基于高效用神经网络的文本分类方法

吴玉佳,李 晶,宋成芳,常 军
(武汉大学计算机学院,湖北武汉 430072)

摘 要: 现有的基于深度学习的文本分类方法没有考虑文本特征的重要性和特征之间的关联关系,影响了分类的准确率.针对此问题,本文提出一种基于高效用神经网络(High Utility Neural Networks,HUNN)的文本分类模型,可以有效地表示文本特征的重要性及其关联关系.利用高效用项集挖掘(Mining High Utility Itemsets,MHUI)算法获取数据集中各个特征的重要性以及共现频率.其中,共现频率在一定程度上反映了特征之间的关联关系.将 MHUI 作为 HUNN 的挖掘层,用于挖掘每个类别数据中重要性和关联性强的文本特征.然后将这些特征作为神经网络的输入,再经过卷积层进一步提炼类别表达能力更强的高层次文本特征,从而提高模型分类的准确率.通过在 6 个公开的基准数据集上进行实验分析,提出的算法优于卷积神经网络(Convolutional Neural Networks,CNN),循环神经网络(Recurrent Neural Networks,RNN),循环卷积神经网络(Recurrent Convolutional Neural Networks,RCNN),快速文本分类(Fast Text Classifier,FAST),分层注意力网络(Hierarchical Attention Networks,HAN)等 5 个基准算法.

关键词: 数据挖掘;关联规则;高效用项集;自然语言处理;文本分类;神经网络
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2020)02-0279-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.02.008

High Utility Neural Networks for Text Classification

WU Yu-jia,LI Jing,SONG Cheng-fang,CHANG Jun
(*School of Computer Science*,*Wuhan University*,*Wuhan*,*Hubei* 430072,*China*)

Abstract: The existing text classification methods based on deep learning do not consider the importance and association of text features. The association between the text features perhaps affects the accuracy of the classification. To solve this problem,in this study,a framework based on high utility neural networks (HUNN) for text classification were proposed. Which can effectively mine the importance of text features and their association. Mining high utility itemsets(MHUI) from databases is an emerging topic in data mining. It can mine the importance and the co-occurrence frequency of each feature in the dataset. The co-occurrence frequency of the feature reflects the association between the text features. Using MHUI as the mining layer of HUNN,it is used to mine strong importance and association text features in each type,select these text features as input to the neural networks. And then acquire the high-level features with strong ability of categorical representation through the convolution layer for improving the accuracy of model classification. The experimental results showed that the proposed model performed significantly better on six different public datasets compared with convolutional neural networks (CNN),recurrent neural networks (RNN),recurrent convolutional neural networks (RCNN),fast text classifier (FAST), and hierarchical attention networks (HAN).

Key words: data mining; association rule; high utility itemset; natural language processing; text classification; neural networks

1 引言

文本分类是当前自然语言处理(Natural Language Processing,NLP)领域的一个热点的研究问题[1,2].传统的一些算法如支持向量机[3]、朴素贝叶斯[4]、决策树[5]、逻辑回归[6]以及 TF-IDF(Term Frequency-Inverse

收稿日期:2018-10-08;修回日期:2019-10-22;责任编辑:覃怀银
基金项目:国家重点基础研究发展规划(973 计划)项目(No.2012CB719905);国家自然科学基金(No.41201404);中央高校基本科研业务费专项资金项目(No.2042015gf0009)

Document Frequency)[7]等都用于解决文本分类问题.但是,由于这些算法没有使用深度神经网络,缺乏很好的特征检测和提取的能力,在文本分类任务中,准确率低于基于深度学习的模型[8,9]. Kim[2]在 2014 年提出一种基于 CNN 的文本分类模型,它是一种浅层神经网络[2,10]. CNN 在语义分析[11]、信息检索[12]、情感分类[13]和个性化推荐[14]等各种任务中都取得了成功.在此基础上,Johnson[15]设计了一个多达 9 层的神经网络. Tang[13]则考虑用户个人偏好和产品整体质量的情况下解决情感分类问题. FAST 的目的是提高算法的运算速度[16].另一种常用的神经网络是 RNN[17],它在提取特征的过程中,可以利用上下文信息. Siwei[18]提出了一种 RCNN 网络,结合了 CNN 和 RNN 的特点.注意力机制可以很直观的解释每个文本特征在文档中的重要性[19].

