DOI:10.7544/issn1000-1239.2015.20140107
52(1):16-26,2015

贝叶斯机器学习前沿进展综述

朱 军 胡文波
(智能技术与系统国家重点实验室(清华大学) 北京 100084)
(清华信息科学与技术国家实验室(筹) 北京 100084)
(清华大学计算机科学技术系 北京 100084)
(dcszj@mail.tsinghua.edu.cn)

Recent Advances in Bayesian Machine Learning

Zhu Jun and Hu Wenbo
(*Sate Key Laboratory of Intelligent Technology and Systems* (*Tsinghua University*), *Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology*, *Beijing* 100084)
(*Department of Computer Science and Technology*, *Tsinghua University*, *Beijing* 100084)

Abstract With the fast growth of big data, statistical machine learning has attracted tremendous attention from both industry and academia, with many successful applications in vision, speech, natural language, and biology. In particular, the last decades have seen the fast development of Bayesian machine learning, which is now representing a very important class of techniques. In this article, we provide an overview of the recent advances in Bayesian machine learning, including the basics of Bayesian machine learning theory and methods, nonparametric Bayesian methods and inference algorithms, and regularized Bayesian inference. Finally, we also highlight the challenges and recent progress on large-scale Bayesian learning for big data, and discuss on some future directions.

Key words Bayesian machine learning; nonparametric methods; regularized methods; learning with big data; big Bayesian learning

摘 要 随着大数据的快速发展,以概率统计为基础的机器学习在近年来受到工业界和学术界的极大关注,并在视觉、语音、自然语言、生物等领域获得很多重要的成功应用,其中贝叶斯方法在过去20多年也得到了快速发展,成为非常重要的一类机器学习方法.总结了贝叶斯方法在机器学习中的最新进展,具体内容包括贝叶斯机器学习的基础理论与方法、非参数贝叶斯方法及常用的推理方法、正则化贝叶斯方法等.最后,还针对大规模贝叶斯学习问题进行了简要的介绍和展望,对其发展趋势作了总结和展望.

关键词 贝叶斯机器学习;非参数方法;正则化方法;大数据学习;大数据贝叶斯学习

中图法分类号 TP181

机器学习是人工智能及模式识别领域的共同研究热点,其理论和方法已被广泛应用于解决工程应用和科学领域的复杂问题.2010年的图灵奖获得者为哈佛大学的Leslie Valliant教授,其获奖工作之一是建立了概率近似正确(probably approximate correct, PAC)学习理论;2011年的图灵奖获得者为加州大学洛杉矶分校的Judea Pearl教授,其主要贡献为建立了以概率统计为理论基础的人工智能方法,其研究成果促进了机器学习的发展和繁荣.

机器学习的一个重要分支是贝叶斯机器学习.

收稿日期:2014-10-13;**修回日期**:2014-11-22
基金项目:国家"九七三"重点基础研究发展计划基金项目(2013CB329403,2012CB316301);国家自然科学基金项目(61322308,61332007)

贝叶斯方法最早起源于英国数学家托马斯·贝叶斯在1763年所证明的一个关于贝叶斯定理的一个特例[1].经过多位统计学家的共同努力,贝叶斯统计在20世纪50年代之后逐步建立起来,成为统计学中一个重要的组成部分[2-3].贝叶斯定理因为其对于概率的主观置信程度[4]的独特理解而闻名.此后由于贝叶斯统计在后验推理、参数估计、模型检测、隐变量概率模型等诸多统计机器学习领域方面有广泛而深远的应用[5-6].从1763年到现在已有250多年的历史,这期间贝叶斯统计方法有了长足的进步[7].在21世纪的今天,各种知识融会贯通,贝叶斯机器学习领域将有更广阔的应用场景,将发挥更大的作用.

1　贝叶斯学习基础

本节将对贝叶斯统计方法进行简要的介绍[5]:主要包括贝叶斯定理、贝叶斯模型的推理方法、贝叶斯统计学的一些经典概念.

1.1　贝叶斯定理

用Θ表示概率模型的参数,D表示给定的数据集.在给定模型的先验分布$p_0(\Theta)$和似然函数$p(D|\Theta)$的情况下,模型的后验分布可以由贝叶斯定理(也称贝叶斯公式)获得[2]:

$$p(\Theta \mid D) = \frac{p_0(\Theta)p(D \mid \Theta)}{p(D)}, \tag{1}$$

其中$p(D)$是模型的边缘似然函数.

贝叶斯定理已经广为人知,这里介绍一种与贝叶斯公式等价但很少被人知道的表现形式,即基于优化的变分推理:

$$\min_{q(\Theta)\in P} \mathrm{KL}(q(\Theta) \parallel p_0(\Theta)) - \mathbb{E}_q[\log\ p(D \mid \Theta)], \tag{2}$$

其中P为归一化的概率分布空间.可以证明,式(2)中的变分优化的最优解等价于式(1)中的后验推理的结果[8].这种变分形式的贝叶斯定理具有两方面的重要意义:1) 它为变分贝叶斯方法[9](variational Bayes)提供了理论基础;2) 提供了一个很好的框架以便于引用后验约束,丰富贝叶斯模型的灵活性[10].这两点在后面的章节中将具体阐述.

1.2　贝叶斯机器学习

贝叶斯方法在机器学习领域有诸多应用,从单变量的分类与回归到多变量的结构化输出预测、从有监督学习到无监督及半监督学习等,贝叶斯方法几乎用于任何一种学习任务.下面简要介绍较为基础的共性任务.

1) 预测.给定训练数据D,通过贝叶斯方法得到对未来数据x的预测[5]:

$$p(x \mid D) = \int_{\Theta} p(x,\Theta \mid D) = p(x \mid \Theta,D)p(\Theta \mid D). \tag{3}$$

需要指出的是,当模型给定时,数据是来自于独立同分布的抽样,所以$p(x|\Theta,D)$通常简化为$p(x|\Theta)$.

2) 模型选择.另一种很重要的贝叶斯方法的应用是模型选择[11],它是统计和机器学习领域一个较为基础的问题.用M表示一族模型(如线性模型),其中每个元素Θ是一个具体的模型.贝叶斯模型选择通过比较不同族模型的似然函数来选取最优的:

$$p(D \mid M) = \int_{\Theta} p(D \mid \Theta)p(\Theta \mid M). \tag{4}$$

当没有明显先验分布的情况下,$p(\Theta|M)$被认为是均匀分布.通过式(4)的积分运算,贝叶斯模型选择可以避免过拟合.

