基于概率语义分布的短文本分类

马成龙 [1] 颜永红 [1, 2]

摘 要 在短文本分类中, 面对特征稀疏的短文本, 如何充分利用文本中的每一个词语成为关键. 本文提出概率语义分布模型的思想, 首先通过查询词矢量词典, 将文本转换为词矢量数据; 其次, 在概率语义分布模型的假设下利用混合高斯模型对无标注的文本数据进行通用背景语义模型训练; 利用训练数据对通用模型进行自适应得到各个领域的目标领域语义分布模型; 最后, 在测试过程中, 计算短文本属于领域模型的概率, 得到最终的分类结果. 实验结果表明, 本文提出的方法能够从一定程度上利用短文本所提供的信息, 有效降低了对训练数据的依赖性, 相比于支持向量机 (Support vector machine, SVM) 和最大熵分类方法性能相对提高了 17.7 %.

关键词 短文本分类, 词矢量, 语义分布, 高斯混合模型

引用格式 马成龙, 颜永红. 基于概率语义分布的短文本分类. 自动化学报, 2016, **42**(11): 1711−1717

DOI 10.16383/j.aas.2016.c150268

Short Text Classification Based on Probabilistic Semantic Distribution

MA Cheng-Long[1] YAN Yong-Hong[1, 2]

Abstract In short text classification, it is critical to deal with each word because of data sparsity. In this paper, we present a novel probabilistic semantic distribution model. Firstly, words are transformed to vectors by looking up word embeddings. Secondly, the universal background semantic model is trained based on unlabelled universal data through mixture Gaussian models. Then, target models are obtained by adapting the background model for each domain training data. Finally, the probability of the test data belonging to each target model is calculated. Experimental results demonstrate that our approach can make best use of each word and effectively reduce the influence of training data size. In comparison with the methods of support vector machine (SVM) and MaxEnt, the proposed method gains a 17.7 % relative accuracy improvement.

Key words Short text classification, word embedding, semantic distribution, Gaussian mixture model

Citation Ma Cheng-Long, Yan Yong-Hong. Short text classification based on probabilistic semantic distribution. *Acta Automatica Sinica*, 2016, **42**(11): 1711−1717

近年来, 随着社交网络和电子商务的飞速发展, 微博、Twitter、即时信息、商品评价等短文本形式的文字充斥着互联网. 这些短文本包含了用户的潜在需求、兴趣点、意图倾向等, 如何能够从这些短文本中获取信息从而更好地为用户提供服务成为关键. 然而, 这些短文本通常都有长度限制, 如微博字数限制在 140 字以内, 短消息限制在 70 字以内, 如何能够从只言片语中挖掘出目标信息成为了一大挑战. 在使用传统的向量空间模型 (Vector space model, VSM) 将短文本数字向量化时, 该向量会很稀疏[1], 特别是在测试阶段, 由于训练数据的不充分, 会造成很多有用特征因未被模型捕获过而被忽略的情况, 因此使用传统的文本分类方法将导致分类结果不理想.

为了充分利用短文本所蕴含的信息, 已有很多相关研究. 一种方案是计算短文本之间的相似性, 文献 [2] 提出使用外部数据作为一个桥梁, 如果预测文档和训练文档同时和某一外部文档相似, 那么领域标签信息也应该一样, 但搜集的外部数据必须和实验数据相关; 文献 [3] 提出使用搜索引擎返回的结果

收稿日期 2015-05-19 录用日期 2016-05-03
Manuscript received May 19, 2015; accepted May 3, 2016
国家高技术研究发展计划 (863 计划) (2015AA016306), 国家重点基础研究发展计划 (973 计划) (2013CB329302), 国家自然科学基金 (11461141004, 61271426, 11504406, 11590770, 11590771, 11590772, 11590773, 11590774), 中国科学院战略性先导科技专项 (XDA06030100, XDA06030500, XDA06040603) 和新疆维吾尔自治区科技重大专项 (201230118-3) 资助
Supported by National High Technology Research Program of China (863 Program) (2015AA016306), National Basic Research Program of China (973 Program) (2013CB329302), National Natural Science Foundation of China (11461141004, 61271426, 11504406, 11590770, 11590771, 11590772, 11590773, 11590774), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06030100, XDA06030500, XDA06040603), and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (201230118-3)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190 2. 新疆民族语音语言信息处理实验室 乌鲁木齐 830011
1. The Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190 2. Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumchi 830011

