异构分类器堆叠泛化及其在恶意评论检测中的应用

吕 品[1],于文兵[2],汪 鑫[1],计春雷[1],周曦民[3]
(1. 上海电机学院电子信息学院 上海 201306;2. 上海电机学院文理学院 上海 201306;
3. 上海超级计算中心 上海 201203)

摘 要: 恶意评论检测是预防社会媒体平台给用户带来负面影响的一项重要工作,是自然语言处理的重要领域之一.为解决单分类器实现恶意评论检测时模型精度不稳定、boosting 集成模型精度较低的问题,提出一种异构分类器堆叠泛化的方法.该方法用深度循环神经网络将多标签的恶意评论分类问题转变为二类分类,防止了模型精度不稳定;用堆叠泛化集成时单个分类器 GRU(Gated Recurrent Unit)和 NB-SVM(Naïve Bayes-Support Vector Machine)在模型结构和分类偏差上的差异性,改善了模型精度.在维基百科恶意评论数据集上的对比实验证明:提出的方法优于 boosting 集成,说明堆叠泛化异构分类器实现恶意评论检测是可行且有效的.

关键词: 堆叠泛化;恶意评论;循环神经网络;NB-SVM;词嵌入
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)10-2228-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.10.026

Stacked Generalization of Heterogeneous Classifiers and Its Application in Toxic Comments Detection

LÜ Pin[1], YU Wen-bing[2], WANG Xin[1], JI Chun-lei[1], ZHOU Xi-min[3]
(1. *Shcool of Electronic Information Engineering, Shanghai Dianji University, Shanghai* 201306, *China*;
2. *Shcool of Arts and Sciences, Shanghai Dianji University, Shanghai* 201306, *China*;
3. *Shanghai Supercomputer Center, Shanghai* 201203, *China*)

Abstract: Toxic comment detection is an important work to prevent the negative impact of social media platform on users, and it is also one of the important fields of natural language processing. In order to solve the problems of unstable model accuracy and low accuracy of boosting ensemble model when an individual classifier detects toxic comments, a stack generalization with heterogeneous classifiers is proposed. In this method, the classification problem of multi-label toxic comments is transformed into binary categories by using deep recurrent neural network, which prevents the model accuracy from being unstable. Individual classifiers called GRU (Gated Recurrent Unit) and NB-SVM (Naïve Bayes-Support Vector Machine) are used during stacked generalization in order to embody the differences on model structure and classification deviation of individual classifiers, the goal is to improve the model accuracy. Experimental results on Wikipedia toxic comments show that the proposed method has better than boosting ensemble, which reports that stacked generalization of heterogeneous classifiers is feasible and effective for toxic comments detection.

Key words: stacked generalization; toxic comments; recurrent neural network; NB-SVM

1 引言

随着信息技术的快速发展和智能终端的普及,互联网在人们日常生活中的地位日益凸显,人们更倾向于借助互联网表达对事物的观点和看法[1].然而,互联网上也暗藏着一些给 Web 用户造成危害的人身攻击、网络骚扰和欺凌行为的恶意评论[4].据维基百科基金统计,经历过网络骚扰的网民,有 54% 的人已减少了参与发表网络评论[4].2014 Pew Report 显示,73% 的成年网民目睹了一些用户被骚扰,40% 的网民

收稿日期:2019-03-10;修回日期:2019-05-28;责任编辑:郭游
基金项目:上海市教育科学研究项目(No. C17014);上海电机学院计算机科学与技术优势学科(No. 16YSXK04)

亲身经历过骚扰[5]. 为防止恶意评论给网民造成的负面影响,恶意评论检测已成为科研工作者与工业界关注的焦点.

