基于LDA的多粒度主题情感混合模型

欧阳继红[1,2],刘燕辉[1,2],李熙铭[1,2],周晓堂[1,2]
(1.吉林大学计算机科学与技术学院,吉林长春 130012;2.符号计算与知识工程教育部重点实验室,吉林长春 130012)

摘 要: 主题情感混合模型(Reverse-Joint Sentiment/Topic Model;Joint Sentiment/Topic Model)能够有效地同时抽取文档的主题和情感信息,在情感分析领域受到广泛的关注,因为没有考虑整体分布与局部分布的关系,导致分类效果不佳且不稳定.本文同时考虑两个粒度上的情感/主题分布——文档级和局部,提出多粒度的主题情感混合模型(MG-R-JST;MG-JST).MG-R-JST/MG-JST在文档级分布和局部分布的共同作用下生成单词的情感/主题;使用吉布斯采样进行模型推理,并给出了推理过程;在MR与MDS数据集上进行实验,实验结果表明本文算法分类效果优于主题情感混合模型,且稳定性更好.

关键词: LDA;主题情感混合模型;情感分析;多粒度
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)09-1875-06
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.09.029

Multi-Grain Sentiment/Topic Model Based on LDA

OUYANG Ji-hong[1,2], LIU Yan-hui[1,2], LI Xi-ming[1,2], ZHOU Xiao-tang[1,2]
(1. *College of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*;
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin* 130012, *China*)

Abstract: The topic and sentiment unification model (Reverse-Joint Sentiment/Topic Model; Joint Sentiment/Topic Model) can effectively extract information of topic and sentiment simultaneously and receives wide attention in the field of sentiment analysis, because it does not consider the relationship between the overall distribution and local distribution, so the classification performance is not good and stable. This paper proposed the multi-grain topic and sentiment unification model (MG-R-JST; MG-JST) by taking into account both grains on sentiment/topic distribution—document-level and local-level. MG-R-JST/MG-JST generated the sentiment/topic of words on the effect of the document-level and local-level distribution. we used Gibbs sampling for model inference and showed the process. Experiments on the dataset of MR and MDS demonstrate the effectiveness of the proposed method, and the classification performance is better and more stable than the topic and sentiment unification model.

Key words: LDA; topic and sentiment unification model; sentiment analysis; multi-grain

1 引言

随着互联网的普及与发展,越来越多的用户在网上发表评论,表达对某个事件或产品的观点和看法.如何有效地分析海量的评论信息,了解大众舆论的观点走向和情感倾向具有重要的现实意义[1].情感分析(sentiment analysis)可以对主观性文档进行分析和处理,挖掘文档背后隐含的情感信息,是信息检索和自然语言处理领域研究的热点问题.

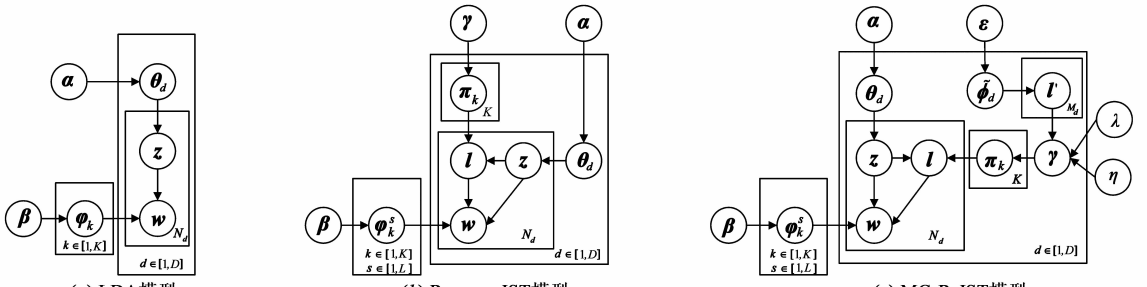
情感分类是情感分析的主要任务之一,通过分析文档,来识别其情感倾向——主要包含正向(褒义)和负向(贬义)情感[1].传统的机器学习算法应用于情感分类的时候,分类效果并不是十分稳定[2]:该类算法大都是有监督学习算法,需要标注过的训练样本,而训练样本有时难以获取;情感特征词的确定依赖于领域信息,使得该类算法领域适应性比较差[3].

