文章编号：2095-2163(2022)09-0032-05 中图分类号：TP391.41 文献标志码：A

融合主题竞争关系的短文本分类方法

潘智勇，赵 港
（北华大学 计算机科学技术学院，吉林 吉林 132013）

摘 要： 网络短文本以短小、快速等特点，现已成为重要的数据资源。但因其内容短小，不利于主题模型等自然语言处理算法从中提取有效的特征表达，严重限制了算法的应用。针对短文本数据特点，本文结合词汇对主题模型和 K 竞争自编码模型数据处理方式，建立表达词汇间关系的“词汇对”表达数据，将竞争关系引入主题表达，突出重点主题；以全连接结构建立主题全局关系，弥补主题模型忽略词汇关系和主题关系的不足，有效增强主题特征的表达能力。实验结果表明，本文方法的分类准确率明显高于传统主题模型。

关键词： 短文本分类；主题模型；K 竞争关系；全连接结构

Short-text classification with topic competitive relationship

PAN Zhiyong, ZHAO Gang
(School of Computer Science and Technology, Beihua University, Jilin Jilin 132013, China)

【Abstract】 Internet short texts have already become the most important data resource as the texts are short and spread widely and rapidly. The short texts usually have a few words, which makes it difficult to extract the effective features for natural language processing algorithms, such as topic model, and limits the applications of the models. For the characteristics of short-text data, this paper combines the data processing of biterm topic model and K - competitive autoencoder to propose a new method. The method uses the biterms to express word relationship, introduces competitive relationship into the topic features, and builds the global relationship of topics by fully-connected layers. Therefore the method highlights the key topics, overcomes the limitation of ignoring the word relationship and the topic relationship, and enhances the representative ability of topic features. The experimental results of short-text classification on two standard datasets (20newsgroup and Reuters-21578) show that the method outperforms the traditional topic models.

【Key words】 short-text classification; topic model; K-competitive relationship; fully-connected layer

0 引 言

随着网络信息数据量的快速增长，以微博、Twitter 和博客等为代表的网络短文本已成为重要的数据资源。与此同时，对这些网络短文本的信息处理也得到了更为广泛的关注[1-5]，而数据表达特征的有效性，将直接决定着分类的准确率。由于文本中含有大量的同义词和多义词，传统基于词频统计的文本处理方法[6-7]将会受到一定影响，限制模型应用。以隐狄利克雷分配（Latent Dirichlet Allocation，LDA）[8]模型为代表的主题模型基于词汇与主题的共现关系，以主题作为底层特征和上层语义之间的中层特征，有效地克服了同义词和多义词的影响。针对短文本数据，Zhang 等人[9]结合词汇及隐主题作为新的词汇来学习短文本的向量表达，提高了文档的表达能力。刘爱琴等人[10]利用 LDA 模型分析短文本，提取主题词，并以主题与词汇的共现矩阵对短文本进行分类。Yan 等人[11]提出词汇对主题模型（biterm topic model，BTM），该模型基于短文本数据中词与词共现关系，提取“词汇对”作为文档的基本特征，从而建立语料级的词汇共现关系，克服忽略词汇关系的不足。BTM 通过“词汇对”与主题的共现关系提取主题特征，在短文本数据分类问题上得到了较好的应用。Yang 等人[12]利用词汇共现关系和类别词汇的相似性，提出种子词汇对主题模型（Seeded Biterm Topic Model，SBTM），并且利用附加用户信息，提出种子推特词汇对主题模型（Seeded Twitter Biterm Topic Model，STBTM）。但

基金项目： 吉林省教育厅科学技术项目（JJKH20190645KJ）；吉林省科技发展计划项目（20210203050SF）；吉林市科技发展计划杰出青年人才培养专项（20200104075）。

