计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(11):213−219 [doi: 10.15888/j.cnki.csa.006071]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

同主题词短文本分类算法中 BTM 的应用与改进①

刘泽锦, 王　洁

(北京工业大学 信息学部 多媒体与智能软件技术北京市重点实验室, 北京 100124)

摘　要: 为解决大规模短文本语料库主题模型参数 K 较大导致求解慢的问题, 本文提出 FBTM 模型, 将 BTM 中单个词对采样复杂度由 O(K) 降低 O(1). 针对短文本词语稀疏、描述能力弱的特点, 提出一种结合同主题词对与 FBTM 的短文本分类算法, 首先使用 FBTM 进行主题建模, 将一段滑动窗口内的同主题词对作为特征扩充到原文本中, 然后使用 FBTM 主题分布作为另一部分文本特征. 对特征扩展后的 Weibo 语料库进行分类实验, 结果显示该方法显著提高了分类性能.

关键词: 滑动窗口词对; 快速双词主题模型 (FBTM); 采样; 特征扩展; 短文本分类

引用格式: 刘泽锦,王洁.同主题词短文本分类算法中 BTM 的应用与改进.计算机系统应用,2017,26(11):213–219. http://www.c-s-a.org.cn/1003-3254/6071.html

Application and Improvement of BTM in Short Text Classification Algorithm of the Same Topic

LIU Ze-Jin, WANG Jie

(Beijing Municipal Key Laboratory of Multimedia and Intelligent Software, Faculty of Information, Beijing University of Technology, Beijing 100124, China)

Abstract: In order to solve the problem of large-scale short-text corpus topic model parameter K, the FBTM model is proposed to reduce the sampling complexity from O (K) to O (1). Aiming at the short spelling of short text and the weak description ability, this paper proposes a short text classification algorithm with biterm with the same topic and FBTM. Firstly, we use FBTM to model the text, and extend the same topic biterm in a sliding window as feature in the original text. Then, we use the FBTM topic distribution as another part of the text feature. The results show that this method has significantly improved the classification performance of Weibo corpus.

Key words: sliding window biterm; fast biterm topic model; sampling; feature expension; short text classification

随着互联网络等社交媒体的兴起, 微博、短信、用户 Query 和评论等短文本信息层出不穷, 数量多, 发布速度快, 导致互联网中短文本的规模飞速增长[1-3]. 短文本信息通常涵盖了人们对社会现象的看法与观点, 因此在舆情调查、热点话题的挖掘发现、新词发现、话题识别等领域有着重要的应用, 越来越受到人们的关注[4].

短文本自身包含词汇个数少、信息描述的能力弱, 传统的长文本挖掘方法对于短文本处理的效果不甚理想. 常见短文本分类方法是进行特征扩展, Banerjee 等[5]将 IR 引擎 Lucence 和维基百科文本相结合, 在短文本原有的特征上添加 Lucence 返回的维基百科数据; Beitzel 等[6]在搜索会话中使用近邻查询及其对应点击的 URL(网址) 作为上下文信息, 进而扩展短文本信息并分类; Wang 等[7]提出基于强特征词库的短文本分类方法, 该方法使用隐含狄利克雷分配 (Lattent Dirichlet Allocation, LDA) 和信息增益 (IG) 模型建立辅助特征词表, 使用该特征词表完成短文本的特征扩展; Kalogeratos 等[8]结合传统的向量空间模型提出上下文向量空间模型, 来对短文本进行特征扩展; Zhang 等[9]使用搜索引

① 收稿时间: 2017-03-02; 修改时间: 2017-03-23; 采用时间: 2017-03-27

擎得到外部语料库, 使用外部语料库训练 LDA 模型, 并且在短文本上进行主题推断, 最后结合词向量完成分类. 可见现有的短文本分类方法通常需要引入外部知识库或者外部语料库, 但外部库获取难度大, 且特征扩展效果受外部语料或知识库影响, 导致效率偏低. 吕超镇等[10]提出一种无需外部语料的短文本特征扩展方法, 该方法采用 LDA 提取文本的主题分布, 将主题特征融合到短文本中进行特征扩展. LDA 模型可以通过捕获文档级的词语共现信息, 得到文档的主题分布, 但对于短文本, 其文档级别的词语共现频率十分稀疏, 导致 LDA 的特征扩展效果有限[11].