关于贝叶斯统计和贝叶斯学习更为详细的内容,有些论文和教材有更进一步的说明[2,5,10,12].

2　非参数贝叶斯方法

在经典的参数化模型中模型的参数个数是固定的,不会随着数据的变化而变化.以无监督的聚类模型为例,如果能通过数据本身自动学习得到聚类中心的个数,比参数化模型(如K均值、高斯混合模型等)根据经验设定一个参数要好得多;这也是非参数模型一个较为重要的优势.相比较参数化贝叶斯方法,非参数贝叶斯方法(nonparametric Bayesian methods)因为其先验分布的非参数特性,具有描述数据能力强的优点[13],非参数贝叶斯方法因此在2000年以后受到较多关注[14].例如具有未知维度的隐式混合模型[15]和隐式特征模型[16]、描述连续函数的高斯过程[17]等.需要强调的是非参数化贝叶斯方法并不是指模型没有参数,而是指模型可以具有无穷多个参数,并且参数的个数可以随着数据的变化而自适应变化,这种特性对于解决大数据环境下的复杂应用问题尤其重要,因为大数据的特点之一是动态多变.下面将主要针对其中的一些较为重要的模型和推理方法进行简要介绍.

2.1　狄利克雷过程

狄利克雷过程(Dirichlet process, DP)是统计学家Ferguson于1973年提出的一个定义在概率测度Ω上的随机过程[18],其参数有集中参数$\alpha>0$和基底

概率分布 G_0,通常记为 $G\sim DP(\alpha,G_0)$. 狄利克雷过程得到的概率分布是离散型的,因此非常适合构建混合模型,例如,Antoniak 于 1974 年通过给每个数据点增加一个生成概率,构造了一个狄利克雷过程混合模型(Dirichlet process mixture, DPM)[15],即

$$x_i \sim p(x \mid \theta_i), \tag{5}$$

其中,$\theta_i\sim G, i\in[N]$是生成每个数据点概率分布的参数,比如高斯分布的均值和协方差等,N 为数据点的个数.

与狄利克雷过程等价的一个随机过程是中国餐馆过程(Chinese restaurant process, CRP)[19]. 中国餐馆过程是定义在实数域上的具有聚类特性的一类随机过程,也因为其特有的较好展示特性而被经常使用. 如图 1 所示,在中国餐馆过程中,假设有无限张餐桌和若干客人;其中第 1 名顾客选择第 1 张餐桌,之后的顾客按照多项式分布选择餐桌,其中选择每张餐桌的概率正比于该餐桌现在所坐的人数,同时以一定概率(正比于参数 α)选择一个没人的餐桌. 可以看到,当所有的客人选择完毕餐桌,我们可以按照餐桌来对客人进行一个划分. 这里,每张餐桌代表一个聚类,每个客人代表一个数据点.

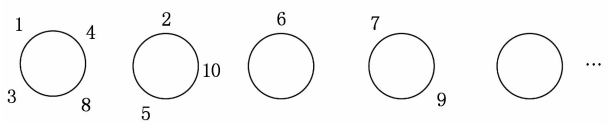


Fig. 1 Illustration of the formation of Chinese restaurant process[16].

图 1 中国餐馆过程的生成过程[16]

可以证明所有的聚类点参数 θ 可以通过式(6)得到:

$$p(\theta_1,\cdots,\theta_N \mid \alpha,G_0) = \int \left(\sum_{i=1}^{N} p(\theta_i \mid G)\right) \mathrm{d}P(G \mid \alpha,G_0), \tag{6}$$

将狄利克雷混合模型中的 G 积分即可得到中国餐馆过程,这也说明了两个随机过程的关系. 这种简洁的表述也很有利于马尔可夫蒙特卡洛方法的采样[20].

另一种构造性的狄利克雷过程的表述是截棍过程(stick breaking construction)[21]. 具体地说,将一根单位长度的棍,第 k 次切割都按照剩下的长度按照贝塔分布的随机变量,按比例切割:

$$\beta_k \sim Beta(1,\alpha),\ \pi_k = \beta_k \prod_{j=1}^{k-1}(1-\beta_j), \tag{7}$$

即如图 2 所示,对于一根长度为单位 1 的棍,第 1 次切割 β_1 长度,以后每次切割都切割剩下部分的 β_k 比例长度. 狄利克雷过程的截棍表述是变分推理的基础[22].

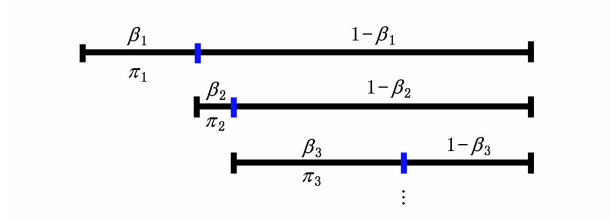


Fig. 2 Illustration of stick breaking construction[23].

图 2 截棍过程示意图[23]

2.2 印度自助餐过程

与混合模型中每一个数据点只属于一个聚类不同,在特征模型中每一个数据点可以拥有多个特征,这些特征构成了数据生成的过程. 这也符合实际情况中样本数据点有多个属性的实际需求. 经典的特征模型主要有因子分析(factor analysis)、主成分分析(principal component analysis)[24-25]等. 在传统的特征模型中,特征的数目是确定的,这给模型的性能带来一定限制. 印度自助餐过程(indian buffet process, IBP)是 2005 年提出的[26],因其非参数特性能从数据中学习得到模型中的特征个数,使得模型能够更好地解释数据,已经在因子分析、社交网络链接预测等重要问题中应用[27-29].

以二值("0"或"1")特征为例,假设有 N 个数据点,所有数据点的特征向量组成一个特征矩阵,IBP 的产生式过程可以形象地类比为 N 个顾客到一个无穷多个餐品的自助餐馆进行选餐的过程,用"1"表示选择,"0"表示不选择,具体描述如图 3 所示的方法进行:

1) 第 1 名顾客选择 K_1 个餐品,其中 $K_1\sim Possion(\alpha)$;

2) 第 2 名及以后的顾客有两种情况:

① 对于已经被选过的餐品,按照选择该餐品的人数成正比的概率选择该餐品;

② 选择 K_i 个未被选过的餐品,其中 $K_i\sim Possion\left(\frac{\alpha}{n}\right)$.

