基于共享背景主题的 Labeled LDA 模型

江雨燕,李 平,王 清
(安徽工业大学管理科学与工程学院,安徽马鞍山 243002)

摘 要: 隐藏狄利克雷分配(Latent Dirichlet Allocation,LDA)模型被广泛应用于文本分析、图像识别等领域.但由于LDA及其扩展模型多为无监督学习模型,无法将其应用于分类任务中.本文通过研究文档标记与LDA模型中主题的映射关系,提出一种新的Labeled LDA模型(Shared Background Topics Labeled LDA,SBTL-LDA).在SBTL-LDA模型中每个标记除了存在若干个独享的局部主题外,还存在若干个共享的背景(Background)主题,这样可以有效分析不同标记所含主题之间的依赖关系,而文档标记被映射为局部主题和共享主题的组合,因此SBTL-LDA模型可以有效提升文档标记判别的准确性.同时SBTL-LDA模型还可以看成是一种半监督聚类模型,在对文档进行聚类分析的过程中模型可以有效的利用文档的标记信息提升文档聚类效果.实验证明SBTL-LDA模型能够有效解决PLDA模型中主题之间的相似性和依赖关系,具有良好的多标记判别能力,并且具有优于LDA、PLDA模型的文档聚类效果.

关键词: 隐藏狄利克雷分配; 文本分析; 多标记学习; 半监督聚类
中图分类号: TN911.23 **文献标识码:** A **文章编号:** 0372-2112 (2013) 09-1794-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.09.020

Labeled LDA Model Based on Shared Background Topics

JIANG Yu-yan, LI Ping, WANG Qing
(*School of Management Science and Engineering, Anhui University of Technology, Ma'anshan, Anhui* 243002, *China*)

Abstract: LDA (Latent Dirichlet Allocation) is widely used in text analysis and images processing. However, LDA and most of its modifications are unsupervised learning models, which are not appropriate for classification especially multi-label classification problem. Through the study on the multi-label documents and LDA models, this paper proposes a new Labeled LDA model, namely Shared Background Topics Labeled LDA(SBTL-LDA). In this new model, each label has not only a set of local topics, but also has several background (global) topics. Experienmental results show that SBTL-LDA can decrease the affect of similarities and dependence between different topics and because the label of document is mapped as a combination of local topics and shared topics, so it has a high accuracy when learning from multi-Labeled documents. In addition, this model can be viewed as a semi-supervised clustering model which can utilize the information of labels and outperfom other models.

Key words: latent Dirichlet allocation; text analysis; multi-label learning; semi-supervised clustering

1 引言

随着网络技术的发展,网络数据量越来越巨大.如何对网络数据尤其是文本数据进行有效的分析已经逐渐成为机器学习、数据挖掘领域学者们研究的重点.近年来概率主题模型被成功应用于各种需要大规模文本分析的领域,包括自然语言处理、文本分析[1]、信息检索、数据挖掘等领域.在对文档进行分析的过程中,概率主题模型可以通过参数学习,获得由所有词表示的主题(其中不同主题所包含词的分布是不同的).同时可以获得文档在主题层次的低维表示,从而实现文本数据的有效降维以及文档的聚类分析.概率主题模型因其良好的性能,不仅在文本分析领域被广泛应用,在图像、生物学信息发现,以及其它需要对离散数据进行降维的非文本数据处理中的应用也不断发展[2].

藏隐狄利克雷分配(Latent Dirichlet Allocation, LDA)[3]是一个被广泛应用的概率主题模型,其经常被应用于文本、图像分析等领域.但是LDA及其扩展模型[4,5]大多为无监督学习模型,无法将其应用于监督学习的多标记判别中.而且LDA模型经常产生无法解释的主题,因此无法将产生的主题与实际应用的主题相关联.

收稿日期:2013-01-25;修回日期:2013-06-03;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.71172219);安徽省自然科学研究项目省级重点项目(No.KJ2011Z039,No.KJ2013A053)