作者简介： 潘智勇（1980-），男，博士，实验师，硕士生导师，主要研究方向：机器学习；赵 港（1996-），男，硕士研究生，主要研究方向：机器学习。

通讯作者： 潘智勇 Email：zhiyong0432@126.com

收稿日期： 2022-04-26

是，上述模型均基于主题独立性假设，利用词汇或词汇对与主题的共现关系提取主题特征，忽略了主题之间的关系，影响了主题特征的表达准确性。

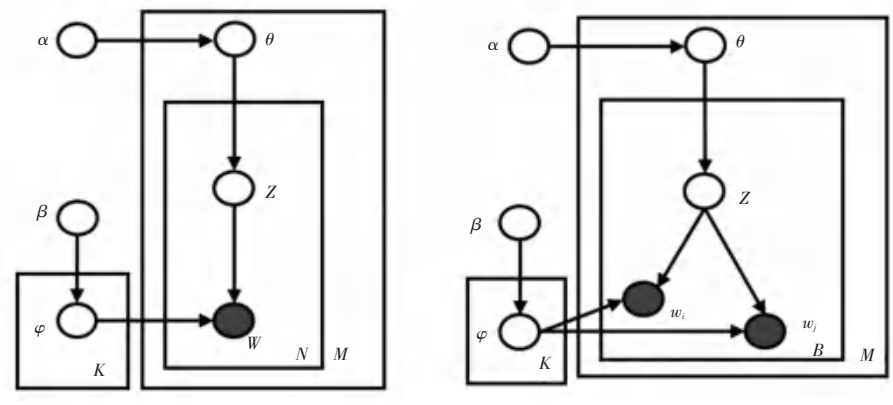
以卷积神经网络（Convolutional Neural Networks，CNN）模型[13]为代表的深度学习算法，在自然语言处理和图像处理领域均取得了较好的应用。CNN 通过多层神经网络提取表达局部特征的神经元，在全连接层（fully-connected layer）建立各层神经元之间的全局关系，从而提高特征表达能力。作为中层特征的主题为隐变量，在表达文档的过程中存在冗余的问题。Chen 等人[14]提出 Kate 模型，将竞争关系引入到神经元的主题获取过程，使神经元的主题更具有区分度，同时降低全连接层模型参数，提高了主题学习效率。但上述模型以词汇作为局部特征，易受到同义词和多义词的影响。

本文融合 BTM 词汇对表达和面向文本的 K 竞争自编码（K - competitive autoencoder for text，Kate）主题竞争关系，并利用全连接层构建起文档中主题之间的全局关系，提出 K 竞争全连接主题网络模型（K - competitive fully-connected topic network，KFTN）。KFTN 以“词汇对”表达文档数据，降低了短文本对主题表达的影响，引入了主题竞争关系，增强主题特征的表达能力，建立起语料级的词汇关系和主题间的全局关系，从而提高短文本分类的准确率。

1　相关模型背景

1.1　词汇对主题模型

由于短文本文档词汇数量较少，词汇特征过于稀疏，同时隐狄利克雷分配（LDA）模型基于词汇独立性假设，限制了 LDA 模型提取主题特征的准确性。词汇对主题模型（BTM）基于文档中共现的词汇构建无向词汇对，利用词汇对与主题的后验概率，对文档中主题特征进行采样。BTM 构建的词汇对融入了词汇的共现关系，解决了短文档的稀疏性问题。LDA 与 BTM 模型的概率图模型如图 1 所示。



（a）LDA 的概率图模型　　（b）BTM 的概率图模型

图 1　LDA 和 BTM 的概率图模型

Fig. 1　Graphical models of LDA and BTM

从图 1 中可以看出，LDA 中相互独立的主题产生相互独立的词汇，而 BTM 以相互独立的主题产生词汇对（w_i 和 w_j）。因此，与 LDA 主题采样不同，BTM 的主题采样过程为：

$$p(z \mid Z_{\neg b}, B, \alpha, \beta) \propto \frac{n_{m,b}^{k} + \alpha}{\sum_{k=1}^{K} n_{m,b}^{k} + K\alpha} \cdot \frac{(n_{k}^{w_i} + \beta)(n_{k}^{w_j} + \beta)}{\left(\sum_{w} n_{k}^{w} + V\beta\right)^{2}} \tag{1}$$

