33]理论研究表明,最优化流行性和同质性可更好地解释许多真实网络的演变.相比单社区模型,MMSB 模型可更好地拟合实际网络,但不能解释节点流行性.a-MMSB 模型在 MMSB 模型基础上提出,可以更好地解释节点的同质性.AMP 模型[15]扩展 a-MMSB 模型考虑节点的流行性,其假设网络生成过程如下:

1) 从 $N(\mu_0,\sigma_0^2)$ 分布为每个社区 k 的链接程度β_k抽样;

2) 为每个节点 $i\in N$;

3.1) 从 $Dir(\alpha)$抽样隶属度向量θ_i;

3.2) 从 $N(0,\sigma_1^2)$ 抽样节点的流行性 l_i;

3) 按照如下步骤为每对节点(i,j)确定是否产生链接:

3.1) 为边对(i,j)的发射链接节点 i 指派社区 $z_{i\to j}\sim Mult(\theta_i)$;

3.2) 为边对(i,j)的接收链接节点 j 指派社区 $z_{i\leftarrow j}\sim Mult(\theta_j)$;

3.3) 根据 $\log it^{-1}(z_{i\to j},z_{i\leftarrow j},l,\beta)$分布确定 i 和 j 是否产生链接.

通过求解 AMP 模型的后验分布 $p(z,\beta,\theta,l\mid X,\alpha,\mu_0,\sigma_0^2,\sigma_1^2)$ 来估计社区的隐含结构,利用类似 a-MMSB 模型的随机变分推理参数估计算法学习模型局部变分参数和全局变分参数,该算法复杂度与 a-MMSB 参数估计算法一样为 $O(cMK)$(c 为常数),可用来执行社区发现和链接预测任务,也可计算节点的流行度.实验结果表明,基于 AMP 模型的链接预测结果优于基于 MMSB 模型的结果.

3.3 aHDPR模型

SBM,MMSB,a-MMSB 和 AMP 模型要求给定社区个数,但许多情形下社区个数未知,且随着网络的增长而增大.针对此问题,Kim 等人提出关系数据的贝叶斯非参数关系模型 aHDPR[18],允许社区个数未知、每个节点隶属多个社区,该模型基于层次狄利克雷过程 HDP,允许社区个数随着观测节点的增加而变化.已有的 HDP 模型都采用马尔科夫蒙特卡洛采样学习参数,aHDPR 模型也属于 HDP 模型,是第 1 个利用随机变分推理算法估计模型参数的 HDP 模型、基于随机抽样增量式的更新全局变分参数.

aHDPR 模型针对无向同配结构网络建模,定义一个全局 DP(dirichlet process)描述每个社区的相关参数.令 β_k 表示每个社区的期望大小,其可通过如下 stick-breaking 过程构造:

$$\beta_k=\nu_k\prod_{l=1}^{k-1}(1-\nu_l),\nu_k\sim Beta(1,\gamma) \tag{18}$$

节点混合社区隶属度 π_i 从基度量 β 抽样,$\pi_i\sim DP(\alpha,\beta)$,其中,$\alpha$ 是社区强度参数,边 x_{ij} 的产生过程为:

1) 根据节点 i 的混合社区隶属度 π_i 为其指派社区 $z_{i\to j}\sim categorical(\pi_i)$;
2) 根据节点 j 的混合社区隶属度 π_j 为其指派社区 $z_{i\leftarrow j}\sim categorical(\pi_j)$;
3) 如果 $z_{i\to j}$ 和 $z_{i\leftarrow j}$ 为相同社区 k,(i,j)以概率 ω_k($\omega_k\sim Beta(a,b)$)产生链接;否则,以小概率 ε 产生链接.

为求解 aHDPR 模型的参数,需要计算观测变量、隐含变量参数的联合分布,类似 a-MMSB 和 AMP 模型的参数求解过程,利用随机变分推理求解.需要指出的是:局部变分参数和全局变分参数的维度是无限的,需要定义一个有限的维度 K,指定一个初始较大的 K,然后在算法迭代过程中裁剪节点数较少的社区.该模型的准确率优于 a-MMSB 模型,并为非参数贝叶斯方法的推理算法提供了一种新的求解方案.但是,该模型生成链接的假设没有考虑节点的流行性及其他特性,还有待完善模型和求解过程.

