

基于 Labeled-LDA 模型的文本分类新算法

李文波^{1),2)} 孙 乐¹⁾ 张大鲲¹⁾

¹⁾(中国科学院软件研究所 北京 100080)

²⁾(中国科学院研究生院 北京 100049)

摘 要 LDA(Latent Dirichlet Allocation)模型是近年来提出的一种能够提取文本隐含主题的非监督学习模型. 通过在传统 LDA 模型中融入文本类别信息,文中提出了一种附加类别标签的 LDA 模型(Labeled-LDA). 基于该模型可以在各类别上协同计算隐含主题的分配量,从而克服了传统 LDA 模型用于分类时强制分配隐含主题的缺陷. 与传统 LDA 模型的实验对比表明:基于 Labeled-LDA 模型的文本分类新算法可以有效改进文本分类的性能,在复旦大学中文语料库上 $micro_F_1$ 提高约 5.7%,在英文语料库 20newsgroup 的 comp 子集上 $micro_F_1$ 提高约 3%.

关键词 文本分类;图模型;隐含狄利克雷分配;变分推断
中图法分类号 TP18

Text Classification Based on Labeled-LDA Model

LI Wen-Bo^{1),2)} SUN Le¹⁾ ZHANG Da-Kun¹⁾

¹⁾(Institute of Software, Chinese Academy of Sciences, Beijing 100080)

²⁾(Graduate University of the Chinese Academy of Sciences, Beijing 100049)

Abstract LDA(Latent Dirichlet Allocation) is a recently proposed model which extracts latent topics from text data. In this paper, Labeled-LDA is proposed to enhance the traditional LDA to integrate the class information. Based on Labeled-LDA, a new algorithm is introduced to figure out the latent topics' quantities of each class synergistically. In such a way, Labeled-LDA model avoids compulsive allocation behaviors of the traditional LDA when it is used as a component in classification frame. Experiments on fudan corpus and the comp subset of 20newsgroup corpus show the new method can improve text classification effectiveness; On $micro_F_1$ measure, it approaches an improvement of 5.7% on fudan corpus and 3% on the comp subset of 20newsgroup corpus.

Keywords text classification; graphical model; Latent Dirichlet Allocation (LDA); variational inference

1 引 言

随着信息技术的发展,各类信息资源的存量和增长都呈现海量特征,其中文本数据始终占据重要地位. 如何有效地管理和使用这些文本信息成为当

前的迫切需求,这促进了自动文本分类技术的迅速发展和广泛应用^[1-2].

文本分类研究的核心内容主要包括分类模型和文本表示两个部分. 近年来,文本分类研究的大量工作集中在分类模型方面,基本上是引入和改进机器学习领域的相关成果^[3],得到了 KNN、SVM、

AdaBoost 等高效的分类模型,在分类性能和可用性方面都比之前的知识工程范式有了显著的进步,使得文本分类进入基本可以实用的阶段.与分类模型方面的工作相比,文本表示及其对分类性能的影响的研究相对较少,长期以来文本表示的主要方法是直接采用向量空间模型 VSM(Vector Space Model),这类方法有些基本的变体:词袋模型 BOW(Bag Of Words)和信息检索中的各种词权重计算方法(如 $tf*idf$).

一些新的研究力图通过语言学和统计两种途径对文本表示方法进行拓展.语言学方面的一些研究^[4-5]尝试引入丰富的语言学特征来提高分类的性能,但效果并不理想,而且由于需要比较复杂的语言处理而降低了系统的效率,从而影响了实用性.统计方法的基本思路是挖掘文本的主题信息,典型代表是由 Deerwester 和 Dumais 等人提出的隐含语义索引(LSI)^[6]方法及其概率化改进版 PLSI^[7].LSI 系列方法在文本分类^[8-9]中的应用得到了深入的研究,其降维作用较为显著,但最终的分类性能往往会受损.另外,由于这类模型的参数空间和训练数据呈正比,不利于对大规模或动态增长的语料库进行建模.针对这些问题,研究者借鉴近年发展起来的概率图模型理论和方法,提出了一系列主题模型(Topic Models),主要是以 LDA(Latent Dirichlet Allocation)^[10]为代表的系列模型.

本文针对 LDA 模型应用到文本分类中存在的问题,提出一种改进的 LDA 模型——Labeled-LDA(附加类别标签的 LDA),将类别信息融入传统的 LDA 模型,进而支持文档在全部类别的隐含主题上进行协同分配,有效克服了传统 LDA 模型必须在单个类别中强制分配隐含主题而影响分类性能的问题.