主题模型(Latent Dirichlet Allocation, LDA)[4~6]是一种能有效捕捉文档隐含主题的无监督学习方法,可以在一定程度上改善情感分类在不同领域的适应性[7].近年来,研究者们提出一些基于LDA的情感分类模型.主题情感混合模型(Reverse-Joint Sentiment/Topic Model, Reverse-JST; Joint Sentiment/Topic Model, JST)[3,8,9]同时考虑

收稿日期:2014-04-08;修回日期:2014-08-11;责任编辑:梅志强
基金项目:国家自然科学基金(No.61170092, No.61133011, No.61103091)

了单词的主题和情感,在 LDA 基础上加入情感层,可以对文档中不同主题与情感中的单词分布进行分析.Dependency-Sentiment-LDA[3]考虑了相邻单词间的局部依赖性——顺连和转折关系.Dynamic Joint Sentiment-Topic Model[10]则主要检测不同时段的主题与情感的变化.

Reverse-JST/JST 通过不同的主题数目来分析不同粒度下的主题与情感的分布关系,但其只考虑了单词的局部情感/主题分布,分类效果及其稳定性容易受局部的否定表达和主题数目的影响.分类效果是在不同主题数目上的分类准确率.稳定性是同一个主题数目上分类准确率的波动性.由于主题情感混合模型是概率统计的方法,对模型参数的计算属于近似的估计方法,每次对参数的估计会有所不同.基于此分析,本文考虑整体分布与局部分布的关系,提出多粒度的主题情感混合模型(MG-R-JST; MG-JST),力求改善主题情感混合模型的分类效果及其稳定性.



(a) LDA模型 (b) Reverse-JST模型 (c) MG-R-JST模型

图1

2 Reverse-JST 与 JST 模型

LDA 是一种应用于学习文档隐含主题的贝叶斯模型.如图 1(a)所示,LDA 假设在可见的文档层和单词层中存在一个隐含的主题层,每篇文档是主题的多项分布,每个主题则是单词的多项分布.

Reverse-JST 假设文档中存在隐含的情感信息,通过在 LDA 中引入情感层,扩展成一个四层的贝叶斯网络.Reverse-JST 中,每篇文档是主题的多项分布,每个主题是情感多项分布,而每个主题情感对是单词的多项分布.图 1(b)是 Reverse-JST 的概率图模型,其文档的生成过程如下:

(1)对于每个主题情感对(k,s)

(ⅰ)生成主题情感对的单词分布 $\boldsymbol{\varphi}_k^s \sim \mathrm{Dir}(\boldsymbol{\beta})$

(2)对于每篇文档 d

(ⅰ)生成主题分布 $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$

(ⅱ)对于每个主题 k

①生成主题的局部情感分布 $\boldsymbol{\pi}_{d,k} \sim \mathrm{Dir}(\boldsymbol{\gamma})$

(ⅲ)对于该文档中的每个单词 $\boldsymbol{w}_i$

①生成主题 $z \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$

②生成局部情感 $l \sim \mathrm{Multinomial}(\boldsymbol{\pi}_{d,z})$

③生成单词 $\boldsymbol{w}_i \sim \mathrm{Multinomial}(\boldsymbol{\varphi}_z^l)$

JST 同样在 LDA 中引入情感层,不同在于 JST 将每篇文档表示为情感的多项分布,每个情感是主题的多项分布.JST 先生成每篇文档的情感分布,再生成每个情感的局部主题分布,最后根据情感主题对中的单词分布生成每个单词.表 1 总结了本文所涉及符号的定义.

表 1 符号及说明

D	文档数目	z	文档集单词的主题分配序列
K	主题数目	l	文档集单词的情感分配序列
L	情感数目	$\boldsymbol{\varepsilon}$	文档级情感分布的 Dirichlet 参数
$\boldsymbol{w}$	文档集单词序列	$\boldsymbol{\varphi}_k^s$	主题情感对(k,s)中的单词分布
l'	文档级情感分配序列	$\tilde{\boldsymbol{\phi}}_d$	文档 d 的文档级情感(主题)分布
z'	文档级主题分配序列	$\boldsymbol{\gamma}$	情感分布的 Dirichlet 参数
$\boldsymbol{\theta}_d$	文档 d 的主题分布	$\boldsymbol{\alpha}$	主题分布的 Dirichlet 参数
$\boldsymbol{\pi}_k$	主题 k 的情感分布	$\boldsymbol{\beta}$	单词分布的 Dirichlet 参数

3 本文算法

3.1 MG-R-JST 与 MG-JST 模型的生成过程

由于主观性文档中一些局部否定表达的存在,可能会使情感词表达出相反的情感色彩,出现情感偏移的现象.Reverse-JST/JST 基于词袋假设,在生成单词的局部情感/主题时,由于没有考虑文档的整体分布,当情感偏移或主题数目较多时(每个主题对应的样本数目则相应地减少),其对局部情感/主题估计的不确定性会增加,从而影响到分类效果及其稳定性.