与中国餐馆过程类似,印度自助餐过程也有其对应的截棍过程[30]. 这里不再赘述,仅列出其构造性表述如下:

$$\nu_j \sim Beta(\alpha,1),\ \pi_k = \prod_{j=1}^{k}\nu_j. \tag{8}$$

但是,与中国餐馆过程的截棍过程不同的是棍的长度之和并不为 1. 印度自助餐过程也有其对应的采样方法和变分优化求解方法[16,30-31].

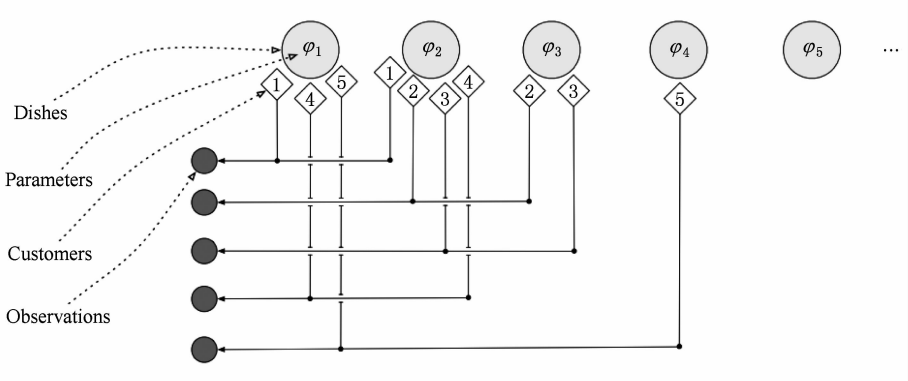


Fig. 3　Illustration of indian buffet process[13].

图 3　印度自助餐过程示意图[13]

2.3　应用及扩展

贝叶斯方法特别是最近流行的非参数贝叶斯方法已广泛应用于机器学习的各个领域,并且收到了很好的效果[32].这里简要提出几点应用和扩展;对于大规模贝叶斯学习的相关应用将在第 5 节介绍,也可查阅相关文献[13-14,33].

经典的非参数化贝叶斯方法通常假设数据具有简单的性质,如可交换性或者条件独立等;但是,现实世界中的数据往往具有不同的结构及依赖关系.为了适应不同的需求,发展具有各种依赖特性的随机过程得到了广泛关注.例如,在对文本数据进行主题挖掘时,数据往往来自不同的领域或者类型,我们通常希望所学习的主题具有某种层次结构,为此,层次狄雷克利过程(hierarchical Dirichlet process, HDP)[34]被提出,可以自动学习多层的主题表示,并且自动确定主题的个数.另外,具有多个层次的 IBP 过程也被提出[35],并用于学习深层置信网络的结构,包括神经元的层数、每层神经元的个数、层间神经元的连接结构等.其他的例子还包括具有马尔可夫动态依赖关系的无限隐马尔可夫模型[36]、具有空间依赖关系的狄雷克利过程[37]等.

另外,对于有监督学习问题,非参数贝叶斯模型最近也受到了广泛的关注.例如,社交网络数据建模和预测是一个重要的问题,近期提出的基于 IBP 的非参数化贝叶斯模型[27,29]可以自动学习隐含特征,并且确定特征的个数,取得很好的预测性能.使用 DP 混合模型同时作聚类和分类任务也取得了很好的结果[38].

3　贝叶斯模型的推理方法

贝叶斯模型的推理方法是贝叶斯学习中重要的一环,推理方法的好坏直接影响模型的性能.具体地说,贝叶斯模型的一个关键性的问题是后验分布通常是不可解的,使得式(3)和式(4)中的贝叶斯积分也是不可解的.这时,就需要一些有效的推理方法.一般而言,主要有两类方法:变分推理方法(variational inference)和蒙特卡洛方法(Monte Carlo methods).这两类方法都在贝叶斯学习领域有广泛的应用,下面分别介绍这两类方法.

3.1　变分推理方法

变分法是一种应用较广的近似优化方法[39-40],在物理、统计学、金融分析、控制科学领域解决了很多问题.在机器学习领域,变分方法也有较多应用:通过变分分析,可以将非优化问题转化成优化问题求解,也可以通过近似方法对一些较难的问题进行变分求解[41].

在变分贝叶斯方法中,给定数据集 D 和待求解的后验分布 $p(\Theta|D)$,变分方法界定其后验分布的近似分布为 $q(\Theta)$.运用杰森不等式,可以得到对数似然的一个下界(evidence lower bound, ELBO).

$$\log p(D) \geqslant \mathbb{E}_q[\log(p(\Theta,D)] - \mathbb{E}_q[\log(q(\Theta))]. \tag{9}$$

通过最大化该对数似然下界:

$$\max_q \mathbb{E}_q[\log(p(\Theta,D)] - \mathbb{E}_q[\log(q(\Theta))]. \tag{10}$$

或者最小化 $q(\Theta)$ 和 $p(\Theta|D)$ 之间的 KL 散度,就可以完成优化求解的过程.因此,变分推理的基本思想是将原问题转化成求解近似分布 $q(\Theta)$ 的优化问题,结合有效的优化算法来完成贝叶斯推理的任务[22,42-43].

很多时候,模型 Θ 中往往有一些参数 θ 和隐变量 h.这时变分问题可以通过变分期望最大化方法求解(variational EM algorithm):通过引入平均场假设(mean-field assumption)$q(\theta,h)=q(\theta)q(h)$,可以迭代进行 EM 算法[44].

3.2 蒙特卡洛方法

蒙特卡洛方法是一类通过利用模拟随机数对未知的概率分布进行估计;当未知分布很难直接估计或者搜索空间太大、计算太复杂时,蒙特卡洛方法就成为重要的推理和计算方法[45-46]. 例如,贝叶斯机器学习通常需要计算某个函数在某种分布(先验或者后验)下的期望,而这种计算通常是没有解析解的. 假设 $p(\Theta)$是一个概率分布,目标是计算如下积分:

$$I \triangleq \int \phi(\Theta) p(\Theta) \mathrm{d}\Theta. \tag{11}$$

蒙特卡洛方法的基本思想是使用如下估计来近似 I:

$$\hat{I}_{MC} \triangleq \frac{1}{N} \sum_{i=1}^{N} \phi(\Theta^i), \tag{12}$$

其中 Θ^i 是从 p 中得到的采样. 根据大数定律,在采样数目足够多时,蒙特卡洛方法可以很好地估计真实期望.