近几年,一些新的 LDA 模型被提出从而实现了 LDA 模型的监督学习机制. Supervised LDA(sLDA)[6]通过在生成过程中添加一个由主题混合模型中产生的响应(repsonse)变量来实现文档的标记判别和回归分析,但是 sLDA 模型是一种只针对单标记文档的监督学习算法,无法将其应用于多标记的学习任务中. Daniel Ramage 等人在 2009 年提出了 Labeled LDA 模型[7]通过将标记直接映射成为主题来实现文档的多标记判定,但是忽略了人工添加的标记(label)与计算机识别的主题之间的差异性,从而导致模型与文档数据的拟合不足,泛化能力较差. Daniel Ramage 等人在 2011 年提出了 PLDA 模型[8](Partially Labeled Dirichlet Allocation),通过将标记映射成为多个主题的组合来实现文档多标记判别和文档聚类. 通过标记与主题之间的映射,PLDA 实现了人工标记与计算机识别的主题之间的关联,使得 PLDA模型的泛化能力在大多数情况下要优于 Labeled LDA 模型. PLDA 模型同时可以看成是一种有效利用标记信息来提升数据聚类效果的半监督学习算法,在模型训练过程中通过对文档进行基于主题层次的采样,实现了文档的聚类,而这个过程中由于使用标记信息,因此聚类效果要优于 LDA 模型. 但是随着主题数量的增加 PLDA 模型中主题的独立性会受到影响,不同标记所对应的主题可能会存在一定的相似性和依赖关系,从而导致模型预测能力和聚类效果下降. 例如存在两个文档标记分别为“机器学习”和“数据挖掘”,对于介绍“频繁项集挖掘算法”的文档可以很容易的将其归类为“数据挖掘”标记下的“关联分析主题”. 然而对于介绍“隐含狄利克雷分配模型”的文档,其既可能同时属于“机器学习”和“数据挖掘”标记,由于在 PLDA 模型中不存在既属于“机器学习”,又属于“数据挖掘”的主题,因此造成模型不能够很好的拟合文档以及泛化能力较差等问题. 本文针对上述问题提出了一种新的基于文档生成过程的 Labeled LDA 模型(Shared Background Topics Labeled LDA,SBTL-LDA). 该模型与 PLDA 模型相似,通过将标记映射成为多个主题的组合来实现文档的多标记学习. 为了消除模型不同标记下主题的相似性,在该模型中,每个标记除了存在若干个独享的局部主题外,还存在若干个共享的背景(全局)主题. 例如在“机器学习”和“数据挖掘”标记上添加一个共享的主题,这个主题既属于“机器学习”又属于“数据挖掘”. 这样可以使模型生成的主题具有更强的独立性和可区分性. 实验结果表明 SBTL-LDA 模型能够有效解决 PLDA 模型中主题之间的相似性和依赖关系,具有良好的多标记判别能力. 同时 SBTL-LDA 模型可以看成是一种半监督聚类模型,实验结果验证了模型具有优于 LDA、PLDA 模型的文档聚类效果.

2　SBTL-LDA 模型

2.1　SBTL-LDA 模型的生成过程

假设存在 l 个文档标记(label) $L=\{1,2,3,\cdots,l\}$,一个文档 d 所拥有的标记为 $\boldsymbol{\Lambda_d}$,$\boldsymbol{\Lambda_d}$ 是一个 $|L|$ 维的向量,其中每个元素 $\boldsymbol{\Lambda_d}(\boldsymbol{l})\in\{0,1\}$,1 表示 d 中存在标记 l,0 表示文档 d 中不存在标记 l. 每个标记 l 被分配了 K_l 个主题 $Z_l=\{z_{l1},z_{l2},\cdots,z_{lK_l}\}$,每个 Z_l 中的主题仅有一个对应的标记 l,这里称 Z_l 的所包含的主题为局部主题(local topic). 除了局部主题 Z_l,还存在 K_b 背景主题 $Z'=\{z'_1,\cdots,z'_{K_b}\}$,$Z'$ 中的主题可以对应所有的 L 中的标记. 局部主题和背景主题都是关于词汇表 V 中所有词的多项分布,并且它们都由以 η 为参数的狄利克雷先验分布产生. 文档标记 L 的分布是由以 γ 为参数的狄利克雷先验分布产生,模型生成过程的概率图模型如图 1 所示. 其中 μ 变量表示 $\boldsymbol{\Lambda_d}$ 由以 μ 为参数的分布产生,在模型推导过程中由于每个文档的 $\boldsymbol{\Lambda_d}$ 变量是已知的,因此 μ 参数在整个模型的求解过程中不起任何作用,这里添加变量 μ 只是为了便于理解模型的结构.