基于上述考虑, 本文提出一种无需外部语料的短文本特征扩展与分类方法. 首先在 BTM 的基础上改进并提出 FBTM 模型, 将 BTM 中单次采样复杂度由 O(K) 降低到 O(1), 并且给出了单个词语对应主题的求解算法; 接着使用 FBTM 模型对短文本进行主题建模, 将文本中一小段滑动窗口内的两个主题相同的词语组成同主题词对, 同主题词对代表了词项特征的协同作用, 可弥补短文本描述能力弱的不足, 加强词项特征的相互影响, 克服短文本词项特征的稀疏性; 然后用 VSM 表示文本, 使用卡方检验进行特征选择; 最后将 FBTM 主题分布作为词语特征的补充并进行分类. FBTM 通过对词对进行建模加强文档级词语的共现效果, 克服 LDA 的缺陷, 适用于短文本主题挖掘.

1 双词主题模型

双词主题模型 (Biterm Topic Model)[12,13]是 Yan 等提出的针对短文本的主题模型, 通过对滑动窗口内的词对建模来增强文档级的词语共现效果, 此外将整个语料库的词对看做是一个主题分布生成, 使用更丰富的内容来推断主题, 得到整个语料库的主题分布, 图 1 为 BTM 的概率图模型.

w_i, w_j为两个词语构成的词对, φ为主题-词语分布, θ为整个语料库中所有词对的主题分布, z 为对θ采样得到的主题, $|B|$是 BTM 语料库中词对的总数. BTM 的建模过程如下:

1) 对于所有 K 个主题, 采样词语-主题分布φ_z $Dir(\beta)$

2) 为整个训练集合得到主题分布θ $Dir(\alpha)$

3) 对集合中的每个词对:

A. 采样得到主题$z \sim Multi(\theta)$

B. 采样得到词对$w_i, w_j \sim Multi(\varphi_z)$

一般通过 Gibbs 采样对模型求解, 得到每个词对对应的主题:

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) = (n^k_{\neg i} + a_k)\frac{n^{w_i}_{k,\neg i} + \beta}{n_{k,\neg i} + \overline{\beta}} \cdot \frac{n^{w_j}_{k,\neg i} + \beta}{n_{k,\neg i} + \overline{\beta}}$$

其中, $\neg i$表示去除第 i 项, $n^k_{\neg i}$表示去除词语 i 时语料库中主题 k 的计数, $n^w_{k,\neg i}$表示去除当前项 i 时主题 k 下词语 w 的计数, $n_{k,\neg i}$表示去除词语 i 时主题 k 的计数汇总. 主题参数 K 一般根据语料库选取, α, β为对称超参数, $\overline{\beta} = V \cdot \beta$, 其中 V 为词表大小. BTM 每次迭代需对 $|B|$个词对进行多项式分布采样, 得到词对的主题, 单次采样的时间复杂度为 O(K). 文档下主题分布的计算方式为: $p(z|d) = \sum_b p(z|b)p(b|d)$, 具体可参照文献[12,13].

