最大熵判别主题模型的高效学习算法

陈键飞[1] 朱 军[1]

摘 要 现有的有监督主题模型训练算法的时间复杂度一般线性于主题数量,限制了其大规模应用. 基于此种情况,文中提出最大熵判别潜在狄利克雷分配(MedLDA)有监督主题模型的高效学习算法. 算法为坐标下降算法,训练分类器的迭代次数少于 MedLDA 已有的蒙特卡洛算法. 算法还利用拒绝采样及高效的预处理技术,将训练的时间复杂度从线性于主题数量降至亚线性于主题数量. 在多个文本数据集上的对比实验表明,相比原有的蒙特卡洛算法,文中算法在训练速度上有大幅提升.

关键词 有监督主题模型, 坐标下降算法, 吉布斯采样算法, 拒绝采样算法

引用格式 陈键飞,朱 军. 最大熵判别主题模型的高效学习算法. 模式识别与人工智能, 2019, 32(8): 736-745.

DOI 10.16451/j.cnki.issn1003-6059.201908007 **中图法分类号** TP 181

Efficient Learning Algorithm for Maximum Entropy Discrimination Topic Models

CHEN Jianfei[1], ZHU Jun[1]

ABSTRACT Time complexity of the existing supervised topic model training algorithms is generally linear to the number of topics and therefore their large-scale application is limited. To solve this problem, an efficient learning algorithm for maximum entropy discrimination of latent Dirichlet allocation (MedLDA) supervised subject model is proposed in this paper. The proposed algorithm is based on coordinate descent, and the number of iterations of training classifiers is less than that of the existing Monte Carlo algorithm for MedLDA. The algorithm also makes use of rejection sampling and efficient preprocessing technique to reduce the time complexity of training from linear to sub-linear with respect to the number of topics. The comparison experiments on multiple text corpora show that the proposed algorithm makes a great improvement in training speed compared with the existing Monte Carlo algorithm.

Key Words Supervised Topic Models, Coordinate Descent, Gibbs Sampling, Rejection Sampling

Citation CHEN J F, ZHU J. Efficient Learning Algorithm for Maximum Entropy Discrimination Topic Models. Pattern Recognition and Artificial Intelligence, 2019, 32(8): 736-745.

主题模型[1]是重要的文本特征学习工具. 这类模型可以从文本数据集中发现主题,即语义相关联的词群,并将文档分解为主题的分布. 由于其良好的特征学习效果和可解释性,主题模型及其变体[2-6]被广泛应用于信息检索[7]、推荐系统[8]、数据可视化[9]、在线广告[10]、可解释学习[11]等领域.

传统的主题模型是无监督的,如潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)[1]. 它们只利用文本本身的信息,而未利用额外的标签. 在实际应用中,存在着大量有标注文本数据集,如在线广告中

收稿日期:2019-05-12;录用日期:2019-07-15
Manuscript received May 12, 2019;
accepted July 15, 2019
国家自然科学基金重点国际合作项目(No. 61620106010)、北京市自然科学基金重点专题项目(No. L172037)资助
Supported by National Natural Science Foundation of China(No. 61620106010), Beijing Natural Science Foundation(No. L172037)
本文责任编委 陈松灿
Recommended by Associate Editor CHEN Songcan
1. 清华大学 计算机科学与技术系 北京 100084
1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084

使用的点击通过率数据集[10]等. 有监督主题模型是主题模型的一个变体. 通过联合建模文本内容和标签的概率分布,有监督主题模型不但能够进行文本分类,还可以利用监督信息学到更好的主题表示. 典型的有监督主题模型包括有监督 LDA(Supervised LDA, sLDA)[12]、判别式 LDA(Discriminative LDA, DiscLDA)[13]、最大熵判别 LDA(Maximum Entropy Discrimination LDA, MedLDA)[14-15]等.

为了使主题模型能够处理海量的文本数据集,需要更快、时间复杂度更低的训练算法. 利用主题模型的稀疏性,学者们提出许多高效算法,包括 SparseLDA[16]、AliasLDA[17]、F+LDA[18]、LightLDA[19]、WarpLDA[20]等. 然而,由于有监督主题模型使用非共轭的先验,这些高效算法不能直接应用于有监督主题模型. 对于有监督主题模型,已有的蒙特卡洛算法[21]的时间复杂度是线性于主题数量的,远高于 LDA 的亚线性、甚至常数级的时间复杂度. 有监督主题模型训练算法的高时间复杂度限制其推广到多主题、多类别的数据集上.

本文针对典型的有监督主题模型 MedLDA,提出高效学习算法. 基于正则化贝叶斯学习[22], MedLDA 融合最大间隔准则和贝叶斯学习. 由于结合两者的优势,相比 sLDA 和 DiscLDA, MedLDA 的分类器只依赖于很少一部分支持向量,在提升模型鲁棒性的同时使更高效的训练成为可能. 本文利用 MedLDA 这一性质,降低训练过程中学习分类器和采样话题的时间复杂度,提升总体的训练效率. 具体地,本文提出 MedLDA 的坐标下降算法,减少分类器训练的迭代次数,并提出拒绝性采样算法,将采样每个词的时间复杂度从线性于主题数量降至亚线性于主题数量. 最后提出利用支持向量稀疏性的预处理算法,将预处理每篇文档的时间复杂度从线性于类别数量降至亚线性于类别数量. 在 20NewsGroups、Wiki、RCV1 等常用数据集上的实验表明,本文算法的训练速度是 MedLDA 现有的蒙特卡洛算法[21]的 4 倍.

1　相关知识

MedLDA 是一个有监督主题模型,同时处理如下 2 个任务:1)对文档主题进行建模;2)利用文档的主题分布预测其标签. 模型的输入为 D 篇文档构成的数据集 $(W, \boldsymbol{Y})$,其中 $W = \{\boldsymbol{w}_d\}_{d=1}^{D}$ 为文档内容,每篇文档 $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ 由 N_d 个词构成,$w_{dn} \in \{1,2,\cdots,V\}$ 为其中第 n 个词在 V 个词构成的词汇表中的序号. $\boldsymbol{Y}$ 为一个 $D \times L$ 的标签矩阵,其中 $y_{dl} \in \{-1, 1\}$ 表示第 d 篇文档的第 l 个标签. 这里描述一个多任务学习的场景,即每个文档有 L 个标签. 特别地,若只需要对文档进行二分类,则 $L = 1$;若需要对文档进行多分类,则每篇文档 d 有且仅有一个 l,使得 $y_{dl} = \pm 1$.