其中，B 是词汇对的集合；$Z_{\neg b}$ 是除当前词汇对 b 以外的主题集合；$n_{m,b}^{k}$ 是文档 m 中主题 k 产生的词汇对 b 的总数；n_{k}^{w} 是主题 k 产生的词汇 w 的总数；超参 α 和 β 避免模型过拟合。

1.2　面向文本的 K 竞争自编码模型

面向文本的 K 竞争自编码（K - competitive autoencoder for text，Kate）模型在隐层编码过程中，以前 K 个权重绝对值较大的神经元作为关键主题，并将其它神经元的权重分别转移到正负权重较大的神经元后置 0，进一步增大关键主题权重。经过引入竞争关系，使重点主题更加突出，增强主题特征稀疏性，降低其它主题的影响。

Kate 模型 K 竞争层结构[14]如图 2 所示。图 2 中，K 竞争层 h_1 和 h_6 分别为正负权重最大神经元，其它神经元按正负权重分别相加后，以超参 α 为系数增加 h_1 和 h_6 权重。

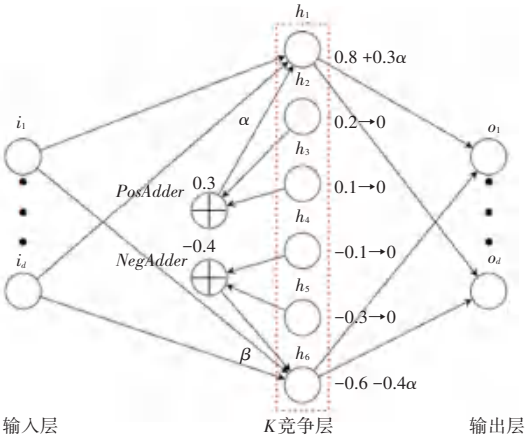


图 2 Kate 模型 *K* 竞争层结构图

Fig. 2 The architecture of *K*-competitive layer

2 *K* 竞争全连接主题网络模型

本文结合 BTM“词汇对”表达和 Kate 主题竞争关系,提出 *K* 竞争全连接主题网络模型(KFTN)。该模型以融合词汇共现关系的“词汇对”表达短文本,提取主题作为初始主题特征。将初始主题特征引入竞争关系,增强主题特征稀疏性,突出关键主题作用,并以全连接结构建立主题间全局关系,提高主题特征表达的准确性。KFTN 文档处理主要结构如图 3 所示。这里对 KFTN 主要部分拟展开阐释分述如下。

(1)主题初始采样。KFTN 利用词汇共现关系提取无向“词汇对”表达文档(参见图 3 中同一颜色表示一组词汇对),建立起主语料级的词汇共现关系,从而使文档由词汇表达转为“词汇对”表达。利用式(1)采样计算后验概率,提取具有一定中层语义的主题,作为初始主题特征。