来衡量两个词语之间的相似度, 但是需要等待搜索引擎返回结果, 比较耗时, 不利于在线实时应用; 文献 [4] 提出使用固定的资源维基百科作为知识库进行搜索. 另一种解决方案是在短文本稀疏特征的基础上扩展相关语义特征, 文献 [5] 提出使用 Lucene[6] 对维基百科建立索引, 在原有特征基础上增加 Lucene 返回的搜索结果作为额外特征; 文献 [7] 提出使用短文本隐藏的主题作为额外特征集, 在相关数据上使用 LDA (Latent Dirichlet allocation)[8] 获得主题模型, 针对短文本首先进行推理得到主题特征, 与原始特征融合用于训练和分类. 上述研究都是基于利用外部相关数据对原始文本进行相似度估计或者特征扩展, 并且取得了不错的效果, 但是对外部数据的相关性要求较高, 而这些相关数据通常是根据领域知识, 人工干预下进行收集的, 在实际应用中获取相关领域的外部数据有时比较困难. 上述方法最终将文本转换为空间向量, 统计特征的共现权重, 简单来说是一种计数原理. 随着神经网络模型在自然语言处理中的广泛应用, 文献 [9] 提出将词矢量作为输入特征, 利用卷积神经网络进行模型训练. 为了得到句子层级的矢量表示, 文献 [10] 提出将变长文本训练为固定维度的段落矢量 (Paragraph vector) 的概念, 文献 [11] 提出动态卷积神经网络, 不依赖于句法解析树, 而是利用动态 k-max pooling 提取全局特征.

基于文献 [7], 为了摆脱对外部相关数据的过度依赖, 本文从句子语义层面出发, 深度挖掘短文本所表达的语义. 本文利用词矢量作为输入特征表征语义. 词矢量是指将词语映射成空间中的一个低维实数向量, 向量之间的距离描述了词与词之间的语义关系, 语义相近的词语在空间中成群出现, 提高了文字表示的泛化能力. 为了更好地利用词矢量, 本文提出了概率语义分布模型, 利用词矢量来表征语义分布, 在一定程度上避免了数据的稀疏性问题, 实验结果表明, 本文所提出的方法准确率相对于传统的分类器提高了 17.7 %.

本文结构如下: 第 1 节简要介绍连续空间词矢量, 第 2 节描述了本文提出的概率语义分布模型, 第 3 节介绍了在概率语义分布模型的假设下, 本文提出了一种基于通用语义背景模型的短文本分类方法, 第 4 节为实验及结果分析, 第 5 节给出总结.

1 连续空间词矢量

近几年, 越来越多的学者开始关注利用低维实数向量来表征一个词、短语或者句子. 例如, LSA (Latent semantic analysis)[12] 和 LDA 模型将文本映射成主题模型里的一个低维向量. 随着神经网络的广泛应用, 人们可以利用神经网络对大规模语料进行语言模型训练, 同时能够得到描述语义和句法关系的词矢量. 其中, 文献 [13] 提出的 Skip-gram 模型便是一种能够高效得到词矢量的训练模型, 通过训练无标注语料将每个词映射成低维实数向量, 每一维都代表了词的浅层语义特征[14]. 同时, 文献 [15] 发现上述模型训练得到的词矢量能够通过余弦距离描述词与词之间的语义和句法关系, 并且相同的余弦距离表征了同样关系, 例如, 向量 “Man” 与向量 “King” 之间的距离近似于向量 “Woman” 与向量 “Queen” 之间的距离. 因此, 本文利用词矢量上述特性, 结合短文本的特点, 提出了概率语义分布模型, 应用于短文本分类中.

2 概率语义分布模型

不同于传统的文本分类算法, 本文认为短文本是在贝叶斯框架下各个领域里的一个抽样. 本文假设短文本数据产生于一个概率语义分布模型, 不同领域数据来自于不同的语义分布模型, 并且我们可以利用已知的文本数据去估计这些模型. 得到这些模型之后, 对于新的测试数据, 计算来源于各个模型的概率, 根据贝叶斯原理选择类别标签作为预测结果.