上面描述的是蒙特卡洛方法的基本原理,但实际过程中 p 的采样并不是很容易就可以得到,往往采用其他的方法进行,常用的方法有重要性采样(importance sampling)、拒绝采样(rejection sampling)、马尔可夫蒙特卡洛方法(Markov Chain Monte Carlo, MCMC)等. 前两者在分布相对简单时比较有效,但是对于较高维空间的复杂分布效果往往不好,面临着维数灾难的问题. 下面重点介绍 MCMC 方法,它在高维空间中也比较有效.

MCMC 方法的基本思想是构造一个随机的马尔可夫链,使得其收敛到指定的概率分布,从而达到推理的目的[47]. 一种较为常用的 MCMC 方法是 Metropolis-Hastings 算法[48](MH 算法). 在 MH 算法中,通过构造一个从 Θ_t 状态到 Θ_{t+1} 状态的转移规则:

1) 根据 $q(\Theta|\Theta_t)$从旧的状态采样中得到一个新的状态采样;

2) 计算接受概率:

$$A(\Theta,\Theta_t) \triangleq \min\left(1, \frac{\tilde{p}(\Theta') q(\Theta_t \mid \Theta')}{\tilde{p}(\Theta_t) q(\Theta' \mid \Theta_t)}\right); \tag{13}$$

3) 从 0-1 均匀分布中采样得到 $\gamma \sim Uniform[0,1]$. 若 $\gamma < A(\Theta,\Theta_t)$,则接受采样 $\Theta_{t+1} \leftarrow \Theta'$,否则拒绝采样 $\Theta_{t+1} \leftarrow \Theta_t$.

另一种常用的 MCMC 方法是吉布斯采样(Gibbs sampling)[46,49],它是 MH 算法的一种特例,吉布斯采样已广泛应用在贝叶斯分析的推理中. 吉布斯采用是对多变量分布中每一个变量在其他已经观察得到采样的变量已知的条件下依次采样,更新现有的参数,最后收敛得到目标后验分布. 假设需要采样的多元分布为 $p(\theta_1,\theta_2,\cdots,\theta_d)$,即每次选出一个维度 j:$1 \leqslant j \leqslant d$,其中 d 是多元分布的维度;随后从条件概率分布 $p(\theta_j|\theta_1,\cdots,\theta_{j-1},\theta_{j+1},\cdots,\theta_d)$对 θ_j 进行采样.

有很多贝叶斯模型都采用了 MCMC 的方法进行推理,取得了很好的效果[20,30,50]. 除此之外,还有一类非随机游走的 MCMC 方法——Langevin MCMC[51] 和 Hybrid Monte Carlo[52]. 这一类方法往往有更快的收敛速度,但是表述的复杂程度较大,因此受欢迎程度不及吉布斯采样,但是,最近在大数据环境下发展的基于随机梯度的采样方法非常有效,后文将会简要介绍.

4 正则化贝叶斯理论及应用举例

在第 2 节中提到了贝叶斯方法的两种等价表现方式,一种是后验推理的方式,另一种是基于变分分析的优化方法,其中第 2 种方式在近年有了较大发展. 基于这种等价关系,我们近年来提出了正则化贝叶斯(regularized Bayesian inference, RegBayes)理论[10]:如图 4 所示,在经典贝叶斯推理过程中,后验分布只能从两个维度来获得,即先验分布和似然函数;而在正则化贝叶斯推理中,后验推理转化成一种变分优化的方式,通过引入后验正则化,为贝叶斯推理提供了第 3 维自由度,极大地丰富了贝叶斯模型的灵活性. 在 RegBayes 理论的指导下,我们系统研究了基于最大间隔准则的判别式贝叶斯学习以及结合领域知识的贝叶斯学习等,取得了一系列的成果[10,53-55].

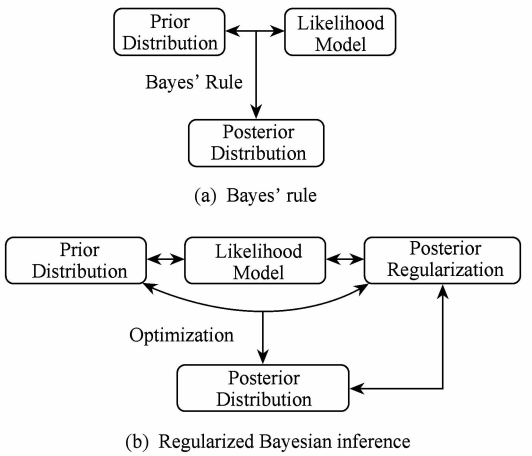


Fig. 4 Two different schemes of Bayesian inference.

图 4 两种不同的贝叶斯推理的方式

正则化贝叶斯推理的基本框架可以简述如下,在式(2)的基础上,引入后验正则化项,考虑领域知识或者期望的模型属性:

$$\inf_{q(\Theta)\in} \mathrm{KL}(q(\Theta)\,\|\,\pi(\Theta)) - \mathbb{E}_q[p(D\mid\Theta)q(\Theta)] + \Omega(q(\Theta)), \quad (14)$$

其中 $\Omega(q(\Theta))$是一个凸函数. 在运用 RegBayes 解决具体问题时需要回答下面 3 个问题:

问题 1. 后验正则化从何而来. 后验正则化是一个通用的概念,可以涵盖任何期望影响后验分布的信息. 比如,在有监督学习任务(如图像/文本分类)中,我们期望后验分布能够准确地预测,这种情况下我们可以将分类错误率(或者某种上界)作为优化目标,通过后验正则化引用到学习过程中,典型的例子包括无限支持向量机[38](infinite SVM)、无限隐式支持向量机[56](infinite latent SVM)、最大间隔话题模型[57](maximum margin supervised topic model, MedLDA)等,这些方法均采用了最大间隔原理,在贝叶斯学习过程中直接最小化分类错误率的上界(即铰链损失函数),在测试数据上取得显著的性能提升.

另外,在一些学习任务中,一些领域知识(如专家知识或者通过众包方式收集到的大众知识)可以提供数据之外的一些信息,对提高模型性能有很大帮助. 在这种情况下,可以将领域知识作为后验约束,与数据一起加入模型中,实现高效贝叶斯学习. 需要指出的是大众知识往往存在很大的噪音,如何采取有效的策略过滤噪音实现有效学习是问题的关键. 在这方面,我们提出了将使用逻辑表达的领域知识鲁棒地引入贝叶斯主题模型,实现了更优秀的模型效果[58].