本文第 2 节回顾了相关的研究工作;第 3 节简要介绍传统 LDA 模型并分析其应用于文本分类时存在的问题;第 4 节论述我们提出的 Labeled-LDA 模型及基于 Labeled-LDA 隐含主题分配的文本分类算法;相关实验及分析在第 5 节给出;最后第 6 节是总结.

2 相关工作

主题模型是当前文本表示研究的主要范式,LDA 模型是其典型代表.LDA 模型较之 LSI/PLSI 等模型有着突出的优点:首先 LDA 模型是全概率

生成模型,因此具有清晰的内在结构,并且可以利用高效的概率推断算法进行计算;再者 LDA 模型参数空间的规模与训练文档数量无关,因此更适合处理大规模语料库.LDA 模型已经在机器学习的诸多领域^[10]以及信息检索^[11]中得到应用.另外,有研究者指出在有监督学习环境下 LDA 模型往往表现欠佳^[12],具体到文本分类中也有初步的研究^[9,13],表明该模型对文本分类任务是有效的,但性能并不特别突出.针对 LDA 模型存在的问题,研究人员提出了一些更有力的主题模型.

Blei 等提出了一种 CTM(Correlated Topic Models)模型^[14].该模型的关键之处在于引入逻辑斯蒂-正态分布(Logistic-Normal distribution)取代了 LDA 模型中使用的狄利克雷分布(Dirichlet distribution),用以刻画文档集合的隐含主题.逻辑斯蒂正态分布有 2 组参数分别是均值向量和协方差矩阵:均值向量的作用类似于 LDA 模型中使用的狄利克雷参数,即用以表示隐含主题的相对强弱;而协方差矩阵描述的是每对隐含主题之间的关联程度,这个结构信息在 LDA 模型中是没有的,实际上在 LDA 模型中隐含主题之间可以认为是一种简单的线性结构.利用 CTM 不仅可以分析文本集合的隐含主题构成,而且还可以考察隐含主题之间的联系,这种联系可以用无向图来表示成一种 2 维平面结构.

Li 等提出的 PAM(Pachinko Allocation Model)模型^[13],其核心思想是用有向无环图(DAG)来描述文档中隐含主题之间的结构.该结构原则上可以是任意的,但通常的 PAM 模型采用层次结构.考虑现实应用中的大规模文本数据集合,其隐含主题结构按照层次结构组织是非常自然的,比起平面结构和线性结构的假设更加合理.该研究表明 PAM 较 CTM 和 LDA 具有更好的文本表示能力,在文本分类方面的实验也表明 PAM 优于 LDA.

这些改进方法最基本的思路是通过对文本集合中隐含主题的结构进行更加深入的挖掘从而实现模型的提升.但是这种改进思路的着眼点在于对文本集合的结构本身的精化,而没有考虑向模型中引入其它丰富信息的办法,并且缺少针对特定类型任务(如分类任务)的改进措施.本文的研究正是从这一角度出发,提出将文本数据的类别信息引入到 LDA 模型中,建立 Labeled-LDA 模型以改善分类性能.

3 LDA 模型在文本分类中的应用

3.1 LDA 模型的基本思想

LDA 模型是一种对文本数据的主题信息进行建模的方法. 如图 1(a) 所示, 它假设文档集合(顶部大圆)可以分成若干隐含主题(底部小圆), 而这些隐含主题拓扑结构是线性的, 进一步利用概率推断算法可以将单个文档表示为这些隐含主题特定比例的混合. 如图 1(b) 所示, LDA 模型是典型的有向概率图模型, 具有清晰的层次结构, 依次为文档集合层、文档层和词层. LDA 模型由文档集合层的参数 (α, β) 确定, α 反映了文档集合中隐含主题间的相对强弱, β 刻画所有隐含主题自身的概率分布. 随机变量 θ 表征文档层, 其分量代表目标文档中各隐含主题的比重. 在词层, z 表示目标文档分配在每个词上的隐含主题份额, w 是目标文档的词向量表示形式.