假设训练数据包含一系列的短文本文档, $D=\{d_1,\ d_2,\ d_3,\cdots,d_n\}$, d_i 表示一条短文本, 共 n 条训练数据, 分别属于 $C=\{c_1,\ c_2,\ c_3,\cdots,c_m\}$, c_j 为领域标记, 共 m 个领域. 本文假设同一领域短文本文档产生于同一个语义分布模型 (模型参数为 λ). 一条短文本数据 d_i 的产生, 首先根据先验概率 $p(c_j|\lambda)$ 选择语义分布模型, 然后根据该领域模型的模型参数 $p(d_i|c_j;\ \lambda)$ 产生文档 d_i. 因此文档 d_i 的产生概率为 $p(d_i|\lambda)$:

$$p(d_i|\lambda)=\sum_{j=1}^{m}p(c_j|\lambda)p(d_i|c_j;\lambda) \tag{1}$$

类似于一元语言模型, 认为短文本中词与词之间是互相独立的, 不依赖于前文信息, d_{ik} 表示短文本 d_i 中位置为 k 的单词, $|d_i|$表示文本中单词的个数, 则有

$$p(d_i|c_j;\lambda)=\prod_{k=1}^{|d_i|}p(d_{ik}|c_j;\lambda) \tag{2}$$

假设已通过训练数据计算得到模型参数 $\hat{\lambda}$, 针对测试数据, 可以分别计算各个分布模型产生该数

据的概率. 根据贝叶斯原理, 由式 (1) 和 (2) 得到

$$p(c_j|d_i;\hat{\lambda}) = \frac{p(c_j|\hat{\lambda})p(d_i|c_j;\hat{\lambda})}{p(d_i|\hat{\lambda})} = \frac{p(c_j|\hat{\lambda})\prod_{k=1}^{|d_i|}p(d_{ik}|c_j;\hat{\lambda})}{\sum_{l=1}^{|C|}p(c_l|\hat{\lambda})\prod_{k=1}^{|d_i|}p(d_{ik}|c_l;\hat{\lambda})} \tag{3}$$

根据上述提出的概率语义分布模型假设, 本文认为可以选择合适的模型去近似描述每个领域内的词语分布. 由于混合高斯模型能够描述任意形状的概率分布, 因此本文选用混合高斯模型. 由于训练数据的不充分, 直接使用混合高斯模型进行多高斯训练时会产生欠拟合, 因此本文在混合高斯模型的基础上提出了一种基于通用语义背景模型的短文本分类方法.

3　基于通用语义背景模型的短文本分类

在实际应用中, 由于自然语言表达的灵活性, 获取足够多的标注数据是一件费时费力的事情, 如何能够充分利用已有数据进行短文本分类成为关键. 在图像处理、说话人识别系统中, 高斯混合−通用背景模型[16−17] 便是一种能够在训练数据不足的情况下, 由一个通用的背景模型根据少量的训练数据自适应到目标模型上, 并且取得了很好效果. 因此, 借鉴于高斯混合−通用背景模型, 在概率语义分布模型的假设下, 首先利用混合高斯构建通用概率语义背景分布模型, 然后根据训练数据自适应得到目标领域概率语义分布模型, 如图 1 所示.

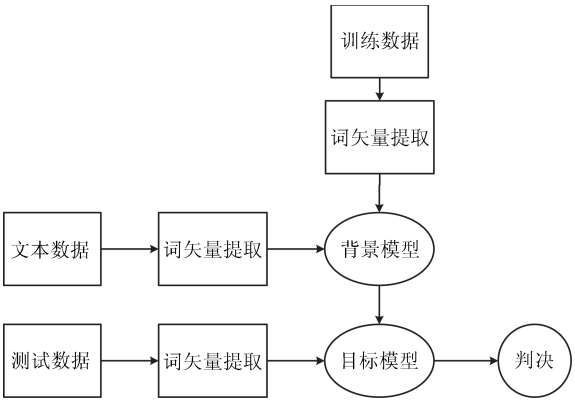


图 1　基于通用语义背景模型的短文本分类

Fig. 1　Short text classification based on universal semantic background model

3.1　词汇特征

在连续空间词矢量表示中, 通过向量之间的空间距离来表征词与词之间的特定关系, 并且文献 [18] 指出从大量无标记文本数据训练得到的词矢量要比随机初始化的矢量性能要好. 在短文本分类中, 我们应该首先训练得到词矢量. 然而, 词矢量的训练通常需要耗费很长时间, 并且已有许多学者将训练好的词矢量进行了开源. 本文的实验直接使用文献 [19] 提供的词矢量词典, 该词典是利用大概十亿单词数量的谷歌新闻数据训练得到的维度为 300 的词矢量.