问题 2. 先验分布、似然函数以及后验正则化之间有何关系. 先验分布是与数据无关的,基于先验知识的概率分布不能反映数据的统计特性;似然函数则是基于数据产生的概率分布,反映了数据的基本性质,通常定义为具有良好解析形式的归一化的概率分布. 而后验正则化项同样是利用数据的特性来定义的,但是,它具有更广泛灵活的方式,不受归一化的约束,因此,可以更方便准确地刻画问题的属性或者领域知识,如问题 1 中所举的最大间隔学习以及领域知识与贝叶斯统计相结合等示例. 甚至可以证明,一些后验分布不可以通过贝叶斯定理得到,但是可以通过后验正则化得到[10]. 因此,RegBayes 是比经典贝叶斯方法更灵活更强大的方法.

问题 3. 如何求解优化问题. 虽然正则化贝叶斯具有极强的灵活性,其学习算法仍然可以使用变分方法或者蒙特卡洛方法进行求解,具体的求解方法请阅读相关论文. 下面介绍的大数据贝叶斯学习理论和算法均可以应用到快速求解正则化贝叶斯模型[55],这也是目前的研究热点.

5　大数据贝叶斯学习

随着互联网技术的发展,研究面向大数据的机器学习理论、算法及应用成为当前研究的热点[59],得到学术界和工业界的广泛关注. 贝叶斯模型有较好的数据适应性和可扩展性,在很多经典问题上都取得了很好的效果,但是,传统贝叶斯模型的一个较大的问题在于其推理方法通常较慢,特别是在大数据背景下很难适应新的模型的要求. 因此,如何进行大规模贝叶斯学习方法是学术界的重要挑战之一. 可喜的是近期在大数据贝叶斯学习(big Bayesian learning, BigBayes)方面取得了显著的进展. 下面简单介绍在随机算法及分布式算法方面的进展,并以我们的部分研究成果作为示例. 表 1 所示为对目前的若干前沿进展简要总结:

Table 1　The Summary of the Recent Methods for BigBayes

表 1　大规模贝叶斯学习的前沿进展总结

Methods	Application Examples	References
Stochastic Learning and Online Learning	SGLD, SHMC, Online BayesPA	Ref [55, 59, 61-62, 65-66]
Distributed Learning	gCTM	Ref [67-69, 72]
Hardware Acceleration	Parallel Inference on GPU	Ref [75-78]

5.1　随机梯度及在线学习方法

当数据量较大时精确的算法往往耗时较长,不能满足需要. 一类常用的解决方案是采用随机近似算法[60-61]. 这类算法通过对大规模数据集的多次随机采样(random subsampling),可以在较快的时间内收敛到较好的结果. 这种思想已经在变分推理和蒙特卡洛算法中广泛采用,简要介绍如下.

在变分推理方面,如前所述,其核心是求解优化问题,因此,基于多次随机降采样的随机梯度下降算法成为很自然的选择. 具体地说,随机梯度下降算法(stochastic gradient descent, SGD)[62]每次随机选取一个数据子集,并用该子集上计算的梯度估计整个数据集上的梯度,对要求解的参数进行更新:

$$\omega_{t+1} = \omega_t - \gamma_t \nabla_\omega Q(z_t, \omega_t), \quad (15)$$

其中 Q 是待优化的目标函数，z_t 是数据的第 t 个子集. 值得注意的是，欧氏空间中的梯度并非最优的求解变分分布的方向；对于概率分布的寻优，自然梯度往往取得更快的收敛速度[63]. 近期的主要进展包括随机变分贝叶斯方法[61]以及多种利用模型特性的快速改进算法[64].

在蒙特卡洛算法方面，可以将随机梯度的方法用于改进对应的基于梯度的采样算法，如随机梯度朗之万动力学采样方法(stochastic gradient langevin dynamics, SGLD)[65]、随机梯度哈密尔顿蒙特卡洛(stochastic Hamiltonian Monte Carlo, SHMC)[66]. 这些算法加快了蒙特卡洛采样的速度、有较好的效果.

例 1. 为了适应动态流数据的处理需求，基于在线学习的大规模贝叶斯推理算法也成为近期的研究热点，主要工作包括流数据变分贝叶斯[67]等. 我们近期提出了在线贝叶斯最大间隔学习(online Bayesian passive-aggressive learning, Online BayesPA)框架，显著提高了正则化贝叶斯的学习效率，并且给出了在线学习后悔值的理论界[55]. 在 100 多万的维基百科页面数据上的部分实验结果如图 5 所示，可以看出，基于在线学习的算法比批处理算法快 100 倍左右，并且不损失分类的准确率.

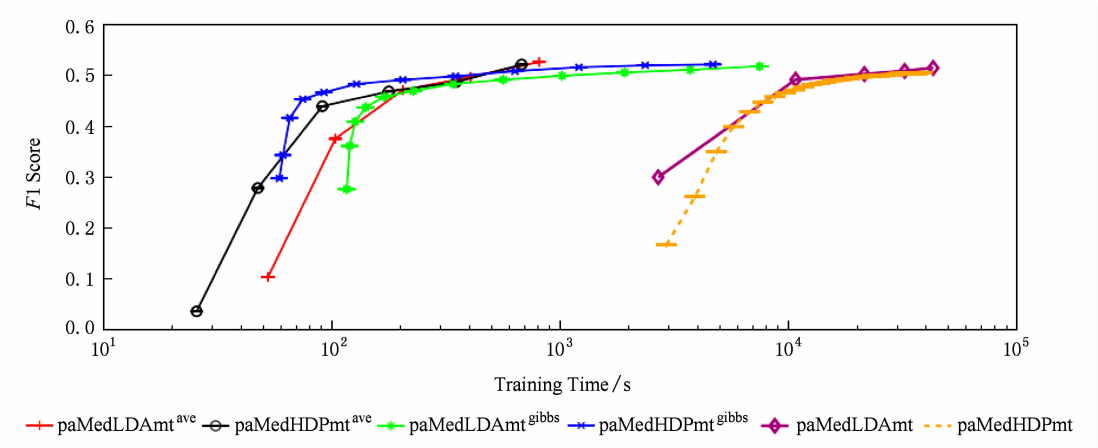


Fig. 5 Comparison between the online BayesPA methods and batch methods[55].