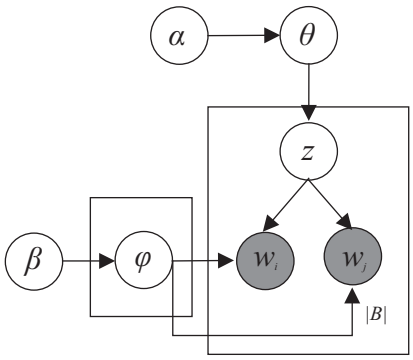

图 1 BTM 主题模型

2 快速双词主题模型

对于海量短文本, 主题参数 K 设置足够大, 才能得到精确的主题模型, 通常在 LDA 与 BTM 中使用 Gibbs 采样求解, 单次采样复杂度为 O(K), Yan 等结合 Metropolis-Hastings 算法与 Alias 表[14], 提出了 AliasLDA 算法, 可将 LDA 单次采样复杂度由 O(K) 降低到 O(1). 但在 BTM 中, 由于对词对进行建模, 致使 AliasLDA 并不适用, 并且 BTM 中并没给出单个词语对应的主题分布 $p(z = k|w, d)$的计算方法. 因此本文在 BTM 的基础上进一步改进并提出快速双词主题模型 (Fast Biterm Topic Model, FBTM), 使得单次采样复杂度由 O(K) 降低到 O(1). 并且 FBTM 还给出了语料库中单个词语对应的主题分布$p(z = k|w, d)$的计算与采样算法.

AliasLDA 中的 Metropolis-Hastings 算法是一种马尔可夫链蒙特卡罗模拟 (MCMC) 方法, 用来获取难以直接采样的目标分布样本. 对于目标分布$p(x)$, Metropolis-Hastings 使用一个易于采样的建议分布 (proposal dis

tribution)$q(x)$, 要求$q(x)$尽可能与$p(x)$相似, 之后以接受率π_t接受$q(x)$的样本, $q(x)$与$p(x)$越相似, 接受率π_t越高, 若$q(x)=p(x)$, 则接受率为$\pi_t=1$, 即直接对$p(x)$采样. 经过若干次迭代, 便可确保最终样本 x 服从目标分布 $p(x)$. 对于初始样本 i, 采样建议分布$j\sim q(j|i)$, 其中接受率π_t为:

$$\pi_t=\min\left\{1,\frac{p(j)q(i|j)}{p(i)q(j|i)}\right\}$$

BTM 中目标分布$p(x)$为复杂度是 O(K) 的多项分布$p(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})$, 便于采样的建议分布$q(x)$有两个选择, 第一个为:

$$q_b(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})=n^k_{\neg i}+a_k$$

第二个为:

$$q_w(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})=\frac{n^{w_i}_{k,\neg i}+\beta}{n_{k,\neg i}+\overline{\beta}}\cdot\frac{n^{w_j}_{k,\neg i}+\beta}{n_{k,\neg i}+\overline{\beta}}$$

两个建议分布与$p(x)$成正比, 但是 BTM 中涉及两个词语组成的词对进行采样, 所以需要对 AliasLDA 进行改进. 下面给出 FBTM 的采样算法, 为了便于说明, 首先表 1 中列出算法所需符号.

表 1 FBTM 算法中的符号表示

n^w_k: 主题k下词语w的计数
n^k: 语料库中词对所属主题k的计数
B: 语料库词对索引表, 其值为对应Alias表
$\widehat{B}$: 语料库中所有文档组成的词对集合
K: 主题数目参数
MH_Steps: Metropolis-Hastings算法步长

FBTM 算法求解过程中会直接保存每个词对对应的主题, 因此建议分布$q_b(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})$直接便可以 O(1) 的复杂度采样, 只需随机选取[1...|B|]中一个随机值对应的词对的主题即为样本. $q_w(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})$是一个关于词对所属主题的多项分布, 首先为每个词对建立一个 hash 索引表 B, 索引表 B 的 key 为词对, value 为词对需建立的 Alias 表. 在采样阶段, 针对索引表 B 中每个 key 对应的 value 构建 Alias 表, 构建过程如图 2 所示, 其复杂度为 O(K). 由于建议分布$q_w(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})$在接下来 K 次采样过程中只会产生轻微改变, 因此通过重复使用该 Alias 表 K 次, 便可以均摊 O(1) 的时间复杂度采样. 综上 FBTM 中改进的 AliasLDA 采样算法如下.