1.1　正则化贝叶斯推理

MedLDA 为正则化贝叶斯推理框架[22]的一个具体例子. 正则化贝叶斯推理在贝叶斯推理基础上,引入正则化项,使模型更灵活. 具体地,令 $\boldsymbol{\Theta}$ 为模型参数、$\mathbb{D}$ 为数据集,则给定先验 $p(\boldsymbol{\Theta})$ 和似然度 $p(\mathbb{D} \mid \boldsymbol{\Theta})$,由贝叶斯公式,模型参数的后验应为

$$p(\boldsymbol{\Theta} \mid \mathbb{D}) = \frac{p(\boldsymbol{\Theta})p(\mathbb{D} \mid \boldsymbol{\Theta})}{p(\mathbb{D})}.$$

贝叶斯推理的过程可以使用一个等效的求解变分分布 $q(\boldsymbol{\Theta})$ 的变分问题描述:

$$\min_{q(\boldsymbol{\Theta})} \mathrm{KL}(q(\boldsymbol{\Theta}) \parallel p(\boldsymbol{\Theta} \mid \mathbb{D})).$$

正则化贝叶斯推理在贝叶斯基础上,引入后验正则化项 $\Omega(q(\boldsymbol{\Theta}); \mathbb{D})$,并定义如下优化问题:

$$\min_{q(\boldsymbol{\Theta})} \mathrm{KL}(q(\boldsymbol{\Theta}) \parallel p(\boldsymbol{\Theta} \mid \mathbb{D})) + c\Omega(q(\boldsymbol{\Theta}); \mathbb{D}).$$

后验正则化项的引入大幅加强贝叶斯推理的灵活性,通过合理选择正则化项,可以实现基于最大间隔准则限制提升预测性能[14]、基于逻辑规则约束利用专家知识[23]、或是限制稀疏性[24]等功能. 具体到 MedLDA,模型分为两部分:LDA 部分由后验分布描述,用于建模文档的分布 $p(\boldsymbol{W})$;分类器部分使用正则化项实现,用于根据文档预测 $\boldsymbol{Y}$.

1.2　潜在的狄利克雷分配

LDA 部分学习 K 个主题,可以使用一个产生式过程描述如下:

1)对每个主题 $k = 1,2,\cdots,K$,生成

$$\phi_k \sim \mathrm{Dir}(\beta, V);$$

2)对每篇文档 $d = 1,2,\cdots,D$,生成

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha, K);$$

3)对每篇文档 d 的每个位置 n,生成主题分配 $z_{dn} \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$,生成词 $w_{dn} \sim \mathrm{Mult}(\phi_{z_{dn}})$.

其中:$\mathrm{Dir}(\alpha, K)$ 为 K 维的对称狄利克雷分布,集中度参数为 α; $\mathrm{Mult}(\boldsymbol{\theta})$ 为离散分布. 记 $\boldsymbol{Z} = \{\boldsymbol{z}_d\}_{d=1}^{D}$, $\boldsymbol{z}_d = \{z_{dn}\}_{n=1}^{N_d}$, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_d\}_{d=1}^{D}$, $\boldsymbol{\phi} = \{\phi_k\}_{k=1}^{K}$,该产生式过程定义一个联合分布 $p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \alpha, \beta)$,根据共轭性[25],可以解析地积分掉 $(\boldsymbol{\theta}, \boldsymbol{\phi})$,得到 $p(\boldsymbol{W}, \boldsymbol{Z} \mid \alpha, \beta)$.

为了符号的简洁,在下文中均省略超参 (α, β). LDA 的训练需要求出隐变量 $\boldsymbol{Z}$ 的后验分布

$p(\boldsymbol{Z}\mid\boldsymbol{W})$,这可以理解为如下变分问题:

$$\min_{q(\boldsymbol{Z})}\mathrm{KL}(q(\boldsymbol{Z})\parallel p(\boldsymbol{Z}\mid\boldsymbol{W})). \tag{1}$$

1.3 分类器

为了分类,MedLDA 学习一组分类器. 每个分类器接受的输入是文档的主题分布

$$\bar{\boldsymbol{z}}_d:=\frac{1}{N_d}\sum_{n=1}^{N_d}\boldsymbol{e}_{z_{dn}},$$

其中 $\boldsymbol{e}_k$ 为第 k 维为 1、其它维为 0 的 K 维单位向量. 分类器 $\boldsymbol{\eta}$ 对于第 l 个任务的输出为 $\boldsymbol{\eta}_l^{\mathrm{T}}\bar{\boldsymbol{z}}_d$. MedLDA 希望学习分类器的分布 $q(\boldsymbol{\eta})$,使期望分类器 $E_{q(\boldsymbol{\eta})q(\boldsymbol{Z})}[\boldsymbol{\eta}_l^{\mathrm{T}}\bar{\boldsymbol{z}}_d]$ 的间隔最大. 分类器权重的先验分布为 $p(\boldsymbol{\eta})=\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$.

简写 $q_{\boldsymbol{Z}}=q(\boldsymbol{Z})$,$q_{\boldsymbol{\eta}}=q(\boldsymbol{\eta})$,结合 LDA 部分的损失式(1) 和最大间隔分类器的损失,MedLDA 定义的优化问题:

$$\begin{aligned}&\min_{q_{\boldsymbol{\eta}},q_{\boldsymbol{Z}},\xi}\mathcal{L}(q_{\boldsymbol{\eta}},q_{\boldsymbol{Z}})+2c\sum_{l=1}^{L}\sum_{d=1}^{D}\xi_{dl},\\&\text{s.t. }y_{dl}E_{q_{\boldsymbol{\eta}},q_{\boldsymbol{Z}}}[\boldsymbol{\eta}_l^{\mathrm{T}}\bar{\boldsymbol{z}}_d]\geqslant l-\xi_{dl},\xi_{dl}\geqslant 0,\forall d,l,\end{aligned} \tag{2}$$

其中 ξ_{dl}、l 分别为最大间隔学习的松弛变量和间隔,

$$\begin{aligned}\mathcal{L}(q_{\boldsymbol{\eta}},q_{\boldsymbol{Z}})=&E_{q_{\boldsymbol{Z}}}[\ln q(\boldsymbol{Z})-\ln p(\boldsymbol{W},\boldsymbol{Z})]+\\&E_{q_{\boldsymbol{\eta}}}[\ln q(\boldsymbol{\eta})-\ln p(\boldsymbol{\eta})].\end{aligned}$$

MedLDA 已有的算法包括变分推理算法[14] 和蒙特卡洛算法[21,26]. 两者相比,蒙特卡洛算法因为近似更少,得到的模型质量优于变分推理算法. 但是两种算法的时间复杂度均正比于主题数量,这限制 MedLDA 应用于大规模数据集.

2 最大熵判别 LDA 的坐标下降算法

本节中提出 MedLDA 的坐标下降算法,相比原有的蒙特卡洛算法[21],坐标下降算法无需在每次优化分类器时将其优化至收敛,因此可大幅减少优化分类器的迭代次数.