基于上述想法,我们提出多粒度的主题情感混合模型(MG-R-JST; MG-JST).MG-R-JST/MG-JST 考虑两个粒度上的情感/主题分布——文档级和局部,局部分布的生成受到文档级分布的影响.相对于局部分布,文档级分布更能代表文档的整体情感倾向,MG-R-JST/MG-JST 通过文档级分布生成局部分布,可以增加对局部情感/主题估计的准确性,从而提升分类效果并增强其稳定性.

首先介绍 MG-R-JST 模型.MG-R-JST 同时考虑文档级情感和局部情感分布,其概率图模型如图 1(c)所示,MG-R-JST 先生成文档级情感分布 $\tilde{\boldsymbol{\phi}}_d$,在该分布下采样 M_d 次得到文档级情感计数,应用公式(1)计算局部情感分布的先验 $\boldsymbol{\gamma}$[11],然后如 Reverse-JST 所示,生成每篇文档的主题分布,再生成每个主题的局部情感分布,最后根据主题情感对中的单词分布生成每个单词.

$$\boldsymbol{\gamma}=\left[\lambda*\frac{\widetilde{N}_{d,1}}{M_d}+\eta,\cdots,\lambda*\frac{\widetilde{N}_{d,s}}{M_d}+\eta,\cdots,\lambda*\frac{\widetilde{N}_{d,L}}{M_d}+\eta\right] \tag{1}$$

其中,$\widetilde{N}_{d,s}$表示文档 d 中采样 M_d 次时文档级情感 s 的计数,$1\leqslant s\leqslant L$.λ 为权重因子,λ 越大,表示局部情感分布越倾向于文档级情感分布,反之亦然.η 为平滑参数,用来平滑文档级情感采样中的小样本问题.

与 Reverse-JST 不同,MG-R-JST 依据文档级情感分布生成局部情感分布的先验,由于每篇文档的文档级情感分布都是不同的,所以局部情感分布的先验也不再是固定的.MG-R-JST 中文档的生成过程总结如下:

(1)对于每个主题情感对(k,s)

(ⅰ)生成主题情感对的单词分布 $\boldsymbol{\varphi}_k^s\sim \mathrm{Dir}(\boldsymbol{\beta})$

(2)对于每篇文档 d

(ⅰ)生成文档级情感分布 $\tilde{\boldsymbol{\phi}}_d\sim \mathrm{Dir}(\boldsymbol{\varepsilon})$

(ⅱ)采样 M_d 次文档级情感,得到文档级情感计数

(ⅲ)根据公式(1)计算局部情感分布的先验 $\boldsymbol{\gamma}$

(ⅳ)生成主题分布 $\boldsymbol{\theta}_d\sim \mathrm{Dir}(\boldsymbol{\alpha})$

(ⅴ)对于每个主题 k

①生成主题的局部情感分布 $\boldsymbol{\pi}_{d,k}\sim \mathrm{Dir}(\boldsymbol{\gamma})$

(ⅵ)对于该文档中的每个单词 $\boldsymbol{w}_i$

①生成主题 $z\sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$

②生成局部情感 $l\sim \mathrm{Multinomial}(\boldsymbol{\pi}_{d,z})$

③生成单词 $\boldsymbol{w}_i\sim \mathrm{Multinomial}(\boldsymbol{\varphi}_z^l)$

与 MG-R-JST 不同,MG-JST 则同时考虑文档级主题和局部主题分布,MG-JST 先生成文档级主题分布,采样 M_d 次得到文档级主题计数后,应用式(2)计算局部主题分布的先验 $\boldsymbol{\alpha}$,然后如 JST 所示,生成每篇文档的情感分布,再生成每个情感的局部主题分布,最后根据情感主题对中的单词分布生成每个单词.

$$\boldsymbol{\alpha}=\left[\lambda*\frac{\widetilde{N}_{d,1}}{M_d}+\eta,\cdots,\lambda*\frac{\widetilde{N}_{d,k}}{M_d}+\eta,\cdots,\lambda*\frac{\widetilde{N}_{d,K}}{M_d}+\eta\right] \tag{2}$$

其中,$\widetilde{N}_{d,k}$表示文档 d 中采样 M_d 次时文档级主题 k 的计数.