3.4 MMTM模型和PTM模型

a-MMSB,AMP,aHDPR 模型都假设网络表示单元为边,在网络结构发现任务中并不需要将边对作为输入,且会引起时间和空间的浪费.Ho 等人提出的 MMTM[21]模型假设网络由三角形模体集合构成,根据定义 1 可知抽样三角形模体个数为 $O(N)$级,为设计快速算法奠定基础.MMTM 模型假设三角模型生成过程如下:

1) 抽样 3 个社区 $x,y,z(x<y<z)$的概率 $B_{xyz}\sim Dir(\lambda)$;
2) 抽样每个节点 $i\in N$ 隶属度向量 $\theta_i\sim Dir(\alpha)$;
3) 对每个三元组(i,j,k):
   3.1) 为三元组(i,j,k)的每个节点指派社区,$s_{i,jk}\sim Discrete(\theta_i),s_{j,ik}\sim Discrete(\theta_j),s_{k,ij}\sim Discrete(\theta_k)$;
   3.2) 根据节点指派和 B_{xyz} 确定(i,j,k)产生什么类型的三角形模体.

MMTM 模型通过吉布斯采样对三角模体的节点进行社区采样,利用最大似然估计模型参数,复杂度为 $O(N\delta^2K^3)$.该类模型虽然可以通过三角形模体抽样减少计算的对象,但每次迭代计算三角形模体的相关局部变分参数比计算边的复杂度大.主要原因是:根据三角形模体相关节点社区指派生成不同类型三角形模体涉及的组合情形太复杂,当社区个数较多时,算法效率很低;且通过吉布斯采样得到稳定的马尔科夫链耗时较长.Yin 等人提出 PTM 模型[22],通过 3 种策略改进 MMTM 模型:简化三角形模体 3 个节点的社区交互模式;抽样部分三角形模体减少算法计算对象;利用随机变分推理和并行技术学习模型参数.MMTM 模型中的社区链接模式参数 B 包含 K^3 个参数.PTM 将 B 分组,该参数的数量减少为 $O(K)$个.参数 B 分为 3 类:

1) 若三角形 3 个节点指派都为 x,则三角形可划分为 2 类,生成概率由 B_{xxx} 决定($\forall x,B_{xxx}\in\Delta^1$);
2) 若三角形 2 个节点指派为 x,则三角形可划分为 3 类,生成概率由 B_{xx} 决定($\forall x,B_{xx}\in\Delta^2$);

3) 如果三角形的 3 个节点指派都不同,则三角形可划分为 2 类,生成概率由 B_0 决定($B_0 \in \Delta^1$).

PTM 模型的三角形模体生成过程与 MMTM 模型类似,仅在步骤 3.2)生成三角形模体存在差异.

链接模式参数的简化为设计有效的推理算法提供了条件;基于三角形模体表示抽样,通过减少算法计算的对象减少算法运行时间;估计算法采用随机变分推理和并行技术求解隐含变量,通过改进算法收敛速度提高其运行效率;3 种策略共同降低了参数估计算法的复杂度,该模型参数估计复杂度为 $O(N\delta^2K)$,是目前为止概率模型中参数学习算法较快的方法.该模型的主要目的是学习节点的隶属度,利用隶属度获得网络的重叠社区划分结果,但没有对节点的流行性和同质性建模.

4 大规模网络结构发现方法比较与分析

为了定性定量地分析与比较大规模网络结构发现的不同参数估计方法和模型的性能,我们首先对在线 EM 算法、随机变分推理技术及典型方法进行定性比较,表 1 比较两种技术,表 2 比较典型方法.然后选择具有代表性的算法,在大规模人工网络数据和不同规模实际网络数据上进行准确率和运行效率两方面的对比实验.