算法 1. FBTM 模型采样算法

1) 初始化阶段

初始化变量: $n^w_k,n^k,B,\widehat{B}$

对于所有文档$m\in[1,M]$

对于文档 m 中的词语, 构建词对索引表 B 与词对集合表$\widehat{B}$

对所有词对$b\in\widehat{B}$, 其中 b 由词语w_i,w_j构成:

随机采样主题: $k\sim Random(1,K)$

增加文档-主题计数: n^k+1

增加主题-词语计数: $n^{w_i}_k+1$

增加主题-词语计数: $n^{w_j}_k+1$

2) 采样阶段

循环执行直到结束:

对于$b\in\widehat{B}$:

对当前 b 的主题分配$\widehat{k}$进行-1 操作:

即$n^{\widehat{k}}-1,n^{w_i}_{\widehat{k}}-1,n^{w_j}_{\widehat{k}}-1$

令新主题$k=\widehat{k}$

对于$l=1...MH_Steps$:

在 B 中查找 b 对应的 Alias 表$B(b)$, 若 Alias 表为空或可用次数为 0, 则重新生成 b 的 Alias 表

采样新主题$j\sim B(j|k;b)$

若$a\sim U(0,1)<\min\left\{1,\frac{p(j)q_w(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})}{p(k)q_w(z_i=j|\overrightarrow{z}_{\neg i},\overrightarrow{w})}\right\}$

那么$k\leftarrow j$

采样$j\sim q_b(j|k)$

若$a\sim U(0,1)<\min\left\{1,\frac{p(j)q_b(z_i=k|\overrightarrow{z}_{\neg i},\overrightarrow{w})}{p(k)q_b(z_i=j|\overrightarrow{z}_{\neg i},\overrightarrow{w})}\right\}$

那么$k\leftarrow j$

对采样得到新的主题 k, 进行+1 操作

即$n^k-1,n^{w_i}_k-1,n^{w_j}_k-1$

此外在 BTM 中只给出了短文本主题分布的计算方法$p(z|d)$, 并没给出文档 d 中词语 w 对应的主题分布$p(z|w,d)$的计算方法. 由于本文需进行同主题词对扩充, 因此需计算$p(z|w,d)$, 并得到该分布得到其样本即为单个词语对应的主题. 待采样阶段完成后, BTM 中 K 个主题下词语对应的分布计算如下:

$$p(w|z=k)=\frac{n^w_{z=k}+\beta}{\sum\limits_w n^w_{z=k}+\overline{\beta}}$$

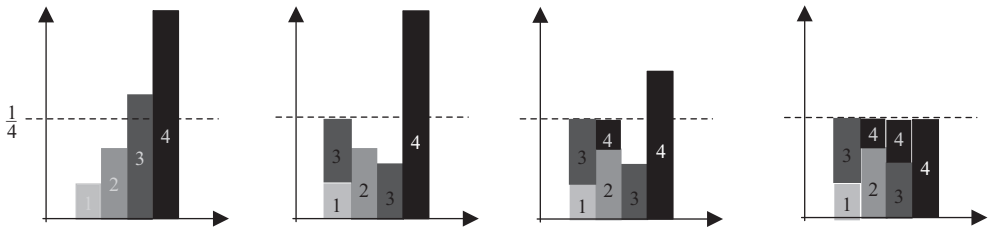

图 2 Alias Table 的构造过程

根据贝叶斯公式, 可得文档 d 中词语 w 对应的主题分布:

$$p(z=k|w,d)=\frac{p(z=k|d)p(w|z=k)}{\sum_{k=1}^{K}p(z=k|d)p(w|z=k)}\propto p(w|z=k)p(z=k|d)$$

得到词语 w 的主题分布 $p(z=k|w,d)$, 对该分布采样即可得到词语对应的主题, $p(z=k|w,d)$ 为一个多项分布, 本文使用的采样算法如算法 2 所示.

算法 2. 词语对应主题的采样算法

输入: 长为 K 的数组 p, 文档 d

对于文档 d 中的每个词语 w

计算 $p(z=k|w,d)$ 并赋值到数组 p

对于 $i=2,...,K$:

$p[i]=p[i-1]+p[i]$

生成 $a\sim U(0,p[k])$

对于 $k=1,...,K$:

如果 $a<p[k]$, 则 w 对应的主题为 k

3 特征扩展与分类

第 2 节给出 FBTM 模型, 本节结合 FBTM 模型提出一种短文本分类方法, 无需引入额外语料库或者字典, 充分利用短文本自身信息. 该方法首先将文本中一小段滑动窗口内的同主题词对作为特征添加到原文本中, 同主题词对特征可在一定程度上克服稀疏性, 弥补短文本描述能力弱的不足, 然后将 FBTM 主题分布作为另一部分特征, 主题特征可对不同类别文本带来一定的区分性, 有助于分类.