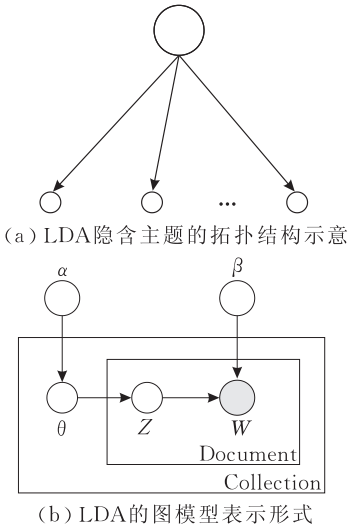


图 1 LDA 模型

构建并使用 LDA 模型的核心问题是隐含变量分布的推断^[15], 也即获得目标文档内部隐含主题的构成信息 (θ, z) . 记 (γ, φ) 为 (θ, z) 的后验分布参数, 则有更新方程^[10] 式(1)^①:

$$\varphi_{ni} \propto \beta_{rw(n)} \exp\{\Psi(\gamma_i)\}, \gamma_i = \alpha_i + \sum_{n=1}^N \varphi_{ni} \quad (1)$$

得到隐含变量分布后即可估计 LDA 模型的参数, 如式(2)所示: 其中 M 是文档总数, d 是文档标识, $\varphi(d)^*_{ni}$ 是由期望步骤得到的最优 φ_{ni} 值.

$$\alpha = \alpha - H(\alpha)^{-1} g(\alpha), \beta_{ij} = \sum_{d=1}^M \sum_{n=1}^N \varphi(d)^*_{ni} w(d)^j_n \quad (2)$$

另外, 还可以利用隐含变量的推断结果计算目

标文档的生成概率值 $p_{LDA}(x|\alpha, \beta)$.

3.2 LDA 模型在文本分类中的应用和不足

LDA 是非监督学习模型, 本身不能直接用于分类, 需要嵌入到合适的分类算法中. LDA 本身是生成模型, 所以与生成型分类算法集成是一种自然的选择. 生成型分类算法的实质是对(对象, 类别)的联合分布 $p(x, c)$ 进行建模, 相应分类任务可表示为如下决策规则:

$$\hat{c} = \arg \max_c p(x|c) p(c) \quad (3)$$

其核心部件是类条件概率 $p(x|c)$, 它表示类别 c “生成”文档 x 的概率, 此处将 LDA 用作类条件概率就得到 LDA 分类器, 如式(4):

$$\hat{c} = \arg \max_c p_{LDA}(x|\alpha_c, \beta_c) p(c) \quad (4)$$

训练阶段, 要为每个类别的文档独立训练相应的子 LDA 模型 (α_c, β_c) . 这样, 类内共享一组主题, 而类间的主题是隔离的. 预测阶段如图 2 所示: 要用每个子 LDA 模型对目标文档 x 进行生成, 以求得各类别上相应的生成概率值 $p_{LDA}(x|\alpha_c, \beta_c)$, 最后根据式(4)即可判定目标文档的类别. 这个过程中需要推断各子 LDA 模型的隐含变量 θ_c (对应后验分布参数是 γ_c), z_c 情况也类似. 但这种分类方法会产生以下问题:

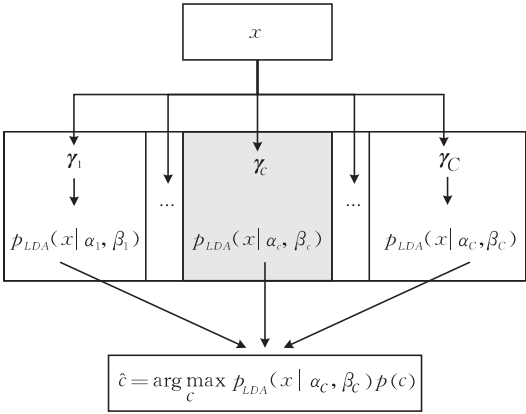


图 2 LDA 模型用于分类

假定目标文档 x 的类别是 c . 那么当用第 c 类的子 LDA 模型生成文档 x 时得到的 γ_c 就是 x 在类 c 上的隐含主题的恰当分配(图 2 中阴影), 这是因为文档 x 和类别 c 中的所有文档共享相同的主题. 但是当用其他类 $c' \neq c$ 的子 LDA 模型生成文档 x 时, 虽然也可以得到相应的 $\gamma_{c'}$, 但这是一种强制的分配, 因为文档 x 所讨论的是类别 c 的隐含主题, 而不是类别 c' 的隐含主题, 这就导致在这些类别上生成

① 式(1)中的 Ψ 表示函数 $\log \Gamma$ 的一阶导函数.