目前,恶意评论检测不仅包括识别评论中的在线骚扰、网络欺凌、仇恨性语言、辱骂性语言以及攻击性语言,也包括恶意评论的分类. 这些检测方法大致分为 2 类:浅层学习和特征工程. 例如,Wulczyn 等[4]采用众包方法在获取高质量的人工标注恶意评论数据集后,运用逻辑回归和多层感知机学得两种分类器,研究发现这两种分类器对恶意评论分类的结果与人工标注几乎一样;Dawei 等[6]基于评论的局部特征、情感特征和上下文特征,训练支持向量机,以期改进在线骚扰识别精度;Ravi 等[7]利用朴素贝叶斯、支持向量机和逻辑斯蒂回归等三种机器学习方法,识别在线评论中潜在的恶意语言;Dadvar 等[8]提出利用评论内容的特征、网络恶意语言的特征和用户特征构建支持向量机,判断一篇评论是否具有恶意;Chen 等[7]提出利用词汇的句法特征识别在线评论中潜在的"恶意"词语. Djuric 等[10]则采用分阶段的方法实现"恶意"语言的检测. 第一阶段采用 paragraph2vec 学习评论中词的分布式表示,得到一个低维文本嵌入;第二阶段利用第一阶段得到的嵌入构建一个二元分类器,以此判断一篇评论是否具有恶意. 尽管以上方法均能实现恶意评论的检测,但由于它们将恶意评论的检测视为两类分类问题,忽视了一篇评论可能同时属于多个类别,导致模型的检测精度不稳定. 与这些方法不同,在本文的方法中,每一篇评论可属于多个类别,是一个典型的多标签问题. 具体来说,在利用深度神经网络实现多标签任务时,为每一个类别设定一个合适的阈值,将连续的 sigmoid 输出转变为二类分类.

为解决模型精度不稳定的问题,一些研究者们尝试了深度学习方法. 例如,Spiros 等[11]运用卷积神经网络研究了包含 6 个类别(恶意、严重恶意、淫秽、威胁、侮辱、仇恨)的公开评论数据集 Wikipedia Talk Pages,将恶意评论的分类问题视为多标签分类问题,获得了比浅层学习更好的分类性能和稳定性. Betty 等[12]也基于这个数据集,研究了梯度提升决策树这种集成方法在恶意评论分类中的应用,实验表明集成方法优于浅层学习和深度神经网络. 已有研究工作一方面表明了深度学习方法在捕获评论中词与词之间的依赖性、重要短语和复杂的上下文信息方面具有强大的能力,另一方面也表明了集成方法优于单个模型. 但是,由于恶意评论中有非常特殊和罕见的词汇,致使 Betty 提出的 boosting 方法[12]的性能评测指标 F1 低于 0.8. 与此集成方法不同的是,本文采用的是堆叠泛化集成,通过多响应线性回归方法结合了泛化能力强且分类分歧较大的 2 个个体分类器,F1 值达到了 0.86.

我们的贡献主要有 2 方面:1)提出了一个异构分类器的堆叠泛化方法,此方法集成了深度神经网络与浅层学习,既充分利用了深度神经网络强大的特征学习能力,也充分利用了组件分类器在模型结构和分类偏差上的差异程度,弥补了已有的 boosting 方法中基分类器受样本扰动的不足,保证了集成分类的效果;2)在维基百科评论数据集上验证了不同基分类器的多样性,发现个体分类器 NB-SVM 可产生强不相关预测,这表明异构集成能进一步提高集成分类的精度. 与现有的 boosting 方法[12]相比,我们的方法取得了较好的性能.

2　异构分类器堆叠泛化

堆叠泛化是利用元学习算法学习多个模型预测结果的一种集成技术[13,14],由于具有能最小化多个个体分类器的泛化误差率的良好特性,近年来在流行病学预测、森林覆盖率的环境检测、推荐系统等领域、Kaggle 竞赛和 Netflix 竞赛中有着广泛应用[15,16]. 堆叠泛化框架如图 1 所示.