3.2 MG-R-JST 与 MG-JST 模型的推理

MG-R-JST 与 MG-JST 中的参数包含 $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varepsilon},\lambda,\eta$,隐含变量包括 $\boldsymbol{z},\boldsymbol{l},\boldsymbol{l}',\tilde{\boldsymbol{\phi}},\boldsymbol{\varphi},\boldsymbol{\theta},\boldsymbol{\pi}$.对于 MG-R-JST,根据文档的生成过程,其联合似然函数为:

$$\begin{aligned}&P(\boldsymbol{w},\boldsymbol{z},\boldsymbol{l},\boldsymbol{l}'\mid\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varepsilon})=\\&P(\boldsymbol{l}'\mid\boldsymbol{\varepsilon})P(\boldsymbol{z},\boldsymbol{l}\mid\boldsymbol{l}',\boldsymbol{\alpha},\boldsymbol{\gamma})P(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{l},\boldsymbol{\beta})\end{aligned} \tag{3}$$

式(3)中,若对 $\boldsymbol{z},\boldsymbol{l},\boldsymbol{l}'$同时估计涉及到多个变量,其联合概率通常难以计算,因此我们采用吉布斯采样(Gibbs sampling)对隐含变量估计,先对 l' 进行估计,更新先验 $\boldsymbol{\gamma}$ 后,再对 $\boldsymbol{z},\boldsymbol{l}$ 进行估计.

给定变量 $\boldsymbol{l},\boldsymbol{l}'$ 条件独立于 $\boldsymbol{z}$,根据如下公式可以计算 $l_i'=s$ 的概率.

$$P(\boldsymbol{l}_i'=s\mid\boldsymbol{l},\boldsymbol{l}'_{-i},\boldsymbol{\gamma},\varepsilon)\propto\frac{\prod_{s=1}^{L}\Gamma(N_{s,d}+\boldsymbol{\gamma}_s)}{\prod_{s=1}^{L}\Gamma(\boldsymbol{\gamma}_s)}\frac{\widetilde{N}_{d,s}^{(-i)}+\varepsilon}{M_d^{(-i)}+L\varepsilon} \tag{4}$$

其中,$N_{s,d}$表示文档 d 中情感 s 出现的次数,$\boldsymbol{\gamma}_s$ 为 $\boldsymbol{\gamma}$ 的第 s 个分量,上标 $-i$ 表示除去位置 i 外的序列.

对 $\boldsymbol{l}'$ 估计后经计算确定参数 $\boldsymbol{\gamma}$,此时 $\boldsymbol{z},\boldsymbol{l}$ 条件独立于 $\boldsymbol{l}'$,根据如下公式计算 $\boldsymbol{z}_i=k,\boldsymbol{l}_i=s$ 的概率.

$$\begin{aligned}&P(\boldsymbol{z}_i=k,\boldsymbol{l}_i=s\mid\boldsymbol{w},\boldsymbol{z}_{-i},\boldsymbol{l}_{-i},\alpha,\beta,\boldsymbol{\gamma})=\\&\frac{N_{t,k,s}^{(-i)}+\beta}{N_{k,s}^{(-i)}+V\beta}\frac{N_{k,s,d}^{(-i)}+\boldsymbol{\gamma}_s}{N_{k,d}^{(-i)}+\sum\boldsymbol{\gamma}_s}\frac{N_{k,d}^{(-i)}+\alpha}{N_d^{(-i)}+K\alpha}\end{aligned} \tag{5}$$

其中,$N_{t,k,s}$表示单词 t 在主题 k 和情感 s 中出现的次数,$N_{k,s,d}$表示文档 d 中 k 和 s 出现的次数,$N_{k,d}$表示文档 d 中 k 中出现的次数,$N_{k,s}$表示 k 和 s 在文档集中出现的次数,N_d 表示文档 d 中单词的个数.

在单次吉布斯采样过程中,需根据式(4)对 $\boldsymbol{l}'$ 估计,计算文档级情感计数后用式(1)更新参数 $\boldsymbol{\gamma}$,再根据公式(5)对主题和情感进行估计,可以看到公式(4)中需对复杂度较高的伽马函数 $\Gamma(x)$ 进行频繁的计算,为确保计算精度的同时提高运算速度,本文采用快速估计的方法[11]:假定局部情感变量 $\boldsymbol{l}$ 的值直接传递到文档级情感变量 $\boldsymbol{l}'$,把 $\boldsymbol{l}$ 的每个值看成 $\boldsymbol{l}'$ 的待估计值,便有文档级情感变量 $\boldsymbol{l}'$ 等于局部情感变量 $\boldsymbol{l}$,在该假设下,有 $M_d=N$ 成立.