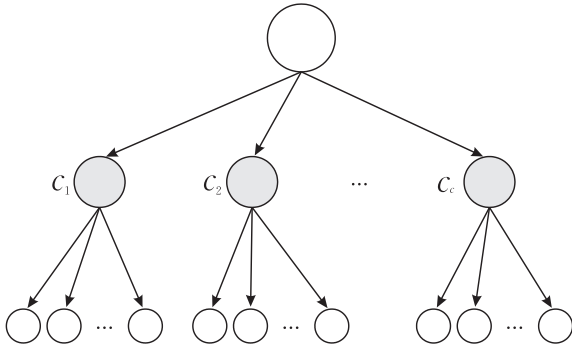
概率的计算产生偏差而降低类分类性能。

4 基于 Labeled-LDA 的分类算法

如上所述,传统 LDA 模型应用于文本分类时存在的问题主要是目标文档在不属于自己的类别上进行生成时就会发生隐含主题的强制分配,针对这个缺陷我们提出 Labeled-LDA 模型以克服这种强制分配行为。

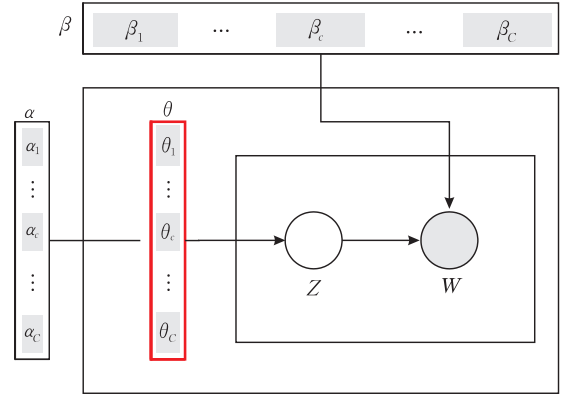
4.1 Labeled-LDA 模型

如图 3(a)所示,Labeled-LDA 模型所要刻画的文档集合的结构是:文档集合(顶部大圆)中的所有文档被按类别划分(中间阴影圆),每一类又讨论若



(a) LDA 的隐含主题拓扑结构示意图

干个隐含主题(底部小圆)。而传统的 LDA 模型所能建模的文档集合的结构是:集合中的所有文档(直接)讨论一组隐含主题(见图 1(a))。这样,Labeled-LDA 模型比标准的 LDA 模型多了一层,即文档类别层。在实现上如图 3(b)所示,这种结构采用了给隐含主题附加类别标签的策略,将类别信息嵌入模型。相应的模型学习算法将(所有类别的)所有隐含主题融合到一个单一模型中,并标明每个隐含主题所属的类别。由于没有另行引入表示类别的随机变量,所以没有增加模型的计算复杂度。附加了类别标签后,模型的各个参数和随机变量与传统的 LDA 模型相比都有变化:



(b) Labeled-LDA 的图模型表示形式

图 3 Labeled-LDA 模型

在模型的文档集合级别,模型参数 α 和 β 按类别被分为 C 组,即 $\alpha = (\alpha_1, \dots, \alpha_c, \dots, \alpha_C)$, $\beta = (\beta_1, \dots, \beta_c, \dots, \beta_C)$ 。参数对 (α_c, β_c) 就定义了类别 c 的那组隐含主题,即用类别标签对每个隐含主题进行标记。向量 α_c 的元素 α_{ci} 表示类别 c 的第 i 个隐含主题的先验概率;矩阵 β_c 的元素 β_{cij} 表示类别 c 的第 i 个隐含主题生成词 j 的概率。进一步,模型中的隐含随机变量,包括文档级的 θ 以及词一级的 z ,也与传统的 LDA 模型不同。目标文档 x 的隐含概念分配向量 θ 也被按类别分为 C 组 $\theta = (\theta_1, \dots, \theta_c, \dots, \theta_C)$,其子向量 θ_c 表示文档 x 在类别 c 的那组隐含主题上的分配额,标量 θ_{ci} 对应文档 x 在类别 c 的第 i 个隐含主题上的分配额。随机变量 z 的情形与 θ 类似,不再赘述。

另外,对于文档 w (即文档 x 的词向量表示形式),在训练阶段要给它附带类别标签,而传统 LDA 模型由于是非监督模型,其训练阶段是不能给 w 附带类别标签的。