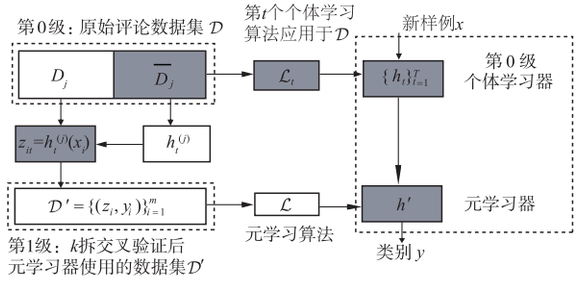

图1　堆叠泛化框架

堆叠泛化的基本思想是:先利用原始数据集训练出 0 级的个体学习器,然后,生成一个新数据集训练 1 级元分类器. 新数据的获得过程如下:首先,将原始训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ 随机划分为 k 个大小相似的集合 $\{D_1,D_2,\cdots,D_k\}$. 令 D_j 和 $\overline{D}_j=D\backslash D_j$ 分别表示第 k 折的测试集和训练集. 对于 T 个不同的个体学习算法,$h_t^{(j)}$ 表示使用第 $t(t\in T)$ 个个体学习算法在训练集 $\overline{D}_j$ 上学得的个体分类器. 对于 D_j 中的每个样本 x_i,令 $z_{it}=h_t^{(j)}(x_i)$ 表示第 t 个个体学习器在第 i 个样本上的预测,则由 x_i 所产生的用于元学习器的训练样本为 $z_i=\{z_{i1},z_{i2},\cdots,z_{iT}\}$,$z_i$ 中的每一个样本的标记为 y_i. 整个交叉验证过程结束之后,从这 T 个个体学习器中产生的、用于元学习器学习的训练集为 $\mathcal{D}'=\{(z_i,y_i)\}_{i=1}^{m}$. 学得元分类器 h' 之后,就可以利用它对 0 级的多个个体学习器的预测结果进行集成.

在图 1 所示的堆叠泛化框架中,本文选用了 2 种不同个体分类器:深度神经网络和浅层学习器 NB-SVM,

主要考虑了这两种不同类型个体分类模型本身的准确性和模型之间存在的差异. 本节首先介绍了两种不同个体分类器的基本思想,之后,介绍了这两种异构分类器的集成方法.

2.1 基于深度神经网络的个体学习算法

堆叠泛化框架中选用的基于深度神经网络的个体学习算法是循环神经网络. 本文选用了4种不同的循环神经网络,LSTM(Long Short Term Memory Network)、双向LSTM、双向GRU(Gated Recurrent Unit)和带有注意层的GRU.

LSTM将输入的恶意评论作为一个词序列,利用嵌入层将词的独热编码转换为嵌入表示. 设置dropout为0.1随机丢弃输入词,以提高网络的鲁棒性. 为处理词嵌入序列,本文使用的LSTM层包含50个单元. 使用了2个全连接层,一个全连接层使用ReLU激活函数实现多标签分类,另一个使用sigmoid激活函数实现二类分类.

采用双向LSTM和GRU的目的是弥补相距较远的词依赖可能发生的错误. 与标准的LSTM相比,双向LSTM使用了2个LSTM层以处理正确的输入序列和反相的输入序列. 然后,求两层输出的平均值. 类似地,双向GRU也由两个堆叠的GRU层构成. 本文为每一个GRU层设置64个单元. 网络结构中其余的部分,如dropout的设置、全连接层的设置与LSTM相同.

Gao等[17]指出具有上下文学习能力的LSTM在恶意评论检测应用中具有较好的效果. Yang等[18]提出词级的注意力机制能够表达词对句子意思的贡献程度,句子级注意力机制能够表达某一句子对文档分类的重要性. 因此,本文在双向GRU中引入注意力层的目的是关注较长评论中的暗含有恶意词语的特定区域,以提高恶意评论检测的精度.

2.2 基于浅层学习的NB-SVM个体学习算法

Wang等[19]研究表明文本分类中常用的基线模型是朴素贝叶斯模型NB(Naïve Bayes)和支持向量机模型SVM(Support Vector Machine),并且这两种模型的性能依赖于模型所使用的特征. 他们发现:1)词的Bigram特征在情感分类任务上能获得较好的性能;2)利用NB的log-count率作为SVM的特征,模型NB-SVM的性能在不同数据集上都优于NB或SVM. 因此,堆叠泛化框架中基于浅层学习的个体学习算法选用了NB-SVM(Naïve Bayes-Support Vector Machine).