Table 1 Contrasts of parameter estimation methods for structure detection based on probabilistic models on massive networks

表 1 大规模网络结构发现概率模型参数估计方法比较

	随机变分推理	在线 EM 算法
解决问题	贝叶斯参数估计算法变分推理的运行效率低	最大似然估计算法 EM 算法运行效率低
迭代原理	抽样整个数据子集估计模型参数噪音自然梯度, 基于随机梯度算法估计模型参数	估计新数据缺失数据, 基于该值和上次迭代参数更新模型参数
算法类型	全局变分参数的随机梯度算法	模型参数的随机梯度算法
关键	根据抽样数据子集计算噪音梯度	构造参数增量更新公式

Table 2 Contrasts of classical probabilitic approaches for structure detection on massive networks

表 2 大规模网络结构发现典型概率方法比较

概率方法	文献	建模对象	结构	节点属性	参数估计策略	复杂度
a-MMSB	[14,15]	边	社区	同质性	随机变分推理	$O(MK)$
AMP	[16]	边	社区	同质性和流行性	随机变分推理	$O(MK)$
aHDPR	[18]	边	社区	同质性	随机变分推理	$O(MK)$
PTM	[22]	三角形模体	社区	同质性	随机变分推理	$O(N\sigma^2K)$
在线 CEM	[19]	边	块结构	异质性和同质性	在线 CEM	$O(N^2K)$
在线 SAEM	[20]	边	块结构	异质性和同质性	在线 SAEM	$O(N^2K^2)$
在线变分 EM	[20]	边	块结构	异质性和同质性	在线变分 EM	$O(N^2K^2)$

4.1 定性比较和分析

表 1 将基于概率模型的大规模网络结构发现参数估计方法从多方面进行比较,具体分析如下:

1) 随机变分推理是贝叶斯估计算法,在线 EM 算法属于最大似然估计算法.随机变分推理和在线 EM 算法最初是用来处理独立同分布的大数据集上的参数估计,最近被应用到大规模网络数据上;
2) 随机变分推理的原理是每次迭代抽样整个数据集合的子集更新局部变分参数,利用局部变分参数计算噪音自然梯度,基于随机梯度算法更新全局变分参数.该算法通过每次迭代的抽样来减少算法的计算量.在线 EM 算法每次迭代估计新产生数据的缺失数据,基于缺失数据和上次迭代的参数更新模型参数;
3) 两类算法都通过每次迭代减少局部变量的计算来降低算法复杂度,研究证明:将每次新访问的样本看作抽样样本,在线 EM 算法可看作随机梯度算法;将每次抽样子集看作在线访问样本,随机梯度算法也可看作在线算法.随机变分推理是全局变分参数的随机梯度算法,在线 EM 算法是模型参数的随机梯度算法.

表 2 给出典型概率方法的比对:建模对象指网络单元;结构指网络结构(社区结构指对紧密链接子图,块结构

指网络的任意结构);节点属性指节点链接相关因素(同质性指同类节点链接,异质性指异类节点链接,流行性指节点易与流行节点产生链接);复杂度指参数估计方法复杂度,其中,M 表示边数,K 表示社区个数,σ表示节点最大度.

由表 2 中的方法可知:

1) 大规模网络结构发现方法设计要考虑模型网络结构的假设、建模对象表示、参数估计等因素.PTM 模型采用三角形模体表示网络,简化网络结构假设,利用随机变分推理技术和并行技术求解模型参数,通过多个方面提高模型参数估计算法.该方法的性能优于仅从单方面改进的方法;
2) 在对网络结构存在先验的条件下,可以通过模型假设简化模型,为提高模型参数求解算法效率提供条件.a-MMSB,AMP,aHDPR,PTM 都简化模型的节点链接模式,使模型对假设类型建模,降低算法复杂度,但也限制了模型识别的结构,在对网络结构有此先验的条件下,可以采用这种简化模型的思想;
3) 简单随机块模型用在线变分 EM、在线 CEM、在线 SAEM 算法提高参数估计效率,在线变分 EM 算法被证明准确率最高.该类算法可识别的结构更多,但复杂度高于随机变分推理方法;
4) 网络结构发现建模中需要考虑节点的不同属性,如异质性、同质性、流行性.异质性表示节点与异类节点产生链接,同质性表示节点与同类社区节点产生链接,流行性指度大的节点产生链接的可能性更大.目前的模型大多只考虑节点的同质性,在实际网络建模过程中需要考虑节点的不同属性;
5) 非参数贝叶斯模型 aHDPR 可以执行社区个数未知条件下的社区发现任务,利用随机变分推理求解模型参数.但该模型仅能发现紧密结构的社区,节点的属性建模考虑不全.该模型为第 1 个采用随机变分推理算法的非参数贝叶斯模型,在运行效率和准确率上还需要进一步提高.