2.1 简化的优化问题

式(2) 为一个凸优化问题,使用拉格朗日乘子法求解. 引入拉格朗日乘子 $\boldsymbol{\mu}$、$\boldsymbol{r}$,则式(2) 的拉格朗日函数为

$$\begin{aligned}L(q_{\boldsymbol{\eta}},q_{\boldsymbol{Z}},\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{r})=&\\\mathcal{L}(q_{\boldsymbol{Z}},q_{\boldsymbol{\eta}})+2c\sum_{l=1}^{L}\sum_{d=1}^{D}\xi_{dl}-\sum_{l=1}^{L}\sum_{d=1}^{D}r_{dl}\xi_{dl}-&\\\sum_{l=1}^{L}\sum_{d=1}^{D}\mu_{dl}(y_{dl}E_{q_{\boldsymbol{\eta}}}[\boldsymbol{\eta}_l]^{\mathrm{T}}E_{q_{\boldsymbol{Z}}}[\bar{\boldsymbol{z}}_d]-l+\xi_{dl}).&\end{aligned} \tag{3}$$

式(2) 的最优解应为

$$\begin{aligned}&\min_{q_{\boldsymbol{\eta}},q_{\boldsymbol{Z}},\xi}\max_{\boldsymbol{\mu},\boldsymbol{r}}L(q_{\boldsymbol{\eta}},q_{\boldsymbol{Z}},\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{r}),\\&\text{s.t. }\mu_{dl}\geqslant 0,r_{dl}\geqslant 0,\forall d,l.\end{aligned} \tag{4}$$

将上式对 $q_{\boldsymbol{\eta}}$ 求导,可得最优解

$$q^*(\boldsymbol{\eta}_l)=\mathcal{N}(\boldsymbol{\kappa}_l(\boldsymbol{\mu}_l,q_{\boldsymbol{Z}}),\boldsymbol{I}), \tag{5}$$

其中

$$\boldsymbol{\kappa}_l(\boldsymbol{\mu}_l,q_{\boldsymbol{Z}})=\sum_{d=1}^{D}\mu_{dl}y_{dl}E_{q_{\boldsymbol{Z}}}[\bar{\boldsymbol{z}}_d].$$

将式(3) 对 $\boldsymbol{\xi}$ 求导,可得

$$L'_{\xi_{dl}}=2c-\mu_{dl}-r_{dl}=0,$$

结合 $\mu_{dl}\geqslant 0$,$r_{dl}\geqslant 0$,可得

$$0\leqslant\mu_{dl}\leqslant 2c. \tag{6}$$

将式(5)、式(6) 代入式(4),可得简化后的优化问题:

$$\begin{aligned}&\min_{q_{\boldsymbol{Z}}}\max_{\boldsymbol{\mu}}L(q_{\boldsymbol{Z}},\boldsymbol{\mu}),\\&\text{s.t. }0\leqslant\mu_d\leqslant 2c,\forall d,l,\end{aligned} \tag{7}$$

其中

$$\begin{aligned}L(q_{\boldsymbol{Z}},\boldsymbol{\mu})=&E_{q_{\boldsymbol{Z}}}[\ln q(\boldsymbol{Z})-\ln p(\boldsymbol{W},\boldsymbol{Z})]-\\&\frac{1}{2}\sum_{l=1}^{L}\boldsymbol{\kappa}_l(\boldsymbol{\mu}_l,\boldsymbol{q}_{\boldsymbol{Z}})^{\mathrm{T}}\boldsymbol{\kappa}_l(\boldsymbol{\mu}_l,q_{\boldsymbol{Z}})+l\sum_{l=1}^{L}\sum_{d=1}^{D}\mu_{dl}.\end{aligned} \tag{8}$$

2.2 坐标下降算法

式(7) 对应的优化问题可以使用坐标下降法求解,从初始解 $(q_{\boldsymbol{Z}}^{(0)},\boldsymbol{\mu}^{(0)})$ 开始,迭代进行如下操作.

1)给定 $q_{\boldsymbol{Z}}^{(t)}$,$\boldsymbol{\mu}^{(t)}$,优化 $q_{\boldsymbol{Z}}^{(t+1)}$. 由式(8) 求导可得最优的 $q_{\boldsymbol{Z}}$ 应满足

$$q_{\boldsymbol{Z}}\propto p(\boldsymbol{W},\boldsymbol{Z})\exp\left(\sum_{l=1}^{L}\boldsymbol{\kappa}_l(\boldsymbol{\mu}^{(t)},q_{\boldsymbol{Z}})^{\mathrm{T}}\sum_{d=1}^{D}\mu_{dl}y_{dl}\bar{\boldsymbol{z}}_d\right). \tag{9}$$

这不是一个解析解,因为式(9) 的左右两边均有 $q_{\boldsymbol{Z}}^{(t)}$. 为了可解性,本文使用 $\boldsymbol{\kappa}_l(\boldsymbol{\mu}^{(t)},q_{\boldsymbol{Z}}^{(t)})$ 近似式(9) 右边的 $\boldsymbol{\kappa}_l(\boldsymbol{\mu}^{(t)},q_{\boldsymbol{Z}})$.

2)给定 $q_{\boldsymbol{Z}}^{(t+1)}$,优化 $\boldsymbol{\mu}^{(t+1)}$. 由式(7)、式(8),优化 $\boldsymbol{\mu}$ 的问题可拆解成 L 个独立的子问题,其中优化 $\boldsymbol{\mu}_l$ 的子问题为

$$\begin{aligned}&\max_{\boldsymbol{\mu}_l}-\frac{1}{2}\left\|\sum_{d=1}^{D}\mu_{dl}y_{dl}E_{q_{\boldsymbol{Z}}^{(t+1)}}[\bar{\boldsymbol{z}}_d]\right\|^2+l\sum_{d=1}^{D}\mu_{dl},\\&\text{s.t. }0\leqslant\mu_{dl}\leqslant 2c,\forall d,l,\end{aligned} \tag{10}$$

其中期望 $E_{q_{\boldsymbol{Z}}^{(t+1)}}[\bar{\boldsymbol{z}}_d]$ 可以通过从 $q_{\boldsymbol{Z}}^{(t+1)}$ 中采样近似.

式(10)的优化问题与支持向量机的对偶问题具有相同形式,可以使用支持向量机的求解算法,如随机对偶坐标上升法[27-28].

总之,坐标下降算法训练流程对每轮迭代重复如下步骤:

1)训练 LDA. 按式(9)得到 $q_{\mathbf{Z}}^{(t+1)}$,并通过采样近似期望 $E_{q_{\mathbf{Z}}^{(t+1)}}[\bar{\boldsymbol{z}}_d]$.

2)训练分类器. 给定 $E_{q_{\mathbf{Z}}^{(t+1)}}[\bar{\boldsymbol{z}}_d]$,按算法 1 优化 $\boldsymbol{\mu}^{(t+1)}$.

这些步骤与原有的蒙特卡洛算法[21]相似. 主要区别如下:在原有的蒙特卡洛算法中,训练分类器的过程要重复迭代直至 $\boldsymbol{\mu}$ 收敛. 这实际上是浪费的,因为 $q_{\mathbf{Z}}$ 在不断变化,对某个固定的 $q_{\mathbf{Z}}$ 将 $\boldsymbol{\mu}$ 优化至收敛并无意义. 而坐标下降算法解决这个问题,在坐标下降的框架中,只要每轮优化 $\boldsymbol{\mu}$ 的迭代对目标函数 $L(q_{\mathbf{Z}},\boldsymbol{\mu})$ 有提升即可. 故每次仅优化 $\boldsymbol{\mu}$ 固定的轮数,记这个轮数为 i_{svm}. 相比优化至收敛,大幅降低迭代次数. 坐标下降算法的具体实现如算法 1 所示.