假设$\boldsymbol{f}_i \in \mathbb{R}^{|V|}$是训练样本$(x_i, y_i)$的特征计数向量,其中,$y_i \in \{-1, +1\}$,$V$是特征集. $\boldsymbol{f}_i^j$表示在训练样例i中特征V_j出现的次数. 为平滑参数α定义计数向量$\boldsymbol{p} = \alpha + \sum_{i:y_i=1} \boldsymbol{f}_i$和$\boldsymbol{q} = \alpha + \sum_{i:y_i=-1} \boldsymbol{f}_i$. 于是log-count率计算如公式(1)所示.

$$r = \log\left(\frac{p/\|p\|_1}{q/\|q\|_1}\right) \tag{1}$$

定义线性分类器为$y_k = \text{sign}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_k + b)$. 对于模型NB,设定$\boldsymbol{x}_k = \boldsymbol{f}_k, \boldsymbol{w} = \boldsymbol{r}, b = \log(N_+/N_-)$. 其中,$N_+$和$N_-$表示训练集中正例样本数量与负例样本数量. 在本文的NB-SV模型中,取$\boldsymbol{x}_k = \hat{\boldsymbol{f}}_k = \mathbb{I}(\boldsymbol{f}_k > 0)$,$\mathbb{I}$为指示函数. 并且$\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}$和$\hat{\boldsymbol{r}}$均用$\hat{\boldsymbol{f}}^k$计算. 对于模型NB-SVM,设定$\boldsymbol{x}_k = \tilde{\boldsymbol{f}}_k, \tilde{\boldsymbol{f}}_k = \hat{r} \cdot \hat{\boldsymbol{f}}_k$,$\cdot$表示点积运算,$\boldsymbol{w}$和$b$的求解需要最小化公式(2)所示的目标函数.

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + C\sum_i \max(0, 1 - y_i(\boldsymbol{W}^{\mathrm{T}}\hat{\boldsymbol{f}}_i + b))^2 \tag{2}$$

2.3 集成过程

本文基于上述的堆叠泛化框架,用算法的形式描述了2种异构个体分类器的集成过程.

输入:训练集$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$;
0级学习算法$\mathcal{L}_1$(LSTM),$\mathcal{L}_2$(双向LSTM),$\mathcal{L}_3$(双向GRU),$\mathcal{L}_4$(具有注意力层的双向GRU),$\mathcal{L}_5$(NB-SVM);
1级学习算法$\mathcal{L}$(多响应线性回归).
输出:$H(x) = h'(h_i(x), h_5(x)), i = 1, 2, 3, 4.$
步骤:
1. for $t = 1, 2, \cdots, 5$ do
2. $h_t = \mathcal{L}_t(D)$;//分别使用0级学习算法$\mathcal{L}_t$产生0级学习器h_t
3. end for
4. for $k = 1, 2, \cdots, 10$ do //10折交叉验证
5. $D \to D_k, \bar{D}_k (\bar{D}_k = D \backslash D_k)$; //拆分$D$为测试集$D_k$和训练集$\bar{D}_k$
6. for $t = 1, 2, \cdots, 5$ do
7. $h_t^k = \mathcal{L}_t(\bar{D}_k)$; //第$t$个0级个体学习算法在第$k$折训练集上学得分类器$h_t^k$
8. $z_{it} = h_t^k(D_j)$; //第t个0级个体学习算法在第k折测试集上的预测结果z_{it}
9. end for
10. end for
11. for $j = 1, 2, \cdots, 4$ do //10折交叉验证
12. $D_j' = ((z_{ij}, z_{i5}), y_i)$ //$j = 1, 2, 3, 4$,D_j'为1级元学习算法的训练数据集
13. $h_j' = \mathcal{L}(D_j')$; //利用元学习算法学得元学习器h_j'
14. end for

由上述过程可知,总共学习了四种不同的基于深度学习的个体分类器,一种浅层分类器. 然后,将每一种基于深度学习的个体分类器与浅层分类器集成,最终得到4个不同的集成分类器.

3 实验

3.1 数据集

本文的实验据集来自Kaggle竞赛网站. 训练集共159571条评论,6个类别(恶意、严重恶意、淫秽、威胁、侮辱、仇恨). 分析数据集可知,每篇评论在每个类别标

签上都有一个取值0或1,一些评论在多个类别标签上取值为1. 显然,这是一个典型的多标签分类问题. 另外,数据集中含类别标记的评论只有16225条,其余143346条评论没有类别标记,约占原始数据集的90%. 为解决原始数据集类别不平衡对分类模型性能的影响,本文采用了增加词、删除词以及同义词替换的数据增强方法[20],为原始数据集中的三个稀疏类别(严重恶意、威胁、仇恨)增加新的评论样本.