4.2 定量比较和分析

根据表 2 的模型可知,已有大规模网络社区发现模型采用的是生成式概率模型,主要基于 SBM 模型和 MMSB 模型的扩展模型,参数估计采用两类方法:在线 EM 算法和随机变分推理算法.下面将典型方法应用到实际数据上验证其准确性(rand index)与运行效率(时间:s),其中,Rand Index 用来计算算法划分结果与实际划分一致的程度[20].实验环境为:处理器 Intel(R) Xeon(R) E7-4807 1.87GHZ,内存 128G,硬盘 160G,操作系统为 Linux (ubuntu13.1064).程序由 c++实现.实验数据包括 3 组:

1) 人工网络数据用 Mixer 的 R 程序包生成,社区个数为 10,每个社区的节点比率相同,节点数由 200 增长到 6 400,边数由 6K~6M.R 程序包可从 http://stat.genopole.cnrs.fr/software/mixnet 下载;
2) 中小规模实际网络的数据,包括 karate(K=2,N=34,M=78),adjnoun(K=2,N=112,M=425),dolphin(K=2,N=62,M=159),football(K=12,N=115,M=613),political(K=2,N=1490,M=19025),polbook(K=3,N=105,M=441),risk(K=6,N=42,M=83),lemis(K=11,N=77,M=254),其中,K 为类个数,N 为节点数,M 为边数.数据从网页 http://www-personal.umich.edu//~mejn/netdata/下载;
3) 大规模网络数据,从 http://snap.stanford.edu/data/web-Stanford.html 下载.

在线 EM 算法中,在线变分 EM 算法最优;基于随机变分推理的方法中,基于 PTM 模型的算法复杂度最低.选择基于 PTM 模型的算法与在线变分 EM 算法进行对比,所有算法都随机初始化.人工网络上算法对比结果见表 3,结果表明,在线变分 EM 算法的准确率明显高于 PTM 算法.

Table 3 Contrasts of online variational EM algorithm and PTM on synthetic networks

表 3 人工网络上在线变分 EM 算法、PTM 比较

网络		在线变分 EM 算法		PTM	
节点数	边数	Randindex	时间	Randindex	时间
200	6 064	0.996 9	0.47	0.122 3	1.312 0
400	24 028	1.000 0	0.70	0.233 2	3.028 0
800	95 664	1.000 0	2.84	0.134 5	4.855 0
1 600	384 406	1.000 0	21.01	0.549 1	9.640 0
3 200	1 536 480	1.000 0	318.12	0.296 9	22.557 0
6 400	6 146 902	1.000 0	6161.51	0.234 0	64.788 0

中小规模实际网络上的比较如图 2 和图 3 所示,可看出:在线变分 EM 算法准确性和效率比 PTM 模型的随机变分推理算法稍占优势.因此,PTM 模型的随机变分推理算法在中小规模网络优势不明显.PTM 模型的准确率不高的主要原因是:在网络节点较少时,随机初始化后节点大都聚为一类,如果初始人工为每个社区指定正确的节点,结果会更准确,这为将来提高算法性能提供了研究方向.

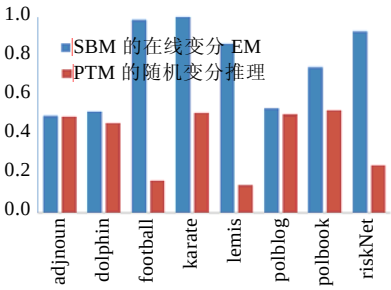

Fig.2 Comparsons of precision on real networks

图 2 实际网络数据上准确性比较

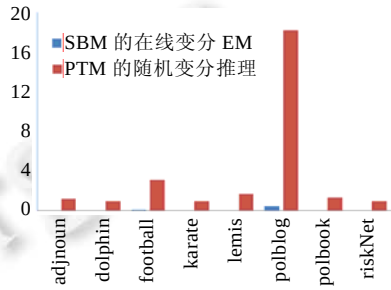

Fig.3 Comparsons of runtime on real networks (s)