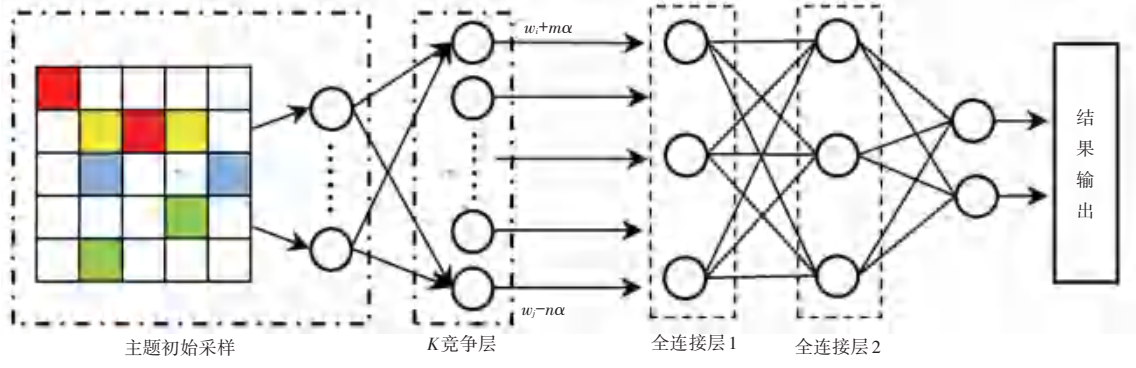


图 3 KFTN 文档处理主要结构图

Fig. 3 The main architecture of KFTN for documents processing

(2) *K* 竞争层。经主题初始采样,KFTN 以主题特征表达文档。但所提取的主题基于独立性假设,忽略了主题之间关系,同时主题特征中还存在一定的噪声。因此,研究中为突出重点主题,增强主题特征稀疏性,降低噪声主题的影响, *K* 竞争层引入竞争机制,对主题特征进行重新编码/解码,保留具有代表性的 *K* 项主题(正负权重各 *K*/2 项),其它主题权重置 0。由图 3 可见,*m*、*n* 分别表示非代表性主题的正负权重和,则正权重代表性主题 *i* 权重由 w_i 重编码为w_i +$m\alpha$,负权重代表性主题 *j* 权重由 w_j 重编码为 w_j + $n\alpha$。其中,α 为权重系数。

(3)全连接层。全连接层以主题全连接结构建立 *K* 竞争层提取的 *K* 项代表性主题,从而构建主题之间全局关系,更准确表达数据。增加全连接层的层数可以提高模型的拟合,但会严重增加模型参数规模。各全连接层主题以线性关系连接:

$$Z_i = W * Z_{i-1} + B \tag{2}$$

其中, Z_i 为各层主题特征;*W* 为权值参数;*B* 为偏置参数。

3 实验分析

本文实验数据来源于 20newsgroup 和 Reuters-21578 两个标准新闻短文本数据集。其中,20newsgroup 由 18 846 篇新闻组成,涉及政治、宗教、计算机科学、体育等 20 类新闻,每篇文档属于一类。Reuters-21578 由路透社新闻报道组成,用以完成信息检索和机器学习等基于语料库的研究。实验根据文档中主题标签,以植物、金融和贸易等 68 类主题词的11 305篇文档为数据集,每篇文档包含一至多个主题词。

实验过程中,选取具有代表性的 4 组主题数(100、200、500 和1 000),以文档主题分布作为

liblinearSVM[15]和 *Softmax* 分类器特征，对比不同主题数的情况下，KFTN 与 LDA 和 BTM 的短文本分类实验准确率。

3.1　20newsgroup 短文本分类实验

为获得更为公平的对比结果，在 20newsgroup 短文本分类中，以 3 次交叉验证的平均分类准确率，对比和评价不同模型。不同模型在 20newsgroup 数据集短文本分类的对比结果如图 4 所示。

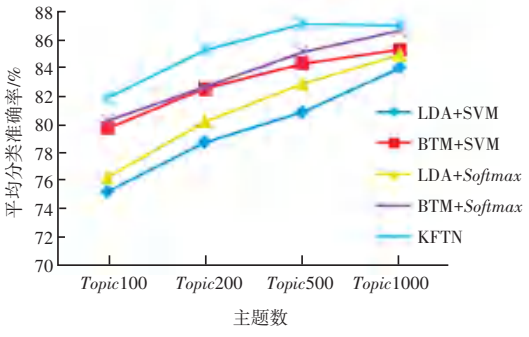


图 4　20newsgroup 短文本分类对比结果

Fig. 4　The comparison results of short－text classification on 20newsgroup

图 4 中，BTM 模型通过“词汇对”建立词汇间的共现关系，克服了词汇特征过于稀疏和忽略词汇关系的不足，其分类准确率高于 LDA 模型。对于相同主题特征，*Softmax* 与线性 SVM 准确率相近，但 *Softmax* 更关注标签与得分的相似度，其准确率略高于线性 SVM。KFTN 以“词汇对”建立底层特征，竞争关系突出重点主题，同时建立主题全局关系，更有效地表达短文本数据，其分类准确率高于其它模型。

