文章编号：2095-2163(2020)09-0073-04 中图分类号：TP391 文献标志码：A

基于 TF-IDF 特征词提取的不平衡文本分类

陈 欢，王忠震
（上海工程技术大学，电子电气工程学院，上海 201620）

摘 要：在文本分类的过程中，文本类别分布不均衡会导致分类准确率下降。针对这个问题，本文提出了一种基于注意力机制的不平衡文本分类方法。首先，利用 TF-IDF 对每个类别的特征词进行词特征提取；其次，将提取到的词特征和原有的文本拼接进行注意力权重分配；最后，使用 softmax 分类。实验在复旦大学开源文本数据集上进行，结果表明本文提出的方法相对于其他对比方法更加稳定，准确率有所提高。

关键词：注意力机制；不平衡文本分类；TF-IDF

Unbalanced text classification based on attention mechanism

CHEN Huan, WANG Zhongzhen
(School of Electrical and Electronic Engineering, Shanghai University of Engineering and Technology, Shanghai 201620, China)

【Abstract】 In the process of text classification, the unbalanced distribution of text categories will lead to the decline of classification accuracy. To solve this problem, this paper proposes an unbalanced text classification method based on attention mechanism. First, TF-IDF is used to extract the feature words of each category, and then the extracted feature and the original text are spliced to allocate the attention weight of the feature words. Finally, softmax is used for classification. The experiment is carried out on the open source text data set of Fudan University. The experiment shows that the method proposed in this paper is more stable and the accuracy is improved compared with other comparison methods.

【Key words】 attention mechanism; unbalanced text classification; TF-IDF

0 引 言

随着 web2.0 时代的到来，我国的网民规模飞速增长，达到了 9.04 亿，网络文本数据也随时间大量累积。对文本分类、整理，发掘文本中的潜在信息成为了研究的热点。然而在实际的网络文本分类过程中，类别分布不均衡制约着文本分类技术的发展。

传统的解决数据类别分布不均衡的方法是通过重采样，数据增强等方法。如张忠林等针对不平衡分类过程中，数据集中存在噪声数据使得边界模糊的现象，提出了将少数样本划分，只对边界样本进行 SMOTE 插值，然后数据清洗，去除噪声数据[1]；蒋华等针对不平衡数据集分类时边界偏移的问题，提出用 ADASYN 和 SMOTE 算法生成小类样本点[2]；史明华等通过使用聚类算法进行聚类，根据类别簇不平衡比的大小对该簇进行相应的处理[3]。

随着深度学习技术的发展，给文本的不平衡分类技术发展提供了新的思路。如陈志等针对训练神经网络模型时参数会被多数类所主导，在损失函数中加入类别标签，强化少数类对模型参数的影响[4]；林怀逸等利用小类别区分的预训练词向量来初始化目标模型，并结合均衡过采样，保持模型在大类别上的精度[5]；万志超等针对文本分布不均衡分类时局限于特征维数过高、数据稀疏、分布不均衡的特点，通过使用有监督的特征选择方法，减少特征词数量，降低特征维度[6]；程艳等提出将不平衡数据划分为若干组均衡数据，使用 CNN 神经网络训练，并使用 EWC 克服 CNN 的灾难性遗忘的缺点[7]；唐焕玲等使用有监督的主题模型 SLDA，建立主题和稀少类别之间的映射，以提高少数类分类的精度[8]；钟将等针对文本特征维度大和训练样本分布不均衡的问题，提出使用 LSA 降维，并利用改进的 KNN 进行文本分类[9]。

综上所述，在进行数据不平衡分类的过程中，主要通过强化类别的边界，去除噪声数据等方法[10]。与其不同的是，在文本分类过程中，解决数据类别分布不均衡的方法，主要有强化类别标签、过采样等方法。因此，本文通过使用 TF-IDF 构建类别特征词，与原有文本拼接来强化各类的类别特征，并使用注意力机制进行文本特征权重分配。