Labeled-LDA 模型的训练采用变分 EM 算法,

如式(5)~(7):

$$\varphi_{nci} \propto \delta(c - l(d)) \beta_{ciw(n)} \exp\{\Psi(\gamma_{ci})\}$$

$$\text{E 步: } \gamma_{ci} = \delta(c - l(d)) \alpha_{ci} + \sum_{n=1}^N \varphi_{nci} \quad (5)$$

$$\alpha_c = \alpha_c - H(\alpha_c)^{-1} g(\alpha_c) \quad (6)$$

$$\text{M 步: } \beta_{cij} = \sum_{d=1}^M \sum_{n=1}^N \varphi(d)_{nci}^* w(d)_n^j \quad (7)$$

在 E 步中,与传统的 LDA 模型不同,Labeled-LDA 模型嵌入了类别信息:相关的量 (γ, φ) 都被附加了类别标签 c 。具体地讲,就是通过引入 δ 函数(狄拉克函数^①),指示模型将目标文档的隐含主题的份额向文档所属类别上进行分配。

在 M 步中,Labeled-LDA 模型的参数估计方程与传统的 LDA 模型的相比也有明显的不同:(1)与 E 步类似相关的量 (α, β) 也都附加了类别标签 c ,以实现参数估计按照类别分组;(2)传统的 LDA 模型中一般是假定 α 是可交换的(即 α 的所有分量取值相同,并且在估计过程中一致地缩放),以表示所有隐含主题是平等的。这对传统的 LDA 模型来说是

适合的,因为没有理由假设某个特定隐含主题的高或低,但是对于 Labeled-LDA 模型则不宜如此:不同类别之间语料数量的大小差异以及语料主题的清晰程度都影响类别间隐含概念的相对强弱.因此我们引入了部分可交换(partial exchangeable)狄利克雷分布,即类别内部的隐含主题间可交换而类别之间的隐含主题不可交换.实现中如式(6)所示,参数 α 的估计不再像传统的 LDA 模型中直接在整个向量上进行,而是在每个子向量 α_c 上独立计算.

Labeled-LDA 模型的训练阶段将类别信息嵌入模型中,实现了将文档集合涉及的所有隐含主题和文档集合的类别体系的关联.进而在预测阶段,由于所有类别的所有隐含主题都存在于模型中,对目标文档的隐含主题进行推断时必有类别与之对应,所以总有合适的类可供目标文档进行隐含概念的分配,这就克服了传统 LDA 模型将目标文档在不对应的类上进行强制分配的缺陷.

4.2 将 Labeled-LDA 用于文本分类

传统 LDA 用于分类时,核心是计算并比较类条件密度 $p_{\text{LDA}}(x|\alpha_c, \beta_c)$. 与此不同,我们提出一种新算法——基于 Labeled-LDA 隐含主题分配的分类算法.对于(多类)文本分类任务,该算法基于以下的基本假设:(1)每类文档讨论若干个主题,类间主题的相关程度低于类内主题;(2)一个具体文档讨论的主题是该文档所属类别的主题集合的子集.在这两个基本假设之上实现如下算法:

首先,通过推断目标文档 x 的隐含变量 (θ, z) 而获得其后验分布参数 (γ, ϕ) ,如式(8):

$$\varphi_{nci} \propto \beta_{ciw(n)} \exp\{\Psi(\gamma_{ci})\}, \gamma_{ci} = \alpha_{ci} + \sum_{n=1}^N \varphi_{nci} \quad (8)$$

然后,将主题份额 γ_{ci} 依类别综合得类别份额 Q_c (K_c :类 c 的隐含主题的数量),如式(9):

$$Q_c = \sum_{i=1}^{K_c} \gamma_{ci} \quad (9)$$

最后,选取被分配最大份额的类别作为目标文档的类别,如式(10):

$$\hat{c} = \arg \max_c Q_c \quad (10)$$

基于 Labeled-LDA 隐含主题分配的分类算法有效地避免了传统 LDA 分类算法的缺陷:因为 Labeled-LDA 模型中关于所有类别的信息都存在,所以对于给定文档必有其所属类别,文档在隐含主题分配的过程中没有被强制限定在任何一个类别上进行计算,而是通过公平竞争自然胜出.

5 实 验

实验使用了复旦大学中文文本分类语料库和 20news-group 英文分类语料库.其中,复旦大学中文文本分类语料库约含 20000 篇文档,分成 20 个类.该语料库是一个不平衡语料库,按数量级大致可以分为 2 个级别:其中有 11 个类是小类(每个类别的训练集和测试集中的文档数量都小于 100),9 个类是大类.在我们的实验中,复旦大学中文文本分类语料库的训练集和测试集按照 1:1 的比例划分.语料库 20newsgroup 的 comp 子集,共 5 个类,每类大约含 1000 个文档,训练集和测试集是按照 3:1 的比例划分,这同文献[13]中的实验设置相同以便比较.