图 5 基于 BayesPA 的在线学习算法与基于批处理算法的比较[55]

5.2 分布式推理算法

另一种适用于大规模贝叶斯学习问题的算法是基于分布式计算的[68]，即部署在分布式系统上的贝叶斯推理算法. 这类算法需要仔细考虑算法的实际应用场景，综合考量算法计算和通信的开销，设计适合于不同分布式系统的推理算法.

一些算法中的部分参数之间不需要交换信息，只需要计算得到最后结果汇总即可；对于这类问题，只需要对原算法进行适当优化，部署在系统上即可有较好的效果. 但是，还有更多算法本身并不适合并行化处理，这就意味着算法本身需要修改，使得其可以进行分布式计算，这也是大规模贝叶斯学习的研究热点之一，并且已经取得很多重要进展，包括分布式变分推理[67]和分布式蒙特卡洛方法[69]等.

例 2. 以主题模型为例，经典的模型使用共轭狄利克雷先验，可以学习大规模的主题结构[70]，但是，不能学习主题之间的关联关系. 为此，使用非共轭 Logistic-Normal 先验的关联主题模型(correlated topic model, CTM)[71]被提出. CTM 的缺点是其推理算法比较困难，已有的算法只能处理几十个主题的图结构学习. 为此，笔者课题组近期提出了 CTM 的分布式推理算法[72]，可以处理大规模的数据集，学习上千个主题之间的图结构. 该算法的部分结果如表 2 所示，其中 D 表示数据集大小，K 表示主题个数. 由表 2 可以看出分布式推理算法(即 gCTM)极大地提高了模型可以承载的数据量(如 600 万的维基百科网页)和更多的主题个数(如1 000). 这个项目的代码及更多信息已经公布，读者可以自行浏览[73].

在上述大规模主题图结构的学习基础上，进一步开发了“主题全景图”(TopicPanorama)可视化界面，它可以将多个主题图结构进行融合，并且以用户友好的方式展现在同一个界面上，如图 6 所示，其中每个节点代表一个主题，节点之间的边代表相关联关系，边的长度代表关联强度，所用数据集为微软、谷歌、雅虎等 3 个 IT 公司相关的新闻网页. 该可视化工具具有多种交互功能，用户可以使用放大或缩小功能对主题图的局部进行仔细查看，同时，也可以修改图的结构并反馈给后台算法进行在线调整. 多位领域专家一致同意该工具可以方便分析社交媒体数据. 更多具体描述参见文献[74].

Table 2　Contrast of the Efficiency Between gCTM and vCTM[72]

表 2　vCTM 和 gCTM 之间的效率比较[72]

Datasets	$10^{-3}\times D$	K	vCTM/min	gCTM/min
NIPS	1.2	100	114	8.9
20NG	11	200	960	9
NYTimes	285	400	N/A*	30
Wiki	6 000	1 000	N/A*	1 020

Note: * not finished within 1 week.

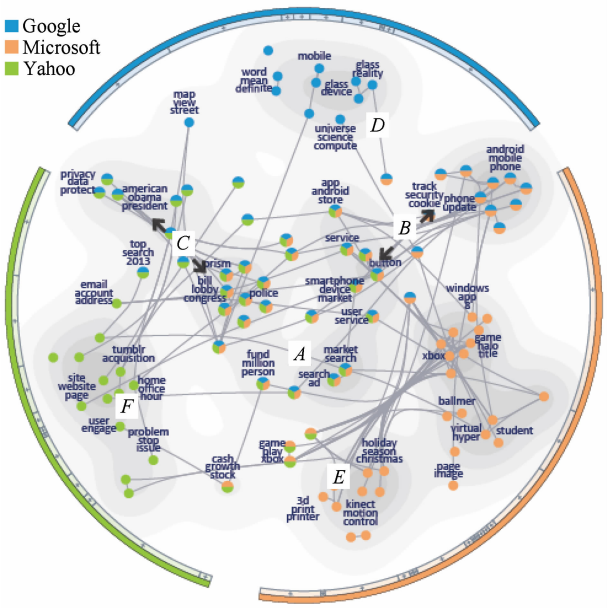


Fig. 6　A Demonstration of the Results of Correlational Topic Model[74].

图 6　关联话题模型结果展示[74]

5.3　基于硬件的加速

随着硬件的发展,使用图形处理器(graphics processing units, GPU)、现场可编程逻辑门阵列(field-programmable gate array, FPGA)等硬件资源对贝叶斯学习方法进行加速也是最近兴起的研究热点.例如,有研究者利用 GPU 技术对话题模型的变分方法[75]和 MCMC 算法[76-77]进行加速,还有一些研究者利用 FPGA 对蒙特卡洛算法[78]进行加速.利用强大的硬件设备,搭配适当的模型和算法架构,可以起到事半功倍的效果.

6　总结与展望

贝叶斯统计方法及其在机器学习领域的应用是贝叶斯学习的重要研究内容.因为贝叶斯理论的适应性和可扩展性使得贝叶斯学习得到广泛的应用.非参数贝叶斯方法和正则化贝叶斯方法极大地发展了贝叶斯理论,使其拥有更加强大的生命力.

近年来,大数据贝叶斯学习成为人们关注的焦点,如何加强贝叶斯学习的灵活性以及如何加快贝叶斯学习的推理过程,使其更加适应大数据时代的挑战成为人们考虑的问题.在这一时期许多新的方法和理论将被提出,贝叶斯学习也与其他许多方面的知识相结合,如并行计算、数据科学等,产生很多新的成果.可以预想,贝叶斯学习肯定会有更多更新更好的成果,也会在将来有更广泛的应用.