此外,观察发现原数据集中绝大多数评论中的唯一词比例大于50%,因此将包含唯一词的比例小于30%的评论视为"噪音"评论,并对这些评论中的重复词进行了只保留一份的处理,以避免它们对词嵌入和网络训练产生不利影响.

3.2　循环神经网络的输入层预处理

实验中使用的4种不同循环神经网络均利用了词嵌入方法保留评论文本中词的分布式语义,因此需要对网络的输入进行预处理. 首先,删除评论文本中所有的数字、标点符号、空格和停用词;然后,设置词汇表大小为20000,每篇评论的长度为100(99%的置信区间表明,每篇评论的句子长度为65),对长度不足100的评论进行0值填充,将每篇评论转化为一个等长的词索引向量.

由于用大规模语料训练好的词嵌入能有效捕获训练数据集中的遗失信息,因此本文实验选取了最常用的2种词嵌入模型GloVe和word2vec. GloVe有50,100,200和300等4种维度,本文选取了维度为50的GloVe. 当用等长的词向量创建嵌入矩阵时,对不在GloVe模型中的词采用了随机初始化的方式. 实现word2vec嵌入时,选用了在Google新闻数据集上预训练好的嵌入模型Google News-vectors-negative 300. bin. gz.

表1中所有模型均使用了2个全连接层. 第一个全连接层包含50个神经元,使用激活函数ReLU,第二个全连接层使用的激活函数sigmoid,选择sigmoid函数的原因是它对输入的微小改变非常敏感,适合于二类分类. 最后,编译上述4个模型时,使用了二元交叉熵为损失函数,Adma为参数求解优化器;在训练时,设置batch_size = 32,迭代次数为30,10%的训练数据作为验证集数据.

3.3　循环神经网络的结构与超参设置

由于词汇表的大小为20000,每条评论长度为100,每一个词的嵌入向量维度为50,因此,4个不同循环神经网络的嵌入层大小均为[20000 * 50]. 实验中使用的模型结构与超参设置如表1所示.

表1　4个不同循环神经网络的模型参数值

模型	特殊层	全连接层
LSTM	1层LSTM,单元个数为50;dropout为0.1	50个神经元
双向LSTM	2层LSTM,每层50个单元,共100;dropout和循环dropout均为0.1	50个神经元
双向GRU	2层GRU,每层GRU单元个数为64,共128个;dropout和循环dropout均为0.1	50个神经元
双向GRU + Attention	2层GRU,每GRU单元个数为64,共128个;	50个神经元

3.4　NB-SVM模型参数的设置

关于数据集的不平衡问题与"噪音"问题,NB-SVM模型采用了与上述4种神经网络模型相同办法. NB-SVM在特征选择时,首先,删除文本中所有数字、标点符号、空格以及停用词;然后,用基于bigram的词袋模型表示每篇评论,构建文档-词矩阵. 训练NB-SVM模型时,采用了与文献[17]一样的参数,即平滑参数$\alpha = 1$,正则化系数$C = 1$,插值参数$\beta = 0.25$.

3.5　异构分类器堆叠集成性能分析

为验证集成异构分类器优于单个个体分类器,首先,以分类的精确度、查准率、召回率、F1-score和AUROC等为评价标准,分析了5个不同的单分类器在恶意评论上的分类效果,如表2所示. 对于循环神经网络,还分别列出了使用不同词嵌入模型的性能值. 观察表2可知:1)所有的深度学习方法均优于浅层学习方法NB-SVM;2)当循环神经网络使用Glove嵌入时,其效果要好于word2vec;3)使用注意力机制的循环神经网络的性能最好. 分析其中的原因,可能是:1)循环神经网络能有效地表达句子中词与词之间的依赖关系,词嵌入模型能表达词的分布式语义,而NB-SVM采用的词袋模型忽略了词与词之间的关系;2)基于统计的Glove模型可能比基于预测的word2vec更合适于本文的应用;3)相比于其它的个体分类器,引入注意力机制的循环神经网络在表示一篇评论时,由于能根据词所在的上下文判断该词的重要程度,对分类决策有很大的帮助作用.