对于文档级情感分布 $\tilde{\boldsymbol{\phi}}_d$,有 $\tilde{\boldsymbol{\phi}}_d\sim \mathrm{Dir}(\varepsilon)$,在上述假设下,$P(\tilde{\boldsymbol{\phi}}_d\mid\boldsymbol{l}'_d,\varepsilon)=\mathrm{Dir}(\tilde{\boldsymbol{\phi}}_d\mid\boldsymbol{l}_d+\varepsilon)$,$\boldsymbol{l}'_d,\boldsymbol{l}_d$ 分别表示文档 d 中文档级情感 l' 和局部情感 l 的序列,利用狄利克雷分布的期望公式,情感 s 出现的概率估计如下:

$$\tilde{\boldsymbol{\phi}}_{s,d}=\frac{N_{s,d}+\varepsilon}{N_d+L\varepsilon} \tag{6}$$

$\tilde{\boldsymbol{\phi}}_{s,d}$可以看成公式(1)中 M_d 次采样时对 $\widetilde{N}_{d,s}/M_d$ 的近似估计,此时即可根据如下公式对 $\boldsymbol{\gamma}$ 的各个分量 $\boldsymbol{\gamma}_s$ 直接进行估计,从而避免了式(4)的复杂计算.

$$\boldsymbol{\gamma}_s=\lambda\cdot\frac{N_{s,d}+\varepsilon}{N_d+L\varepsilon}+\eta \tag{7}$$

MG-R-JST 模型快速估计的单次吉布斯采样迭代过程总结如下:

(1)设定文档级情感序列等于局部情感序列,即 $\boldsymbol{l}' = \boldsymbol{l}$;

(2)根据式(7),更新局部情感分布的先验 $\boldsymbol{\gamma}$;

(3)根据式(5),更新单词的主题 z 和情感 l;

经过吉布斯采样后,MG-R-JST 中对 $\tilde{\boldsymbol{\phi}}, \boldsymbol{\varphi}, \boldsymbol{\theta}, \boldsymbol{\pi}$ 估计如下:

$$\tilde{\boldsymbol{\phi}}_{s,d} = \frac{N_{s,d} + \varepsilon}{N_d + L\varepsilon} \tag{8}$$

$$\boldsymbol{\varphi}_{t,k,s} = \frac{N_{t,k,s} + \beta}{N_{k,s} + V\beta} \tag{9}$$

$$\boldsymbol{\theta}_{k,d} = \frac{N_{k,d} + \alpha}{N_d + K\alpha} \tag{10}$$

$$\boldsymbol{\pi}_{k,s,d} = \frac{N_{k,s,d} + \boldsymbol{\gamma}_s}{N_{k,d} + \sum \boldsymbol{\gamma}_s} \tag{11}$$

表 2 为使用吉布斯采样对 MG-R-JST 进行推理的伪代码描述.

表 2 MG-R-JST 中吉布斯采样的过程

```
1. Initialize all variables and hyperparameters
2. For i = 1 to max iterations
3.   For d = 1 to D
4.     Count the values of the variable ℓ and update the priori r by Eq.7
5.     For n = 1 to N_d
6.       Read a word from the documentd
7.       Calculate the probability of
         assigning the word to topic-sentiment labels according to Eq.5
8.       Sample the topic-sentiment label for the word
9.     End for
10.  End for
11. End for
```

与 MG-R-JST 相比,MG-JST 同时考虑文档级主题和局部主题分布,故推理过程为:先用式(12)估计文档级主题变量 z',再根据式(13)估计变量 $z, \boldsymbol{l}$.

$$P(z_i' = k \mid z, z'_{-i}, \boldsymbol{\alpha}, \varepsilon) \propto \frac{\prod_{k=1}^{K} \Gamma(N_{k,d} + \boldsymbol{\alpha}_k)}{\prod_{k=1}^{K} \Gamma(\boldsymbol{\alpha}_k)} \frac{\widetilde{N}_{d,k}^{(-i)} + \varepsilon}{M_d^{(-i)} + K\varepsilon} \tag{12}$$

$$P(z_i = k, \boldsymbol{l}_i = s \mid \boldsymbol{w}, \boldsymbol{z}_{-i}, \boldsymbol{l}_{-i}, \boldsymbol{\alpha}, \beta, \gamma) \propto \frac{N_{t,s,k}^{(-i)} + \beta}{N_{s,k}^{(-i)} + V\beta} \frac{N_{s,k,d}^{(-i)} + \boldsymbol{\alpha}_k}{N_{s,d}^{(-i)} + \sum \boldsymbol{\alpha}_k} \frac{N_{s,d}^{(-i)} + \gamma}{N_d^{(-i)} + L\gamma} \tag{13}$$