20newsgroup 数据集由 20 类新闻组成，主题作为文档的中层特征，并不能直接表示新闻类别。同时随着主题数的增加，主题特征的表达能力也得到增强，分类准确率得到提高。但过高维度的主题特征会造成特征过于稀疏，增加模型参数规模，影响模型学习的效率和应用。因此，当主题数达到 500 时，KFTN 分类准确率趋于稳定。

3.2　Reuters－21578 短文本分类实验

Reuters－21578 数据集中文档为多标签文档，因此本文通过 3 次交叉验证方法，以 $micro\_f_1$ 和 $macro\_f_1$ 对比和评价不同模型。表 1 为不同模型在 Reuters－21578 数据集短文本分类的实验对比结果。

表 1　Reuters－21578 短文本分类对比结果

Tab. 1　The comparison results of short－text classification on Reuters－21578

模型	*Topic*100		*Topic*200		*Topic*500		*Topic*1000	
	$macro\_f_1$	$micro\_f_1$	$macro\_f_1$	$micro\_f_1$	$macro\_f_1$	$micro\_f_1$	$macro\_f_1$	$micro\_f_1$
LDA+SVM	0.14	0.71	0.15	0.69	0.19	0.67	0.22	0.68
BTM+SVM	0.18	0.77	0.24	0.78	0.31	0.79	0.32	0.78
LDA+Softmax	**0.26**	0.75	0.36	0.77	0.45	0.81	0.46	0.82
BTM+Softmax	0.25	0.77	0.34	0.81	0.43	0.83	0.45	0.84
KFTN	0.24	**0.79**	**0.38**	**0.83**	**0.47**	**0.85**	**0.49**	**0.85**

采用 *Softmax* 分类器交叉熵作为损失函数，衡量标签与得分的相似度，更有利于多标签分类，因此其分类准确率略高于线性 SVM 分类器准确率。基于“词汇对”表达方法的 BTM 模型对于短文本的表达能力优于基于词频的 LDA 模型，在各个主题数下，分类准确率都高于 LDA 模型。由于 $macro\_f_1$ 易受到识别性高的类别影响，LDA 主题特征基于主题独立性假设，更易提取识别性高的类别特征。因此，在 *Softmax* 分类过程，LDA 模型的 $macro\_f_1$ 值略高于 BTM 模型。但 LDA 和 BTM 均忽略了词汇关系和主题关系，影响了主题特征表达。KFTN 模型融合词汇关系和主题全局竞争关系，提取的主题特征更为准确有效，因此分类准确率高于 LDA 和 BTM 模型。

4　结束语

针对主题模型等算法处理短文本数据的不足，从短文本数据特点展开研究，本文提出 K 竞争全连接主题网络模型(KFTN)。通过构建“词汇对”表达和引入主题权重竞争，建立词汇语料级关系和主题全局关系，突出重点主题的特征表达，降低了噪声对主题特征的影响。KFTN 克服了主题模型忽略词汇关系和主题关系的不足，增强了主题特征的表达能力，提高了短文本分类的准确性。

参考文献

[1] 陈亚茹, 陈世平. 融合自注意力机制和 BiGRU 网络的微博情感分析模型[J]. 小型微型计算机系统, 2020, 41 (08): 1590-1595.

[2] 牛雪莹, 赵恩莹. 基于 Word2Vec 的微博文本分类研究[J]. 计算机系统应用, 2019, 28(08): 256-261.

[3] JIANG Zhiying, GAO Bo, HE Yanlin, et al. Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports[J]. Mathematical Problems in Engineering, 2021, 2021 (6): 1-30.

[4] HELMSTETTER S, PAULHEIM H. Weakly supervised learning for fake news detection on Twitter[C]// 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona, Spain: IEEE, 2018: 274-277.

[5] 刘德喜, 聂建云, 万常选, 等. 基于分类的微博新情感词抽取方法和特征分析[J]. 计算机学报, 2018, 41 (7): 1574-1597.