在深度学习处理 NLP 的任务中,词向量表示是一个重要的步骤. Word2vec[20]正好用于处理词向量的表示问题.另一种广泛使用的词向量表示工具是 Glove[21].现有基于深度学习的算法中,其性能很大程度上取决于所选择特征[22].因此,许多工作的重点都是研究如何设计特征[23,24],或者从神经网络中学习特征[25~27].再通过使用词向量表示工具进行预训练,用于获取特征之间的距离关系.但是利用词向量工具依然无法获取文本特征的共现频率.共现频率指的是两个强关联词在同一个样本中同时出现的频率,共现频率高的关联词往往具有更强的类别表达能力[28~30].

因此,如何挖掘共现频率高的文本特征,提升神经网络的分类性能是一个值得研究的问题.本文通过引入高效用项集挖掘算法,挖掘重要性和关联性强的文本特征,再利用卷积操作进一步提取类别表达能力更强的高层次特征,以提高文本分类模型的准确率.通过在不同类型的数据集上进行对比实验,与 5 个基准算法进行比较,提出的方法获得了较好的性能提升.

2 高效用神经网络模型框架

为了解决现有方法存在的挑战,本文提出一种基于高效用神经网络的文本分类方法.高效用神经网络是在卷积神经网络的基础上,加入一个挖掘层,用于对文本特征进行提炼,从而提升模型的准确率.

2.1 卷积神经网络

基于深度学习的文本分类模型中,卷积神经网络是一种性能良好并且训练速度较快的浅层神经网络模型[2].这种方法的主要思路是利用卷积提取文本的局部特征,它的网络结构包括输入层、卷积层、池化层、全连接层和输出层.然后再利用词向量表示工具[20,21]进行训练,并且在训练过程中,保持这些词向量不变,而只学习网络的其他一些参数.其中,卷积层和最大池化层是 CNN 网络的特征提取层.

特征提取包括两个步骤:第一步是使用卷积操作提取特征;第二步是对这些特征使用最大池化进行信息聚合.样本中的每一个词用 $i_k \in R^m$ 表示,它表示句子中第 k 个词的 m 维词向量.一个包含 k 个词的句子 T 表示方法如式(1)所示.

$$T = \{i_1, i_2, \cdots, i_k\} \tag{1}$$

一个卷积滤波器表示为 $\boldsymbol{w}_n \in R^{hk}$,它表示一个在 h 个单词的窗口产生一个新的特征.特征 y_n 是对句子 T 使用式(2)产生的特征.

$$y_n = f(\boldsymbol{w} \cdot T + b) \tag{2}$$

其中,$b \in R$ 是偏置项,$f(\cdot)$是一个非线性双曲正切函数.

使用不同窗口大小的 n 个滤波器对句子 T 进行卷积操作,以获取多个特征.然后得到一组特征图 $\boldsymbol{s} = (y_1, y_2, \cdots, y_n)$.为了降低网络的复杂度,同时保留最显著的特征,需要对特征图进行最大池化操作,如式(3)所示.

$$q = \max(\boldsymbol{s}) \tag{3}$$

2.2 高效用项集挖掘

MHUI 作为高效用神经网络模型框架的挖掘层,用于对文本特征进行提炼.高效用项集不同于传统的频繁项集挖掘算法[28~30],它不仅考虑一个词在句子中是否出现,并且能表示词出现的次数.高效用项集挖掘是针对频繁项集挖掘算法的一些不足所提出的新方法[28].

在高效用神经网络模型框架中,对于每一个类别的文本,使用高效用项集挖掘算法分别获取此类文本中表达能力强的特征,然后将这些特征作为神经网络的输入,再使用卷积操作进一步提取特征,从而提升神经网络的分类能力.其中,效用值是高效用项集算法表示词的重要性和关联性的量化指标.特征的效用值越高,则特征的重要性和关联性越强.例如,在著名的搜狗语料库中,“苹果-商店”在“科技”类中会频繁共现,而“苹果-果实”在“教育”类中会频繁共现.如果一个未知类别的文本出现了关联词“苹果-商店”,那么它具有极高的概率是属于“科技”类,而仅凭单词“苹果”则难以判断其具体的类别.因此,共现频率对分类结果的具有重要影响.下面,简单介绍如何挖掘高效用用项集.