经过吉布斯采样后,MG-JST 对 $\boldsymbol{\varphi}$ 的估计如式(9)所示,对 $\tilde{\boldsymbol{\phi}}, \boldsymbol{\theta}, \boldsymbol{\pi}$ 估计如下:

$$\tilde{\boldsymbol{\phi}}_{k,d} = \frac{N_{k,d} + \varepsilon}{N_d + K\varepsilon} \tag{14}$$

$$\boldsymbol{\theta}_{s,k,d} = \frac{N_{s,k,d} + \boldsymbol{\alpha}_k}{N_{s,d} + \sum \boldsymbol{\alpha}_k} \tag{15}$$

$$\boldsymbol{\pi}_{s,d} = \frac{N_{s,d} + \gamma}{N_d + L\gamma} \tag{16}$$

4 实验

进行情感分类时,需计算文档中各情感标签出现的概率 $P(l \mid d)$,Reverse-JST 与 MG-R-JST 中 $P(l = s \mid d) = \sum_{k=1}^{K} \boldsymbol{\theta}_{k,d} \cdot \boldsymbol{\pi}_{k,s,d}$,JST 与 MG-JST 中 $P(l = s \mid d) = \boldsymbol{\pi}_{s,d}$,若负向词出现的概率等于正向词的概率,进行随机预测(本文算法则利用文档级分布进行分类,此时 MG-R-JST 中 $P(l = s \mid d) = \tilde{\boldsymbol{\phi}}_{s,d}$,MG-JST 中 $P(l = s \mid d) = \sum_{k=1}^{K} \tilde{\boldsymbol{\phi}}_{k,d} \cdot \boldsymbol{\theta}_{s,k,d}$),若负向词出现的概率大于正向词的概率,该文档是负向的,否则是正向的.

4.1 实验设定

数据集 我们选择情感分析的公用数据集 MR(Movie Reviews),MDS(Multi-Domain Sentiment).MR 共有 2000 个文档,包含正负向文档各 1000 个,MDS 包含 book,dvd,electronic,kitchen4 个类别的评论,共含 4000 个正向文档和 4000 个负向文档.本文对二者进行预处理:先去除数据集中的标点,数字和非字母符号以及停止词,再对单词进行标准词干化(Porter Stemming Algorithm).通用情感词典 MPQA 标注了 8221 个常见的情感词汇,本文对其进行标准词干化,再去除处理前后情感极性相反的词,共 4581 个词汇.

对比算法 本文采用 Reverse-JST、JST、Dependency-Sentiment-LDA 作为对比算法.根据文献中的讨论[9],参数设置如下:(1)α、β 为对称狄利克雷先验,$\alpha = 50/K$,$\beta = 0.01$;(2)L 为 3,包括正向,负向,中性.$\boldsymbol{\gamma}$ 为非对称狄利克雷先验,Reverse-JST 中 $\boldsymbol{\gamma} = (0.01, 0.012, 0.01)$,JST 中 $\boldsymbol{\gamma} = (0.01, 1.8, 0.01)$. Dependency-Sentiment-LDA 在 Reverse-JST 上考虑连词时单词前后的依赖关系,其参数设置与 Reverse-JST 相同. Dynamic Joint Sentiment-Topic Model 在 JST 上主要考虑主题与情感随时间的变化趋势,使用的数据集也是随时间变化的,而本文主要集中在情感分类的效果上,故不与其对比.

4.2 抽样结果示例

为了直观显示本文考虑文档级分布的有效性,表 3、表 4 列出了 K 为 10 时各算法在 MR 上抽样结果.其中的单词表为正负情感标签中词的分布按概率递减排序的结果.

表 3 中,MG-R-JST 中 good,love,great,bad,kill 等情感词的位置比 Reverse-JST 更靠前些,并抽取到 star,effect,horror 等情感词.表 4 中,MG-JST 比 JST 抽取到 funni,effect,fight,horror 等更直观的情感词.总体来看,本文算法抽取到的情感词较主题情感混合模型更强烈与丰富.

表 3　Reverse-JST 与 MG-R-JST 抽样结果比较

Reverse-JST	正	film, movi, time, make, charact, good, stori, anim, plai, life, love, scene, year, peopl, work, great, show
	负	film, movi, alien, plot, comedi, war, make, scene, bad, action, kill, plai, scream, charact, murder, killer, laugh
MG-R-JST	正	film, time, good, charact, movi, make, stori, star, love, life, plai, great, perform, work, effect, back, live
	负	movi, film, plot, scene, charact, bad, action, war, kill, comedi, plai, make, end, thing, gui, horror, problem