图 3 实际网络数据上运行时间比较(s)

为测试 SBM 模型的在线变分 EM 算法和 PTM 的随机变分推理在稀疏大规模网络上运行效率,选择 Stanford 大学 Leskovec 研究组提供的 3 个大网络,测试不同社区个数下两个算法的运行时间,结果见表 4.在线变分 EM 算法在大规模网络上的运行效率很低,如在 Brightkite 网络上,社区个数为 64 时,经过 5 天也没有运行结束,该时间已经不在用户容忍范围,其他配置运行时间更久;而 PTM 的随机变分推理算法可在用户容忍时间范围内给出划分结果.主要原因是在线变分 EM 算法复杂度为 $O(N^2K^2)$,而 PTM 的随机变分推理算法的复杂度为 $O(NK)$.

Table 4 Comparsons of runtime on real large networks

表 4 实际大网络上运行时间比较

网络	节点数	社区数	边数	在线变分 EM	PTM 随机变分推理
Brightkite	58 228	64	428 156	–	2 061.72
Brightkite	58 228	300	428 156	–	10 419.48
Slashdot	82 168	100	948 464	–	8 618.96
Slashdot	82 168	300	948 464	–	24 312.56
webGraph_20	281 903	100	2 312 497	–	20 540.14

4.3 大规模网络结构发现模型设计原则

根据上述理论和实际应用的对比分析,给出设计大规模网络结构发现模型时需要考虑的因素:

1) 有效的网络结构发现概率模型要从表示、模型假设、推理及参数学习多个角度考虑.网络的建模对象表示要简单,模型关于网络对象生成过程假设要简单,参数学习算法要采用大规模数据处理技术;
2) 模型链接生成过程要考虑多个方面:节点的各种属性,如流行性(popularity)、同质性(homogeneity)、异质性(heterogeneity)等;不同类型链接模式,如紧密子图、多分结构、层次结构、星型结构等;链接的分布类型,如泊松分布、伯努利分布、多项式分布;
3) 大规模网络结构发现模型考虑结合节点内容属性及其他辅助信息提高算法性能;
4) 非参数贝叶斯模型 aHDPR 利用层次狄利克雷过程对网络的生成过程建模,不需要设定类的个数,因此,大规模网络结构发现要考虑类个数未知的情形.

这里仅根据已有的少量大规模网络结构发现模型总结了该类模型设计原则,在当今信息冗余的大数据时代,还需要探索更多的因素来提高大规模网络结构发现方法的性能.

5 未来研究课题和展望

在线社交网站产生了大量先验极少的网络数据,对这些数据建模、发现其潜在结构,可为多领域产生具大经济效益.基于概率模型的方法成为网络结构发现的主流方法,但该类方法存在计算瓶颈问题.文中对最近几年出现的一些大规模网络结构发现方法的技术与典型方法的背景、原理、特点等进行总结分析,并从理论和实际应用进行比较分析.分析结果表明:仅有的大规模社区发现概率方法还不能从性能、准确性上满足实际需求,该领域的研究还处于初期阶段,未来需研究的核心课题还有许多,如:

1) 融合多种大规模数据处理技术提高网络结构发现方法的有效性.现有方法主要从一个方面提高网络发现模型效率,以后研究需要从模型假设、表示、推理及学习各个层面利用大数据处理技术;
2) 根据实际网络特性,设定更符合实际网络特征的生成过程假设.网络链接生成过程建模涉及节点的属性、类链接模式、链接的分布等假设,需要折中模型泛化能力和简单特性来对网络生成过程建模;
3) 已有的网络结构发现方法大多假设社区个数已知,以后的研究应该侧重社区个数未知时的网络结构发现任务.层次狄利克雷过程为解决此问题提供了很好的技术,但还需要继续结合实际问题细化模型假设;
4) 将网络链接信息作为建模的一方面信息,与其他信息融合构建更有效的网络结构发现概率模型.目前的快速网络结构方法主要对网络的链接信息建模,已有研究表明:链接信息只是网络的单面信息,考虑融入其他辅助信息建模,可更有效地处理大规模网络数据潜在结构发现问题;
5) 已有的典型方法还停留在研究初期阶段,如何将这些模型应用到在线社交媒体的各种服务中,用服务来评价模型的优劣,是亟待解决的问题;
6) 大规模网络结构发现概率模型的研究缺乏公开数据集合,将来需要一些在线社交媒体提供有标定结果的测试数据集,以便研究者检验模型和算法的优劣.