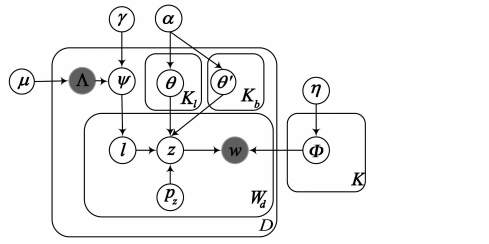


图1　SBTL-LDA的概率图模型

生成文档时首先确定要生成的文档 d 的标记 $\boldsymbol{\Lambda_d}$,然后根据所选择的标记生成标记所对应的分布 $\boldsymbol{\psi}\sim$ Dirichlet(γ)以及标记对应的局部主题和背景主题的分布 $\boldsymbol{\theta}\sim$ Dirichlet(α). 生成第 i 个词 w(word)时,首先根据分布 $\boldsymbol{\psi}$ 选择词对应的标记 l,然后选择对应的主题 Z_l 或全局主题 Z',这里假设文档中词选择 Z_l 的概率为 p_z,那么选择 Z' 的概率为 1-p_z. 添加一个标识变量 $t\in\{0,1\}$,$t=1$ 表示词是由局部主题 Z_l 中的主题产生,$t=0$ 表示词由背景主题 Z' 中的主题产生. 首先由分布 Bernoulli(p_z)中产生值 t,若 $t=1$ 则选择 l 中所对应的主题,若 $t=0$ 则从 Z' 中选择主题. 最后由选择的主题 z_i 来产生 word. 整个文档的生成过程如表 1 所示.

表 1　SBTL-LDA 模型的生成过程

- For each topic $k\in\{1,\cdots,Z_l,1,\cdots,Z'\}$:
  - Generate $\boldsymbol{\varphi}_k=(\varphi_{k,1},\varphi_{k,2},\cdots,\varphi_{k,V})^{\mathrm{T}}\sim\mathrm{Dir}(\cdot\mid\eta)$
- For each document d:
  - Select labels of document $\boldsymbol{\Lambda_d}$
  - Generate $\psi^{(d)}=(\psi_{l_1},\cdots,\psi_{l_{M_d}})\sim\mathrm{Dir}(\cdot\mid\gamma,\boldsymbol{\Lambda_d})$

- Generate local topic distribution
  $\boldsymbol{\theta}_{(d)} = (\theta_{z_{l_1}}, \cdots, \theta_{z_{lK_l}}) \sim \mathrm{Dir}(\cdot \mid \alpha, \boldsymbol{\Lambda_d})$, where $\boldsymbol{\Lambda}(\boldsymbol{l}) = 1$
- Generate background or global topic distribution
  $\boldsymbol{\theta'}_{(d)} = (\theta_1, \cdots, \theta_{\mid K_b \mid}) \sim \mathrm{Dir}(\cdot \mid \alpha, \boldsymbol{\Lambda_d})$
- For each word i in$\{1, \cdots, N_d\}$:
  Generate $l_i \sim \mathrm{Multinomial}(\psi_{l_1}, \cdots, \psi_{l_{M_d}})$
  Generate $t_i \sim \mathrm{Bernoulli}(p_z)$
  If($t_i = 1$)
    Generate $z_i \sim \mathrm{Multinomial}(\boldsymbol{\theta}_{(d)})$
    Generate $w_i \sim \mathrm{Mutinomial}(\varphi_{z_i,1}, \cdots, \varphi_{z_i,V})$
  Else
    Generate $z'_i \sim \mathrm{Multinomial}(\boldsymbol{\theta'}_{(d)})$
    Generate $w_i \sim \mathrm{Mutinomial}(\varphi_{z'_i,1}, \cdots, \varphi_{z'_i,V})$

2.2 模型学习与推理

由于 SBTL-LDA 模型的参数非常复杂,通常不能够精确的计算参数的后验概率.近年来许多 LDA 参数后验概率估计方法被提出,其中包括 EM[9]或者变分 EM 方法[3], expectation propagation(EP)方法[10]以及蒙特卡洛采样方法等.本文使用 collapsed Gibbs sampling[11]方法和马尔科夫链蒙特卡洛方法(MCMC)[12]估计 SBTL-LDA 模型中的参数.

根据文档生成模型及参数设置可知模型的联合似然函数为

$$p(\boldsymbol{w},\boldsymbol{z},\boldsymbol{l},\boldsymbol{t}\mid\boldsymbol{\Lambda},\alpha,\eta,\gamma,p_z)=p(\boldsymbol{w}\mid\boldsymbol{z},\eta)p(\boldsymbol{z},\boldsymbol{l},\boldsymbol{t}\mid\boldsymbol{\Lambda},\alpha,\gamma,p_z) \quad (1)\quad (2)$$

其中式(1)$p(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{\eta})=\int p(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{\Phi})p(\Phi\mid\eta)\mathrm{d}\Phi$ 与标准 LDA[12]中相同,根据假设得出

$$p(w\mid z,\eta)=\int p(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{\Phi})p(\Phi\mid\boldsymbol{\eta})\mathrm{d}\Phi=\prod_{z=1}^{K}\frac{\Delta(\boldsymbol{n}_z+\boldsymbol{\eta})}{\Delta(\boldsymbol{\eta})} \quad (3)$$

其中 $\boldsymbol{n}_z=\{n_z^{(v)}\}_{v=1}^{V}$.

