自动化学报
ACTA AUTOMATICA SINICA

基于扩散方法的分布式随机变分推断算法

付维明 秦家虎 朱英达

Distributed Stochastic Variational Inference Based on Diffusion Method

FU Wei-Ming, QIN Jia-Hu, ZHU Ying-Da

在线阅读 View online: https://doi.org/10.16383/j.aas.c200445

基于扩散方法的分布式随机变分推断算法

付维明 [1] 秦家虎 [1] 朱英达 [1]

摘 要 分布式网络上的聚类、估计或推断具有广泛的应用, 因此引起了许多关注. 针对已有的分布式变分贝叶斯 (Variational Bayesian, VB) 算法效率低, 可扩展性差的问题, 本文借用扩散方法提出了一种新的分布式随机变分推断 (Stochastic variational inference, SVI) 算法, 其中我们选择自然梯度法进行参数本地更新并选择对称双随机矩阵作为节点间参数融合的系数矩阵. 此外, 我们还为所提出的分布式 SVI 算法提出了一种对异步网络的适应机制. 最后, 我们在伯努利混合模型 (Bernoulli mixture model, BMM) 和隐含狄利克雷分布 (Latent Dirichlet allocation, LDA) 模型上测试所提出的分布式 SVI 算法的可行性, 实验结果显示其在许多方面的性能优于集中式 SVI 算法.

关键词 分布式算法, 随机变分推断, 扩散方法, 异步网络, 主题模型

引用格式 付维明, 秦家虎, 朱英达. 基于扩散方法的分布式随机变分推断算法. 自动化学报, 2021, **47**(1): 92−99

DOI 10.16383/j.aas.c200445

Distributed Stochastic Variational Inference Based on Diffusion Method

FU Wei-Ming[1] QIN Jia-Hu[1] ZHU Ying-Da[1]

Abstract Clustering, estimation, or inference in distributed networks has received considerable attention due to its broad applications. Considering that existing distributed variational Bayesian (VB) algorithms have the weaknesses of low efficiency and poor scalability, this paper proposes a new distributed stochastic variational inference (SVI) algorithm by borrowing the diffusion method, where the natural gradient method is used for the update of local parameters, and a symmetric and doubly stochastic matrix is applied for the fusion of local parameters. In addition, an adaptation mechanism is introduced in the proposed distributed SVI algorithm for use in asynchronous networks. The feasibility of the proposed distributed SVI algorithm is demonstrated with the Bernoulli mixture model (BMM) and the latent Dirichlet allocation (LDA) model. Experimental results show that the proposed distributed SVI algorithm outperforms the centralized one in many aspects.

Key words Distributed algorithm, stochastic variational inference (SVI), diffusion method, asynchronous network, topic mode

Citation Fu Wei-Ming, Qin Jia-Hu, Zhu Ying-Da. Distributed stochastic variational inference based on diffusion method. *Acta Automatica Sinica*, 2021, **47**(1): 92−99

在大数据时代, 数据通常会被分布式地存储在多个节点上, 例如传感器网络[1−3] 和分布式数据库[4] 中等, 其中每个节点只拥有部分数据. 考虑到单个节点的存储容量有限以及保护数据隐私或安全的需求[5−6], 通常无法将所有数据都发送给一个中心节点, 然后利用集中式的方法处理这些数据, 因此开发高效的算法对分布式存储的数据进行挖掘已成为当前一个重要的研究方向[7−12].

变分贝叶斯 (Variational Bayesian, VB) 推断[13] 是一种功能强大的数据挖掘技术, 被广泛用于解决实际问题, 如识别文档主题[14−15], 对数据进行聚类和密度估计[16] 以及预测未知数据[17] 等. 近年来, 研究者们已提出很多分布式的 VB 算法[3, 18−20], 然而在大多数这些算法的每步迭代中, 都需要基于整个数据集更新全局参数, 这不仅会导致算法计算代价大、效率低, 还会导致算法可扩展性差, 难以扩展到在线学习或者流数据处理的情况.

随机变分推断 (Stochastic variational inference, SVI)[15] 的提出使得贝叶斯推断方法在处理海量数据时具有更高的效率和可扩展性. 它借用了随机优化的方法, 根据基于子样本的噪声自然梯度来优化目标函数, 大大减小了每步迭代时所需的存储量和计算量. 目前已有一些研究者将其扩展为分布

收稿日期 2020-06-22 录用日期 2020-09-22
Manuscript received June 22, 2020; accepted September 22, 2020
国家自然科学基金 (61873252, 61922076), 霍英东教育基金会高等院校青年教师基金 (161059) 资助
Supported by National Natural Science Foundation of China (61873252, 61922076), Fok Ying-Tong Education Foundation for Young Teachers in Higher Education Institutions of China (161059)
本文责任编委 杨涛
Recommended by Associate Editor YANG Tao
1. 中国科学技术大学自动化系 合肥 230027
1. Department of Automation, University of Science and Technology of China, Hefei 230027

式版本, 以提高分布式数据的处理效率以及将其应用于分布式数据流的处理[21]. 具体地, 文献 [22] 提出了一种有中心的异步分布式 SVI 算法, 该算法中的中心节点负责收发全局参数, 其余节点并行地更新全局参数. 值得一提的是, 这类有中心的算法往往会存在鲁棒性差, 链路负载不平衡, 数据安全性差等缺点. 在文献 [11] 中, 交替方向乘子方法 (Alternating direction method of multipliers, ADMM)[23] 被用来构造两种无中心的分布式 SVI 算法, 克服了有中心的算法的缺点, 但它们存在每步迭代中全局参数本地更新所需的计算代价大以及不适用于异步网络的缺点.

本文以 SVI 为核心, 借用多智能体一致优化问题中的扩散方法[24], 发展了一种新的无中心的分布式 SVI 算法, 并针对异步网络提出了一种适应机制. 在所提出的算法中, 我们利用自然梯度法进行全局参数的本地更新, 并选择对称双随机矩阵作为节点间参数融合的系数矩阵, 减小了本地更新的计算代价. 最后, 我们在伯努利混合模型 (Bernoulli mixture model, BMM) 和隐含狄利克雷分布 (Latent Dirichlet allocation, LDA) 上验证了所提出的算法的可行性, 实验结果显示所提出的算法在发现聚类模式, 对初始参数依耐性以及跳出局部最优等方面甚至优于集中式 SVI 算法, 这是以往分布式 VB 算法所没有表现出来的.