参　考　文　献

[1] Bayes T. An essay towards solving a problem in the doctrine of chances [J]. London: Philosophical Transactions Royal Society, 1763, 53: 370-418

[2] Mao Shisong. Bayesian Statistics [M]. Beijing: Chinese Statistics Press, 1999 (in Chinese)
(茆诗松.贝叶斯统计[M]. 北京:中国统计出版社, 1999)

[3] Lee P M. Bayesian statistics: An Introduction [M]. New York: John Wiley & Sons, 2012

[4] Savage L J. The Foundations of Statistics [M]. New York: Courier Dover Publications, 1972

[5] Gelman A, Carlin J, Stern H, et al. Bayesian Data Analysis [M]. Boca Raton: CRC Press, 2013

[6] Barber D. Bayesian Reasoning and Machine Learning [M]. Cambridge: Cambridge University Press, 2012

[7] Efron B. Bayes' theorem in the 21st century [J]. Science, 2013, 340(6137): 1177-1178

[8] Zellner A. Optimal information processing and Bayes's theorem [J]. The American Statistician, 1988, 42(4): 278-280

[9] Attias H. Inferring parameters and structure of latent variable models by variational Bayes [C] //Proc of the 15th Conf on Uncertainty in Artificial Intelligence. Stockholm, Sweden: AUAI, 1999: 21-30

[10] Zhu J, Chen N, Xing E P. Bayesian inference with posterior regularization and applications to infinite latent SVMs [J]. Journal of Machine Learning Research, 2014, 15: 1799-1847

[11] Wasserman L. Bayesian model selection and model averaging [J]. Journal of Mathematical Psychology, 2000, 44(1): 92-107

[12] Bishop C M. Pattern Recognition and Machine Learning [M]. Berlin: Springer, 2006

[13] Gershman S, Blei D. A tutorial on Bayesian nonparametric models [J]. Journal of Mathematical Psychology, 2012, 56(1): 1-12

[14] Jordan M I. Dirichlet processes, Chinese restaurant processes and all that [C] //Proc of Tutorial Presentation at the NIPS Conf. Vancouver, Canada: NIPS Foundation, 2005

[15] Antoniak C E. Mixtures of dirichlet processes with applications to Bayesian nonparametric problems [J]. The Annals of Statistics, 1974, 2(6): 1152-1174

[16] Griffiths T L, Ghahramani Z. The Indian buffet process: An introduction and review [J]. Journal of Machine Learning Research, 2012, 12: 1185-1224

[17] Rasmussen C E, Williams C K. Gaussian processes for machine learning [M]. Cambridge: The MIT Press, 2006

[18] Ferguson T. A Bayesian analysis of some nonparametric problems [J]. The Annals of Statistics, 1973, 1(2): 209-230

[19] Pitman J. Combinatorial Stochastic Processes [M]. Berlin: Springer, 2006

[20] Neal R M. Markov chain sampling methods for Dirichlet process mixture models [J]. Journal of Computational and Graphical Statistics, 2000, 9(2): 249-265

[21] Sethuraman J. A constructive definition of Dirichlet priors [R]. Tallahassee: Department of Statistics, Florida State University, 1991

[22] Blei D, Jordan M I. Variational inference for Dirichlet process mixtures [J]. Bayesian Analysis, 2006, 1(1): 121-143

[23] Sudderth E B. Graphical models for visual object recognition and tracking [D]. Cambridge: Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA, 2006

[24] Zhang Yaoting, Fang Kaitai. An Introduction to Multivariant Statistics [M]. Beijing: Science Press, 1998 (in Chinese)
(张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 1998)

[25] Jolliffe I. Principal Component Analysis [M]. New York: John Wiley & Sons, 2005

[26] Griffiths T, Ghahramani Z. Infinite latent feature models and the Indian buffet process [C] //Proc of Advances in Neural Information Processing Systems. Vancouver & Whistier, Canada: NIPS Foundation, 2005

[27] Miller K, Jordan M I, Griffiths T L. Nonparametric latent feature models for link prediction [C] //Proc of Advances in Neural Information Processing Systems. Vancouver & Whistier, Canada: NIPS Foundation, 2009

[28] Paisley J, Carin L. Nonparametric factor analysis with Beta process priors [C] //Proc of the 26th Int Conf on Machine Learning. Montreal, Canada: IMLS, 2009

[29] Zhu J. Max-margin Nonparametric Latent feature models for link prediction [C] //Proc of the 29th Int Conf on Machine Learning. Scotland: IMLS, 2012: 719-726

[30] Teh Y W, Görür Dilan, Ghahramani Z. Stick-breaking construction for the Indian buffet process [C] //Proc of Int Conf on Artificial Intelligence and Statistics. San Juan: The Society for Artificial Intelligence and Statistics, 2007: 556-563

[31] Escobar M D, West M. Bayesian density estimation and inference using mixtures [J]. Journal of the American Statistical Association, 1995, 90(430): 577-588

[32] Hollander M, Wolfe D A, Chicken E. Nonparametric Statistical Methods [M]. Hoboken: John Wiley & Sons, 2013

[33] Hjort N L, Holmes C C, Müller P, et al. Bayesian Nonparametrics [M]. Cambridge: Cambridge University Press, 2010

[34] Teh Y W, Jordan M I, Deal M J. Hierarchical Dirichlet processes [J]. Journal of the American Statistical Association, 2006, 101(476): 1566-1581

[35] Adams R P, Wallach H M, Ghahramani Z. Learning the structure of deep sparse graphical models [C] //Proc of Int Conf on Artificial Intelligence and Statistics. Sardinia: The Society for Artificial Intelligence and Statistics, 2010

[36] Beal M J, Ghahramani Z, Rasmussen C E. The infinite hidden Markov model [C] //Proc of Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS Foundation, 2001: 577-584

[37] Duan J A, Guindani M, Gelfand A E. Generalized spatial Dirichlet process models [J]. Biometrika, 2007, 94(4): 809-825

[38] Zhu J, Chen N, Xing E P. Infinite SVM: A Dirichlet process mixture of large-margin kernel machines [C] //Proc of the 28th Int Conf on Machine Learning. Washington: IMLS, 2011: 617-624

[39] Zhang Gongqing. Lecture Materials on Variational Methods. Higher Education Press [M]. Beijing: Higher Education Press, 2011 (in Chinese)
(张恭庆. 变分学讲义[D]. 北京: 高等教育出版社, 2011)

[40] Struwe M. Variational Methods [M]. Berlin: Springer, 1990

[41] Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference [J]. Foundations and Trends in Machine Learning, 2008, 1(2): 1-305

[42] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[43] Miller K T, Gael J V, Teh Y W. Variational inference for the Indian buffet process [C] //Proc of the Int Conf on Artificial Intelligence and Statistics. Florida: The Society of Artificial Intelligence and Statistics, 2009: 137-144

[44] Palmer J, Kreutz-Delgado K, Rao B D, et al. Variational EM algorithms for non-Gaussian latent variable models [C] //Proc of Advances in Neural Information Processing Systems. Vancouver & Whistier, Canada: NIPS Foundation, 2005: 1059-1066