在本节中,每个项表示一个词.其相关定义如下:某一类别的文档 $D = \{T_1, T_2, \cdots, T_n\}$,包含一组句子.其中,每一个句子 $T_d (1 \leqslant d \leqslant n)$ 包含 k 个词,$T_d = \{i_1, i_2, \cdots, i_k\}$,每个词 i_k 在句子 T_d 中出现的次数为词的效用值 $U(i_k, T_d)$.表 1 为一个具体的示例,文档 D 有 4 个句子,每个句子有若干词,使用大写字母表示.

表 1　文档 D 示例

编号	句子	事务数据表示
T_1	*CACECE*	$(A,1)$ $(C,3)$ $(E,2)$
T_2	*ABAFEF*	$(A,2)$ $(B,1)$ $(E,1)$ $(F,2)$
T_3	*DBDFD*	$(B,1)$ $(D,3)$ $(F,1)$
T_4	*BDCDBE*	$(B,2)$ $(C,1)$ $(D,2)$ $(E,1)$

定义 1　词 i_k 在句子 T_d 中的效用 $U(i_k,T_d)$,定义为 $U(i_k,T_d)=\mathrm{Count}(i_k,T_d)$,其中 $\mathrm{Count}(\cdot)$表示计数. 例如,在表 1 中,$U(\{C\},T_1)=3$.

定义 2　项集 X 在句子 T_d 中的效用 $U(X,T_d)$,定义为 $U(X,T_d)=\sum_{i_k\in X\wedge X\subseteq T_d}U(i_k,T_d)$. 例如,$\mathrm{U}(\{AC\},T_1)=4$.

定义 3　项集 X 在文档 D 中的效用 $U(X)$,定义为 $U(X)=\sum_{X\subseteq T_d\wedge T_d\subseteq D}U(X,T_d)$. 例如,$U(\{AE\})=6$.

定义 4　给定一个用户定义最小效用阈值 E,若 $U(X)\geq E$,则称项集 X 为高效用项集. 例如,在表 1 中,假设 $E=4$,则$\{AE\}$是一个高效用项集,因为它的效用值等于 6.

每个类别的高效用 2-项集组成一个过滤器 F_k,F_k 表示所有高于阈值的高效用 2-项集的集合,计算方法如式(4)所示.

$$F_k=\{X\mid X\in D^k,U(X)\geq E\} \tag{4}$$

其中,k 表类别标签,D^k 表示第 k 类文档集.

高效用项集挖掘是数据挖掘领域的一个重要的研究任务,挖掘高效用项集的方法,可以参考文献[29,30].

2.3　高效用神经网络

在本文中,分别对每个类别的样本数据进行高效用项集挖掘. 共中,对于每一个类别的数据,将数据分为测试集和训练集,仅对训练集进行高效用项集挖掘. 图 1 给出了高效用神经网络框架.

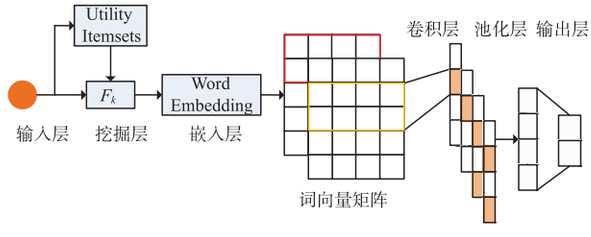

图1　高效用神经网络框架

在图 1 中,模型的输入数据是未经处理的原始文档. 首先对文档数据进行高效用项集挖掘,高于效用阈值的高效用 2-项集是需要保留的项集. 第 k 类文档挖掘到的所有高效用 2-项集组成此类文档的过滤器 F_k. 模型的第 3 层为嵌入层,利用词向量工具对文本数据进行词向量表示,得到词向量矩阵,矩阵的每一行表示一个单词. 然后使用大小不同的卷积核对词向量矩阵进行卷积操作,进一步提取文本的高层次特征,组成特征图. 池化层采用最大池化操作,保留特征中最显著的特征,同时降低网络的复杂度.