对于语料库 D, 短文本 $d\in D$, 文本 d 由词语 $w_1, w_2, ..., w_n$ 构成, d 中词语 w_k 的长度为 $|L|$ 的滑动窗口为:

$$Windows(w_k)=w_k,w_{k+1},...,w_{k+|L|-1},k+|L|-1\le n$$

对于同一窗口内的任意两个词语 w_i, w_j, w_i 在 w_j 之前, 词对 p 的构造方法为:

$$p=w_i\_w_j,\ \ if\ \varphi_{w_i}=\varphi_{w_j}$$

φ_w 为词语 w 所对应的主题, 每个长度为 $|L|$ 的滑动窗口均可产生 $C_{|L|}^2$ 个词对, 根据 FBTM 得到该滑动窗口内每个词语的主题, 将同主题词对到短文本中作为特征补充即可, 特征扩展后的短文本变为 $d=w_1,w_2,...,w_n,p_1,p_2,....$ 扩展过程中可能会产生重复词对, 而短文本中单个词语特征基本出现一次, 因此重复出现的词对仅添加一次即可.

词对特征扩展后之后, 使用 VSM 表示词对与词语, 进行卡方特征选择, 特征选择可以降低特征维度, 提高分类精度. 卡方特征选择中特征 t 对类别 c 的卡方值计算如下:

$$\chi(t,c)=\frac{(AD-BC)^2}{(A+B)(C+D)}$$

这里 A 代表属于类别 c 且包含词项 t 的文档数, B 代表包含词 t 但不属于类别 c 的文档数, C 代表不包含 t 属于类 c 的文档数, D 代表即不包含 t 也不是类 c 的文档的数目. 整体特征扩展与分类框架如图 3 所示, 分类算法如下所述.

输入: 文本集合 D, 词窗口大小 $|L|$, 特征数 N;
输出: 经过特征扩展后的特征集 F.

1. 使用文本集合 D 得到 FBTM 模型 T, 窗口为 $|L|$;
2. 对于训练文本 $d\in D$, 按窗口 $|L|$ 进行同主题词对特征扩展;
3. 对扩展后的文本进行卡方特征选择, 每个类别选择前 N 的特征, 添加至 F;
4. 使用 TF-IDF 方法计算词语与词对特征的权重;
5. 使用模型 T 对文档进行主题推断, 将主题特征添加至特征集 F.
6. 对于特征扩展后的文本, 本文采用支持向量机 (support vector machines, SVM) 方法来进行试验.

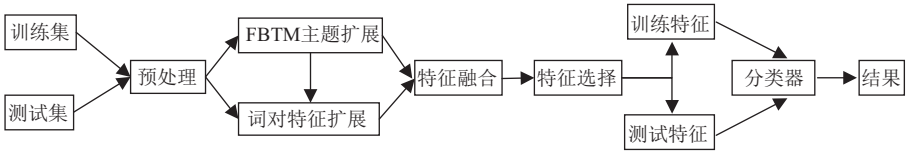

图 3 特征扩展与分类流程

5 实验结果及分析

5.1 实验数据及评价标准

实验采用 NLPIR 平台提供的新浪微 Weibo 数据集, 共记 7774 条数据, 分为财经, 美食, 房产, 教育, 汽车, 体育 6 个类别, 经过 Hanlp 分词, 去除停用词后文本的平均长度约为 13, 数据集情况如表 2 所示.