式(2)$p(\boldsymbol{z},\boldsymbol{l},\boldsymbol{t}\mid\boldsymbol{\Lambda},\boldsymbol{\alpha},\gamma,p_z)$的推导过程如下,对$p(\boldsymbol{z},\boldsymbol{l},\boldsymbol{t}\mid\boldsymbol{\Lambda},\boldsymbol{\alpha},\gamma)$进行如下分解

$$p(\boldsymbol{z},\boldsymbol{l},\boldsymbol{t}\mid\boldsymbol{\Lambda},\alpha,\gamma,p_z)=p(\boldsymbol{z}\mid\boldsymbol{l},\boldsymbol{t},\alpha)p(\boldsymbol{l}\mid\boldsymbol{\Lambda},\gamma)p(\boldsymbol{t}\mid\boldsymbol{p}_z) \quad (4)\quad (5)\quad (6)$$

其中式(4)可以分解为如下形式,

$$p(\boldsymbol{z}\mid\boldsymbol{l},\boldsymbol{t},\alpha)=\int p(\boldsymbol{z}\mid\boldsymbol{l},\boldsymbol{t},\Theta,)\ p(\Theta\mid\boldsymbol{t},\boldsymbol{\alpha})\mathrm{d}\Theta \quad (7)$$

根据假设可以得到,

$$\begin{aligned}&p(\boldsymbol{z}\mid\boldsymbol{l},\boldsymbol{t},\Theta)\\&=\prod_{d=1}^{D}\prod_{i=1}^{W_d}p(z_{d,i}\mid l_{d,i},\theta_{d,l_{d,i}})^{t_{d,i}}p(z'_{d,i}\mid l_{d,i},\theta'_{d,i})^{1-t_{d,i}}\\&=\prod_{d=1}^{D}\prod_{i=1}^{W_d}(\theta_{d,l_{d,i},z_{d,i}})^{t_{d,i}}(\theta'_{d,i})^{1-t_{d,i}}\end{aligned} \quad (8)$$

接下来推导式(5)$p(\boldsymbol{l}\mid\boldsymbol{\Lambda},\gamma)$,由于 label 的分布服从以 $\boldsymbol{\gamma}$ 为参数的狄利克雷分布,可以得到

$$p(\boldsymbol{l}\mid\boldsymbol{\Lambda_d},\boldsymbol{\gamma})=\prod_{d=1}^{D}\prod_{j\in\Lambda_d}\frac{\Delta(n_{d,:,:,\cdot}+\boldsymbol{\gamma}}{\Delta(\boldsymbol{\gamma})} \quad (9)$$

根据伯努利分布的性质可以得到式(6)等于

$$p(\boldsymbol{t}\mid p_z)=(p_z)^{n_{t=1}}(1-p_z)^{n_{t=0}} \quad (10)$$

根据式(3)、(8)、(9)、(10)得到,当 $t=1$ 时,词由标记相对应的主题产生,

$$\begin{aligned}&p(l_{d,i}=j,z_{d,i}=k,t_{d,i}=1\mid l_{\neg d,i},z_{\neg d,i},w_{d,i}=v;\alpha,\gamma,\eta)\\&\propto I[j\in\boldsymbol{\Lambda}_d\wedge k\in 1..K_j]\left(\frac{n^{(\neg d,i)}_{\cdot,j,k,v}+\eta}{n^{(\neg d,i)}_{\cdot,j,k,\cdot}+V\eta}\right)\cdot\left(\frac{n^{(\neg d,i)}_{d,j,\cdot,\cdot}+\gamma}{n^{(\neg d,i)}_{d,\cdot,\cdot,\cdot}+\sum_{j'\in\boldsymbol{\Lambda}_d}\gamma}\right)\cdot\left(\frac{n^{(\neg d,i)}_{d,j,k,\cdot}+\alpha}{n^{(\neg d,i)}_{d,j,\cdot,\cdot}+K\alpha}\right)p_z\\&\propto I[j\in\boldsymbol{\Lambda}_d\wedge k\in 1..K_j]\left(\frac{n^{(\neg d,i)}_{\cdot,j,k,v}+\eta}{n^{(\neg d,i)}_{\cdot,j,k,\cdot}+V\eta}\right)(n^{(\neg d,i)}_{d,j,k,\cdot}+\alpha)p_z\end{aligned} \quad (11)$$