本文其余部分安排如下: 第 1 节介绍集中式 SVI 算法; 第 2 节介绍本文所提出的分布式 SVI 算法并给出了一种针对异步网络的适应机制; 第 3 节展示在 BMM 和 LDA 模型上的实验结果; 第 4 节对本文工作进行总结.

1　随机变分推断

1.1　模型介绍

SVI 基本模型包含以下这些量: 数据集 $\boldsymbol{x}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$, 局部隐藏变量 $\boldsymbol{y}=\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_N\}$, 全局隐藏变量 $\boldsymbol{\beta}$ 以及模型参数 $\boldsymbol{\alpha}$. 模型的概率图如图 1 所示, 其中黑色圆圈代表固定参数, 灰色圆圈代表数据集, 白色圆圈代表隐藏变量, 箭头描述了它们之间的依赖关系. 具体地, $\boldsymbol{\alpha}$ 直接影响 $\boldsymbol{\beta}$, $\boldsymbol{\beta}$ 直接影响局部变量对 $(\boldsymbol{x}_n,\boldsymbol{y}_n)$. 我们假设全局隐藏变量 $\boldsymbol{\beta}$ 的先验分布属于指数族分布且具有如下形式:

$$p(\boldsymbol{\beta};\boldsymbol{\alpha})\propto\exp\left(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{u}(\boldsymbol{\beta})-A(\boldsymbol{\alpha})\right)\tag{1}$$

其中, $\boldsymbol{u}(\boldsymbol{\beta})$ 表示自然参数, $A(\boldsymbol{\alpha})$ 表示归一化函数; 不同局部变量对 $(\boldsymbol{x}_n,\boldsymbol{y}_n)$ 之间相互独立且其分布也属于指数族分布, 具体形式如下:

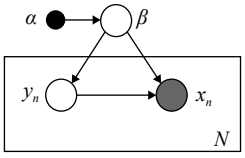


图 1　本文考虑的模型的概率图表示

Fig. 1　The graphic model considered in this paper

$$p(\boldsymbol{x}_n,\boldsymbol{y}_n|\boldsymbol{\beta})\propto\exp\left(\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{\beta})f(\boldsymbol{x}_n,\boldsymbol{y}_n)\right)\tag{2}$$

其中 $f(\boldsymbol{x}_n,\ \boldsymbol{y}_n)$ 表示自然充分统计量; 此外, 还假设上述两个指数族分布满足共轭条件关系[25], 以使后验分布与先验分布的形式相同. 我们的目标是根据观测到的数据集来估计局部隐藏变量的分布, 即其后验分布 $p(\boldsymbol{y},\boldsymbol{\beta}|\boldsymbol{x})$.

1.2　平均场变分推断

平均场变分推断是一种用一个可以因式分解的变分分布去近似后验分布的方法. 在上一节介绍的模型基础上, 我们可以用变分分布 $q(\boldsymbol{y},\boldsymbol{\beta})$ 来近似 $p(\boldsymbol{y},\boldsymbol{\beta}|\boldsymbol{x})$, 并假设该变分分布满足以下条件:

$$q(\boldsymbol{y},\boldsymbol{\beta})=q(\boldsymbol{\beta};\boldsymbol{\lambda})\prod_n q(\boldsymbol{y}_n;\boldsymbol{\phi}_n)\tag{3}$$

$$q(\boldsymbol{\beta};\boldsymbol{\lambda})\propto\exp\left(\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{u}(\boldsymbol{\beta})-A(\boldsymbol{\lambda})\right)\tag{4}$$

$$q(\boldsymbol{y}_n;\boldsymbol{\phi}_n)\propto\exp\left(\boldsymbol{\phi}_n^{\mathrm{T}}\boldsymbol{u}(\boldsymbol{y}_n)-A(\boldsymbol{\phi}_n)\right)\tag{5}$$

其中, $\boldsymbol{\lambda}$ 和 $\boldsymbol{\phi}=\{\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,\cdots,\boldsymbol{\phi}_N\}$ 是变分参数. 此时需要最小化 $q(\boldsymbol{y},\boldsymbol{\beta})$ 和 $p(\boldsymbol{y},\boldsymbol{\beta}|\boldsymbol{x})$ 之间的 Kullback-Leibler (KL) 散度来让 $q(\boldsymbol{y},\boldsymbol{\beta})$ 逼近 $p(\boldsymbol{y},\boldsymbol{\beta}|\boldsymbol{x})$, 这等价于最大化

$$L(\boldsymbol{\lambda},\boldsymbol{\phi})=\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{\beta};\boldsymbol{\alpha})}{q(\boldsymbol{\beta};\boldsymbol{\lambda})}\right]+\sum_{n=1}^{N}\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{x}_n,\boldsymbol{y}_n|\boldsymbol{\beta})}{q(\boldsymbol{y}_n;\boldsymbol{\phi}_n)}\right]\tag{6}$$

其中, $\mathbb{E}_q[\cdot]$ 表示在分布 $q(\boldsymbol{y},\boldsymbol{\beta})$ 下的期望函数, $L(\boldsymbol{\lambda},\boldsymbol{\phi})$ 是对数证据 $\ln p(\boldsymbol{x})$ 的一个下界, 被称为 Evidence lower bound (ELBO)[15]. 基于 $q(\boldsymbol{y},\boldsymbol{\beta})$) 可分解的假设, 最大化 $L(\boldsymbol{\lambda},\boldsymbol{\phi})$ 可以利用坐标上升法[26] 通过交替更新 $\boldsymbol{\lambda}$ 和 $\boldsymbol{\phi}$ 来实现. 下文讨论的 SVI 以上述平均场变分推断方法为基础.