3.2　高斯混合模型

高斯混合模型 (Gaussian mixture model, GMM) 作为一种通用的概率模型, 只要高斯数足够大, 便能有效地模拟多维矢量的连续概率分布, 因而很适合去表征语义分布. 高斯混合模型是一系列高斯分布的加权组合. 一个由 M 个高斯分量组成的高斯混合密度函数是 M 个高斯密度函数的线性加权和:

$$p(d_i|\lambda) = \sum_{k=1}^{M} w_k p_k(d_i) \tag{4}$$

上式中 λ 为 GMM 模型参数, $p_k(d_i), k = 1, \cdots, M$ 是高斯分量密度函数. $w_k, k = 1, \cdots, M$ 是各个高斯分量的权重, 满足 $\sum_{k=1}^{M} w_k = 1$. 每个高斯分量的概率密度函数公式 $p_k(d_i)$ 表示如下:

$$\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(d_i-\mu_k)^{\mathrm{T}}\Sigma_k^{-1}(d_i-\mu_k)\right\} \tag{5}$$

这里 μ_k 是第 k 个高斯分量的均值矢量, Σ_k 为相应的协方差矩阵, D 是特征矢量的维度. 这样, GMM 模型便可以由以下参数集合表示:

$$\lambda = \{w_k, \mu_k, \Sigma_k\}, \quad k = 1, 2, \cdots, M \tag{6}$$

使用 GMM 对概率语义分布建模主要基于两个出发点: 1) GMM 的高斯分量能够描述一定词矢量的分布; 2) 线性加权的高斯密度函数可以逼近任意形状的概率分布, 因此选用 GMM 对语义分布进行描述.

3.3　最大后验模型自适应

利用高斯混合模型在无标注文本数据上训练得到通用概率语义背景分布模型, 再用带有标记的训练数据进行模型自适应得到目标模型. 最大后验概率 (Maximum a posteriori, MAP) 是一种典型的贝叶斯估计, 它首先计算训练数据相对于通用背景模型的各个统计量, 然后用一个相关系数将通用背景模型参数与相关统计量联合, 得到目标模型. 给定通用背景模型: $\lambda = \{w_k, \mu_k, \Sigma_k\}$, $k = 1, 2, \cdots, M$, 以及某一特定领域内的短文本训练数据 $D_{c_j} = \{d_{c_1}, \cdots, d_{c_i}, \cdots, d_{|c_j|}\}$, 对每一条训练数

据计算其在各高斯分量上的占有率, 即后验条件概率:

$$p(k|d_{c_i}) = \frac{w_k p_k(d_{c_i})}{\sum_{j=1}^{M} w_j p_j(d_{c_i})} \tag{7}$$

然后便可计算出与权重相关的零阶统计量 n_k, 与均值相关的一阶统计量 $E_k(d)$ 以及与协方差矩阵相关的二阶统计量 $E_k(d^2)$:

$$n_k = \sum_{c_i=1}^{|c_j|} p(k|d_{c_i}) \tag{8}$$

$$E_k(d) = \frac{1}{n_k} \sum_{c_i=1}^{|c_j|} d_{c_i} p(k|d_{c_i}) \tag{9}$$

$$E_k(d^2) = \frac{1}{n_k} \sum_{c_i=1}^{|c_j|} d_{c_i}^2 p(k|d_{c_i}) \tag{10}$$

用以上计算得到的统计量对通用背景模型的各个高斯分量的权重、均值和协方差进行自适应, 得到新的模型参数:

$$w_k^* = \left[\frac{\alpha_k^w n_k}{T} + (1 - \alpha_k^w) w_k \right] \gamma \tag{11}$$

$$\mu_k^* = \alpha_k^m E_k(d) + (1 - \alpha_k^m) \mu_k \tag{12}$$

$$\sigma_k^{2*} = \alpha_k^\nu E_k(d^2) + (1 - \alpha_k^\nu)(\sigma_k^2 + \mu_k^2) - (\mu_k^*)^2 \tag{13}$$

其中 γ 用来平衡高斯分量的权值, 以保证更新后各分量的权值和为 1. $\{\alpha_k^w, \alpha_k^m, \alpha_k^\nu\}$ 是调整新旧模型参数平衡的自适应系数, 通常使用同一个自适应系数. 为了能够确定上述参数, 本文在训练集上使用 5 折交叉验证来确保参数的可靠性.