当 $t=0$ 时,词由背景主题产生,

$$\begin{aligned}&p(l_{d,i}=j,z_{d,i}=k,t_{d,i}=0\mid l_{\neg d,i},z_{\neg d,i},w_{d,i}=v;\alpha,\gamma,\eta)\\&\propto I[j\in\boldsymbol{\Lambda}_d\wedge k\in 1..K_b]\left(\frac{n^{(\neg d,i)}_{\cdot,j,k,v}+\eta}{n^{(\neg d,i)}_{\cdot,j,k,\cdot}+V\eta}\right)\\&\cdot\left(\frac{n^{(\neg d,i)}_{d,j,\cdot,\cdot}+\gamma}{n^{(\neg d,i)}_{d,\cdot,\cdot,\cdot}+\sum_{j'\in\boldsymbol{\Lambda}_d}\gamma}\right)\cdot\left(\frac{n^{(\neg d,i)}_{d,j,k,\cdot}+\alpha}{n^{(\neg d,i)}_{d,j,\cdot,\cdot}+K\alpha}\right)(1-p_z)\end{aligned} \quad (12)$$

3 试验分析

3.1 聚类

为了验证 SBTL-LDA 在文档聚类问题上的性能,分别使用了 tmc2007 据集、enron 数据集、bibtex 数据集,对模型进行了测试*. tmc2007 数据集为 SIAM 数据挖掘国际会议竞赛数据,其中包含航空飞行时的 22 个安全问题的 28596 份报告,试验中使用了其中的含有 3510 个文档,词数为 49060 个. enron 数据包含了 Enron 公司员工的 1702 个内部电子邮件数据,其中包含 53 个标记例如"公司策略"、"政策建议"等,词数为 1001 个. bibtex 数据集中含有 7395 个文档,其中词数为 1836 个,标记数为 159 个,在实验中使用了其中的 2563 个文档.

在训练时,为了保证模型收敛每次训练迭代次数 1000 次.为了验证聚类效果,采用 perplexity 来度量拟合程度:

$$\mathrm{perlexity}=\exp\left(-\frac{\sum_{m=1}^{M}\log p(\boldsymbol{w_m}\mid\boldsymbol{M})}{\sum_{m=1}^{M}N_m}\right) \quad (13)$$

其中, $\log p(\boldsymbol{w_m}\mid\boldsymbol{M})=\sum_{v=1}^{V}n_m^{(t)}\log\left(\sum_{k=1}^{K}\beta_{k,v}\cdot\theta_{m,k}\right)$.

---

* 试验代码已上传至 https://github.com/liping079094256/LDA-Models.git

试验中还使用了斯坦福大学的 Topic Model Toolkit 开源工具库. http://nlp.stanford.edu/software/tmt/tmt-0.4/

试验中分别对 SBTL-LDA、PLDA、LDA 模型进行了测试,每次保证取相同的 topic 数,图 2 中显示了随着 topic 数目的增加,三个模型的 perplexity 值的变化.

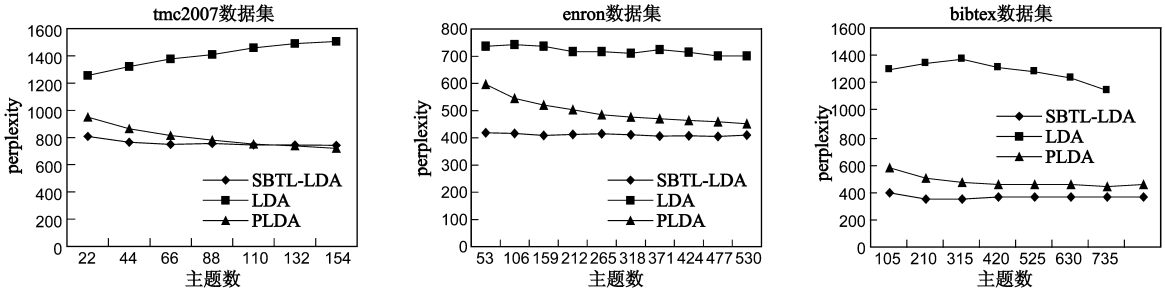


图2　在三个数据集下SBTL-LDA、PLDA、LDA的测试结果

在标准 LDA 模型中文档的聚类可以看成是在用户设定的主题上进行的聚类操作,尽管相对于其它方法 LDA 具有较好的效果,但由于没有利用文档标记信息,因此还有一定的缺陷.在大多数情况下,对文档标记的有效利用可以提高文档聚类的效果.PLDA 和 SBTL-LDA 模型可以看成是一种半监督的聚类模型,文档聚类过程中通过使用文档标记来促进文档主题的生成,因此效果要明显优于标准 LDA 模型.