其次,将具有最好分类性能的个体分类器(具有注意力机制的双向GRU + Glove)与NB-SVM进行了堆叠集成,实验结果也列于表2中. 从表2可知,集成方法比最好的个体分类器在F1-score上要高1%.

表 2 不同个体学习器在恶意评论数据集上实现分类的性能比较

模型	性能评价指标				
	Accuracy	Precision	Recall	F1-score	AUROC
NB-SVM(Bigram)	0.887	0.702	0.834	0.752	0.945
LSTM(word2vec)	0.899	0.714	0.854	0.778	0.961
LSTM(Glove)	0.932	0.745	0.842	0.791	0.982
双向 LSTM(word2vec)	0.899	0.714	0.863	0.781	0.975
双向 LSTM(Glove)	0.928	0.743	0.846	0.791	0.981
双向 GRU(word2vec)	0.911	0.726	0.866	0.790	0.974
双向 GRU(Glove)	0.922	0.735	0.852	0.789	0.983
双向 GRU + 注意力机制(word2vec)	0.923	0.738	0.876	0.801	0.986
双向 GRU + 注意力机制(Glove)	0.933	0.748	0.871	0.805	0.987
集成方法:双向 GRU + 注意力机制(Glove)、NB-SVM	0.952	0.837	0.880	0.858	0.988

3.6 个体分类器的多样性分析

用相关系数度量了集成分类器中个体分类器的多样性,即分析了集成异构分类器比集成同质分类器在稀疏类别的检测上更具优势. 表 4 列出了不同集成方法在三个稀疏类别上的 F1-score 和相关系数. 由于本文的恶意评论检测最终将多标签分类问题转换为二类分类问题,因此,对预处理后的数据集中的每一个样本,需要得到两个个体分类器的预测结果列联表(如表 3 所示). 表 3 中值 a 表示个体分类器 h_1 和 h_2 均预测为恶意的评论样本数据数量. 值 b,c,d 含义由此类推. 显然,值 a,b,c,d 之和为给定的测试集样本总数. 表 4 中两个体分类器的相关系数 $\boldsymbol{\rho}_{12}$ 用公式(3)计算. $\boldsymbol{\rho}_{12} \in [-1,1]$ 的值越大,则个体分类器之间的相关性越强. 观察表 4 可知,相关系数低的个体分类器,集成后的性能要比相关系数高的分类器集成效果好. 这说明深度神经网络与浅层学习的异构集成效果优于 2 种深度神经网格的同质集成.

$$\rho_{12} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} \tag{3}$$

表 3 个体分类器在某一类别上预测结果列联表

		个体分类器 h_1	
	预测结果	恶意	非恶意
个体分类器 h_2	恶意	a	c
	非恶意	b	d

3.7 不同元学习算法对堆叠泛化性能的影响

元学习算法体现了不同个体学习器的结合策略,能在很大程度上矫正不同学习器之间的偏差. 为观察本文所使用的元学习算法多响应线性回归 MLR(Multi-response Linear Regression)的效果,增加了投票法 voting 和 IBk(Instance-based k)方法[21]作为基线法. 实验仍基于 10 折交叉验证得到集成分类器的错误率,然后分别计算成对 t-检验和分类精度的相对改进,用于比较不同元学习算法的分类性能,如表 5 所示. 在评价分类器 C_1 相对分类器 C_2 在分类精度上的改进时,使用了公式 $1 - error(C_1)/error(C_2)$. 另外,还使用了成对 t-检验表示不同分类器在分类性能上的不同统计显著性. 表 4 给出了置信水平为 95% 时两个分类器的性能,+ 表示分类器 C_1 显著优于分类器 C_2,-表示分类器 C_2 显著优于分类器 C_1.

由表 5 可知,MLR 相对 Voting 集成策略有 21.5% 的分类精度改进,相对于 IBk 集成策略有 4.6% 的集成改进,且均显著优于 voting 和 IBk. 这表明用多响应线性回归作为元分类学习算法集成单个最优分类器能显著改进分类的性能.