表 2 语料库样本

编号	类别	数量
1	财经	1678
2	美食	1553
3	房产	453
4	教育	1756
5	汽车	879
6	体育	1455

试验采取了准确率 (precision)Pr、召回率 (recall)Re、F1 值 (F1-measure) 作为评价标准, 计算如下:

$$\mathrm{Pr}=\frac{TP}{TP+FP},\mathrm{Re}=\frac{TP}{TP+FN},F1=\frac{2\times\mathrm{Pr}\times\mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}}$$

5.2 实验结果

随机从语料库抽取 2000 篇文本作为主题模型的实验语料, 设置参数$\alpha=K/50,\beta=0.01$, 迭代次数为 100, 实验结果如图 4 所示.

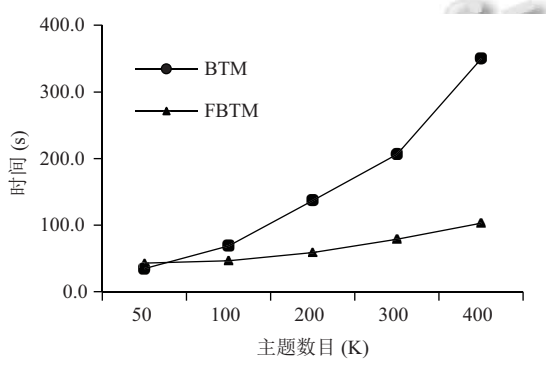

图 4 采样时间随主题数的变化

可见当主题数目 K 达 50, BTM 与 FBTM 的求解时间基本相同, 而当$K>50$以后, FBTM 完成采样的时间远远少于 BTM, 而且随着 K 的增大, 这一差距会越来越明显, 可见 FBTM 方法适用于大规模语料库中主题数 K 偏大的情况.

主题模型质量的衡量方法可采用相关度得分[15] (coherence score), 使用每个主题 z 对应多项分布概率最大的前 T 个词语, $V^{(z)}=(v_1^{(z)},...,v_T^{(z)})$, 则主题 z 的相关度得分为:

$$C(z,V^{(z)})=\sum_{t=2}^{T}\sum_{l=1}^{t}\log\frac{D(v_t^{(z)},v_l^{(z)})+1}{D(v_l^{(z)})}$$

整体相关度得分即对 K 个主题相关度得分取平均: $\frac{1}{K}\sum_z C(z,V^{(z)})$, 表 3 为对比 FBTM, BTM 与 LDA 的主题质量的实验, 由数据可见 FBTM 主题质量基本与 BTM 持平, 两者效果均好于 LDA, 文本长度越短, LDA 效果越差, BTM 与 FBTM 得益于直接对词语共现来建模, 解决了短文本中词语稀疏的问题.

表 3 相关度得分对比

T	LDA	BTM	FBTM
5	-27.7	-24.7	-24.9
10	-123.3	-115.3	-116.3
15	-280.8	-274.0	-274.8
20	-532.4	-507.3	-506.4
25	-850.2	-812.2	-815.7
30	-1244.0	-1203.3	-1209.4

表 4 展示了设置 K=3 时 BTM 与 FBTM 主题分布下的概率由大到小的 top 主题词, 由于两个模型的相关度得分基本持平, 其 top 主题词基本类似, 可见 FBTM 加快了采样速度, 且与 BTM 模型的主题质量基本相同.

表 4 FBTM 与 BTM 的 Top 主题词

K	BTM	FBTM
1	房产 市场 房地产 城市 楼市 住宅	房价 市场 开发商 调控 政策 楼市
2	旅游 滑雪 温泉 景区 雪山 世界	旅游 活动 旅行 滑雪 温泉 景区
3	汽车 车型 车展 奥迪 车辆 全新	汽车 车型 品牌 发动机 车展 利率

实验使用 LibSVM 对不同短文本特征扩展方法进

行分类对比, 实验中设置滑动窗口参数|L| = 10, 主题模型参数 K=12, 卡方特征数 3000, 基准方法参照为传统的向量空间模型 VSM 与文献[10]方法 (VSM+LDA), 综合比较准确率, 召回率以及 F1 值的实验结果如图 5 所示.