最后一层是输出层,输出是样本的类别概率分布. 它的输入是经过池化操作后的这些特征图,使用全连接网络传递给输出层,最后得到输出类别的概率分布. 在输出层中,使用 Softmax 函数构建分类器,如式(5)所示.

$$\mathrm{Softmax}_j=\frac{\exp(q)}{\sum_{j=1}^{c}\exp(q_j)} \tag{5}$$

其中,c 表是类别,q 是全连接层的输入.

本文提出的文本分类模型 HUNN 的源代码已经上传至 https://github.com/wyjwhucs/HUNN.

3　实验结果与分析

3.1　实验数据集与基准算法

根据词向量表示方法的不同,设置 3 组实验分别采用了 One-hot、Word2vec 和 Glove 词向量编码. 并且在这些实验中,通过使用高效用项集挖掘,设置不同的阈值来过滤表达能力弱或者对分类无意义的特征.

实验采用了 6 个基准数据集,包括:MR,MPQA,SUBJ,SST1,SST2,TREC 等文本分类领域的真实数据集[2].

基准算法包括:CNN[2],FAST[16],HAN[19],RNN[17],RCNN[18]. 实验使用 10% 的数据作为测试数据,词向量维度为 128/300. 使用 2,3,4,5 等四种大小的卷积核,数量为 128.

3.2　实验分析

表 2 ~ 4 给出本文提出算法和基准算法在 6 个不同类型公开数据集上的实验结果,最优结果使用加粗表示.

表 2 是 One-hot 编码下的实验结果,提出的 HUNN 模型的性能全面超过其它 5 个基准算法. 其中,在 MR 数据集上,HUNN 的准确率高于基准算法 2% 以上. 在 MPQA 数据集上,HUNN 的准确率为 85%,高于其它算法 1% 左右. 在 SST1 数据集上,HUNN 算法在这个数据集上的准确率为 42.4%,而其它几个基准算法都低于 40%. 在 SST2 数据集上,HUNN 的准确率为 81%,其它几个数据集都不高于 80%. 在 TREC 数据集上,HUNN 的准确率高于其它算法 2% 以上.

在使用 Word2vec 词向量工具的情况下,实验结果如表 3 所示. HUNN 在 MR 数据集上获得了 80.9% 的准确率,较在 One-hot 编码下提升了 2 个百分点,并优于 5 个基准算法. 在 MPQA 数据集上,HUNN 和 RNN 的准确

率皆为88.3%,优于其它几个算法.在SUBJ数据集上,HUNN的准确率为93.9%,RCNN为93.2%,剩下的4个算法都低于93%.在SST1数据集上,HUNN的准确率为44.1%,优于其它5个基准算法.在SST2数据集上,CNN和HUNN的准确率为84.4%,优于其它4个算法.在TREC数据集上,HUNN和RNN的准确率皆为88.4%,略高于另外4个基准算法.

表2 One-hot编码下各模型的实验结果

Model	MR	MPQA	SUBJ	SST1	SST2	TREC
CNN	74.5	81.3	90.4	38.8	79.5	84.4
FAST	65.2	69.6	83.3	26.5	68.7	46
HAN	74.9	80.4	90.1	36.4	77.2	78.6
RNN	75.1	84.2	90.6	38.1	75.8	79.7
RCNN	76.1	82.6	90.6	39.1	78.9	84.3
HUNN	**78.9**	**85**	**92.1**	**42.4**	**81**	**87**

表3 Word2vec编码下各模型的实验结果

Model	MR	MPQA	SUBJ	SST1	SST2	TREC
CNN	77.6	88.1	92.7	42.3	**84.4**	88.2
FAST	70.3	68.4	86.4	34.1	73.1	39.9
HAN	71.3	85.7	91.7	42.1	83.5	55.1
RNN	78.5	**88.3**	92.5	32.1	79.8	**88.4**
RCNN	79.3	87.6	93.2	44	83.5	88.2
HUNN	**80.9**	**88.3**	**93.9**	**44.1**	**84.4**	**88.4**