本文的主要工作如下：

(1)利用 TF-IDF 给文本中词赋权的方法进行类别关键词提取。将数据集划分为若干个平衡的子

作者简介：陈 欢(1992-)，男，硕士研究生，主要研究方向：文本分类、情感分析。

收稿日期：2020-06-11

数据集,输入到 TF-IDF 模型进行类别关键词提取。

(2)将训练集和测试集输入到 word2vec 词嵌入模型进行词向量训练,得到 TF-IDF 提取到的关键词和原有的文本数据拼接后的词向量表达,输入到注意力机制模型训练和权重分配,最终进行文本分类。

1 基础理论

1.1 TF-IDF 特征权重计算方法

TF-IDF 是一种用于信息检索的文本加权技术,在文本信息检索的过程中,通过对文本赋予不同的权重,从而判断与检索词的关系,提高检索的准确率和召回率。

TF-IDF 的具体思想可以表述为:在一篇文章中,如果一个词在该篇文章中出现的频率较高,而在语料集的其它文章中出现的频率较低,则该词更能代表该篇文章。其中,TF 表示词频;IDF 表示包含该词的文档数目。在数学上可以表示为公式(1)、(2)、(3)。

$$TF - IDF = tf_{i,j} * idf_i, \tag{1}$$

$$tf_{i,j} = \frac{n_{i,j}}{|j|}, \tag{2}$$

$$idf_i = \log \frac{|D|}{df_i + 1}. \tag{3}$$

其中,$n_{i,j}$ 表示词语 i 在文档 j 中的频率;$|j|$ 表示文档 j 中词的总数;$|D|$ 表示语料集中文档的总数;df_i 表示语料集中包含词语 i 的文档总数。idf_i 的计算过程中分母加 1 是为了防止违反运算法则的情况出现。

1.2 LDA 文本降维

LDA 模型是一种主题概率模型,将文本表示为文本-主题、主题-词的概率分布。LDA 的概率图模型如图 1 所示。其中,K 表示主题数;D 表示文档数;N 表示一篇文档中词的数目。

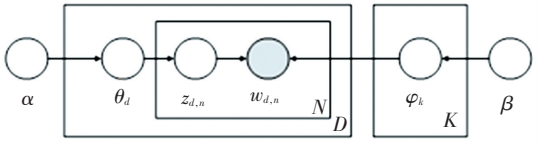

图 1 LDA 概率图模型

Fig. 1 LDA probability diagram model

(1)模型假设文档的主题先验分布服从参数为 α 的 *Dirichlet* 概率分布,其中文档 d 的主题概率分布为 $\theta_d = Dirichlet(\alpha)$。

(2)模型假设主题中的词的先验分布服从参数为 β 的先验概率分布。其中,主题 k 的词概率分布为 $\varphi_k = Dirichlet(\beta)$。

(3)文档 d 中的第 n 个词,从主题分布获得其主题编号概率分布为 $z_{dn} = multi(\theta_d)$。

(4)文档 d 中的第 n 个词分布 w_{dn} 的概率分布为 $w_{dn} = multi(\varphi_{z_{dn}})$。

由于 Dirichlet-multi 是共轭分布,可以利用贝叶斯推断的方法求得后验分布,在得到文档主题,主题词的后验概率分布后,利用 Gibbs 采样的方法获得每个文档的主题分布和每个主题的词分布。

1.3 注意力机制

注意力机制首先被提出用于图像特征提取领域,其次被 Bahdanau 等人推广到自然语言处理领域。其思想可以描述为通过改变模型参数来加强某个输入对输出的影响[11-12]。其中 google 提出的最初注意力计算方法如公式(4)所示,k_s (key)与 v_s (value)一一对应,通过计算 q_t (query)和各个 k_s 的内积,求得与各个 v_s 的相似度,然后加权求和归一化。

$$Attention(q_t, K, V) = \sum_{s=1}^{m} \frac{1}{Z} \exp\left(\frac{< q_t, k_s >}{\sqrt{d_k}}\right) v_s, \tag{4}$$