如果我们固定 $\boldsymbol{\phi}$, 则可以把 $L(\boldsymbol{\lambda},\boldsymbol{\phi})$ 看成是 $\boldsymbol{\lambda}$ 的函数, 此时需要求解

$$L(\boldsymbol{\lambda})=\max_{\boldsymbol{\phi}}L(\boldsymbol{\lambda},\boldsymbol{\phi})\tag{7}$$

常用的方法是对其求 (欧氏) 梯度, 但是用欧氏距离表征不同 $\boldsymbol{\lambda}$ 之间的远近关系是不合理的, 这是因为 $\boldsymbol{\lambda}$ 为变分参数, 我们所关心的是不同的 $\boldsymbol{\lambda}$ 所刻画的分布 $q(\boldsymbol{y},\boldsymbol{\beta})$ 之间的差异, 此时可以引入自然梯度[15],

它表示的是函数在黎曼空间上的梯度. 通过对 $L(\boldsymbol{\lambda},\boldsymbol{\phi})$ 关于 $\boldsymbol{\phi}$ 求自然梯度, 可以将平均场变分推断推广到随机优化的版本, 即随机变分推断. 具体地, 我们定义如下的随机函数

$$L_I(\boldsymbol{\lambda}):=N\max_{\boldsymbol{\phi}_I}\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{x}_I,\boldsymbol{y}_I|\boldsymbol{\beta})}{q(\boldsymbol{y}_I;\boldsymbol{\phi}_I)}\right]+\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{\beta};\boldsymbol{\alpha})}{q(\boldsymbol{\beta};\boldsymbol{\lambda})}\right] \tag{8}$$

其中, I 是均匀取值于 $\{1,\cdots,N\}$ 的随机变量. 易知 $L_I(\boldsymbol{\lambda})$ 的期望等于 $L(\boldsymbol{\lambda})$, 因此每次均匀地选取一个数据点 n 时, $L_n(\boldsymbol{\lambda})$ 给出了 $L(\boldsymbol{\lambda})$ 的一个无偏估计. 根据随机优化理论, 集中式 SVI 的过程由下面两步构成:

1) 均匀地随机选取一个数据点 n, 并计算当前最优的局部变分参数 $\boldsymbol{\phi}_n^*$;

2) 通过

$$\boldsymbol{\lambda}^{t+1}=(1-\rho_t)\boldsymbol{\lambda}^t+\rho_t\left(\boldsymbol{\alpha}+N\mathbb{E}_{\boldsymbol{\phi}_n^*}\left[f(\boldsymbol{x}_n,\boldsymbol{y}_n)\right]\right) \tag{9}$$

更新全局变分参数 $\boldsymbol{\lambda}$.

上述 SVI 算法一次迭代只采样一个数据点, 其也可以被直接扩展成一次采样一个数据批量 (Batch) 的版本, 详见文献 [15].

2 基于扩散方法的分布式 SVI 算法

2.1 问题描述

我们考虑一个由 J 个节点组成的分布式网络, 其中每个节点 i 存储包含 N_i 个数据项的数据集 $\boldsymbol{x}_i=\{\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iN_i}\}$, 于是整个网络上存储的完整数据集为 $\boldsymbol{x}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_J\}$, 总数据项数为 $N=\sum_i N_i$. 假设网络的通讯拓扑是一个无向图 $G=(V,E)$, 其中 $V=\{1,\cdots,J\}$ 是节点集合, $E\subseteq V\times V$ 是边集合, $(i,j)\in E$ 表明信息可以在节点 i 和节点 j 之间直接传输, 记节点 i 的邻居集合为 $B_i=\{j\in V:(j,i)\in E\}$. 此外, 我们还假设 G 是连通的, 即对 $\forall i\neq j$, 存在至少一条路径连接节点 i 和节点 j.

如果记节点 i 的局部隐藏变量为 $\boldsymbol{y}_i=\{\boldsymbol{y}_{i1},\cdots,\boldsymbol{y}_{iN_i}\}$, 记对应的局部变分参数为 $\boldsymbol{\phi}_i=\{\boldsymbol{\phi}_{i1},\cdots,\boldsymbol{\phi}_{iN_i}\}$, 则 ELBO 可以写为

$$L(\boldsymbol{\lambda},\boldsymbol{\phi})=\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{\beta};\boldsymbol{\alpha})}{q(\boldsymbol{\beta};\boldsymbol{\lambda})}\right]+\sum_{j=1}^{J}\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{x}_j,\boldsymbol{y}_j|\boldsymbol{\beta})}{q(\boldsymbol{y}_j;\boldsymbol{\phi}_j)}\right] \tag{10}$$

2.2 算法设计

我们借用多智能体一致优化问题中的扩散方法来发展分布式 SVI 算法. 扩散方法的基本思想是交替执行本地更新和节点间参数融合两个步骤, 从而使所有节点的参数收敛到所希望的全局最优值或者局部最优值.

对于节点 i, 如果定义其局部 ELBO 为

$$L_i(\boldsymbol{\lambda},\boldsymbol{\phi}_i)=\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{\beta};\boldsymbol{\alpha})}{q(\boldsymbol{\beta};\boldsymbol{\lambda})}\right]+J\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{x}_i,\boldsymbol{y}_i|\boldsymbol{\beta})}{q(\boldsymbol{y}_i;\boldsymbol{\phi}_i)}\right] \tag{11}$$

则显然有 $L(\boldsymbol{\lambda},\boldsymbol{\phi})=\dfrac{1}{J}\sum_i L_i(\boldsymbol{\lambda},\boldsymbol{\phi}_i)$, 因此每个节点 i 可以把 $L_i(\boldsymbol{\lambda},\boldsymbol{\phi}_i)$ 作为优化目标, 进行本地更新. 在本地更新步骤中, 我们分批次训练以提高学习效率. 不失一般性, 将 $\boldsymbol{x}_i$ 划分成 M 个子集 $\boldsymbol{x}_i=\{\boldsymbol{x}_{i1}^b,\cdots,\boldsymbol{x}_{iM}^b\}$, 对应地, $\boldsymbol{y}_i=\{\boldsymbol{y}_{i1}^b,\cdots,\boldsymbol{y}_{iM}^b\}$, $\boldsymbol{\phi}_i=\{\boldsymbol{\phi}_{i1}^b,\cdots,\boldsymbol{\phi}_{iM}^b\}$, 定义

$$L_{im}^b(\boldsymbol{\lambda},\boldsymbol{\phi}_{im}^b)=\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{\beta};\boldsymbol{\alpha})}{q(\boldsymbol{\beta};\boldsymbol{\lambda})}\right]+MJ\mathbb{E}_q\left[\ln\frac{p(\boldsymbol{x}_{im}^b,\boldsymbol{y}_{im}^b|\boldsymbol{\beta})}{q(\boldsymbol{y}_{im}^b;\boldsymbol{\phi}_{im}^b)}\right] \tag{12}$$