上述仅是该领域的部分核心课题,在实际应用中还会产生更多的研究内容.该类方法对研究者的程序设计能力要求很高,尤其是处理大规模网络数据.目前出现了一些图处理的平台或框架,如 SNAP,GraphLab,未来的大规模网络结构发现只有基于这些技术,结合工业界的具体应用需求和网络分析目标,才能设计出有用的网络结构发现算法.

References:

[1] Wang YJ, Wong GY. Stochastic blockmodels for directed graphs. Journal of the American Statistical Association, 1987,82(397): 8−19. [doi: 10.1080/01621459.1987.10478385]

[2] Airodi EM, Blei DM, Fienberg SE, Stephen E, Eric XE. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 2008,9(1):1981−2014.

[3] Ren W, Yan GY, Liao XP, Lan X. Simple probabilistic algorithm for detecting community structure, Physical Review E, 2009, 79(3):036111.

[4] Shen HW, Cheng XQ, Guo JF. Exploring the structural regularities in networks. Physical Review E, 2011,84(5):056111. [doi: 10.1103/PhysRevE.84.056111]

[5] Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Physical Review E, 2011,83(11):016107. [doi: 10.1103/PhysRevE.83.016107]

[6] Ball B, Karrer B, Newman MEJ. An efficient and principled method for detecting communities in network. Physical Review, 2011, 84(3):036103. [doi: 10.1103/PhysRevE.84.036103]

[7] Hofman JM, Wiggins CH. Bayesian approach to network modularity. Physical Review Letters, 2008,100(25):258701. [doi: 10.1103/PhysRevLett.100.258701]

[8] Latouche P, Birmele E, Ambroise C. Overlapping stochastic block models with application to the french political blogosphere. The Annals of Applied Statistics, 2011,5(1):309−336. [doi: 10.1214/10-AOAS382]

[9] Chai BF, Yu J, Jia CY, Yang TB, Jiang YW. Combining a popularity-productivity stochastic block model with a discriminative-content model for general structure detection. Physical Review E, 2013,88(3):012807. [doi: 10.1103/PhysRevE.88.012807]

[10] Yang B, Liu JM, Liu D. Characterizing and extracting multiplex patterns in complex networks. IEEE Trans. on Systems, Man, and Cybernetics—Part B: Cybernetics, 2012,42(2):469−481. [doi: 10.1109/TSMCB.2011.2167751]

[11] Nowicki K, Snijders T. Estimation and prediction for stochastic block structures. Journal of American Statistical Association, 2001, 96(455):1077−1087. [doi: 10.1198/016214501753208735]

[12] Daudin J, Picard F, Robin S. A mixture model for random graphs. Journal of Statistical Computation and Simulation, 2008,18(2):179−183. [doi: 10.1007/s11222-007-9046-7]

[13] Ho Q, Parikh AP, Xing EP. A multiscale community blockmodel for network exploration. Journal of the American Statistical Association, 2012,107(499):916−934. [doi: 10.1080/01621459.2012.682530]

[14] Gopalan P, Mimno D, Gerrish SM, Freedman MJ, Blei DM. Scalable inference of overlapping communities. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. Proc. of the 26th Annual Conf. on Neural Information Processing Systems 2012. San Francisco: Inc. Curran Associates, 2012. 2258−2266.

[15] Gopalan PK, Blei DM. Efficient discovery of overlapping communities in massive networks. Proc. of the National Academy of Sciences, 2013,110(36):14534−14539. [doi: 10.1073/pnas.1221839110]

[16] Gopalan PK, Wang C, Blei DM. Modeling overlapping communities with node popularities. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ, eds. Proc. of the 27th Annual Conf. on Neural Information Processing Systems 2013. San Francisco: Inc. Curran Associates, 2013. 2850−2858.