表 4　JST 与 MG-JST 分类结果比较

JST	正	film, time, good, charact, stori, life, love, star, make, great, back, year, plai, perform, live, interest, real
	负	movi, film, scene, make, charact, plot, stori, end, thing, bad, action, plai, comedi, director, watch, work, long
MG-JST	正	film, good, time, star, movi, effect, stori, love, anim, plai, make, charact, year, origin, funni, great, perform
	负	movi, film, action, charact, comedi, alien, war, plai plot, make, scene, bad, kill, black, big, fight, horror

4.3　实验结果

本文算法:参数 ε、η 为 0.01,MG-R-JST 中 λ 为 100,MG-JST 中 λ 为 1000,其他参数设置与 4.1 节相同.本文算法的参数通过实验验证,在 4.3.1 节列出了 MR 上 λ,η 的参数实验,MDS 与之类似,不再赘述.根据 4.1 节中的数据集(MR、MDS)及对比算法的参数设定,对本文算法及对比算法在 K 为 1,10,30,50,100 时进行 10 轮测试,算法每轮迭代次数设为 2000 次,独立运行 5 次,即共独立运行 50 次.吉布斯采样中使用处理后的 MPQA 对情感变量进行初始化.

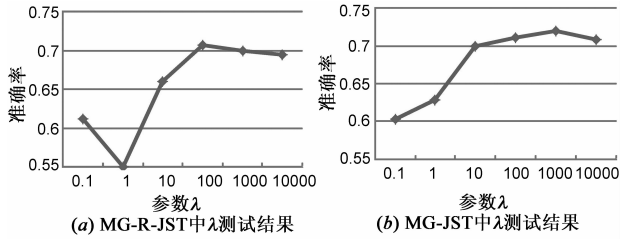
4.3.1　参数实验

本文列出了 MR 上在 K 为 10,对 λ,η 的测试结果.从中图 2 可以看出,当 λ 较大时,分类准确率较好.MG-R-JST 中 λ 为 100 时最优,MG-JST 中 λ 为 1000 时最优.λ 越大,意味着局部分布越倾向于文档级分布,间接地说明对文档级分布采样的有效性.η 作为平滑参数,若 $\eta \gg 1$,则对 γ 影响过大,从而影响到对样本的学习;若 $\eta \ll 1$,则无法起到平滑的作用.确定 λ 为最佳,在 η 为 0.1、0.01、0.001 测试,η 为 0.01 效果较佳.

4.3.2　分类效果

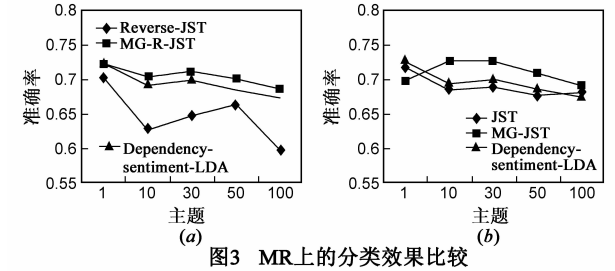
本文中分类准确率为每个 K 值的吉布斯采样运行 10 轮结果的均值,如图 3 与图 4 所示.

MR 仅包含电影评论,主题单一,K 增加使局部分布的样本数量减少,增加了局部情感/主题估计的不准确性,Reverse-JST、JST、Dependency-Sentiment-LDA 上的分类准确率随 K 的增大有所降低.MG-R-JST/MG-JST 较 Reverse-JST、JST 大约提高 2%-6%,当 K 较少或较大时,MG-JST 中的文档级主题分布对局部情感估计的影响可能较小,故较 JST 改善效果不明显;MDS 包含了 4 个类别的评论,主题比较丰富,K 过少难以表示多个类别的丰富信息,而 K 过大也会使各主题上的样本减少,K 为 30 或 50 时 Reverse-JST、JST 效果较佳,而 MG-R-JST/MG-JST 降低了主题数目的影响,提高了分类准确率,多数情况下也优于 Dependency-Sentiment-LDA.



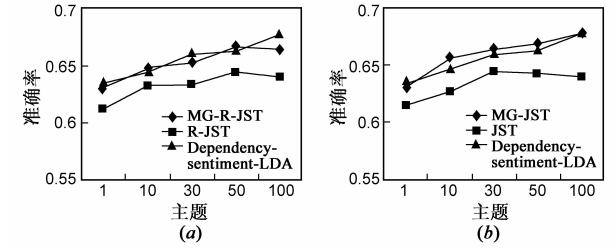
(*a*) MG-R-JST中λ测试结果　(*b*) MG-JST中λ测试结果

图2　最佳λ测试结果



(*a*)　(*b*)

图3　MR上的分类效果比较



(*a*)　(*b*)

图4　MDS上的分类效果比较

总体来看,主题数目影响到 Reverse-JST/JST 对局部情感/主题估计的准确性,依赖于特定连词的 Dependency-Sentiment-LDA 也改善了分类效果,而本文算法从文档级分布的角度出发,不需识别具体连词,降低了局部情感/主题分布对分类效果的影响,提高了不同主题数目上的分类准确率,从而更准确地分析不同粒度上的主题与情感之间的分布关系.