其中,$\sqrt{d_k}$ 为输入词嵌入向量的维度,起到调节因子的作用,使得内积不至于太大。

2 模型描述

2.1 模型框架

基于词嵌入的不平衡文本数据分类框架如图 2 所示。

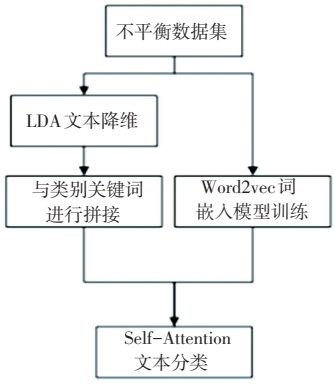

图 2 模型结构

Fig. 2 Model structure

模型主要分为三个部分:首先,使用 TF-IDF 进行类别关键词提取,与原有文本拼接,输入到 word2vec 模型进行词向量训练;其次,使用 TF-IDF 对文本降维,并和类别关键词拼接,并将其用词向量表示;最后,使用 Self-Attention 对词向量表示后的文本进行特征权重分配。

2.2　TF-IDF 类别关键词特征提取

类别关键词特征作为类别之间的区分，具有明显的类别特性。将数据集划分为若干个平衡的子数据集，输入到 TF-IDF 模型，获得子数据集每个文本的 TF-IDF 表示，统计每个类别的 TF-IDF 权值大的词作为类别的关键词特征。

划分为平衡数据集是为了 TF-IDF 在词特征提取的过程中能够有更好的效果。否则可能出现少数类别文章关键词存在本文属于高频，而在语料集中很少出现，就会导致该词的权重过大，但该词并不能代表该类别。

2.3　LDA 文本降维

将文本输入到 LDA 模型，得到每篇文章的主题词分布和主题分布，通过这两个分布可以将文章的主要特征进行表示，从而实现文章的降维。设每篇文章的主题词分布为 $t\_w=[w_1,w_2,\dots,w_N]$，文章主题分布为 $d\_t=\{z_1,z_2,\dots,z_K\}$。通过将两个分布对应相乘，选择结果较大的词作为 LDA 降维后的文本特征，并将其用词向量的形式表示，进行下一步的操作。

2.4　Self-Attention 权重分配

传统的注意力机制通过计算源端的每个词与目标端的每个词之间的依赖关系来更新训练参数，Self-Attention 机制仅通过关注自身信息更新训练参数，不需要添加额外的信息。将前述通过 CBOW 模型得到的融合主题特征的评论文本向量输入到 Self-Attention 层，通过公式(5)计算权重分布。

$$\text{Self-Attention}=softmax\left(\frac{W\cdot W^T}{\sqrt{d_k}}\right)W. \tag{5}$$

2.5　模型分类

将注意力机制编码后获得的文本信息，使用交叉熵作为损失函数，利用 adam 更新网络参数。利用公式(6)求解文本特征向量 γ_x 属于类别 y_x 的概率，n_c 为类别的数目。以公式(7)为损失函数，其目的是通过迭代的更新参数最小化监督标签 g_x 和预测标签之间的交叉熵。

$$y_x=\frac{\exp(\gamma_x)}{\sum_{q=1}^{n\_c}\exp(\gamma_q)}, \tag{6}$$

$$L=-\sum_{n\_a}g_x\log y_x. \tag{7}$$

3　实验分析

3.1　实验数据集

实验数据集采用复旦大学中文文本分类数据集，该数据集分为训练集和测试集两部分，共有 20 个类别，类别数最多的文本有 1 357 篇，最少文本的只有 27 篇。本文选择其中文本较多的 9 个类别进行文本分类。各个类别分布如图 3 所示，其中类别数最多的有 1 357 篇，最少的有 466 篇。