表4 Glove编码下各模型的实验结果

Model	MR	MPQA	SUBJ	SST1	SST2	TREC
CNN	78.9	89.2	93.2	43.8	**85.6**	88.2
FAST	71.1	69.6	88.6	30	71.8	40
HAN	77.2	87.6	91.5	35.8	79.4	67.9
RNN	79.2	88.6	93.6	43.8	83.9	86.9
RCNN	80.5	88.7	**94.1**	44.7	85.3	89.4
HUNN	**80.6**	**89.3**	**94.1**	**45.2**	85.4	**89.7**

表4是Glove编码下HUNN的实验结果.HUNN算法在MR数据集上,它的准确率为80.6%,优于其它5个算法.在MPQA数据集上,HUNN的准确率为89.3%,优于其它算法.在SUBJ数据集上,HUNN和RCNN的准确率皆为94.1%,优于其它算法.在SST1数据集上,HUNN的准确率高于其它算法1%左右.而在SST2数据集上,CNN取得了最好效果,准确率达到了85.6%,而HUNN的准确率为85.4%.在TREC数据集上,HUNN的准确率为89.7%.

3.3 不同的阈值对结果的影响

定义不同大小的阈值E会挖掘到不同数量的高效用项集,从而产生不同数量的文本特征,对实验效果也会产生影响.

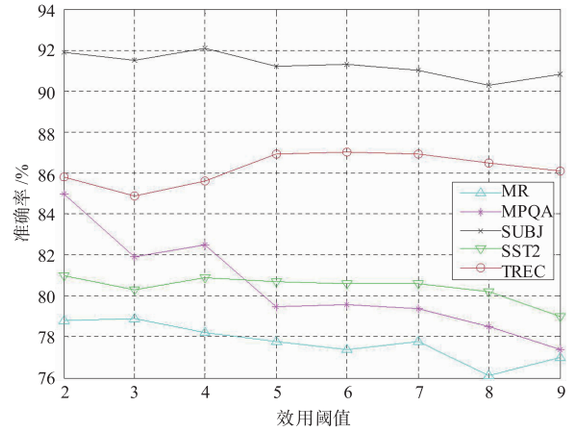

图2 One-hot编码下HUNN在不同阈值下的实验结果

图2给出了HUNN算法在5个数据集上不同阈值的实验结果.可以看到,过高或者过低的效用阈值都会影响实验结果.因为过高的效用阈值可能会过滤掉太多的重要特征,导致准确率降低.过低的效用阈值会让许多效用值不高的特征通过挖掘层,也影响准确率.随着阈值的增加,算法的准确率整体上趋于下降,说明阈值不宜设置过高.

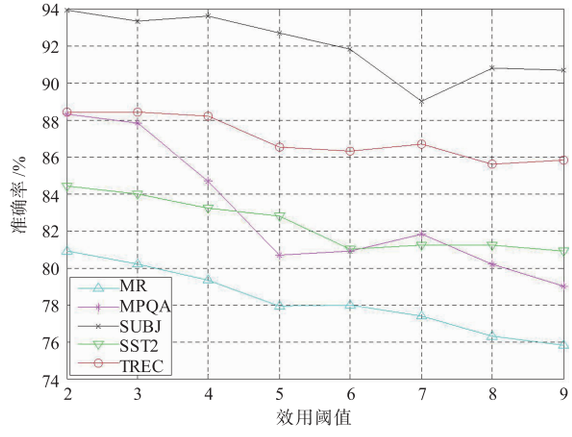

图3 Word2vec编码下HUNN在不同阈值下的实验结果

图3给出了HUNN算法在使用Word2vec工具的情况下,分别在5个数据集上设置不同的阈值情况下的实验结果.可以看到,阈值设置为2时,实验效果最好,随着阈值的提升,准确率会有所降低.

图4给出了HUNN算法在使用Glove工具的情况下,分别在5个数据集上设置不同的阈值情况下的实验结果.可以看到,在TREC数据集上,阈值为4时,实验效果最好.而在其它5个数据集上,阈值设置为2时,实验效果最好,随着阈值的提升,准确率会有所降低.