不难证明 $L_{im}^b(\boldsymbol{\lambda},\boldsymbol{\phi}_{im}^b)$ 是 $L_i(\boldsymbol{\lambda},\boldsymbol{\phi}_i)$ 的一个无偏估计, 因此可以基于子样本 $\boldsymbol{x}_{im}^b$ 更新参数. 令 $\boldsymbol{\phi}_i$ 表示节点 i 的全局变分参数, 则类似于集中式[2] 的情况, 节点 i 中第 t 次本地更新由下列公式描述:

$$\boldsymbol{\phi}_{im}^{b^*}=\arg\max_{\boldsymbol{\phi}_{im}^b}L_{im}^b(\boldsymbol{\lambda}_i^t,\boldsymbol{\phi}_{im}^b) \tag{13}$$

$$\boldsymbol{z}_i^{t+1}=\boldsymbol{\lambda}_i^t+\rho_t\tilde{\nabla}L_{im}^b(\boldsymbol{\lambda}_i^t,\boldsymbol{\phi}_{im}^{b^*}) \tag{14}$$

其中, $\tilde{\nabla}L_{im}^b(\boldsymbol{\lambda},\boldsymbol{\phi}_{im}^{b*})=\boldsymbol{\alpha}-\boldsymbol{\lambda}+MJ\sum_{x_{in}\in x_{im}^b}\mathbb{E}_q[f(\boldsymbol{x}_{in},\boldsymbol{y}_{in})]$ 表示 $L_{im}^b(\boldsymbol{\lambda},\boldsymbol{\phi}_{im}^{b*})$ 关于 $\boldsymbol{\lambda}$ 的自然梯度, $\tilde{\nabla}$ 为自然梯度符号. 我们选择随时间衰减形式的学习率 $\rho_t=(t+\tau)^{-\kappa}$, 其中 $\kappa\in[0.5,1]$, $\tau\geq 0$, 这样一来根据随机优化理论[15], 可以保证基于噪声自然梯度的更新式 (14) 可以使得全局变分参数收敛到 $L_i(\boldsymbol{\lambda},\boldsymbol{\phi}_i)$ 的一个局部最优值.

注意本地更新只能使每个节点的全局变分参数独立地收敛到各自的局部 ELBO 的局部最优值, 我们还要保证每个节点学得的全局变分参数收敛到一致, 即 $||\boldsymbol{\lambda}_i-\boldsymbol{\lambda}_j||\to 0,\ \forall i\neq j$. 由于我们已经假设拓扑图是连通的, 因此只要使 $||\boldsymbol{\lambda}_i-\boldsymbol{\lambda}_j||\to 0$, $\forall(i,j)\in E$ 就可以保证所有节点的全局变分参数都收敛到一致. 为此, 根据扩散方法, 我们在每次本地更新之后, 将每个节点的当前全局变分参数发送给其邻居节点, 然后将当前的全局变分参数与从邻居节点接受到的全局变分参数进行融合. 上述过程可以由下面公式描述:

$$\boldsymbol{\lambda}_i^{t+1}=\sum_{j\in B_i\cup\{i\}}p_{ji}\boldsymbol{z}_j^{t+1} \tag{15}$$

其中, p_{ij} 是融合系数, 我们采用如下的定义

$$p_{ij}=\begin{cases}\dfrac{1}{\max(|B_i|,|B_j|)}, & i\in B_j\\ 0, & i\neq j,\ i\notin B_j\\ 1-\displaystyle\sum_{l=1,l\neq i}^{J}p_{il}, & i=j,\ i\notin B_j\end{cases}\tag{16}$$

事实上, 如上定义的 $[p_{ij}]$ 是一个对称随机矩阵. 当迭代次数很大的时候, ρ_t 变得很小, 则有 $\boldsymbol{z}_i^{t+1}\approx\boldsymbol{\lambda}_i^t$, 分布式 SVI 算法退化成由式 (15) 描述的平均一致性协同过程, 所以 $\boldsymbol{\lambda}_i^t$ 将收敛到所有节点初始参数值的平均值. 这样使得训练结果不会对任何节点的数据分布有偏向性.

2.3　针对异步网络的适应机制

上节所述的分布式 SVI 算法默认是同步执行的, 即所有节点在每个迭代步同步地执行本地更新和参数融合两个步骤. 但是所有节点同步执行需要使用时间同步协议去估计和补偿时序偏移, 这会带来额外的通信负载. 此外, 执行快的节点需要等待执行慢的节点, 这会大大降低算法的执行速度. 为此我们设计了一种机制使所提出的分布式 SVI 算法适应异步通信网络. 具体地, 每个节点额外开辟一块存储区域将邻居节点发送过来的 $\boldsymbol{\lambda}_i^t$ 存储起来. 在每个参数融合步中, 如果在等待一定的时间后收到了来自邻居节点发送过来的 $\boldsymbol{\lambda}_i^t$, 则更新存储区域中的 $\boldsymbol{\lambda}_i^t$ 的值, 然后, 用更新后的 $\boldsymbol{\lambda}_i^t$ 进行本地参数更新; 否则, 直接用存储区域的 $\boldsymbol{\lambda}_i^t$ 值进行本地参数更新. 这样一来, 既可以使所提出的分布式算法以异步方式执行, 又尽可能地保证了算法的性能.

3　实验

这一节我们将所提出的分布式 SVI 算法 (我们称之为异步分布式 SVI) 应用于 BMM 模型和 LDA 主题模型, 并在不同的数据集上测试其性能. 并且将其与集中式 SVI 算法和 dSVB 算法[3] 进行对比, 其中 dSVB 算法被我们以同样的方式扩展成随机的版本以方便比较.