[45] Liu J. Monte Carlo Strategies in Scientific Computing [M]. Berlin: Springer, 2008

[46] Robert C P, Casella G. Monte Carlo Statistical Methods [M]. Berlin: Springer, 1999

[47] Dani G, Lopes H F. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference [M]. Boca Raton: CRC Press, 2006

[48] Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm [J]. The American Statistician, 1995, 49(4): 327-335

[49] Casella G, George E I. Explaining the Gibbs sampler [J]. The American Statistician, 1992, 46(3): 167-174

[50] Teh Y W, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation [C] //Proc of Advances in Neural Information Processing Systems. Vancouver & Whistler, Canada: NIPS Foundation, 2006

[51] Stramer O, Tweedie R L. Langevin-type models II: Self-targeting candidates for MCMC algorithms [J]. Methodology and Computing in Applied Probability, 1999, 1(3): 307-328

[52] Brooks S, Gelman A, Jones G L, et al. Handbook of Markov chain Monte Carlo [M]. London: Chapman and Hall, 2011

[53] Zhu J, Ahmed A, Xing E P. MedLDA: Maximum margin supervised topic models [J]. Journal of Machine Learning Research, 2012, 13(1): 2237-2278

[54] Zhang A, Zhu J, Zhang B. Max-margin infinite hidden Markov models [C] //Proc of the 31st Int Conf on Machine Learning. Beijing: IMLS, 2014: 315-323

[55] Shi T, Zhu J. Online Bayesian Passive-Aggressive Learning [C] //Proc of the 31st Int Conf on Machine Learning. Beijing: IMLS, 2014: 378-386

[56] Zhu J, Chen N, Xing E P. Infinite latent SVM for classification and multi-task learning [C] //Proc of Advances in Neural Information Processing Systems. Granada, Spain: NIPS Foundation, 2011: 1620-1628

[57] Zhu J, Ahmed A, Xing E P. MedLDA: Maximum margin supervised topic models for regression and classification [C] //Proc of the 26th Int Conf on Machine Learning. Montreal, Canada: IMLS, 2009: 1257-1264

[58] Mei S, Zhu J, Zhu X. Robust RegBayes: Selectively incorporating first-order logic domain knowledge into Bayesian models [C] //Proc of the 31st Int Conf on Machine Learning. Beijing: IMLS, 2014

[59] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers [J]. Foundations and Trends in Machine Learning, 2011, 3(1): 1-122

[60] Kushner H J, Yin G G. Stochastic Approximation Algorithms and Applications [M]. Berlin: Springer, 1997

[61] Hoffman M, Blei D, Wang C, et al. Stochastic variational inference [J]. Journal of Machine Learning Research, 2013, 14(1): 1303-1347

[62] Bottou L. Large-scale machine learning with stochastic gradient descent [C] //Proc of COMPSTAT'2010. Limassol: European Regional Section of the IASC, 2010: 177-186

[63] Girolami M, Calderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011, 73(2): 123-214

[64] Mandt S, Blei D. Smoothed Gradients for Stochastic Variational Inference [C] //Proc of Advances in Neural Information Processing Systems. Montreal, Canada, NIPS Foundation, 2014

[65] Welling M, Teh Y W. Bayesian learning via stochastic gradient Langevin dynamics [C] //Proc of the 28th Int Conf on Machine Learning. Bellevue, USA: IMLS, 2011: 681-688

[66] Chen T, Fox E B, Guestrin C. Stochastic gradient Hamiltonian Monte Carlo [C] //Proc of the 31st Int Conf on Machine Learning. Beijing: IMLS, 2014: 210-222

[67] Broderick T, Boyd N, Wibisono A, et al. Streaming Variational Bayes [C] //Proc of Advances in Neural Information Processing Systems. Nevada, USA: NIPS Foundation, 2013: 1727-1735

[68] Langford J, Smola A, Zinkevich M. Slow learners are fast [C] //Proc of Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS Foundation, 2009

[69] Minsker S, Srivastava S, Lin L, et al. Robust and scalable Bayes via a median of subset posterior measures [C] //Proc. of the 31st Int Conf on Machine Learning. Beijing: IMLS, 2014: 567-580

[70] Ahmed A, Aly M, Gonzalez J, et al. Scalable inference in latent variable models [C] //Proc of the 5th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2012: 123-132

[71] Blei D, Lafferty J. Correlated topic models [C] //Proc of Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS Foundation, 2006: 147

[72] Chen J, Zhu J, Wang Z, et al. Scalable inference for logistic-normal topic models [C] //Proc of Advances in Neural Information Processing Systems. Vancouver, Nevada, USA: NIPS Foundation, 2013: 2445-2453

[73] Scalable inference for logistic-normal topic models [EB/OL]. (2013-11-22)[2014-11-16]. http://bigml.cs.tsinghua.edu.cn/~scalable-ctm/

[74] Liu S, Wang X, Chen J, et al. Topicpanorama: A full picture of relevant topics [J]. IEEE Trans on Visualization and Computer Graphics, 2014, 20: 519-530

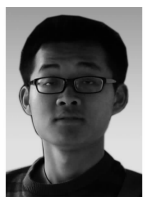
[75] Yan F, Xu N, Qi Y. Parallel inference for latent Dirichlet allocation on graphics processing units [C] //Proc of Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS Foundation, 2009

[76] Suchard M A, Wang Quanli, Chan C, et al. Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures [J]. Journal of Computational and Graphical Statistics, 2010, 19(2): 419-438

[77] Beam A L, Ghosh S K, Doyle J. Fast Hamiltonian Monte Carlo using GPU computing [J/OL]. ArXiv: 1402.4089, preprint, 2014 [2014-11-21]. http://arxiv.org/abs/1402.4089

[78] Chau T, Targett J, Wijeyasinghe M, et al. Accelerating sequential Monte Carlo method for real-time air traffic management [C] //Proc of Int Symp on Highly Efficient Accelerators and Reconfigurable Technologies (HEART). New York: ACM, 2013: 35-40

Zhu Jun, born in 1983. Associate professor and PhD supervisor in Tsinghua University. His current research interests include machine learning, Bayesian statistics, and large-scale learning algorithms and applications.



Hu Wenbo, born in 1992. PhD candidate in Tsinghua University. His current research interests include machine learning and scalable Bayesian learning methods (hwb13@mails.tsinghua.edu.cn).