PLDA 模型将文档标记映射成为若干个主题的组合,这能够有效的利用文档的标记信息.但是也存在一定的问题,当文档标记所对应的主题数不断增加时,文档标记信息不断地被主题所细化,文档不同标记之间的主题可能会存在相似性,这样会导致文档聚类的质量的下降.SBTL-LDA 模型为了避免这一问题的产生,通过添加若干个背景(全局)主题来表示文档不同标记所对应主题之间的相似性,这样可以有效避免相似或相近主题的生成,同时达到提高聚类质量的效果.

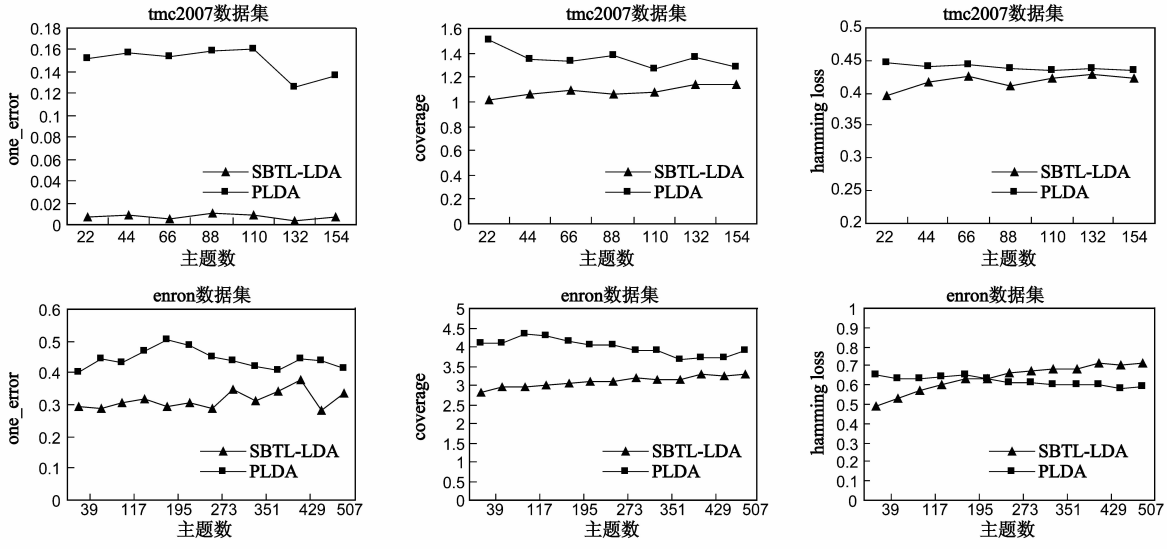
3.2　分类

为了验证 SBTL-LDA 模型在面向分类任务的准确性,采用 tmc2007、enron、bibtex 三个数据集,将模型与 PLDA 进行了比较,分别选取多标记分类模型评价指标中常用的 one-error、coverage、hamming loss 作为准确率判定标准.由于 SBTL-LDA 模型为标记排序(label ranking,LR)算法,无法直接计算出 hamming loss 的值,因此本文中对 hamming loss 进行了一定的变换,新的 hamming loss 的计算方式如下,

$$\text{hamming-loss} = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{|Y_i|} \tag{14}$$

其中 m 表示测试文档的数量,Y_i 表示第 i 个测试文档所包含的标记,$|Y_i|$ 表示第 i 个文档所含标记的数量.首先对第 i 测试文档通过 SBTL-LDA 模型进行排序,然后将前 $|Y_i|$ 个标记 Z_i 作为模型预测的结果,Δ 表示集合 Y_i 和集合 Z_i 中的元素对称不相等,$|Y_i \Delta Z_i|$ 表示两个集合不相等的元素的数量.