3.1　伯努利混合模型

我们考虑具有 K 个成分的混合多变量伯努利模型. 该模型的全局隐藏变量包括: 每个成分 k 的全局隐藏变量 $\boldsymbol{\beta}_k$, 其维度等于数据维度, 每个维度的值表示该维度的数据值属于 "0" 的概率, 以及成分的混合概率 $\boldsymbol{\pi}=\{\pi_1,\cdots,\pi_K\}$, 其中隐藏变量的先验分布形式如下:

$$p(\boldsymbol{\pi})=Dir(\boldsymbol{\pi};\boldsymbol{\alpha})\tag{17}$$

$$p(\boldsymbol{y}_{in}|\boldsymbol{\pi})=\prod_{k=1}^{K}{\pi_k}^{y_{ink}}\tag{18}$$

$$p(\boldsymbol{\beta})=\prod_{k=1}^{K}p(\boldsymbol{\beta}_k)\propto\prod_{k=1}^{K}\prod_{d=1}^{D}\beta_{kd}^{a-1}(1-\beta_{kd})^{b-1}\tag{19}$$

其中, $\boldsymbol{\alpha}=[\alpha]^K$, a 和 b 是固定的超参数, 在 BMM 模型上的实验中, 我们均设置 $\alpha=a=b=1$.

我们将混合多变量伯努利模型应用到 MNIST 数据集上. 在预处理中, 每张图的每个像素根据其像素值被设为 0 或者 1, 然后每张图被展开成 $28\times28=784$ 维的向量. 我们随机生成包含 50 个节点, 166 条边的无向连通网络, 其拓扑结构如图 2 所示, 并将训练数据平均分给 50 个节点, 每个节点包含 1 200 条数据 (整个 MNIST 训练集包含 60 000 条数据). 实验中, 设置 $K=40$, 并设置全局隐藏变量的先验分布为均匀分布.

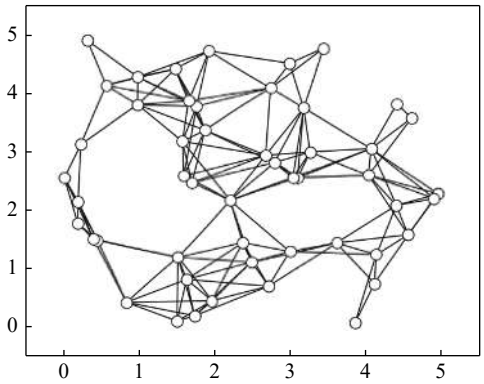
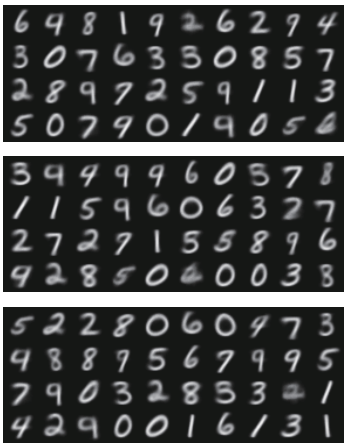


图 2　通信网络拓扑图

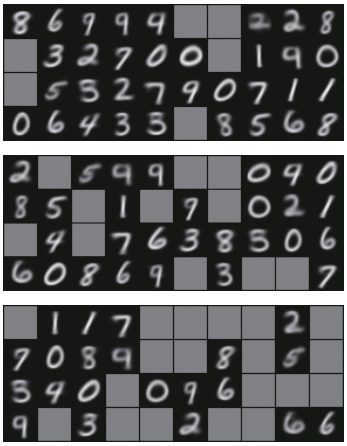
Fig. 2　The topology of the communication network

图 3 展示了所提出的异步分布式 SVI 算法在 $\kappa=0.5$, $\tau=10$ 下, 每份数据分 6 个批次训练 200 个 epoch 得到的聚类中心 (由每个成分 k 的全局隐藏变量 $\boldsymbol{\beta}_k$ 的期望所定义的向量对应的图片) 和相同设置下集中式 SVI 算法得到的聚类中心. 由图 3 可知, 异步分布式 SVI 算法可以充分找到所有潜在的聚类模式, 而集中式 SVI 则往往不能充分找出所有的聚类模式.

在相同设置下多次运行三种算法得到的所有节点估计的 ELBO 的平均值以及相校平均值的偏差演化曲线如图 4 所示, 可以看到异步分布式 SVI 算法相比集中式 SVI 算法能够收敛到更好的值, 并且多次运行得到的结果之间的误差更小, 表现更加稳定. 此外, 异步执行的方式破坏了 dSVB 算法的收敛性, 而异步分布式 SVI 算法对异步网络具有良好



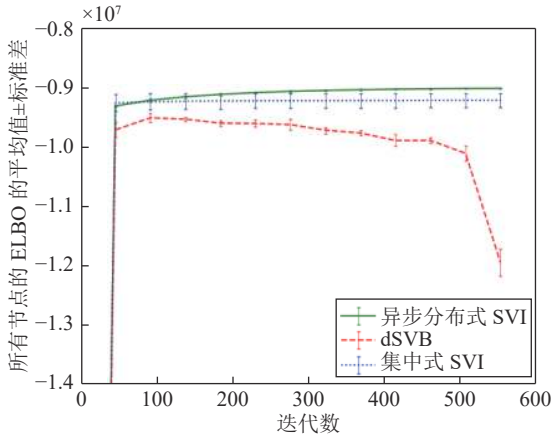
(a) 异步分布式 SVI 算法得到的聚类中心的三个例子
(a) Cluster centers obtained by the asynchronous distributed SVI in three examples



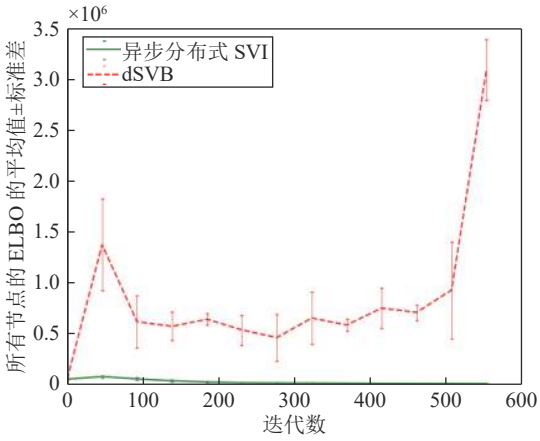
(b) 集中式 SVI 算法得到的聚类中心的三个例子
(b) Cluster centers obtained by the centralized SVI in three examples

图 3 异步分布式 SVI 算法和集中式 SVI 算法得到的聚类中心

Fig. 3 Cluster centers obtained by the asynchronous distributed SVI and the centralized SVI



(a) 节点间 ELBO 平均值的演化
(a) The evolution of the means of the ELBO in nodes



(b) 节点间 ELBO 偏差的演化
(b) The evolution of the deviations of the ELBO in nodes

图 4 异步分布式 SVI 算法、dSVB 算法、集中式 SVI 算法的 ELBO 的平均值和偏差演化

Fig. 4 The evolution of the means and deviations of the ELBO for the asynchronous distributed SVI, the dSVB, and the centralized SVI

的适应性.