算法 1　MedLDA 的坐标下降算法

令 $\boldsymbol{Z}\leftarrow\boldsymbol{Z}^{(0)},\boldsymbol{\mu}\leftarrow\boldsymbol{\mu}^{(0)}$,

对 $l\in\{1,2,\cdots,L\}$,计算 $\boldsymbol{\kappa}_l\leftarrow\sum_d\mu_{dl}y_{dl}\bar{\boldsymbol{z}}_d$

for 每轮迭代 do

　采样 $\boldsymbol{Z}\sim p(\boldsymbol{W},\boldsymbol{Z})\exp\left(\sum_{l=1}^{L}\boldsymbol{\kappa}_l^{\mathrm{T}}\sum_{d=1}^{D}\mu_{dl}y_{dl}\bar{\boldsymbol{z}}_d\right)$

　for 每个任务 $l\in\{1,2,\cdots,L\}$ do

　　计算 $\boldsymbol{\kappa}_l\leftarrow\sum_d\mu_{dl}y_{dl}\bar{\boldsymbol{z}}_d$

　　for 每次迭代 $i\in\{1,2,\cdots,i_{\text{svm}}\}$ do

　　for 每个文档 d,顺序随机 do

　　$g(\mu_{dl})\leftarrow l-y_{dl}\boldsymbol{\kappa}_l^{\mathrm{T}}\bar{\boldsymbol{z}}_d$

　　$\mu_{dl}\leftarrow\mu_{dl}+\dfrac{g(\mu_{dl})}{\|\bar{\boldsymbol{z}}_d\|^2}$

　　$\boldsymbol{\kappa}_l\leftarrow\boldsymbol{\kappa}_l-\mu_{dl}y_{dl}\bar{\boldsymbol{z}}_d$

　　$\mu_{dl}\leftarrow\max\{0,\min\{2c,\mu_{dl}\}\}$

　　$\boldsymbol{\kappa}_l\leftarrow\boldsymbol{\kappa}_l+\mu_{dl}y_{dl}\bar{\boldsymbol{z}}_d$

　　end for

　end for

　end for

end for

3　快速采样算法

本节讨论如何快速采样式(9)中的 $q(\boldsymbol{Z})$,使采样每个词的时间复杂度从线性于主题数量降至亚线性.

3.1　吉布斯采样

可以使用吉布斯采样从 $q(\boldsymbol{Z})$ 中采样,由狄利克雷分布的共轭性[25],得

$$p(\boldsymbol{W},\boldsymbol{Z})\propto\prod_{d=1}^{D}\frac{B(\boldsymbol{C}_d+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha}\mathbf{1})}\prod_{k=1}^{K}\frac{B(\boldsymbol{C}_k+\boldsymbol{\beta})}{B(\boldsymbol{\beta}\mathbf{1})},$$

其中

$$B(\boldsymbol{\alpha})=\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}$$

为多元贝塔函数,$\Gamma(\cdot)$ 为伽马函数,$\mathbf{1}=(1,1,\cdots,1)$,$\boldsymbol{C}_d=\sum_{n=1}^{N_d}\boldsymbol{e}_{z_{dn}}$ 为文档-主题计数向量,$\boldsymbol{e}_{z_{dn}}$ 为第 1 节中定义的单位向量,

$$\boldsymbol{C}_k=\sum_{d=1}^{D}\sum_{n=1}^{N_d}\boldsymbol{I}(z_{dn}=k)e_{w_{dn}}$$

为主题-单词计数向量,其中 $\boldsymbol{I}(z_{dn}=k)$ 为指示器函数,在 $z_{dn}=k$ 时为 1,其余时候为 0. 根据式(9)整理得

$$q(z_{dn}=k\mid\boldsymbol{Z}_{\neg dn})\propto q_{dnk},$$

其中

$$q_{dnk}=(C_{dk}^{\neg dn}+\alpha)\frac{C_{k,w_{dn}}^{\neg dn}+\beta}{\sum_v C_{kv}^{\neg dn}+V\beta}\lambda_{dk},\tag{11}$$

$$\lambda_{dk}=\frac{\exp\Lambda_{dk}}{\sum_k\exp\Lambda_{dk}},$$

$$\Lambda_{dk}=\frac{1}{N_d}\sum_{l=1}^{L}\mu_{dl}y_{dl}\kappa_{lk},$$

$$\kappa_{lk}=\boldsymbol{\kappa}_l(\boldsymbol{\mu}_l^{(t)},q_{\mathbf{Z}}^{(t)})_k,$$

$\neg dn$ 为从相应的集合或计数中去掉 (d,n) 位置的词. 已有的蒙特卡洛算法直接根据式(11)采样,时间复杂度正比于主题数.

下面分析时间复杂度. 按式(5)计算所有 κ_{lk} 需要 $O(DK_dL)$ 的时间,其中

$$K_d=\frac{1}{D}\sum_{d=1}^{L}\|\boldsymbol{C}_d\|_0$$

表示计数向量 $\boldsymbol{C}_d$ 的平均非零元数量. 直接计算 Λ_{dk} 需要 $O(DKL)$ 的时间,直接按式(11)采样 z_{dn} 需要 $O\left(D\bar{N}_dK\right)$ 的时间,其中

$$\bar{N}_d=\frac{1}{D}\sum_{d=1}^{D}N_d$$

为平均文档长度. 故总时间复杂度为

$$O\left(DK\left(L+\bar{N}_d\right)\right),$$

正比于主题数量.

3.2 拒绝采样算法

本文设计拒绝采样算法,使采样 z_{dn} 更高效. 记

$$\boldsymbol{\lambda}_d = (\lambda_{dk})_{k=1}^K,\ \boldsymbol{\Lambda}_d = (\Lambda_{dk})_{k=1}^K.$$

注意到 $\boldsymbol{\lambda}_d$ 实际上是对 $\boldsymbol{\Lambda}_d$ 进行 softmax 变换的结果. 若 $\boldsymbol{\Lambda}_d$ 的每维较接近,则 $\boldsymbol{\lambda}_d$ 的每维应较接近 $1/K$;若 $\boldsymbol{\Lambda}_d$ 的不同维相差很大,则 λ_{dk} 在 $\boldsymbol{\Lambda}_{dk}$ 最大的一维会接近 1,而其它维接近 0. 无论哪种情况,$\boldsymbol{\lambda}_d$ 的大部分维度应当差不多. 拒绝采样算法为 λ_{dk} 设定一个上界

$$\lambda_{dk} \leqslant \varepsilon + \hat{\lambda}_{dk},$$

其中

$$\hat{\lambda}_{dk} = \max\{0, \lambda_{dk} - \varepsilon\}.$$

基于上述观察,在 ε 选取合适时,$\hat{\lambda}_{dk}$ 应当非常稀疏. 记

$$\hat{\phi}_{kv} = \frac{C_{kv}^{\neg dn} + \beta}{\sum_v C_{kv}^{\neg dn} + V\beta},$$

可得

$$q_{dnk} = C_{dk}^{\neg dn}\lambda_{dk}\hat{\phi}_{kv} + \alpha\lambda_{dk}\hat{\phi}_{k,w_{dn}} \leqslant \underbrace{C_{dk}^{\neg dn}\lambda_{dk}\hat{\phi}_{k,w_{dn}}}_{q_{dnk}^1} + \underbrace{\alpha\hat{\lambda}_{dk}\hat{\phi}_{k,w_{dn}}}_{q_{dnk}^2} + \underbrace{\alpha\varepsilon\hat{\phi}_{k,w_{dn}}}_{q_{dnk}^3}. \tag{12}$$