4.3.3　分类稳定性

以 MR 为例,本文计算每个 K 上 10 轮分类准确率的标准差,标准差越小,稳定性越好,结果如表 5 所示.多数情况下,MG-R-JST/JST 的标准差小于 Reverse-JST/JST 与 Dependency-Sentiment-LDA.本文通过考虑文档级分布对局部分布的影响,减弱了对局部情感/主题估计的不确定性,增强了分类稳定性.

表 5 各算法稳定性对比表

K	Reverse-JST	MG-R-JST	JST	MG-JST	Dependency-Sentiment-LDA
1	0.0039	0.0027	0.0034	0.0023	0.0030
10	0.0129	0.0089	0.0112	0.0052	0.0091
30	0.0107	0.0095	0.0039	0.0088	0.0036
50	0.0092	0.0050	0.0089	0.0041	0.0068
100	0.0046	0.0087	0.0146	0.0125	0.0050

5 结语

主题情感混合模型(Reverse-JST;JST)没有考虑整体分布与局部分布的关系,导致分类效果不佳且不稳定.本文同时考虑文档级分布和局部分布,提出多粒度的主题情感混合模型.MG-R-JST/MG-JST 先生成文档级情感/主题分布,通过对文档级情感/主题计数采样来计算局部情感/主题分布的先验,生成局部情感/主题分布,再生成单词.由于直接对 MG-R-JST/MG-JST 中隐含变量进行估计有较高的时间复杂度,本文采用快速估计的方法进行推理.在 MR 数据集上进行实验,实验结果表明本文算法的分类效果优于主题情感混合模型,且稳定性更好.

本文研究了无监督学习的情感分析,当数据含有标签或者部分有标签时,如何利用标签信息改进 MG-R-JST/MG-JST 模型是一个可行的研究方向.

参考文献

[1] 赵妍妍,秦兵,刘挺.文本情感分析[J].软件学报,2010,21(8):1834 – 1848.
Zhao YY, Qin B, Liu T. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834 – 1848. (in Chinese)

[2] Bo Pang, Lillian Lee. Thumbs up? sentiment classification using machine learning techniques[A]. Proceedings of the Conference on Empirical Methods in Natural Language Processing [C]. Philadelphia PA: EMNLP, 2002. 79 – 86.

[3] Fangtao Li, et al. Sentiment analysis with global topics and local dependency[A]. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence[C]. Atlanta, USA: AAAI, 2010. 1371 – 1376.

[4] Blei D M, et al. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993 – 1022.

[5] 江雨燕,李平,王清.基于共享背景主题的 Labeled LDA 模型[J].电子学报,2013,(9):1794 – 1799.
Jang Yu-yan, Li Ping, Wang Qing. Labeled LDA model based on shared background topics[J]. Acta Electronica Sinica, 2013, (9): 1794 – 1799. (in Chinese)

[6] 王李东,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)

[7] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1 – 2): 1 – 135.

[8] Chenghua Lin, Yulan He. Joint sentiment/topic model for sentiment analysis[A]. CKIM[C]. Hong Kong, China: ACM, 2009. 375 – 384.

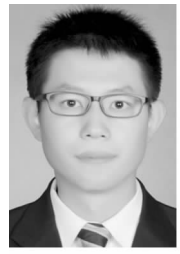
[9] Chenghua Lin, Yulan He, Richard Everson. A comparative study of bayesian models for unsupervised sentiment[A]. Proceedings of the Fourteenth Conference on Computational Natural Language Learning[C]. Uppsala, Sweden: Association for computational linguistics, 2010. 144 – 152.

[10] YulanHe, et al. Dynamic joint sentiment-topic model[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 5(1): 6.

[11] Timothy Rubin, America Chambers. Statistical topic models for multi – label document classification[J]. Journal of Machine Learning Research, 2012, 88(1 – 2): 157 – 208.

作者简介

欧阳继红 女,1964 年生于吉林长春,教授,博导,研究方向:知识工程与专家系统、空间推理和数据挖掘.
E-mail: ouyj@jlu.edu.cn



刘燕辉 男,1989 年生于山东德州,硕士,研究方向:文本分类与情感分析.
E-mail: sdpy_lyh@163.com