为了研究超参数 κ 和 τ 对所提出的分布式 SVI 算法表现的影响, 我们在 $(\kappa=0.5,\tau=1)$, $(\kappa=0.5,\tau=10)$, $(\kappa=0.5,\tau=100)$, $(\kappa=0.75,\tau=10)$, $(\kappa=1,\tau=10)$ 几组参数下进行实验, 所得到的所有节点 ELBO 的平均值的演化曲线见图 5, 可以看到在不同的 (κ,τ) 设置下所提出的异步分布式 SVI 均优于集中式 SVI.

3.2 LDA 主题模型

LDA 主题模型是文档集的概率模型, 它使用隐藏变量对重复出现的单词使用模式进行编码, 由于这些模式在主题上趋于一致, 因此被称为 “主题模型”. 其已经被应用于很多领域, 例如构建大型文档库的主题导航或者辅助文档分类. LDA 模型的贝叶斯网络结构如图 6 所示, 其中变量的说明见表 1.

我们首先在 New York Times 和 Wikipedia 两个数据集上验证异步分布式算法在 LDA 模型上的性能. 首先我们生成一个包含 5 个节点 7 条边的网络, 将每个数据集的文档随机分配给各个节点. 在实验中我们设置 $K=5$, 并以文档集的生成概率的对数 $\ln p(w)=\sum_d\sum_n \ln p(w_{d,n}|\beta_k,y_{d,n})$ 作为评价指标.

图 7 展示了在 $\alpha=0.2$, $\eta=0.2$, $\kappa=0.5$, $\tau=10$, 训练 epoch 取 40, 分布式算法中每个节点的批大小取 10, 集中式算法的批大小取 50 的设置下, 异步分布式 SVI, 集中式 SVI 和 dSVB 以异步方式分别在两个数据集上运行多次得到的 $\ln p(w)$

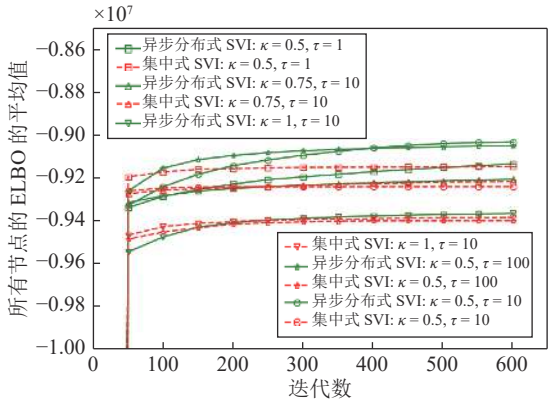


图 5　不同 (κ,τ) 设置下异步分布式 SVI 和集中式 SVI 的 ELBO 的平均值演化

Fig. 5　The evolution of the means of the ELBO for the asynchronous distributed SVI and the centralized SVI under different settings of (κ,τ)

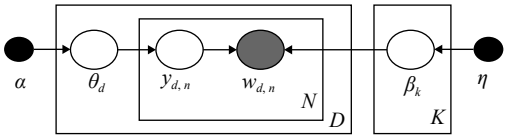


图 6　LDA 模型的贝叶斯网络结构图

Fig. 6　The Bayesian graphic model of LDA

的演化曲线, 可见异步分布式 SVI 算法表现优于另外两种算法. 不同参数设置下异步分布式 SVI 和集中式 SVI 在 New York Times 数据集上收敛时的 $\ln p(w)$ 见表 2, 可见不同设置下异步分布式 SVI 的表现均优于集中式 SVI.

然后我们在复旦大学中文文本分类数据集上测试所提出的异步分布式 SVI 算法. 该数据集来自复旦大学计算机信息与技术系国际数据库中心自然语

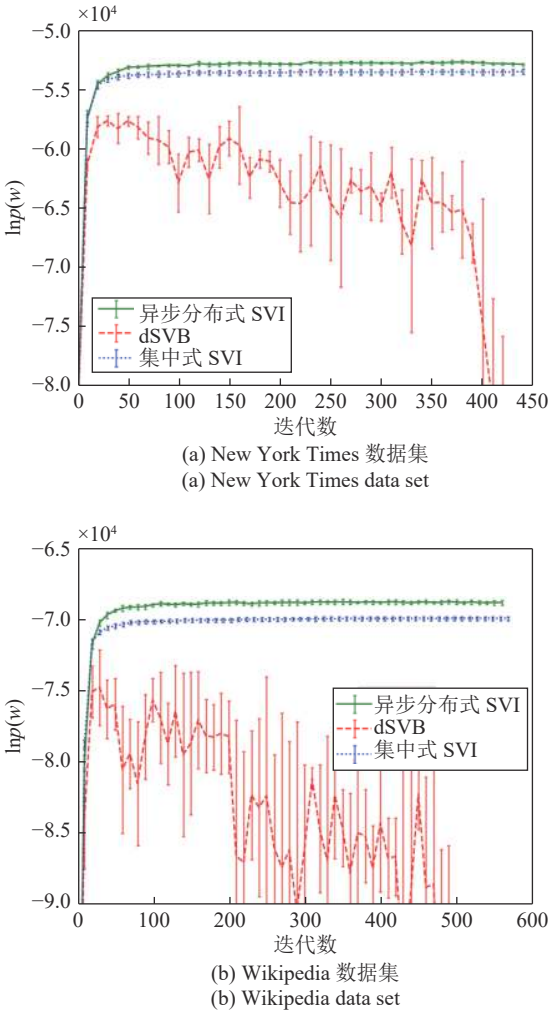


图 7　异步分布式 SVI、集中式 SVI 和 dSVB 在两个数据集上的表现

Fig. 7　Performance of the asynchronous distributed SVI, the centralized SVI, and the dSVB on the two data sets