式(12)可以看作 3 个分布 $q_{dn}^1, q_{dn}^2, q_{dn}^3$ 的混合,三者的混合比例分别正比于 $\sum_k q_{dnk}^1, \sum_k q_{dnk}^2, \sum_k q_{dnk}^3$. 因此可随机按照混合比例选择一个分布,再从对应分布中采样. 注意到,分别由于 $C_{dk}^{\neg dn}$ 和 $\hat{\lambda}_{dk}$ 的稀疏性,q_{dn}^1 和 q_{dn}^2 都是稀疏的. 而 $q_{dnk}^3 \propto \hat{\phi}_{k,w_{dn}}$ 只和 w_{dn} 有关,和具体的文档 d 无关,可以采用和 F+LDA[18] 同样的技巧,按单词序号 w_{dn} 的顺序,而非文档序号 d 的顺序采样每个词,通过 Fenwick 树[18] 在 $O(\ln K)$ 的时间内实现从 q_{dn}^3 中的采样. 最后,需要以概率

$$1 - \frac{q_{dnk}}{C_{dk}^{\neg dn}\lambda_{dk}\hat{\phi}_{kv} + \alpha\hat{\lambda}_{dk}\hat{\phi}_{kv} + \alpha\varepsilon\hat{\phi}_{k,w_{dn}}}$$

拒绝采到的样本. 拒绝概率与参数 ε 的选择有关. 若 $\varepsilon = 0$,拒绝概率为 0,但是 $\hat{\lambda}_{dk}$ 不是稀疏的,因此采样的时间复杂度仍是 $O(K)$. 若 ε 很大,则 $\hat{\lambda}_{dk}$ 很稀疏,但拒绝概率可能很高. 拒绝采样算法的伪代码如算法 2 所示.

算法 2 MedLDA 的拒绝采样算法

```
for 每个词 v ∈ {1,2,…,v} do
    为分布 q^3_dn 建 Fenwick 树
    for 每个满足 w_dn = v 的词项 do
        更新 C_{dz_dn}、C_{z_dn v} 和 Fenwick 树
        通过枚举所有 C_dk > 0 的 k,计算分布 q^1_dn 和混合比例 Σ_k q^1_dnk
        通过枚举所有 λ̂_dk > 0 的 k,计算分布 q^2_dn 和混合比例 Σ_k q^2_dnk
        do
            采样 u ~ Mult(Σ_k q^1_dnk, Σ_k q^2_dnk, Σ_k q^3_dnk)
            若 u = 3,从 Fenwick 树中采样主题 k,
            否则从 q^u_dn 中采样主题 k
        while 样本 k 被拒绝
        z_dn ← k,更新 C_dk、C_kv 和 Fenwick 树
    end for
end for
```

该算法时间复杂度分析如下. 计算 $\sum_k q_{dnk}^1$ 及从 q_{dn}^1 中采样的时间复杂度都是每个词 $O(K_d)$ 的. 计算 $\sum_k q_{dnk}^2$ 及从 q_{dn}^2 中采样的时间复杂度是每个词 $O(K_\lambda)$ 的,其中

$$K_\lambda = \frac{1}{D}\sum_{d=1}^D \|\hat{\boldsymbol{\lambda}}_d\|_0$$

反映 $\hat{\boldsymbol{\lambda}}_d$ 的稀疏性. 维护 Fenwick 树及从 q_{dn}^3 中的采样的时间复杂度为 $O(\ln K)$. 因此,从 $q(\boldsymbol{Z})$ 采样的时间复杂度从 $O\left(D\overline{N}_d K\right)$ 降至

$$O\left(D\overline{N}_d(K_d + K_\lambda + \ln K)\right),$$

亚线性于主题数量.

3.3 预处理算法

如第 3.1 节中所述,直接计算所有 Λ_{dk} 的时间复杂度是 $O(DKL)$. 但是,注意到由于支持向量机的性质,μ_{dl} 是稀疏的,故 $\mu_{dl} = 0$ 的项无需考虑,时间复杂度可降为 $O(DKL_{\mathrm{SV}})$,其中

$$L_{\mathrm{SV}} = \frac{1}{D}\|\boldsymbol{\mu}\|_0$$

表示平均每个文档有的支持向量数量.

表 1 整理蒙特卡洛算法[21] 和本文的高效算法的时间复杂度.

表 1 2 种算法的时间复杂度对比

Table 1 Time complexity comparison of 2 algorithms

步骤	操作	蒙特卡洛算法	高效算法
采样	优化 $\boldsymbol{\mu}$	$O(I_{\mathrm{svm}}DK_dL)$	$O(i_{\mathrm{svm}}DK_dL)$
分类器	计算 $\boldsymbol{\kappa}$	$O(DK_dL)$	$O(DK_dL)$
	采样 $\boldsymbol{Z}$	$O\left(D\overline{N}_dK\right)$	$O\left(D\overline{N}_d(K_d + K_\lambda + \ln K)\right)$
预处理	预处理 $\boldsymbol{\Lambda}$	$O(DKL)$	$O(DKL_{\mathrm{SV}})$

表 1 中蒙特卡洛算法支持向量机的迭代次数记为 I_{svm}. 相比蒙特卡洛算法,本文的高效算法将优化 $\boldsymbol{\mu}$ 的迭代次数从 I_{svm} 降至 i_{svm},并将采样的时间复杂度从线性降至亚线性于主题数量,将预处理的时间复杂度降至亚线性于标签数量.

4　实验及结果分析

在 20NewsGroups(http://qwone.com/~jason/20Newsgroups/20news-bydate-matlab.tgz)、LSHTC2-Wiki-large(简写为 Wiki)、RCV1[29] 数据集上进行实验. 除了 20NewsGroups 是多分类问题以外,另外两个数据集都是多任务问题,数据集的统计信息见表 2. 20NewsGroups 数据集使用原始的训练/测试划分;Wiki、RCV1 数据集随机抽取 5 000 篇文档作为测试集. 由于 Wiki、RCV1 数据集较大,这里从 2 个数据集中各随机选取 20 000 篇训练文档和 2 000 篇测试文档的子集以方便调参,记为 Wiki 子集和 RCV1 子集.

表 2　实验数据集

Table 2　Experimental datasets

名称	文档数量 D	词汇表大小 V	平均文档长度 $\overline{N}_d$	分类任务数 L
20NewsGroups	18774	53485	115	20
Wiki 子集	22000	80442	61	20
RCV1 子集	22000	47168	129	103
Wiki	1105000	915420	61	20
RCV1	804414	285483	122	103

在模型评价方面:多分类问题(20NewsGroups)使用测试准确率作为判断算法收敛情况的标准;多任务问题使用测试集上宏平均的 F1(macro-averaged F1)作为判断算法收敛情况的标准. 宏平均的 F1 将每个二分类任务分别计算的 F1 值取平均得到. 为了表述简洁,将测试准确率和测试集上宏平均的 F1 统称为"测试结果".