4 实验结果与分析

为了验证所提出方法的有效性, 本文利用文献 [7] 提供的短文本数据, 首先验证背景模型和高斯数对分类性能的影响, 其次与基线系统进行比较, 最后验证所提出的方法对训练数据的依赖性.

4.1 实验数据与评价标准

本文选择文献 [7] 提供的网页搜索片段数据作为实验数据, 网页搜索片段数据集是将特定领域词送入谷歌搜索引擎得到的搜索结果片段, 为了保证领域的特定性, 通常选取前 20 ∼ 30 个片段作为引用数据. 例如计算机类, 选取 60 个计算机领域的词语, 分别送入谷歌搜索引擎, 每次抽取搜索结果的前 20 条数据作为训练数据, 则可以得到 1 200 条数据, 数据分布如表 1. 为了区分训练数据和测试数据, 在生成测试数据时所使用的领域词不同于训练数据. 如表 2 所示, 无论是英文单词未经提取词干还是经过提取词干 (Porter stemming)[20] 之后, 都会有超过 40 % 的未登录词 (未登录词通常是指未在词典中出现的词[21]) 出现在测试集中, 这极大地增加了分类的难度.

表 1 网页搜索片段数据分布

Table 1 Statistics of web snippets data

编号	领域	训练数据	测试数据
1	商业	1 200	300
2	计算机	1 200	300
3	文化与艺术	1 880	330
4	教育与科技	2 360	300
5	技术	220	150
6	健康	880	300
7	社会政策	1 200	300
8	体育	1 120	300
共计		10 060	2 280

表 2 未登录词分布

Table 2 Statistics of unseen words

	原始单词	词干
训练数据	26 265	21 596
测试数据	10 037	8 200
未登录词	4 378	3 677
未登录词的比例	43.62 %	44.84 %

在实验过程中, 本文使用精度 (Precision, P)、召回率 (Recall, R)、F1 值和准确率 (Accuracy, A) 作为评价标准.

4.2 实验

4.2.1 参数设置

如何选择背景数据进行通用背景语义模型训练以及不同的背景模型对性能如何影响, 混合高斯模型中的高斯数如何确定, 这些参数都需要通过实验进行验证. 本文选择: 1) 相关数据: 去掉标注的训练数据作为背景数据; 2) 通用数据: 选取语言资源联盟 (Linguistic Data Consortium) 提供的新闻数据[22], 本文仅选取标签 Headline 下的文本; 3) 混合数据: 相关数据和通用数据的混合, 分别作为背景数据进行背景模型训练, 实验结果如图 2 所示.

当我们不断增加高斯数时, 混合高斯能够很好地拟合特征分布, 但是当高斯数过高时, 由于数据的稀缺, 会出现过拟合现象, 正如图 2 中当使用训练数据 1) 进行背景模型训练时, 高斯数达到 256 时无法拟合出混合高斯模型. 在图 2 中, 直接使用无标注的训练数据进行通用背景模型训练, 在低维混合高斯下能够快速地提高分类性能, 但是由于数据有限, 无法进行高维高斯拟合, 高斯数为 128 时准确率达到 78.6 %; 使用通用数据, 由于数据量较大, 能够进行

高维高斯拟合, 并且在高维混合高斯的情况下能够达到直接使用训练数据的分类性能, 高斯数为 8 时准确率达到最高 75.83 %; 当使用无标注的训练数据 + 通用数据时, 高斯数为 16, 短文本分类准确率达到最高值 80 %.

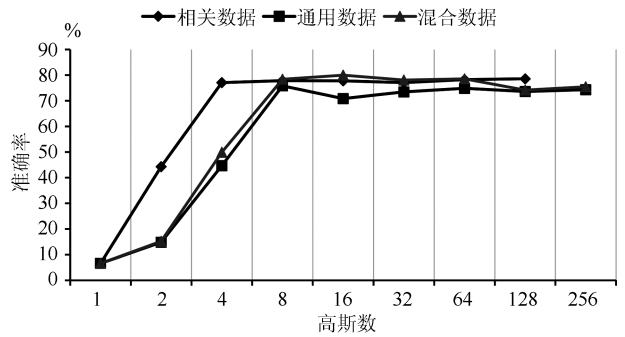


图 2 不同的背景数据和高斯数对分类结果的影响

Fig. 2 Influence of background data and the number of GMM

4.2.2 与基线系统相比

为了验证本文所提方法的有效性, 本文选择以下方法作为基线系统:

1) TF*IDF + SVM/MaxEnt: 特征值采用 TF*IDF 进行计算, 利用支持向量机 (Support vector machine, SVM) 或最大熵 (MaxEnt) 作为分类器.