表 1　LDA 模型变量

Table 1　Variables in LDA model

变量	α	η	K	D	N	θ_d	$y_{d,n}$	$w_{d,n}$	β_k
类型	固定参数	固定参数	固定参数	输入参数	输入参数	局部隐藏变量	局部隐藏变量	单词向量	全局隐藏变量
描述>			主题数	文档数	单词数	决定文档的主题分布	单词所属的主题		决定主题的单词分布
分布						$Dir(\theta_d\|\alpha)$	$Mult(y_{d,n}\|\theta_d)$	$Mult(w_{d,n}\|\beta_k,y_{d,n})$	$Dir(\beta_k\|\eta)$

表 2　不同参数设置下异步分布式 SVI 和集中式 SVI 收敛的值

Table 2　The convergent values of the asynchronous distributed SVI and the centralized SVI under different parameter settings

参数设置	$\alpha=\eta=0.4$ $\kappa=0.5,\tau=1$	$\alpha=\eta=0.8$ $\kappa=0.5,\tau=1$	$\alpha=\eta=0.4$ $\kappa=0.7,\tau=10$	$\alpha=\eta=0.8$ $\kappa=0.7,\tau=10$	$\alpha=\eta=0.4$ $\kappa=1,\tau=100$	$\alpha=\eta=0.8$ $\kappa=1,\tau=100$
异步分布式SVI	—53791.33	—55554.12	—54350.50	—56212.30	—57003.45	—57567.67
集中式SVI	—54198.30	—56327.50	—54776.18	—56721.87	—57805.78	—58191.39

言处理小组, 其由分属 20 个类别的 9 804 篇文档构成, 其中 20 个类别的标签分别为 Art、Literature、Education、Philosophy、History、Space、Energy、Electronics、Communication、Computer、Mine、Transport、Environment、Agriculture、Economy、Law、Medical、Military、Politics 和 Sports. 在预处理步骤中, 我们首先去除了文本中的数字和英文并用语言技术平台 (Language technology plantform, LTP) 的分词模型对文本进行分词处理. 为了减小训练的数据量, 我们只读取每个类别的前 100 篇文档进行训练. 图 8 展示了在 $K = 20$, $\alpha = 0.2$, $\eta = 0.2$, $\kappa = 0.5$, $\tau = 10$, 分布式算法 Batch size (批大小) 取 2, 集中式算法 batch size 取 100 的设置下, 异步分布式 SVI 和集中式 SVI 分别在复旦大学中文文本分类数据集上运行多次得到的 $\ln p(w)$ 的演化曲线, 可以看到异步分布式 SVI 收敛速度慢于集中式 SVI, 但是最终得到的 $\ln p(w)$ 值优于集中式 SVI.

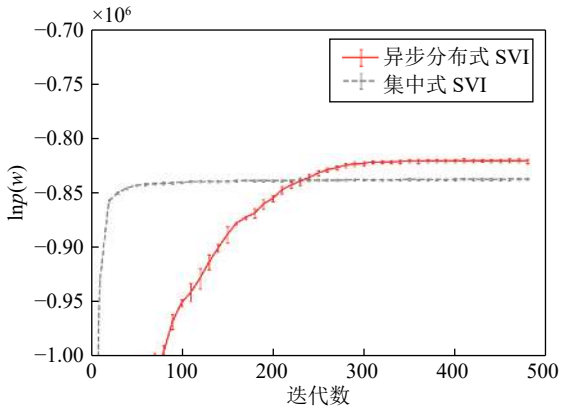


图 8 异步分布式 SVI 和集中式 SVI 在复旦大学中文文本分类数据集上的表现

Fig. 8 Performance of the asynchronous distributed SVI and the centralized SVI on the Chinese text classification data set of Fudan University

图 9 展示了在表 3 所示的超参数组合设置下异步分布式 SVI 和集中式 SVI 在复旦大学中文文本分类数据集上训练 100 个 epoch 得到的 $\ln p(w)$ 的值的对比, 其中横坐标为集中式 SVI 得到的 $\ln p(w)$ 的值, 纵坐标为对应超参数设置下异步分布式 SVI 得到的 $\ln p(w)$ 的值. 可以看到大部分数据点都位于左上方, 表明大部分情况下异步分布式 SVI 都优于集中式 SVI. 并且注意到当 batch size 取 1 时异步分布式 SVI 表现最差, 在 ($\kappa = 0.5, \tau = 1$, batchsize $= 1$) 的设置下其表现不如集中式 SVI. 我们认为这是由于当 batch size 太小时, 分布式 SVI 的收敛速度过慢造成的.

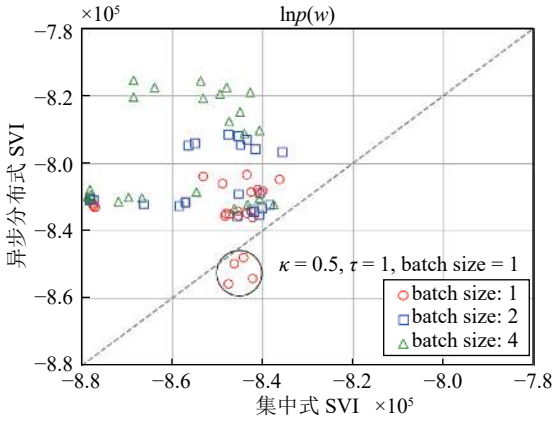


图 9 不同超参数设置下异步分布式 SVI 和集中式 SVI 在复旦大学中文文本分类数据集上表现

Fig. 9 Performance of the asynchronous distributed SVI and the centralized SVI on the Chinese text classification data set of Fudan University under different hyperparameter settings

表 3 超参数取值表

Table 3 The values of hyperparameters

κ	τ	batch size	α	η
0.5	1	1	0.1	0.1
1.0	10	2	0.2	0.2
—	100	4	—	—

4 结论

本文针对无中心的分布式网络, 基于扩散方法提出了一种新颖的分布式 SVI 算法, 其中采用自然梯度法进行本地更新以及采用对称双随机矩阵作为信息融合系数, 并且为其设计了一种针对异步网络的适应机制. 然后将其应用于 BMM 和 LDA 主题模型. 在不同数据集上的实验均表明所提出的算法确实适用于异步分布式网络, 而且其在发现聚类模式和对抗浅的局部最优方面的表现优于集中式 SVI 算法.