在超参方面,狄利克雷先验 $\alpha = 10/K, \beta = 0.01$ 为手工指定,而超参 K, c, ℓ 通过网格搜索得到,选取的值如表 3 所示. 在 MedLDA 的蒙特卡洛算法[21] 中训练支持向量机直到对偶间隙小于 0.1,而本文的高效算法固定 SVM 训练轮数 $i_{SVM} = 3$,阈值 $\varepsilon = 2/K$.

表 3　各数据集选择的超参数

Table3　Selected hyperparameters for different datasets

数据集	主题数量 K	分类器权重 c	间隔大小 l
20NewsGroups	200	3	10
Wiki	400	2.56	128
RCV1	800	1	0.1

4.1　收敛速度及训练时间

为了探究本文算法效率,对比如下 4 种算法的收敛速度及训练算法中各部分耗时:

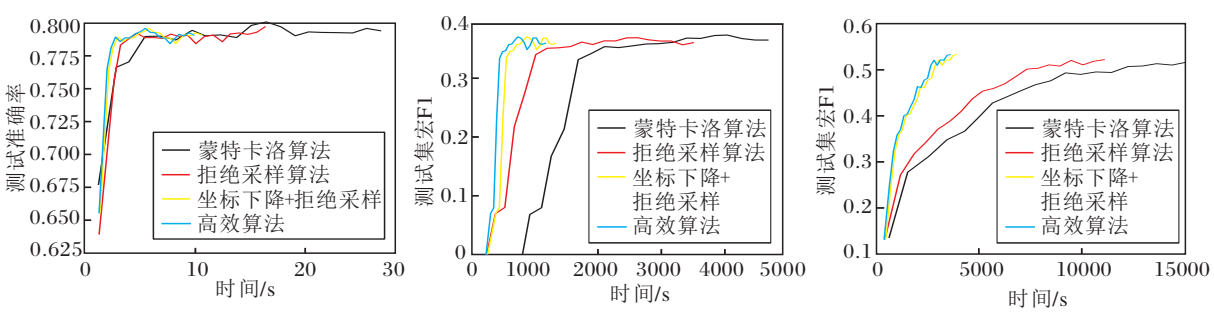
1)蒙特卡洛算法[21].

2)拒绝采样算法. 将蒙特卡洛采样算法中的采样部分替换成 3.2 节中的拒绝采样算法.

3)坐标下降+拒绝采样. 在上一种算法的基础上使用 2.2 节中的坐标下降算法,并令 $i_{SVM} = 3$.

4)高效算法. 本文的完整算法,在上一种算法的基础上使用 3.3 节的预处理算法.

每种算法迭代 20 轮. 图 1 反映测试结果随时间的变化情况. 可以看到,相比蒙特卡洛算法,本文的高效算法在相同时间内得到更好的测试结果.



(a)20NewsGroups　　(b)Wiki　　(c)RCV1

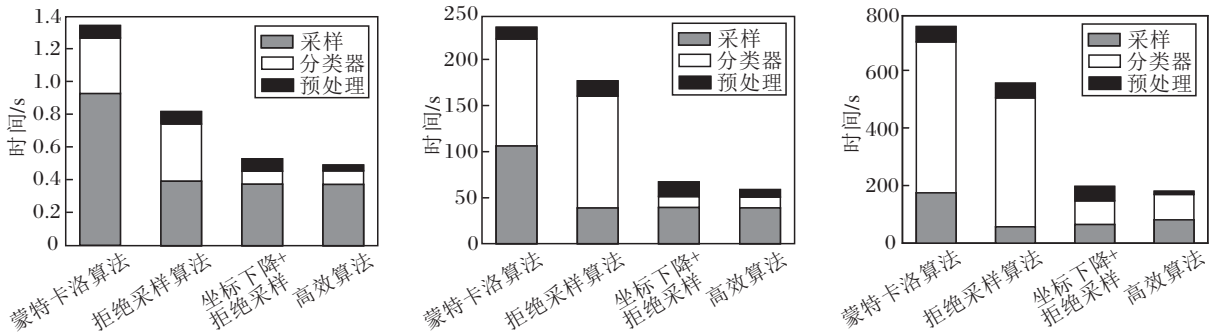
图 1　4 种算法在 3 个数据集上的收敛速度对比

Fig. 1　Convergence rate comparison of 4 algorithms on 3 datasets

为了深入研究采用高效算法后不同部分的用时变化,分别测量表 1 中总结的采样、分类器、预处理部分的运行时间,结果如图 2 所示. 对于类别数量较少、文档长度较长的 20NewsGroups 数据集,时间占比最大的是采样部分. 而对于文档较短的 Wiki 数据集和类别数量较多的 RCV1 数据集,时间占比较大的是分类器部分.

对比各算法运行时间,可以发现采样部分的用时在使用拒绝采样算法后显著下降,而分类器部分的用时在使用坐标下降算法后显著下降,这与表 1 中对时间复杂度的分析一致.

总之,蒙特卡洛算法比本文的高效算法在 20NewsGroups、Wiki 和 RCV1 数据集上每轮迭代的用时分别延至 2.4、4 和 4.17 倍.



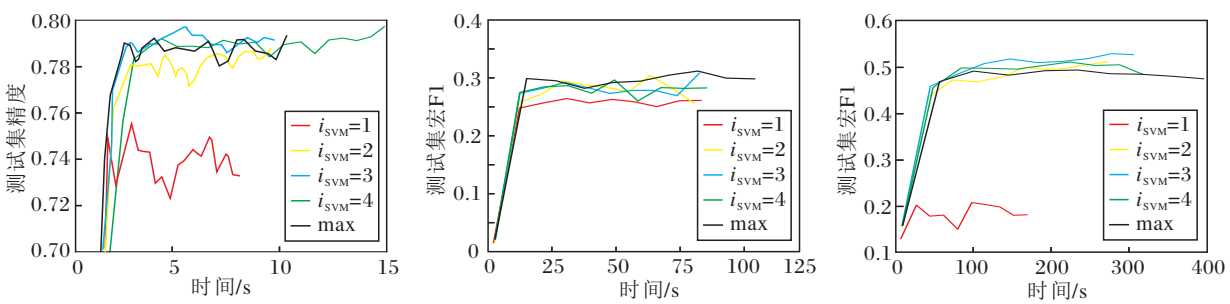
(a)20NewsGroups (b)Wiki (c)RCV1

图 2 4 种算法在 3 个数据集上的训练时间对比

Fig. 2 Training time comparison of 4 algorithms on 3 datasets

4.2 分类器训练轮数的影响

本节验证坐标下降算法分类器训练轮数 i_{svm} 对算法收敛速度的影响. 分别设 $i_{\text{svm}}=1$、2、3、4,以及采用蒙特卡洛算法的策略,迭代直到对偶间隔小于 0.1 时停止(记作 max),得到测试结果如图 3 所示. 可以看到,$i_{\text{svm}}\geqslant 2$ 时算法均能收敛到较优结果,这证明蒙特卡洛算法每次将 $\boldsymbol{\mu}$ 优化至收敛是没有必要的,可以选择较小的 i_{svm} 以提高算法效率. 值得注意的是,"max" 的测试结果在 RCV1 子集数据集上不如 $i_{\text{svm}}=3$ 的好,这可能是因为过度优化分类器导致过拟合到训练数据造成的. 基于这些实验结果,本文推荐选择整体效果最优的 $i_{\text{svm}}=3$.