2) LDA + MaxEnt: 在文献 [7] 中, 利用 LDA 对文本进行主题特征提取, 与文本特征进行合并, 利用 MaxEnt 进行分类模型的训练.

3) Wiki feature + SVM: 对维基百科数据[1] 进行去除网页标签、网页链接等预处理之后, 使用 Lucene 对其建立索引, 对每一条短文本实验数据进行检索. 在检索结果中, 类似文献 [5] 中提出的方法, 将维基百科数据的标题作为额外的文本特征扩充到原始短文本数据中. 不同于文献 [5] 中所描述的聚类任务, 我们将融合后的文本用于短文本分类.

4) Paragraph vector + SVM: 文献 [10] 提出了一种无监督的方法, 利用定长数学向量表征不定长文本. 该模型认为当前词语的选择不仅由上下文决定, 还由隐藏的文本矢量共同决定. 该隐藏文本矢量可以看做为文本的隐藏主题[23].

5) LSTM (Long short term memory): 对文献 [24] 中提出的 LSTM 模型进行修改, 组成结构为单一的 LSTM 层、均值池化层 (Average pooling layer) 和逻辑回归层 (Logistic regression layer), 使其能够进行文本类别预测[23].

[1] http://download.wikipedia.com/enwiki/

[2] http://word2vec.googlecode.com/svn/trunk/

在传统的文本分类方法中, 通常是利用词袋模型 (Bag of words, BoW) 将文本离散化, 计算特征权重, 转换为向量空间模型中的特征权重向量, 每个词被转换为字典中的索引数字. 这种方法降低了计算复杂度, 但是对于未登录词的处理能力大幅度降低.

由于在训练的过程中, 分类模型未捕捉到未登录词对分类结果的贡献能力, 在测试阶段, 未登录词通常会被忽略. 尤其是在该测试集中会出现超过 40 % 的未登录词, 这极大地增加了分类难度. 因此, 在表 3 中传统的文本分类方法 SVM 和 MaxEnt 性能均不是很高. 以维基百科作为搜索库, 利用 Lucene 的搜索结果进行原始短文本扩展, 在一定程度上降低了特征稀疏性, 对分类性能有所提升. 本文的方法利用词矢量将文本向量化, 词矢量体现了一定的语言泛化能力, 充分利用了训练数据里的每一个有用词语, 使得准确率相对传统方法提高了 17.7 %, 并且如表 4 所示每一领域的分类结果 F1 值均优于传统的分类结果. 在 Paragraph vector 和 LSTM 这两种模型中, 都使用到了词矢量, 但都未能有效地捕获到语句中的语义信息.

表 3 与基线系统对比实验结果 (%)

Table 3 Experimental results of the proposed method against other methods (%)

方法	Accuracy
TF*IDF + SVM	66.14
TF*IDF + MaxEnt	66.80
LDA + MaxEnt	82.18
Wiki feature + SVM	76.89
Paragraph vector + SVM	61.90
LSTM	63.00
本文的方法	80.00

文献 [7] 提到的方法需要根据领域知识额外准备大概 470 000 篇维基百科数据, 共计 3.5 GB 的相关数据进行主题模型训练, 增加了收集数据的难度. 本文在使用混合数据时准确率达到 80 %, 略低于文献 [7] 中的 82.18 %, 但是本文有效地避免了收集相关数据的困难. 本文选用维基百科数据, 对其进行去除网页标签、链接等预处理之后, 用于 LDA 主题模型训练和词矢量训练. 在主题模型训练过程中, 主题数目选择为 50、100、200、300、400 等, 在训练集上利用五折交叉验证确定最优主题数. 针对词矢量的训练, 使用开源工具 word2vector[2] 训练得到维度为 300 的词矢量. 在使用相同外部数据的情况下, 本文方法取得 79.93 % 的性能, 略高于基于 LDA + MaxEnt 方法的 79.89 %. 从这一点可以看出, 在使用外部数据进行主题模型训练时, 外部数据与实验数据的相关性, 是影响主题特征贡献能力的