具体实验数据如表 5 所示, 由实验数据可见直接使用 VSM 模型, 文本分类效果 F1 值仅为 75.1%, 当添加词对后, 分类的 F1 值提升了 4.4%, 上升到 79.9%, 得益于滑动窗口词对增加了短文本的描述能力, 进一步添加主题分布特征, 使用 FBTM 主题特征扩展方法, F1 值特征到 83.2%, 效果好于文献[10]方法, FBTM 通过直接对词对建模, 使得主题分布更加准确. 整体实验结果符合预期, 验证了使用 FBTM 对短文本进行主题建模的有效性, 以及对文本进行主题特征扩展与同主题词对特征扩展方法的有效性.

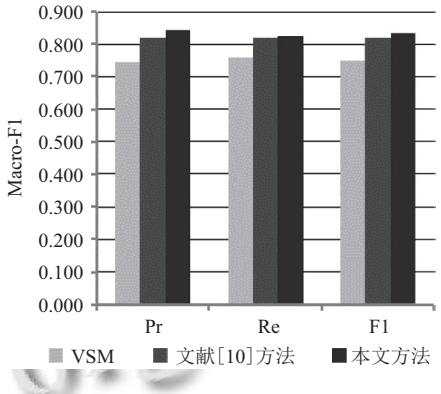

图 5 三种方法分类 F1 值比较

表 5 实验结果对比

	VSM			VSM+词对			文献[10]方法			本文方法		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
体育	0.847	0.567	0.679	0.866	0.777	0.819	0.872	0.868	0.870	0.881	0.918	0.899
教育	0.766	0.868	0.814	0.882	0.865	0.873	0.873	0.870	0.873	0.894	0.879	0.886
房产	0.401	0.857	0.546	0.499	0.857	0.641	0.522	0.853	0.648	0.537	0.881	0.667
汽车	0.505	0.830	0.630	0.624	0.839	0.715	0.777	0.840	0.750	0.682	0.840	0.780
美食	0.790	0.740	0.764	0.905	0.714	0.798	0.918	0.743	0.821	0.922	0.761	0.834
财经	0.674	0.755	0.713	0.818	0.834	0.825	0.904	0.791	0.839	0.942	0.801	0.866
Macro	0.742	0.760	0.751	0.801	0.789	0.795	0.822	0.819	0.820	0.842	0.823	0.832

5.3 实验参数的影响

本节通过实验分析参数对本文方法的影响. 影响分类效果的一个关键因素是主题数目 K, 图 6(a) 展示了模型的 F1 值与 K 值的关系, 当设置 K=10 时, 模型的 F1 值达到最大, 当 K<10 时, 随着 K 的增大, 模型的 F1 值也会增大, 当 K>10 的时候, 模型的 F1 值反而开始下降, 因为主题数 K 太大, 导致 FBTM 模型引入噪音. 另一个比较关键的参数是滑动窗口大小|L| 设置 |L|=2, 其效果等同于同主题的 Bi-Gram 扩展, 这时可以提升分类效果, 但在 FBTM 中|L|=2 得到的仍然是一个稀疏的模型, 并不能准确抓住文档级词语的共现, 导致主题质量效果比较差, 当|L|设置偏大时, 可能会产生无意义的词对, 增加计算复杂度, 导致计算复杂度偏高, 图 6(b) 为通过实验得到|L|的大小, 可见当|L|=6 时得到的 F1 值基本不再上升. 卡方特征选择的数目也会对实验结果产生影响, 由图 6(c) 可见, 特征数为 0 代表不进行特征选择时, 随着特征数目的增加, 其 F1 值不断增大, 当特征数目为 3000, |L|=6, K=12 时 F1 值达到最大, 即对每个类别下取卡方值最大的前 3000 个特征即可达到较好的分类效果.