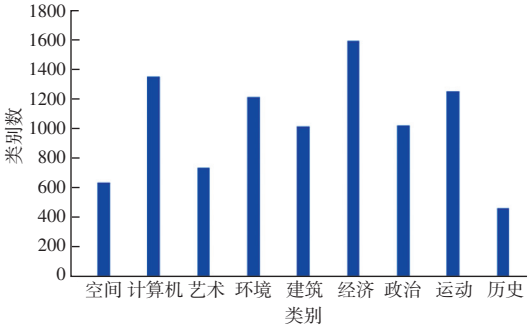

图 3　训练集各类别数据分布

Fig. 3　Data distribution of training sets

3.2　TF-IDF 类别关键词提取

将训练的数据集划分为两个训练集，输入到 TF-IDF模型进行训练。其中，低于 1 000 的数据集进行两次模型训练，高于 1 000 的划分为两个部分输入到模型训练。

TF-IDF 提取到的类别关键词特征示例，见表 1。可以看到，每个领域的特征词都有明显的领域特征，因此与原有文章进行拼接可以加强少数类的类别特征，从而提高文本分类的准确率。

表 1　TF-IDF 提取类别关键词特征示例

Tab. 1　TF-IDF extract category keyword feature example

	类别关键词特征示例
太空	液滴、雷达、图层、起动机、飞机、传感器、可靠性、雷电、图层、飞船、发射
计算机	数据库、服务器、线程、分类器、数据仓库 、机器人、图象、服务器、调度、消息、网络
艺术	歌剧、艺术史、后现代、审美、影视、放映论、神话、小说、演绎、修辞、油画、大众文化、
环境	事故、水库、抽烟、生态、空气质量、采矿、清洁、燃料、实惠、高岭土、水质、固体废物
农业	农业投入、棉田、旅游、农产品、超高产、污泥、胡萝卜、农机、旱区、优质、生态、现代农业
经济	公共财政、合同法、道德、产业、数理经济学、发展经济学、会计信息学、工会、剥削、计划
政治	形式主义、马克思、纪检监察、政治腐败、人口、发展、保守主义、合法性、媒体、行为主义
运动	学校、道德教育、女足、体育老师、素质奥羽、思维、花钱、跆拳道、竞争、交流、女排、射门
历史	巴比伦、勾践、马克思、陈列、资产阶级、西方、澳门、人文竞赛、长征、戏剧、悲剧、新文化

3.3　模型对比分析

目前的分类评价方法评价指标有精确度、召回率和 F1 值，本文也采用这些指标进行分类结果评价。

使用 gensim 库进行 LDA 和 word2vec 词嵌入模

型训练,同时和其它几种基于 LDA 和 word2vec 的模型进行训练,得到准确率对比,见表 2。实验证明了本文提出的方法优于其它的传统方法。

表 2 结果对比分析

Tab. 2 Comparative analysis of results

	准确率/%	召回率/%	F1 值/%
SVM	87.42	90.41	88.91
LDA+SVM	88.54	92.98	90.23
word2vec+SVM	91.43	94.12	92.76
LDA+word2vec+SVM	89.14	88.18	88.64
本文方法	94.24	94.12	94.17

4 结束语

针对文本数据分类不均衡的问题,本文提出使用 TF-IDF 进行类别关键词特征提取,然后输入到注意力机制模型进行文本分类。在复旦大学语料集上证明了本文提出模型优于其它的经典模型,具有更好的分类效果。但本文提出的模型也有一定的不足,如在 TF-IDF 特征词提取的过程中,TF-IDF 不能得到很好的效果,其中有一些词不具有类别代表性,因此需要对其进行人工筛选。

参考文献

[1] 张忠林,曹婷婷. 基于重采样与特征选择的不均衡数据分类算法[J]. 小型微型计算机系统,2020,41(6):1327-1333.

[2] 蒋华,江日辰,王鑫,等. ADASYN 和 SMOTE 相结合的不平衡数据分类算法[J]. 计算机仿真,2020,37(3):254-258,420.

[3] 叶雪梅,毛雪岷,夏锦春, 等. 文本分类 TF-IDF 算法的改进研究[J]. 计算机工程与应用,2019,55(2):104-109,161.

[4] 陈志,郭武. 不平衡训练数据下的基于深度学习的文本分类[J]. 小型微型计算机系统,2020,41(1):1-5.