表 4 SVM、MaxEnt 和本文方法的实验结果

Table 4 Evaluations of SVM, MaxEnt and the proposed method

	SVM			MaxEnt			本文的方法		
领域	P (%)	R (%)	F1	P (%)	R (%)	F1	P (%)	R (%)	F1
社会政策	77.61	52.00	0.6228	70.75	50.00	0.5859	86.36	70.37	0.7755
计算机	73.75	63.67	0.6834	72.26	66.00	0.6899	80.31	87.29	0.8365
教育与科技	41.98	82.00	0.5553	45.93	82.67	0.5905	81.60	68.23	0.7432
体育	85.19	76.67	0.8070	86.08	78.33	0.8202	84.54	89.93	0.8715
健康	89.01	56.67	0.6925	86.94	64.33	0.7395	76.35	85.57	0.8070
技术	76.53	50.00	0.6048	72.84	39.33	0.5108	58.82	93.33	0.7216
商业	70.37	57.00	0.6298	68.05	60.33	0.6396	73.99	67.33	0.7051
文化与艺术	62.27	81.52	0.7060	62.86	78.48	0.6981	88.15	77.85	0.8268

一个重要因素. 因此, 当面对一个新的分类任务时, 文献 [7] 中的方法需要根据领域知识重新挑选大量相关语料进行主题模型训练, 从一定程度来讲, 本文的方法更易实现.

4.2.3 训练数据大小对分类效果的影响

为了验证本文方法对训练数据的依赖性, 本文将训练数据保持原领域数据的分布比例不变平均分成 10 份, 每次增加 1 份进行试验, 在同一测试集上进行测试, 得到 10 组实验结果, 如图 3 所示. 由于 SVM 和 MaxEnt 的分类效果相差不大, 因此仅选择了 MaxEnt 作为基线系统. 随着训练数据的减少, 测试集中未登录词的比重会逐渐加大, MaxEnt 的分类效果变化幅度较大, 对训练数据的依赖性比较大. 在训练数据稀缺的情况下 (仅占原训练数据的 1/10), 本文方法能够将正确率从 47.06 % 提高到 71.54 % (相对提高 52 %). 从另一角度说明如何充分利用词汇信息成为分类的关键, 而这也是本文方法的关键.

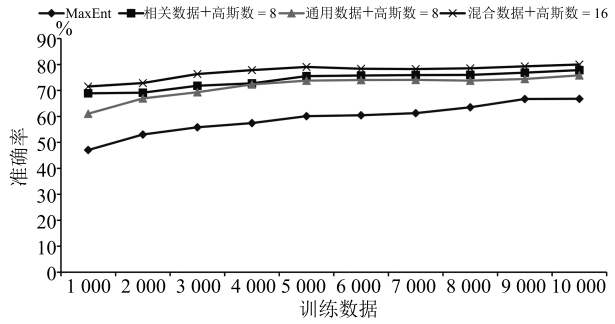


图 3 训练数据大小对分类效果的影响 (1)

Fig. 3 Influence of training set size (1)

为了进一步检验训练数据对本文方法的影响, 本文继续将训练数据数量缩小, 如图 4 所示. 在仅有 100 条训练数据的情况下, 本文所提出的方法准确率能够达到 51.4 %, 高于 MaxEnt 在 1 000 条训练数据下的 47.06 %, 这对于获取训练数据比较困难的应用来说, 可以大大地降低对训练数据的依赖性.

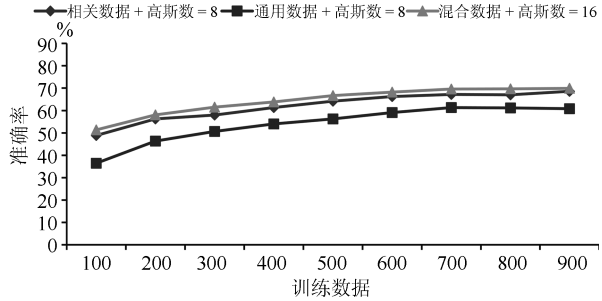


图 4 训练数据大小对分类效果的影响 (2)

Fig. 4 Influence of training set size (2)

5 结论

本文摒弃了传统的文本向量空间表示模型, 提出概率语义分布模型, 认为短文本是来自于概率语义模型的一个抽样, 利用词矢量将文本数字化, 通过无标记数据构建通用语义背景模型, 利用训练数据进行自适应得到目标模型. 实验结果验证了本文所提出方法的可行性, 利用能够表征语义和句法关系的词矢量有效地降低了训练数据不充分所带来的影响, 短文本分类性能明显优于传统的文本分类方法, 降低了对训练数据的依赖性. 虽然本文的实验结果略低于基于主题模型的短文本分类系统的结果, 但明显优于基于 SVM 和最大熵的分类算法, 并且本文的方法无需准备大量的相关数据, 在一定程度上本文方法更易实现.