5.1 复旦中文文本分类语料库上的实验

这部分实验以 SVM 做分类性能参考,将基于传统的 LDA 模型和 Labeled-LDA 模型的文本分类算法的分类性能进行了比较,SVM 分类器使用 LibSVM^[16],核函数使用线性核,其他参数使用默认设置,实验结果如图 4 所示.

从图 4(a)看,基于 Labeled-LDA 隐含主题分配的分类算法的优势表现在两个方面:(1)该算法性能一致高于传统 LDA 模型的性能,当主题数量/类=10 时,Labeled-LDA 相对于传统 LDA 模型的 $micro_F_1$ 有 5.7% 的提高(从 85.1% 提高到 90.8%),而传统 LDA 模型与 SVM 分类器的性能相当(相差小于 1%).(2)传统 LDA 模型的性能与隐含主题的数量基本无关,而 Labeled-LDA 模型的性能随隐含主题数量的增长而提高,这表明基于 Labeled-LDA 隐含主题分配的分类算法可以有效利用模型隐含主题数量扩张带来的增益.

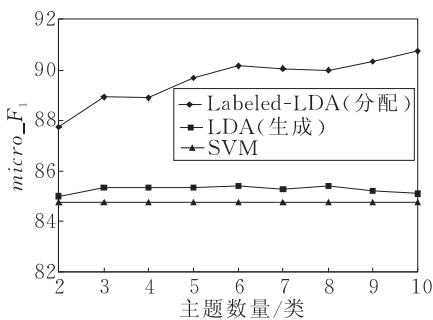
图 4(b)对主题数量/类=10 的情况进行了更进一步的剖析.图中横坐标依据类别所含文档数量升序排列,其中 1~11 是小类,12~20 是大类.基于 Labeled-LDA 隐含主题分配的分类算法的优势表现在:首先,在所有类上该算法性能一致高于传统 LDA 模型(及 SVM)的分类性能.另外,注意到在小类上提升幅度更大,这表明该方法对于处理不平衡分类问题亦有良好的效果.

在曾雪强等的研究中^[17],利用改进的 LSI 模型在复旦大学语料库上进行了测试,其 $micro_F_1$ 性能

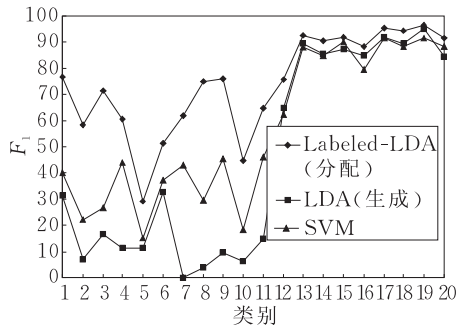
① 狄拉克函数: $\delta(x) = \begin{cases} 1, & x=0 \\ 0, & x \neq 0 \end{cases}$.

指标在 85%~88% 之间. 如前所述, 与 LDA 系列的方法比较, LSI 存在参数规模大和计算复杂度高的问题. 在张启蕊等的研究中^[18], 利用聚合复旦大学语料库上的小类别来抵抗语料分布不平衡. 其宏性能指标($macro_F_1$)得到显著改善, 而 $micro_F_1$ 性能

指标的提高幅度较小(从 85% 提高到 86%), 该方法主要是通过在保持总体分类性能($micro_F_1$)的前提下, 提高小类别上的性能. 而本文的方法首要的是提高总体分类性能($micro_F_1$), 这样就可以自然增大各个类别性能提升的空间.



(a) 基于综合指标 $micro_F_1$ 的比较



(b) 基于各类别上 F_1 的比较 (topics=10)

图 4 在复旦语料库上的性能评估

5.2 20newsgroup 的 comp 子集上的实验

这部分实验以 PAM 模型作为分类性能参考, 如图 5 所示: PAM* 和 LDA* 是引用文献[13]中的结果^①, 同本文的实验结果 Labeled-LDA(分配)和 LDA(生成)比较.