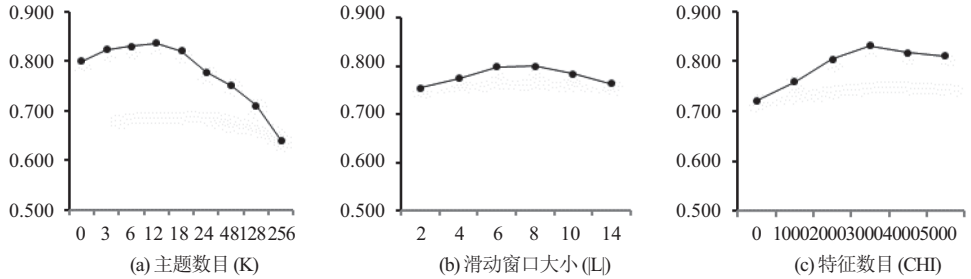

图 6 参数对 F1 值的影响

6 结束语

本文主要有以下两点贡献:

1) 改进 BTM 模型, 将 BTM 中单个词对采样复杂度由 O(K) 降低到 O(1), 并且给出$p(z|w,d)$的计算方式以及词语对应主题的采样算法;

2) 通过对短文本进行滑动窗口划分, 提取窗口内的同主题词对作为特征扩展融入到原始文本特征内, 然后通过 FBTM 模型提取主题分布做进一步特征融合.

通过实验证明, 本文特征扩展算法是有效的, 可以提高文本分类的准确率, 召回率及 F1 值. 进一步考虑结合上下文信息引入更高效特征权重计算方法来实现短文本分类.

参考文献

1 黄九鸣, 吴泉源, 刘春阳, 等. 短文本信息流的无监督会话抽取技术. 软件学报, 2012, 23(4): 735–747.

2 Sriram B, Fuhry D, Demir E, *et al*. Short text classification in twitter to improve information filtering. Proc. of the 33rd International ACM SIGIR Conference on Research and development in Information Retrieval. Geneva, Switzerland. 2010. 841–842.

3 Ferragina P, Scaiella U. TAGME: On-the-fly annotation of short text fragments (by wikipedia entities). Proc. of the 19th ACM International Conference on Information and Knowledge Management. Toronto, ON, Canada. 2010. 1625–1628.

4 王仲远, 程健鹏, 王海勋, 等. 短文本理解研究. 计算机研究与发展, 2016, 53(2): 262–269. [doi: 10.7544/issn1000-1239.2016.20150742]

5 Banerjee S, Ramanathan K, Gupta A. Clustering short texts using wikipedia. Proc. of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, the Netherlands. 2007. 787–788.

6 Beitzel SM, Jensen EC, Frieder O, *et al*. Improving automatic query classification via semi-supervised learning. Proc. of 5th IEEE International Conference on Data Mining. Houston, TX, USA. 2005. 42–49.

7 Wang BK, Huang YF, Yang WX, *et al*. Short text classification based on strong feature thesaurus. Journal of Zhejiang University Science C, 2012, 13(9): 649–659. [doi: 10.1631/jzus.C1100373]

8 Kalogeratos A, Lika A. Text document clustering using global term context vectors. Knowledge and Information Systems, 2012, 31(3): 455–474. [doi: 10.1007/s10115-011-0412-6]

9 Zhang H, Zhong GQ. Improving short text classification by learning vector representations of both words and hidden topics. Knowledge-Based Systems, 2016, 102: 76–86. [doi: 10.1016/j.knosys.2016.03.027]

10 吕超镇, 姬东鸿, 吴飞飞. 基于 LDA 特征扩展的短文本分类. 计算机工程与应用, 2015, 51(4): 123–127.

11 Blei DM. Probabilistic topic models. Communications of the ACM, 2012, 55(4): 77–84. [doi: 10.1145/2133806]

12 Pan YL, Yin J, Liu SP, *et al*. A biterm-based dirichlet process topic model for short texts. International Conference on Computer Science and Service System. 2014.

13 Hoffman MD, Blei DM, Bach F. Online learning for latent dirichl*et al*location. International Conference on Neural Information Processing Systems. Curran Associates Inc. 2010. 856-864.

14 Yan XH, Guo JF, Lan YY, *et al*. A biterm topic model for short texts. Proc. of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil. 2013. 1445–1456.

15 Mimno D, Wallach HM, Talley E, *et al*. Optimizing semantic coherence in topic models. Proc. of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom. 2011.