References

1 Qin J, Fu W, Gao H, Zheng W X. Distributed k-means algorithm and fuzzy c-means algorithm for sensor networks based on multiagent consensus theory. *IEEE Transactions on Cybernetics*, 2016, **47**(3): 772−783

2 Niu Jian-Jun, Deng Zhi-Dong, Li Chao. Distributed scheduling approaches in wireless sensor network. *Acta Automatica Sinica*, 2011, **37**(5): 517−528
(牛建军, 邓志东, 李超. 无线传感器网络分布式调度方法研究. 自动化学报, 2011, **37**(5): 517−528)

3 Hua J, Li C. Distributed variational Bayesian algorithms over sensor networks. *IEEE Transactions on Signal Processing*, 2015,

64(3): 783−798

4 Corbett J C, Jeffrey D, Michael E, et al. Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems*, 2013, **31**(3): 1−22

5 Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. *IEEE Access*, 2016, **4**: 1821−1834

6 Liu S, Pan S J, Ho Q. Distributed multi-task relationship learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada, 2017. 937−946

7 Wang S, Li C. Distributed robust optimization in networked system. *IEEE Transactions on Cybernetics*, 2016, **47**(8): 2321−2333

8 Hong M, Hajinezhad D, Zhao M M. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017. 1529−1538

9 Scardapane S, Fierimonte R, Di Lorenzo P, Panella M, Uncini A. Distributed semi-supervised support vector machines. *Neural Networks*, 2016, **80**: 43−52

10 Aybat N S, Hamedani E Y. A primal-dual method for conic constrained distributed optimization problems. In: Proceedings of Advances in Neural Information Processing Systems, Barcelona, Spain, 2016. 5049−5057

11 Anwar H, Zhu Q. ADMM-based networked stochastic variational inference [Online], available: https://arxiv.org/abs/1802.10168v1, February 27, 2018.

12 Qin J, Zhu Y, Fu W. Distributed clustering algorithm in sensor networks via normalized information measures. *IEEE Transactions on Signal Processing*, 2020, **68**: 3266−3279

13 Qu Han-Bing, Chen Xi, Wang Song-Tao, Yu Ming. Forward affine point set matching under variational Bayesian framework. *Acta Automatica Sinica*, 2015, **41**(8): 1482−1494
(曲寒冰, 陈曦, 王松涛, 于明. 基于变分贝叶斯逼近的前向仿射变换点集匹配方法研究. 自动化学报, 2015, **41**(8): 1482−1494)

14 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003, **3**(Jan): 993−1022

15 Hoffman M D, Blei D M, Wang C, Paisley J. Stochastic variational inference. *The Journal of Machine Learning Research*, 2013, **14**(1): 1303−1347

16 Corduneanu A, Bishop C M. Variational Bayesian model selection for mixture distributions. In: Proceedings of Artificial intelligence and Statistics, Waltham, MA: Morgan Kaufmann, 2001. 27−34

17 Letham B, Letham L M, Rudin C. Bayesian inference of arrival rate and substitution behavior from sales transaction data with stockouts. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2016. 1695−1704

18 Safarinejadian B, Menhaj M B. Distributed density estimation in sensor networks based on variational approximations. *International Journal of Systems Science*, 2011, **42**(9): 1445−1457

19 Safarinejadian B, Estahbanati M E. A novel distributed variational approximation method for density estimation in sensor networks. *Measurement*, 2016, **89**: 78−86

20 Mozaffari M, Safarinejadian B. Mobile-agent-based distributed variational Bayesian algorithm for density estimation in sensor networks. *IET Science, Measurement and Technology*, 2017, **11**(7): 861−870

21 Zhang Yu, Bao Yan-Ke, Shao Liang-Shan, Liu Wei. A multivariate decision tree for big data classification of distributed data streams. *Acta Automatica Sinica*, 2018, **44**(6): 1115−1127
(张宇, 包研科, 邵良杉, 刘威. 面向分布式数据流大数据分类的多变量决策树. 自动化学报, 2018, **44**(6): 1115−1127)

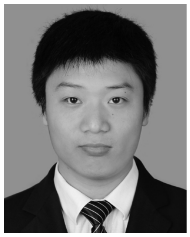
22 Mohamad S, Bouchachia A, Sayed-Mouchaweh M. Asynchronous stochastic variational inference. In: Proceedings of INNS Big Data and Deep Learning Conference, Sestri Levante, Italy: Springer, 2019. 296−308

23 Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 2010, **3**(1): 1−122

24 Chen J, Sayed A H. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 2012, **60**(8): 4289−4305

25 Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008, **1**(1−2): 1−305

26 Wright S J. Coordinate descent algorithms. *Mathematical Programming*, 2015, **151**(1): 3−3



付维明 中国科学技术大学自动化系博士后. 主要研究方向为多智能体系统一致与物理信息系统安全.
E-mail: fwm1993@ustc.edu.cn
(**FU Wei-Ming** Postdoctor in the Department of Automation, University of Science and Technology of China. His research interest covers consensus in multiagent systems and security in cyber-physical systems.)

秦家虎 中国科学技术大学自动化系教授. 主要研究方向为多智能体系统, 复杂动态网络以及物理信息系统. 本文通信作者.
E-mail: jhqin@ustc.edu.cn
(**QIN Jia-Hu** Professor in the Department of Automation, University of Science and Technology of China. His research interest covers multiagent systems, complex dynamical networks, and cyber-physical systems. Corresponding author of this paper.)

朱英达 中国科学技术大学自动化系硕士. 主要研究方向为数据挖掘和分布式算法.
E-mail: xy131237@mail.ustc.edu.cn
(**ZHU Ying-Da** Master student in the Department of Automation, University of Science and Technology of China. His research interest covers data mining and distributed algorithms.)