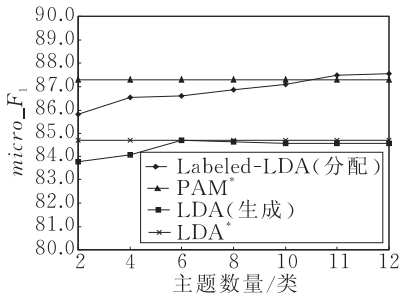


图 5 comp 子集上的性能评估

同上一个实验一样, Labeled-LDA 模型的性能一致高于传统 LDA 模型. 同时, 实验也显示了基于 Labeled-LDA 隐含主题分配的分类算法可以有效利用隐含主题数量扩张带来的增益: 当主题数量/类=12 时, Labeled-LDA 相对于传统 LDA 模型, $micro_F_1$ 有 3% 的提高(从 84.6% 提高到 87.6%). 当主题数量/类>6 时, 传统 LDA 模型的性能趋于稳定, 而 Labeled-LDA 模型的性能随隐含主题数量的增长而平稳提高.

PAM 的核心思想是利用有向无环图(DAG)结构对隐含主题间的关联进行建模以提升模型能力, 而我们的 Labeled-LDA 是通过嵌入类别信息来改进 LDA 模型, 这是两种不同的思路. 上述实验表明这两种方法可以取得相近的分类性能, 这也从侧面

表明 Labeled-LDA 模型的有效性. 另外由于 Labeled-LDA 模型的隐含主题是线性结构, 比 PAM 的 DAG 结构简洁, 所以 Labeled-LDA 在计算效率上比 PAM 更具优势.

6 总 结

本文主要从文本表示方法的角度对改善文本分类性能进行了探索. 我们首先研究了 LDA 模型在文本分类任务中存在的问题: 在计算文档的类生成概率时, 目标文档在其非所属类别上存在强制分配隐含主题的现象, 从而影响了分类性能. 针对这一缺陷, 我们提出了 Labeled-LDA 模型, 该模型为传统的 LDA 增加了建模数据类别的能力. 在此基础上, 通过协同计算隐含主题在全部类别的隐含主题上的分配量实现文本分类, 从而克服了传统的 LDA 在分类过程中将文档强制在单个类别上分配隐含主题的缺陷. 在复旦大学中文文本分类语料库和英文语料库 20newsgroup 的 comp 子集上的实验表明: 基于 Labeled-LDA 的分类算法较传统 LDA 在综合性性能指标 $micro_F_1$ 上分别获得了 5.7% 和 3% 的提高; 再者, Labeled-LDA 以相对 PAM 较低的计算复杂度达到了与其相当的性能.

我们未来拟开展的研究包括: (1) 将 Labeled-

① 文献[13]中的实验采用的是 Accuracy 指标, 本文统一采用 $micro_F_1$, 二者在单标签分类情况下是等价的, 参见文献[19]. 另外由于文献[13]中没有给出主题数量/类, 所以我们用直线画出.

LDA 模型中类别化的隐含主题结构与核方法相结合以进一步提升分类性能;(2) Labeled-LDA 模型并不局限于文本分类,可以应用到其它的监督学习任务中。

致 谢 在此,我们向对本文工作给予建议、帮助的老师 and 同学,尤其是中国科学院软件研究所中文信息处理研究组的黄瑞红同学和冯元勇同学,表示感谢!

参 考 文 献

- [1] Fabrizio Sebastiani. Text categorization//Alessandro Zanzi. Text Mining and its Applications. Southampton, UK: WIT Press, 2005: 109-129
- [2] Su Jin-Shu, Zhang Bo-Feng, Xu Xin. Advances in Machine Learning Based Text Categorization. Journal of Software, 2006, 17: 1848-1859(in Chinese)
(苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展. 软件学报, 2006, 17: 1848-1859)
- [3] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1-47
- [4] Moschitti A, Basili R. Complex linguistic features for text classification: A comprehensive study//McDonald S, Tait J. Proceedings of the ECIR-04. Sunderland: Springer-Verlag. Sunderland, U. K., 2004: 181-196
- [5] Kehagias A, Petridis V, Kaburlasos V G, Fragkou P. A comparison of word- and sense-based text categorization using several classification algorithms. Journal of Intelligent Information Systems, 2003, 21(3): 227-247
- [6] Deerwester S, Dumais S T, Furnas et al. Indexing by latent semantic indexing. Journal of the American Society for Information Science, 1990, 41(6): 391-407
- [7] Thomas Hofmann. Probabilistic latent semantic indexing//Proceedings of the SIGIR. Berkeley, CA, USA, 1999: 50-57
- [8] Schutze H, Hull D A et al. A comparison of classifiers and document representations for the routing problem//Proceed-