测试结果如图 3 所示,从图 3 中可以看出 SBTL-LDA



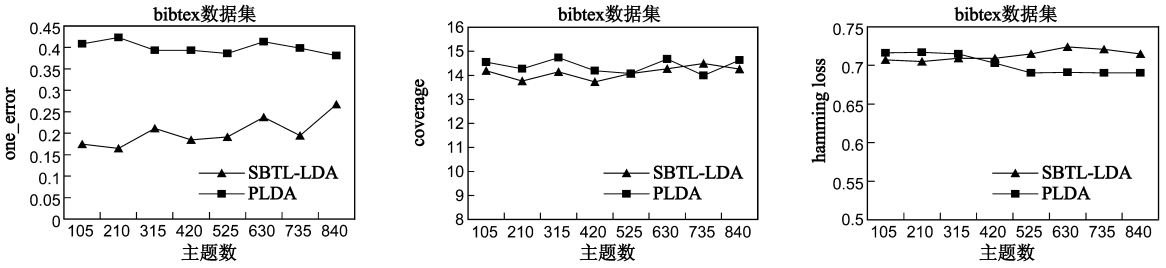


图3 SBTL-LDA与PLDA的分类效果比较

模型与 PLDA 模型相比具有更强的标记判别能力,尤其是在 one-error 和 coverage 评价标准下,SBTL-LDA 模型的错误率要明显低于 PLDA 模型.而在 hamming loss 评价标准下,SBTL-LDA 模型对 tmc2007 数据集的分类效果明显优于 PLDA 模型,当 topic 数目的较大 SBTL-LDA 模型对 enron 和 bibtex 数据集分类的 hamming loss 值逐渐高于 PLDA 模型,但此时的计算代价非常大.因此考虑到计算机运行速度,SBTL-LDA 模型在较低计算代价的情况下具有明显优于 PLDA 模型的分类效果.同时 PLDA 为 Labeled LDA 模型的扩展模型,当 PLDA 每个标记对应的主题数为 1 时,PLDA 转化为 Labeled LDA 模型,因此通过 SBTL-LDA 模型与 PLDA 模型的比较可以同时得到 SBTL-LDA 模型与 Labeled LDA 模型的比较结果.由图 3 的比较结果可以得出,本文提出的模型具有优于 Labeled LDA 和 PLDA 模型多标记判别能力.

本文同时将提出的模型在文本分类等数据集上与其它非概率主题模型算法进行了比较*.其中包括基于 k 近邻方法的 MLKNN[13]算法、BRKNN[14]算法、IBLR_ML[15]算法以及基于决策树的 HMC[16]算法.除了 enron 和 bibtex 数据集外,还是用了 medical 数据集对算法进行测试.Medical 数据集是一个含有 978 个文档的病诊报告集合,词量为 1449 个,包括 45 类疾病编码.试验中采用了两种常用的多标记学习模型评价指标:one-error 和 coverage.SBTL-LDA 模型与其它非概率主题模型的比较如表 2 所示.

在对 enron 和 bibtex 数据集进行分类时,SBTL-LDA 模型的判定能力要明显优于 MLKNN、BRKNN 等四个算法.对 medical 数据集进行分类时 SBTL-LDA 模型的 coverage 值低于 MLKNN、BRKNN、IBLR_ML 和 HMC 算法,而 one_error 的值比 IBLR_ML 算法的值低,但比 MLKNN、BRKNN 和 HMC 的 one_error 值略高.综合以上实验结果,SBTL-LDA 模型具有优良的数据多标记判别能力.

表 2 SBTL-LDA 模型与其它非概率主题模型的比较

算法名称	enron 数据集		bibtex 数据集		medical 数据集	
	one-error	coverage	one-error	coverage	one-error	coverage
SBTL-LDA	0.286	2.954	0.165	13.762	0.424	3.107
MLKNN	0.301	3.237	0.609	24.896	0.352	3.115
BRKNN	0.463	5.428	0.680	27.430	0.423	3.332
IBLR_ML	0.383	3.744	0.828	25.142	0.526	4.280
HMC	0.446	9.199	0.607	31.524	0.220	4.156

3.3 算法效率

在 tmc2007、enron、bibtex 数据集下 LDA、PLDA、SBTL-LDA 模型的平均迭代时间比较如图 4 所示.由于模型复杂度与主题数成线性关系,所以采样时的平均迭代时间随着主题数目的增长而增长.由于 LDA 模型在采样时,要对所有的主题进行采样,而 PLDA 模型只需要对文档所对应的标记的 topic 进行采样,SBTL-LDA 模型只需要对文档对应标记的主题和背景(全局)主题进行采样,因此 LDA 模型的平均迭代时间要明显高于 PLDA 模型和 SBTL-LDA 模型.而 SBTL-LDA 模型采样时的主题数由于加入了背景(全局)主题,所以平均迭代时间要略高于 PLDA 模型.