[6] ALBITAR S, FOURNIER S, ESPINASSE B. An effective TF/IDF-based text-to-text semantic similarity measure for text classification [M]//BENATALLAH B, BESTAVROS A, MANOLOPOULOS Y, et al. Web Information Systems Engineering - WISE 2014. WISE 2014. Lecture Notes in Computer Science. Cham: Springer, 2014, 8786: 105-114.

[7] 刘浩然, 丁攀, 郭长江, 等. 基于贝叶斯算法的中文垃圾邮件过滤系统研究[J]. 通信学报, 2018, 39 (12): 151-159.

[8] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. the Journal of machine Learning research, 2003, 3: 993-1022.

[9] ZHANG Heng, ZHONG Guoqiang. Improving short text classification by learning vector representations of both words and hidden topics[J]. Knowledge-Based Systems, 2016, 102: 76-86.

[10] 刘爱琴, 马小宁. 基于概率主题模型的短文本自动分类系统构建[J]. 国家图书馆学刊, 2020, 29 (06): 102-112.

[11] YAN Xiaohui, GUO Jiafeng, LAN Yanyan, et al. A biterm topic model for short texts[C]// Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil: ACM, 2013: 1445-1456.

[12] YANG Yi, WANG Hong'an, ZHU Jiaqi, et al. Dataless short text classification based on biterm topic model and word embeddings [C]//Twenty-Ninth International Joint Conference on Artificial Intelligence {IJCAI-PRICAI-20}. 2020: 3969-3975.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. Navada, USA: NIPS Foundation, 2012: 1097-1105.

[14] CHEN Yu, ZAKI M J. Kate: K-competitive autoencoder for text [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2017: 85-94.

[15] FAN Rong'en, CHANG Kaiwei, HSIEH C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research, 2008, 9: 1871-1874.

(上接第 31 页)

UKF 算法预测,在碰撞后采用简单物理运动模型预测乒乓球的轨迹。实验证明改进的 UKF 乒乓球轨迹预测算法可以一定程度上提高轨迹预测的精度,为后续乒乓球机器人击球工作的开展提供了保障。

参考文献

[1] CHEN Xiaopeng, HUANG Qiang, ZHANG Weimin, et al. Ping-pong trajectory perception and prediction by a PC based high speed four-camera vision system[C]//Proceedings of the 8th World Congress on Intelligent Control and Automation. Taipei: IEEE, 2011: 1087-1092.

[2] ZHANG Yifeng, ZHAO Yongsheng, XIONG Rong, et al. Spin observation and trajectory prediction of a ping-pong ball[C]// 2014 IEEE International Conference on Robotics & Automation (ICRA). Hong Kong, China: IEEE, 2014: 4108-4114.

[3] LIN H I, HUANG Yichen. Ball trajectory tracking and prediction for a ping-pong robot[C]// 9th International Conference on Information Science and Technology (ICIST). Hulunbuir, China: IEEE, 2019: 222-227.

[4] CONG V D, HANH L D, PHUONG L H. Real-time measurement and prediction of ball trajectory for ping-pong robot[C]//2020 5th International Conference on Green Technology and Sustainable Development (GTSD). Ho Chi Minh City, Vietnam: IEEE, 2020: 9-14.

[5] TEBBE J, KLAMT L, CAO Yapeng, et al. Spin detection in robotic table tennis[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020: 9694-9700.

[6] SUN Jing, TIAN Jiandong, DU Yingkui, et al. Fast ball detection method for ping-pong playing robots[C]// Proceedings of the 2009 Second International Symposium on Information Science and Engineering (ISISE'09). Washington, DC, United States: ACM, 2009: 339-343.

[7] ZHOU Kai, HUANG Yingping, CHEN Enqing, et al. Real-time detection and spatial segmentation of difference image motion changes[J]. IEEE Access, 2020, 8: 144931-144944.

[8] ZHAO Yongsheng, WU Jun, ZHU Yifeng, et al. A learning framework towards real-time detection and localization of a ball for robotic table tennis system[C]//Proceedings of The 2017 IEEE International Conference on Real-time Computing and Robotics. Okinawa, Japan: IEEE, 2017: 97-102.