(a)20NewsGroups (b)Wiki (c)RCV1

图 3 不同 i_{SVM} 对训练时间的影响

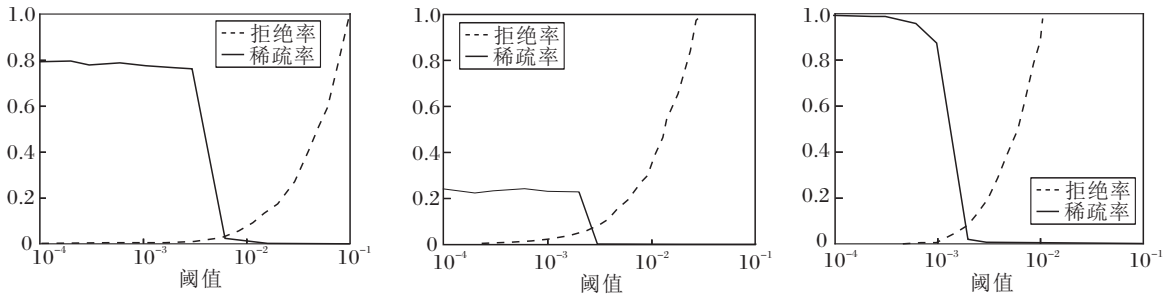
Fig. 3 Effect of different i_{SVM} on training time

4.3 超参数 ε 的选择

在拒绝采样算法中,超参 ε 表示 $\hat{\boldsymbol{\lambda}}_d$ 的稀疏率(即非零元所占比例)与采样算法的拒绝率的一个取舍,若 $\varepsilon=0$,拒绝率为 0,但稀疏率为 1,即 $\hat{\boldsymbol{\lambda}}_d$ 是稠密的,此时算法无法利用稀疏性加速,时间复杂度仍然与主题数成正比. 随着 ε 的增大,稀疏率下降,拒绝率增大. 为了提高算法效率,需要选取适当的 ε.

图 4 和图 5 为 ε 不同时算法的稀疏率、拒绝率及

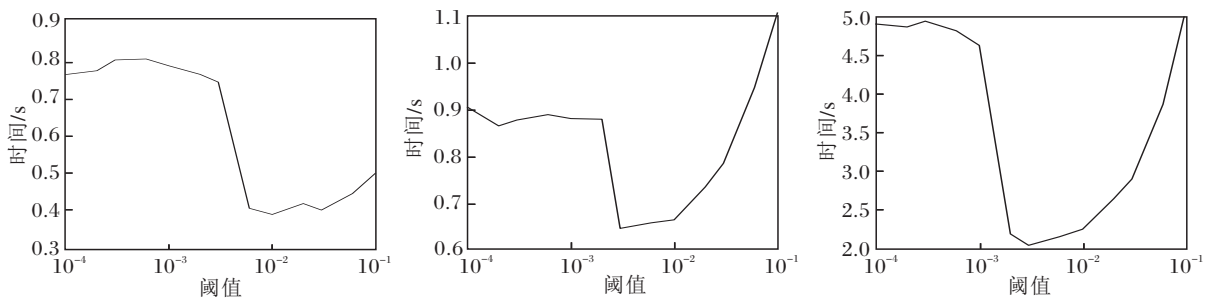
采样时间. 实验表明在 $\varepsilon \in [1/K, 2/K]$ 时,稀疏率和拒绝率都较低,此时采样时间达到最短. 如 3.2 节中所述,这是因为大多数 Λ_{dk} 都很接近,故 $\lambda_{dk} \approx 1/K$,在 ε 刚好大于 $1/K$ 时能较好处理这种情况,故此时稀疏率急剧下降,而拒绝率变化很小. 因此本文推荐使用 $\varepsilon = 2/K$.



(a)20NewsGroup　　(b)Wiki　　(c)RCV1

图 4　ε 对拒绝率和稀疏率的影响

Fig. 4　Effect of ε on rejection rate and sparsity rate



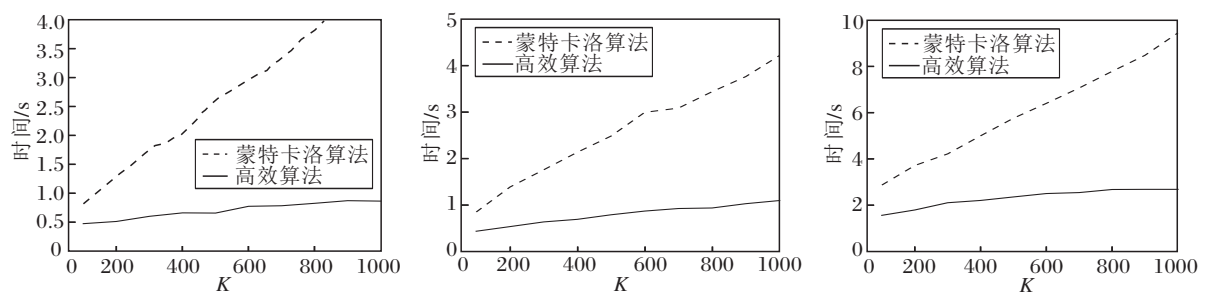
(a)20NewsGroup　　(b)Wiki　　(c)RCV1

图 5　ε 对采样时间的影响

Fig. 5　Effect of ε on sampling time

4.4　训练时间随主题数量的变化

图 6 为训练时间随主题数量 K 的变化情况. 与表 1 中对时间复杂度的分析一致,蒙特卡洛算法的训练时间基本随主题数量线性增长,因此无法适用于主题数多的情况. 而本文的高效算法时间复杂度随主题数量亚线性变化,耗时随主题数增长很慢.



(a)20NewsGroups　　(b) Wiki　　(c)RCV1

图 6　2 种算法的训练时间随主题数量的变化情况

Fig. 6　Training time of 2 algorithms varying with the number of topics

4.5 影响时间复杂度的值

表 1 中使用值如下:文档-主题向量的非零元数量 K_d,λ_{dk} 的非零元数量 K_λ,每篇文档平均的支持向量数量 L_{SV},类别数 L,蒙特卡洛算法训练支持向量机的迭代轮数 I_{svm}. 表 4 为在 3 个数据集上的典型值. 由表可见,稀疏性明显降低时间复杂度,即 $K_d \ll K$, $K_\lambda \ll K$, $L_{\mathrm{SV}} \ll L$. 这些结果表明本文算法的计算量确实远小于蒙特卡洛算法.

表 4 影响时间复杂度的常数取值

Table 4 Values of constants affecting time complexity

数据集	K_d	K_λ	K	L_{SV}	L	I_{svm}
20NewsGroups	11	2	200	1	20	18
Wiki	5	2	400	2	20	53
RCV1	18	25	800	3	103	26

5 结束语

本文针对有监督主题模型训练时间复杂度较高、正比于主题数量的问题,提出 MedLDA 的高效训练算法. 算法基于坐标下降的框架,降低训练分类器的迭代轮数,并通过拒绝采样和高效的预处理将时间复杂度从线性于主题数量降至亚线性. 今后进一步研究方向包括:1)进一步优化支持向量机训练的时间复杂度,降至亚线性于任务数量 L;2)每篇文档分别自适应地调整阈值;3)基于正则化贝叶斯推理框架,推广至更多模型.