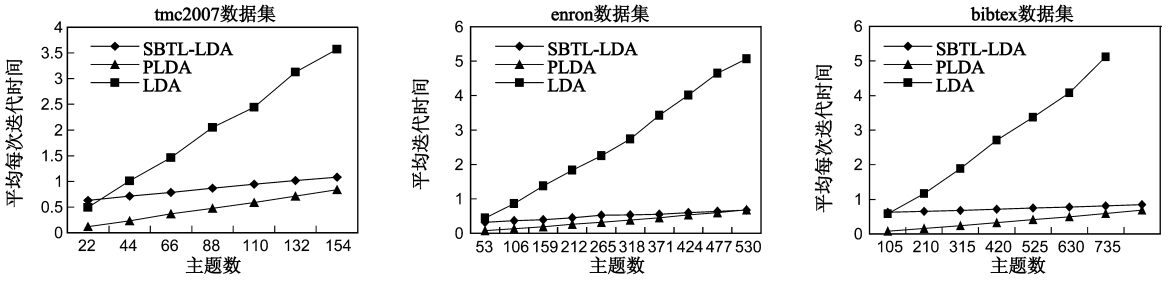


图4 三个数据集下LDA、PLDA、SBTL-LDA算法平均迭代时间

* 试验中使用了多标记学习算法库 Mulan——http://mulan.sourceforge.net/index.html

4 结束语

本文通过研究多标记文档的标记与 LDA 模型中主题的映射关系,提出了一个新的 Labeled LDA 模型,并通过 Gibbs Sampling 方法求解模型参数.实验结果表明该模型具有优良的多标记判别能力和聚类性质.需要指出的是,本文中提出的标记与主题映射关系只是众多映射关系中的一种,还有许多不同的映射方式,这也是监督 LDA 模型研究的一个重要研究方向.

参考文献

[1] 王李冬,魏宝刚,袁杰.基于概率主题模型的文档聚类[J].电子学报,2012,11(11):2346 – 2350.
Wang Li-dong, Wei Bao-gang, Yuan Jie. Document clustering based on probabilistic topic model[J]. Acta Electronica Sinica, 2012,11(11):2346 – 2350.(in Chinese)

[2] 吴永辉,王晓龙,丁宇新, 徐军,郭鸿志.基于主题的自适应、在线网络热点发现方法及新闻推荐系统[J].电子学报,2010,11(11):2620 – 2624.
Wu Yong-hui, Wang Xiao-long, Ding Yu-xin, Xu Jun, Guo Hong-zhi. Adaptive on-line web topic detection method for web news recommendation system[J]. Acta Electronica Sinica, 2010,11(11):2620 – 2624.(in Chinese)

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

[4] Lafferty J D, Blei M D. Correlated topic models[A]. Advances in Neural Information Processing Systems, Proceedings of the 2005 Conference[C]. Vancouver: Bradford Books, 2006. 147 – 155.

[5] Li W, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations[A]. Proceedings of the 23rd International Conference on Machine Learning[C]. New York: ACM, 2006. 577 – 584.

[6] D M Blei, J McAuliffe. Supervised topic models[A]. Advances in Neural Information Processing System[C]. Vancouver, British Columbia Canada: Curran, 2008. 121 – 128.

[7] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[A]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics[C]. Singapore: Springer, 2009. 248 – 256.

[8] Ramage D, Manning C D, Dumais S. Partially labeled topic models for interpretable text mining[A]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2011. 457 – 465.

[9] Hofmann T. Probabilistic latent semantic analysis[A]. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence[C]. Morgan Kaufmann, San Mateo, CA: Morgan Kaufmann Publishers Inc, 1999. 289 – 296.

[10] Minka T, Lafferty J. Expectation-propagation for the generative aspect model[A]. Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence[C]. Morgan Kaufmann, San Mateo, CA: Morgan Kaufmann Publishers Inc, 2002. 352 – 359.

[11] Griffiths T L, Steyvers M. Finding scientific topics[J]. National Academy of Sciences of the United States of America, 2004, 101(Suppl 1): 5228 – 5235.

[12] Griffiths T L, Steyvers M, Blei D M, et al. Integrating topics and syntax[J]. Advances in Neural Information Processing Systems, 2005, 17: 537 – 544.

[13] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038 – 2048.

[14] Spyromitros E, Tsoumakas G, Vlahavas I. An empirical study of lazy multilabel classification algorithms[A]. Proceedings of the 5th Hellenic Conference on Artificial Intelligence[C]. Berlin, Heidelberg: Springer-Verlag, 2008. 401 – 406.

[15] Cheng W, Hüllermeier E. Combining instance-based learning and logistic regression for multilabel classification[J]. Machine Learning, 2009, 76(2 – 3): 211 – 225.

[16] C Vens, J Struyf, L Schietgat, S Dzeroski, H Blockeel. Decision trees for hierarchical multi-label classification[J]. Machine Learning, 2008, 73(2): 185 – 214.

作者简介

江雨燕　女,1966 年生于安徽宣城,安徽工业大学管理科学与工程学院副教授、硕士生导师,主要研究方向为机器学习、CSCW、信息集成.

李　平　男,1987 年生于河北藁城,安徽工业大学管理科学与工程学院硕士研究生,研究方向为机器学习、数据挖掘.

王　清(通信作者)　男,1981 年出生于安徽巢湖,博士.主要研究方向为机器学习、数据挖掘、推荐系统等.
E-mail: wangq@ahut.edu.cn