ings of the SIGIR-95. Seattle, Washington, USA, 1995: 229-237

- [9] Chen L, Tokuda N, Nagai A. A new differential LSI space-based probabilistic document classifier. Information Processing Letters, 2003, 88(5): 203-212
- [10] Blei D, Ng A, Jordan M. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022
- [11] Wei Xing, Croft W Bruce. LDA-based document models for Ad-Hoc retrieval//Proceedings of the SIGIR. Seattle, Washington, USA, 2006: 178-185
- [12] Vijay Krishnan. Shortcomings of latent models in supervised settings//Proceedings of the SIGIR. Salvador, Brazil, 2005: 625-626
- [13] Li Wei, McCallum Andrew. Pachinko allocation: DAG-structured mixture models of topic correlations//Proceedings of ICML-06. Pittsburgh, Pennsylvania, 2006: 577-584
- [14] Blei D, Lafferty J. Correlated topic models. Advances in Neural Information Processing Systems, 2005, 18: 147-154
- [15] Wainwright M J, Jordan M I. A variational principle for graphical models//Haykin S, Principe J, Sejnowski T, McWhirter J eds. New Directions in Statistical Signal Processing: From Systems to Brain. Cambridge, MA: MIT Press, 2005: 155-202
- [16] Chang Chih-Chung, Lin Chih-Jen. LIBSVM: A Library for Support Vector Machines. Taiwan, China, 2001
- [17] Zeng Xue-Qiang, Wang Ming-Wen, Chen Su-Fen. A text classification model based on the latent semantic structure. Journal of South China University of Technology, 2004, 32: 99-102(in Chinese)
(曾雪强,王明文,陈素芬. 一种基于潜在语义结构的文本分类模型. 华南理工大学学报, 2004, 32: 99-102)
- [18] Zhang Qi-Rui, Zhang Ling, Dong Shou-Bin et al. Effects of category distribution in a training set on text categorization. Journal of Tsinghua University, 2005, 45: 1802-1805 (in Chinese)
(张启蕊,张凌,董守斌等. 训练集类别分布对文本分类的影响. 清华大学学报, 2005, 45: 1802-1805)
- [19] Yang Yi-Ming. An evaluation of statistical approaches to text categorization. Information Retrieval, 1999, 1(1-2): 69-90



LI Wen-Bo, born in 1975, Ph. D. candidate. His research interests include information retrieval, text classification and machine learning.

SUN Le, born in 1971, associate professor. His research interests include information retrieval and natural language processing.

ZHANG Da-Kun, born in 1980, Ph. D. candidate. His research interests include machine translation and natural language processing.

Background

This paper focuses on the new text presentation methods and its application in text classification. Classical text presen-

tation methods mainly include vector space model, n-grams, HMM, and etc. These text presentation methods have been

widely used in natural language processing. Recently, a new type of statistical language models, named as topic model, becomes an active research direction of text presentation. The fundamental target of topic models is to explore the latent structure of document by content analysis. The differences among the topic models are mainly at the assumptions of their topic structure, such as linear array of LDA model, DAG of PAM model, complete graph of CTM model and etc. By means of more reasonable topic structure, more expressive topic model can be obtained.

In their research, the authors propose a new topic model, the Labeled-LDA model, which can encode the class information of document into the traditional LDA model. In this way, they obtain a more capable text presentation method which avoids compulsive allocation behaviors of the traditional LDA when it is used in text classification. Based on the Labeled-LDA model, they introduce a new text classification algorithm to figure out the latent topics' quantities of each class synergistically.

This research is supported by the National Natural Science Foundation Program of China under grants (60773027, 60736044) and the National High Technology Research and

Development Program of China (863 Program) (2006AA010108); Researches on the theory, algorithm and implement of statistical language models and their applications in areas of natural language processing and information retrieval, etc. Statistical language models play a fundamental role in the natural language processing. At the same time, information retrieval also takes the language model as one of the most important paradigms.

This research group has worked on many aspects of statistical language models. Related papers have been published on international conferences (COLING-International Conference on Computational Linguistics, IJCNLP- International Joint Conference on Natural Language Processing, AIRS-Asia Information Retrieval Symposium, etc.) and journals (JCIP-Journal of Chinese Information Processing, etc.). In this paper, they study the topic language models and propose the Labeled-LDA model, which integrates the class information into traditional LDA model. Furthermore, they apply the Labeled-LDA model to text classification. Experiments show that this method can enhance performance of text classification.