[5] 林怀逸,刘箴,柴玉梅,等. 基于词向量预训练的不平衡文本情绪分类[J]. 中文信息学报,2019,33(5):132-142.

[6] 万志超,胡峰,邓维斌. 面向不平衡文本情感分类的三支决策特征选择方法[J]. 计算机应用,2019,39(11):3127-3133.

[7] 程艳,朱海,项国雄,等. 融合 CNN 和 EWC 算法的不平衡文本情绪分类方法[J]. 中文信息学报,2020,34(4):92-100.

[8] 唐焕玲,刘艳红,郑涵,等. 融合 SLDA 主题模型的不均衡文本分类方法[J/OL]. 计算机工程与应用:1-11[2020-06-11]. http://kns.cnki.net/kcms/detail/11.2127.TP.20200529.1707.022.html.

[9] 钟将,刘荣辉. 一种改进的 KNN 文本分类[J]. 计算机工程与应用,2012,48(2):142-144.

[10] 李勇,刘占东,张海军. 不平衡数据的集成分类算法综述[J]. 计算机应用研究,2014,31 (5):1287-1291.

[11] 朱张莉,饶元,吴渊,等. 注意力机制在深度学习中的研究进展[J]. 中文信息学报,2019,33(6):1-11.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in neural information processing systems. 2017: 5998-6008.

(上接第 72 页)

[3] 杨玥. 中文文本主题关键短语提取算法研究[D]. 西安: 西安理工大学,2017.

[4] DavidMcHugh, Sarah Shaw, Travis R. Moore, et al. Uncovering themes in personalized learning: Using natural language processing to analyze school interviews [J]. Journal of Research on Technology in Education,2020,52(3):391-402.

[5] HABIBI M, POPESCU-BELIS A. Keyword extraction and clustering for document recommendation in conversations [J]. IEEE/ACM Transactions on Audio Speech and Language Processing ,2015,23 (4) : 746-759.

[6] 牛永洁,田成龙. 融合多因素的 TF-IDF 关键词提取算法研究[J]. 计算机技术与发展,2019,29(7):80-83.

[7] 但唐朋,许天成,张姝涵. 基于改进 TF-IDF 特征的中文文本分类系统[J]. 计算机与数字工程,2020,48(3):556-560.

[8] HORITA K, KIMURA F, MAEDA A. Automatic keyword extraction for wikification of east asian language documents[J]. International journal of computer theory and engineering, 2016, 8 (1): 32-35.

[9] 吴晶晶,荆继武,聂晓峰,等. 一种快速中文分词词典机制[J]. 中国科学院研究生院学报,2009,26(5):703-711.

[10] LV N, LIANG X, CHEN C, et al. A Long Short-Term Memory Cyclic model With Mutual Information For Hydrology Forecasting: A Case Study in the Xixian Basin [J]. Advances in Water Resources, 2020: 103622.

[11] 林雷蕾,杨良,闻立杰,等. 基于信息熵的无标日志划分评价方法[J]. 计算机集成制造系统,2020,26(6):1483-1491.

[12] 张成,褚莹,凌力. 基于安全字典树的关键词密文模糊搜索方案[J]. 微型电脑应用,2018,34(4):33-36.

[13] 张建娥. 基于 TF-IDF 和词语关联度的中文关键词提取方法[J]. 情报科学,2012,30(10):1542-1544,1555.

[14] Willyan D. Abilhoa, Leandro N. de Castro. A keyword extraction method from twitter messages represented as graphs[J]. Applied Mathematics and Computation,2014,240.

[15] 陈伟鹤,刘云. 基于词或词组长度和频数的短中文文本关键词提取算法[J]. 计算机科学,2016,43(12):50-57.

[16] ZHOU Zhuo, QIN Jiaohua, XIANG Xuyu, et al. News Text Topic Clustering Optimized Method Based on TF-IDF Algorithm on Spark [J]. CMC: Computers, Materials & Continua, 2020,62(1):217-231.