[17] Hoffman MD, Blei DM, Wang C, Paisley J. Stochastic variational inference. Journal of Machine Learning Research, 2013,14: 1307−1347.

[18] Kim DI, Gopalan P, Blei DM, Sudderth EB. Efficient online inference for bayesian nonparametric relational models. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ, eds. Proc. of the 27th Annual Conf. on Neural Information Processing Systems 2013. San Francisco: Inc. Curran Associates, 2013. 962−970.

[19] Zanghi H, Ambroise C, Miele V. Fast online graph clustering via Erdös-Rényi mixture. Pattern Recognition, 2008,41(12): 3592−3599. [doi: 10.1016/j.patcog.2008.06.019]

[20] Zanghi H, Picard F, Miele V, Ameroise C. Strategies for online inference of model-based clustering in large and growing networks. The Annals of Applied Statistics, 2010,4(2):687−714. [doi: 10.1214/10-AOAS359]

[21] Ho QR, Yin JM, Xing EP. On triangular versus edge representations—Towards scalable modeling of networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. Proc. of the 26th Annual Conf. on Neural Information Processing Systems 2012. San Francisco: Inc. Curran Associates, 2012. 2141−2149.

[22] Yin JM, Ho QR, Xing EP. A scalable approach to probabilistic latent space inference of large-scale networks. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ, eds. Proc. of the 27th Annual Conf. on Neural Information Processing Systems 2013. San Francisco: Inc. Curran Associates, 2013. 422−430.

[23] Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 2008,1(1-2):1−305.

[24] Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Machine Learning, 1999, 37:183−233. [doi: 10.1023/A:1007665907178]

[25] Amari S. Natural gradient works efficiently in learning. Neural Computation, 1998,10(2):251−276. [doi: 10.1162/ 0899766983000 17746]

[26] Lai TL. Stochastic approximation. The Annals of Statistics, 2003,31(2):391−406. [doi: 10.1214/aos/1051027873]

[27] Hoffman DH, Blei DM, Bach FR. Online learning for latent dirichlet allocation. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds. Proc. of the 24 th Annual Conf. on Neural Information Processing Systems 2010. San Francisco: Curran Associates Group, Inc., 2010. 856−864.

[28] Neal RM, Hinton GE. A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan MI, ed. Proc. of the Learning in Graphical Models. Cambridge: MIT Press, 1999. 335−368. [doi: 10.1007/978-94-011-5014-9_12]

[29] Bottou L. Online algorithms and stochastic approximations. In: Saad D, ed. Proc. of the Online Learning and Neural Networks. Cambridge: Cambridge University Press, 1998.

[30] Hoi S CH, Wang JL, Zhao PL. LIBOL: A library for online learning algorithms. Journal of Machine Learning Research, 2014,15: 495−499.

[31] Cappé O, Moulines E. Online EM algorithm for latent data models. Journal of the Royal Statistical Society B, 2009,71(3):593−613.

[32] Cappé O. Online Expectation-Maximisation. Mixtures: Estimation and Applications, 2011. 31−53.

[33] Papadopoulos F, Kitsak M, Serrano M, Boguna M, Krioukov D. Popularity versus similarity in growing networks. Nature, 2012,489(7417):537−540. [doi: 10.1038/nature11459]

[34] Nepusz T, Petrczi A, Ngyessy L, Bazso F. Fuzzy communities and the concept of bridgeness in complex networks. Physical Review E, 2008,77(1):016107. [doi: 10.1103/PhysRevE.77.016107]

柴变芳(1979－),女,山西运城人,博士生,CCF 学生会员,主要研究领域为复杂网络分析,文本挖掘.

E-mail: chaibianfang@163.com

贾彩燕(1976－),女,博士,副教授,CCF 会员,主要研究领域为复杂网络分析,生物信息学.

E-mail: cyjia@bjtu.edu.cn

于剑(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,图像处理.

E-mail: jianyu@bjtu.edu.cn