References

1 Wang B K, Huang Y F, Yang W X, Li X. Short text classification based on strong feature thesaurus. *Journal of Zhejiang University Science C*, 2012, **13**(9): 649−659

2 Zelikovitz S, Hirsh H. Improving short text classification using unlabeled background knowledge to assess document similarity. In: Proceedings of the 17th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 2000. 1183−1190

3 Bollegala D, Matsuo Y, Ishizuka M. Measuring semantic similarity between words using web search engines. In: Proceedings of the 16th International Conference on World Wide Web. New York, USA: ACM, 2007. 757−766

4 Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 2007. 1606−1611

5 Banerjee S, Ramanathan K, Gupta A. Clustering short texts using Wikipedia. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2007. 787−788

6 Lucene [Online], available: https://lucene.apache.org/, May 3, 2016.

7 Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web. New York, USA: ACM, 2008. 91−100

8 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, **3**: 993−1022

9 Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, 2014. 1746−1751

10 Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR, 2014. 1188−1196

11 Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA: Association for Computational Linguistics, 2014. 655−665

12 Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis. *Discourse Processes*, 1998, **25**(2−3): 259−284

13 Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.

14 Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, 2010. 384−394

15 Mikolov T, Yih W T, Zweig G. Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013. 746−751

16 Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 1995, **17**(1−2): 91−108

17 Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, **10**(1−3): 19−41

18 Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 2011, **12**: 2493−2537

19 Mikolov T, Sutskever I, Chen K, Corrado G S, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 2013 Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2013. 3111−3119

20 Porter M F. An algorithm for suffix stripping. *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann, 1997. 313−316

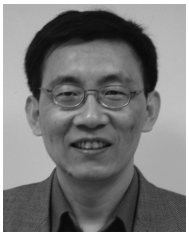
21 Ling G C, Asahara M, Matsumoto Y. Chinese unknown word identification using character-based tagging and chunking. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics, 2003. 197−200

22 Parker R, Graff D, Kong J B, Chen K, Maeda K. English Gigaword Fifth Edition [Online], available: https://catalog.ldc.upenn.edu/LDC2011T07, May 3, 2016.

23 Wang P, Xu B, Xu J M, Tian G H, Liu C L, Hao H W. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, 2016, **174**: 806−814

24 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

马成龙 中国科学院声学研究所博士研究生. 2011 年获得山东大学 (威海) 通信工程学士学位. 主要研究方向为自然语言处理, 口语理解, 情感分析, 深度学习. 本文通信作者.
E-mail: machenglong@hccl.ioa.ac.cn
(**MA Cheng-Long** Ph. D. candidate at the Institute of Acoustics, Chinese Academy of Sciences. He received his bachelor degree from Shandong University, Weihai in 2011. His research interest covers natural language processing, spoken language understanding, sentiment analysis and deep learning. Corresponding author of this paper.)



颜永红 中国科学院声学研究所语言声学与内容理解重点实验室教授. 1990 年在清华大学获得学士学位, 1995 年 8 月于美国俄勒冈研究院 (Oregon Graduate Institute, OGI) 获得计算机科学和工程博士学位. 他曾在 OGI 担任助理教授 (1995 年), 副教授 (1998 年) 和副主任 (1997 年). 主要研究方向为语音处理和识别, 语言/说话人识别和人机界面.
E-mail: yanyonghong@hccl.ioa.ac.cn
(**YAN Yong-Hong** Professor at The Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences. He received his bachelor degree from Tsinghua University in 1990, and Ph. D. degree from Oregon Graduate Institute (OGI), USA. He worked in OGI as assistant professor (1995), associate professor (1998) and associate director (1997) of Center for Spoken Language Understanding. His research interest covers speech processing and recognition, language/speaker recognition, and human computer interface.)