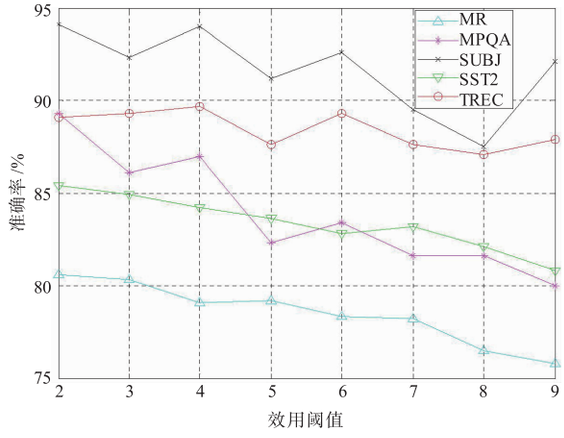

图4　Glove编码下HUNN在不同阈值下的实验结果

4　结论

在本文中,提出了一种基于高效用神经网络的文本分类模型. 针对频繁项集存在的问题,引入了高效用项集. 挖掘重要性和关联性强的文本特征作为神经网络的输入. 通过在 6 个不同的数据集上,和 5 个基准算法进行对比实验. 在 One-hot 编码下,算法的性能优于全部 5 个基准算法. 在使用 Word2vec 工具和 Glove 工具的情况下,也取得了不错的实验效果,在其中 5 个数据集上都取得了最佳的实验结果. 并且,通过设置不同的阈值测试 HUNN 的性能,实验结果表明,阈值设置过高或者过低,算法都不能取得最好的效果,合适的阈值需要通过实验确定.

致谢　感谢审稿专家给本文提出的宝贵意见.

参考文献

[1] 胡小娟,刘磊,邱宁佳. 基于主动学习和否定选择的垃圾邮件分类算法[J] 电子学报,2018,46(1):203 – 209.
HU Xiao-juan,LIU Lei,QIU Ning-jia. A novel spam categorization algorithm based on active learning method and negative selection algorithm[J]. Acta Electronica Sinica, 2018,46(1):203 – 209.

[2] Y Kim. Convolutional neural networks for sentence classification[A]. Conference on Empirical Methods in Natural Language Processing[C]. Doha, Qatar: ACL, 2014. 1746 – 1751.

[3] T Mullen,N Collier. Sentiment analysis using support vector machines with diverse information sources[A]. Conference on Empirical Methods in Natural Language Processing [C]. Barcelona,Spain:ACL,2004. 412 – 418.

[4] S Tan,X Cheng,Y Wang,et al. Adapting naive Bayes to domain adaptation for sentiment analysis[A]. The 31th European Conference on IR Research [C]. Toulouse,France: Springer,2009. 337 – 349.

[5] S Wawre,S Deshmukh. Sentiment classification using machine learning techniques[J]. International Journal of Science and Research,2016,5(4):819 – 821.

[6] A Maas,R Daly,P Pham,et al. Learning word vectors for sentiment analysis[A]. The 49th Annual Meeting of the Association for Computational Linguistics [C] Portland, Oregon,USA:ACL,2011. 142 – 150.

[7] Trstenjak B,Mikac S,Donko D. KNN with TF-IDF based framework for text categorization[J]. Procedia Engineering,2014,69(1):1356 – 1364.

[8] Johnson R,Zhang T. Deeppyramid convolutional neural networks for text categorization[A]. 55nd Annual Meeting of the Association for Computational Linguistics[C]. Vancouver,Canada:ACL,2017. 562 – 570.

[9] Ammar Ismael Kadhim. Survey on supervised machine learning techniques for automatic text classification[J]. Artificial Intelligence Review,2019,52(1):273 – 292.

[10] R Johnson and T Zhang. Effective use of word order for text categorization with convolutional neural networks [A]. Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies [C]. Denver, Colorado, USA: NAACL,2015. 103 – 112.

[11] W Yih,X He,C Meek. Semantic parsing for single-relation question answering[A]. 52nd Annual Meeting of the Association for Computational Linguistics[C]. Baltimore, Maryland,USA:ACL,2014. 643 – 648.

[12] Y Shen,X He,J Gao,et al. Learning semantic representations using convolutional neural networks for web search [A]. Proceedings of the 23rd International Conference on World Wide Web [C]. Seoul, Korea: ACM, 2014. 373 – 374.

[13] D Tang,B Qin,T Liu. Learning semantic representations of users and products for document level sentiment classification[A]. 53rd Annual Meeting of the Association for Computational Linguistics [C]. Beijing, China: ACL, 2015. 1014 – 1023.