参 考 文 献

[1] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] BLEI D M, LAFFERTY J D. Correlated Topic Models[C/OL]. [2019-04-21]. http://people.ee.duke.edu/~lcarin/Blei2005CTM.pdf.

[3] BLEI D M, LAFFERTY J D. Dynamic Topic Models // Proc of the 23rd International Conference on Machine Learning. Berlin, Germany: Springer, 2006. DOI: 10.1145/1143844.1143859.

[4] CHEN J F, ZHU J, LU J, *et al*. Scalable Training of Hierarchical Topic Models. Proceedings of the VLDB Endowment, 2018, 11(7): 826-839.

[5] BLEI D M, GRIFFITHS T L, JORDAN M I, *et al*. Hierarchical Topic Models and the Nested Chinese Restaurant Process[C/OL]. [2019-04-21]. https://people.eecs.berkeley.edu/~jordan/papers/lda-crp.pdf.

[6] MIAO Y S, YU L, BLUNSOM P. Neural Variational Inference for Text Processing // Proc of the 23rd International Conference on Machine Learning. Berlin, Germany: Springer, 2016: 1727-1736.

[7] WEI X, CROFT W B. LDA-Based Document Models for AD-HOC Retrieval // Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2006: 178-185.

[8] WANG C, BLEI D M. Collaborative Topic Modeling for Recommending Scientific Articles // Proc of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2011: 448-456.

[9] LIU S X, WANG X T, LIU J F, *et al*. TopicPanorama: A Full Picture of Relevant Topics // Proc of the IEEE Conference on Visual Analytics Science and Technology. Washington, USA: IEEE, 2014: 183-192.

[10] WANG Y, ZHAO X M, SUN Z L, *et al*. Towards Topic Modeling for Big Data[C/OL]. [2019-04-21]. https://arxiv.org/pdf/1405.4402v1.pdf.

[11] DONG Y P, SU H, ZHU J, *et al*. Improving Interpretability of Deep Neural Networks with Semantic Information // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2017: 4306-4314.

[12] MCAULIFFE J D, BLEI D M. Supervised Topic Models // PLATT J C, KOLLER D, SINGER Y, *et al*., eds. Advances in Neural Information Processing Systems 20. Cambridge, USA: The MIT Press, 2008: 121-128.

[13] LACOSTE-JULIEN S, SHA F, JORDAN M I. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification // KOLLER D, SCHUURMANS D, BENGIO Y, *et al*., eds. Advances in Neural Information Processing Systems 21. Cambridge, USA: The MIT Press, 2009: 897-904.

[14] ZHU J, AHMED A, XING E P. MedLDA: Maximum Margin Supervised Topic Models. Journal of Machine Learning Research, 2012, 13: 2237-2278.

[15] ZHU J, CHEN N, PERKINS H, *et al*. Gibbs Max-Margin Topic Models with Data Augmentation. Journal of Machine Learning Research, 2014, 15: 1073-1110.

[16] YAO L M, MIMNO D, MCCALLUM A. Efficient Methods for Topic Model Inference on Streaming Document Collections // Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2009: 937-946.

[17] LI A Q, AHMED A, RAVI S, *et al*. Reducing the Sampling Complexity of Topic Models // Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2014: 891-900.

[18] YU H F, HSIEH C J, YUN H, *et al*. A Scalable Asynchronous Distributed Algorithm for Topic Modeling // Proc of the 24th International Conference on World Wide Web. New York, USA: ACM, 2015: 1340-1350.

[19] YUAN J H, GAO F, HO Q R, *et al*. LightLDA: Big Topic Models on Modest Computer Clusters // Proc of the 24th International Conference on World Wide Web. New York, USA: ACM, 2015: 1351-1361.

[20] CHEN J F, LI K W, ZHU J, *et al*. WarpLDA: A Cache Efficient O(1) Algorithm for Latent Dirichlet Allocation. Proceedings of the

VLDB Endowment, 2016, 9(10): 744-755.

[21] JIANG Q X, ZHU J, SUN M S, *et al*. Monte Carlo Methods for Maximum Margin Supervised Topic Models // BENGIO S, WALLACH H M, LAROCHELLE H, *et al*., eds. Advances in Neural Information Processing Systems 25. Cambridge, USA: The MIT Press, 2012: 1592-1600.

[22] ZHU J, CHEN J F, HU W B, *et al*. Big Learning with Bayesian Methods. National Science Review, 2017, 4(4): 627-651.

[23] MEI S K, ZHU J, ZHU X J. Robust RegBayes: Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models // Proc of the 31st International Conference on Machine Learning. Berlin, Germany: Springer, 2014: 253-261.

[24] KOYEJO O, GHOSH J. Constrained Bayesian Inference for Low Rank Multitask Learning // Proc of the 29th Conference on Uncertainty in Artificial Intelligence. Bellevue, USA: AUAI Press, 2013: 341-350.

[25] GRIFFITHS T L, STEYVERS M. Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(s1): 5228-5235.

[26] ZHENG X, YU Y L, XING E P. Linear Time Samplers for Supervised Topic Models Using Compositional Proposals // Proc of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2015: 1523-1532.

[27] HSIEH C J, CHANG K W, LIN C J, *et al*. A Dual Coordinate Descent Method for Large-Scale Linear SVM // Proc of the International Conference on Machine Learning. Berlin, Germany: Springer, 2008: 408-415.

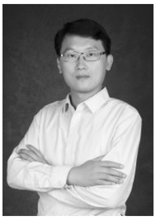
[28] SHALEV-SHWARTZ S, ZHANG T. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. Journal of Machine Learning Research, 2013, 14: 567-599.

[29] LEWIS D D, YANG Y M, ROSE T G, *et al*. RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, 2004, 5: 361-397.

作者简介

陈键飞(通讯作者),博士,主要研究方向为大规模机器学习、概率推理、主题模型. E-mail:chenjian14@ mails. tsinghua. edu. cn.

(**CHEN Jianfei** (Corresponding author), Ph. D. His research interests include large-scale machine learning, probabilistic inference and topic models.)



朱　军(通讯作者),博士,教授,主要研究方向为机器学习. E-mail: dcszj @ mail. tsinghua. edu. cn.

(**ZHU Jun**(Corresponding author), Ph. D., professor. His research interest includes machine learning.)

* *

(上接 735 页)

十、奖励委员会主席

孙富春(清华大学教授)

魏　平(西安交通大学副教授)

十一、大会秘书组

组长:李　杰(西安交通大学)、张　楠(中国自动化学会)、周　馨(中国认知科学学会)、张　璇(西安交通大学)

成员:王　坛、吕爱英、马海彬、叩　颖、唐竹青、周秋硕、王文琪、刘怡兰、安　妮、胡月雯、杨　勐、刘龙军、张　玥、姚慧敏、孙海燕、李逸妍

十二、会务组联系方式

联系人:叩　颖、王文琪

联系电话:010-62522248　18612978071

联系邮箱:caa@ ia. ac. cn

会议网址:http://www. caa. org. cn/cchi2019