[14] Batmaz Z,Yurekli A,Bilge A,et al. A review on deep learning for recommender systems:challenges and remedies[J]. Artificial Intelligence Review,2018,52(1):1 – 37.

[15] Zhang X,Zhao J,Lecun Y. Character-level convolutional networks for text classification[A]. Advances in Neural Information Processing Systems [C]. Montreal,Canada: MIT Press,2015. 649 – 657.

[16] Joulin A,Grave E,Bojanowski P,et al. Bag of tricks for efficient text classification[A]. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics [C]. Valencia, Spain: EACL,

2017. 427 – 431.
[17] P Liu, X Qiu, X Huang. Recurrent neural network for text classification with multi-task learning [A]. The 26th International Joint Conference on Artificial Intelligence[C]. Melbourne, Australia: Morgan Kaufmann, 2017. 1480 – 1489.
[18] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[A]. Conference of the Association for the Advancement of Artificial Intelligence [C]. Austin, Texas, USA: AAAI, 2015. 2267 – 2273.
[19] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification[A]. Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies[C]. San Diego, California, USA: NAACL, 2016. 1480 – 1489.
[20] T Mikolov, I Sutskever, K Chen, et al. Distributed representations of words and phrases and their compositionality [A]. Advances in Neural Information Processing Systems [C]. Lake Tahoe, Nevada, USA: MIT Press, 2013: 3111 – 3119.
[21] J Pennington, R Socher, C Manning. Glove: global vectors for word representation [A]. Conference on Empirical Methods in Natural Language Processing [C]. Doha, Qatar: ACL, 2014. 1532 – 1543.
[22] Pedro Domingos. A few useful things to know about machine learning[J]. Communications of the ACM, 2012, 55 (10): 78 – 87.
[23] S Kiritchenko, X Zhu, S Mohammad. Sentiment analysis of short informal texts [J]. Journal of Artificial Intelligence Research, 2014, 50(1): 723 – 762.
[24] L Qu, G Ifrim, G Weikum. The bag-of-opinions method for review rating prediction from sparse text patterns[A]. 23rd International Conference on Computational Linguistics[C]. Beijing, China: ACM, 2010. 913 – 921.
[25] R Socher, A Perelygin, J Wu, et al. Recursive deep models for semantic compositionality over a sentiment treebank [A]. Conference on Empirical Methods in Natural Language Processing [C]. Seattle, Washington, USA: ACL, 2013. 1631 – 1642.
[26] N Kalchbrenner, E Grefenstette, P Blunsom. A convolutional neural network for modelling sentences [A]. 52rd Annual Meeting of the Association for Computational Linguistics[C]. Baltimore, Maryland, USA: ACL, 2014. 655 – 665.
[27] Quoc V, T Mikolov. Distributed representations of sentences and documents[A]. The 31st International Conference on Machine Learning [C]. Beijing, China: ACM, 2014. 1188 – 1196.
[28] C. F. Ahmed, S. K. Tanbeer, B. S. Jeong et al. Efficient tree structures for high utility pattern mining in incremental databases [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(12): 1708 – 1721.
[29] M. C. Liu, J. F. Qu. Mining high utility itemsets without candidate generation[A]. Proc of the 12th International Conference on Information and Knowledge Management [C], Maui, HI, USA, ACM, 2012. 55 – 64.
[30] 黄坤,吴玉佳,李晶. 基于差集的高效用项集挖掘方法[J]. 电子学报,2018,46(8):1804 – 1814.
HUANG Kun, WU Yu-jia, LI Jing. Mining high utility itemsets using diffsets[J]. Acta Electronica Sinica, 2018, 46(8): 1804 – 1814.

作者简介

吴玉佳 男,1986 年 11 月出生,湖北广水人,武汉大学博士生,计算机应用技术专业. 研究方向:数据挖掘,自然语言处理,深度学习.
E-mail: wuyujia@ whu. edu. cn

李 晶(通信作者) 男,1967 年 7 月生,湖北武汉人,武汉大学教授,博士生导师. 研究方向:数据挖掘,多媒体技术,人工智能.
E-mail: leejingcn@ whu. edu. cn