表 4 个体分类器多样性度量

	类别	F1-score	两个个体分类器的相关系数
双向 GRU + Attention 与 LSTM 集成	严重恶意	0.73	0.96
	威胁	0.71	0.94
	仇恨	0.72	0.94
双向 GRU + Attention 与 NB-SVM 集成	严重恶意	0.80	0.65
	威胁	0.78	0.78
	仇恨	0.79	0.67

表 5　不同元学习算法的性能比较(10 折交叉验证的平均值/95% 的置信度)

	MLR		Voting		IBk	
	相对改进	显著性	相对改进	显著性	相对改进	显著性
MLR	0	0	21.5%	11% +/5% -	4.6%	5% +/2% -
Voting	-27.5%	5% +/11% -	0	0	-21.5%	7% +/10% -
IBk	-4.8%	2% +/5% -	17.7%	10% +/7% -	0	0

4　结论

本文提出了一种应用于恶意评论检测的异构分类器堆叠泛化集成方法. 集成时选用的个体分类器分别是循环神经网络 GRU 和传统的浅层学习算法 NB-SVM,目的是分析个体分类器的准确性与多样性对集成方法的影响. 实验结果表明,单个体分类器准确性越高,相关性越小,集成后的分类器性能越高. 此外,本文还分析了三种不同的集成策略多响应回归 MLR、投票法 Voting 和基于示例的近邻法 IBk 对集成分类器性能的影响,实验结果表明多响应回归的集成策略最优. 本文正是利用多响应回归集成 GRU 和 NB-SVM,获得的分类器在性能评价指标 F1 上的值为 0.86,优于目前在相同数据集上的 boosting 集成方法.

参考文献

[1] 周孟,朱福喜. 基于情感标签的极性分类[J]. 电子学报,2017,45(4):1018 - 1024.
ZHOU Meng, ZHU Fu-xi. Polarity classification based on sentiment tags[J]. Acta Electronica Sinica, 2017, 45(4): 1018 - 1024. (in Chinese)

[2] 卜湛,伍之昂,曹杰,朱桂祥. 在线评论情感计算与博弈预测[J]. 电子学报,2015,43(12):2530 - 2535.
BU Zhan, WU Zhi-ang, CAO Jie, ZHU Gui-xiang. Affective computing and game theory based prediction for online reviews[J]. Acta Electronica Sinica, 2015, 43(12): 2530 - 2535. (in Chinese)

[3] 翟延东,王康平,等. 一种基于 WordNet 的短文本语义相似性算法[J]. 电子学报,2012,40(3):617 - 620.
ZHAI Yan-dong, WANG Kang-ping, et al. An algorithm for semantic similarity of short text based on word net[J]. Acta Electronica Sinica, 2012, 40(3): 617 - 620. (in Chinese)

[4] Ellery Wulczyn, Nithum Thain, Lucas Dixon. Ex machina: Personal attacks seen at scale [A]. Proceedings of the 26th International Conference on World Wide Web[C]. Perth: Australia, ACM, 2017. 1391 - 1399.

[5] Maeve Duggan. Online harassment. Pew Research Center [OL]. 2014. http://www.pewinternet.org/2014/10/22/.

[6] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, Lynne Edwards. Detection of harassment on web 2.0[A]. In Proceedings of the Content Analysis in the WEB 2.0 Workshop at WWW2009[C]. Madrid: Spain, WWW, 2009. 1 - 7.

[7] Prashant Ravi. Detecting Insults in Social Commentary [OL]. https://www.overleaf.com/articles/detecting-insults-in-social-commentary/gkvrrwryjxhr, 2019.

[8] Maral Dadvar, Dolf Trieschnigg, Ordelman Roeland, Jong Franciska. Improving cyberbullying detection with user context[A]. ECIR 2013 Lecture Notes in Computer Science[C]. Springer, Berlin, Heidelberg, 2013. vol 7814. 693 - 696.

[9] Ying Chen, Yilu Zhou, Sencun Zhu, Heng Xu. Detecting offensive language in interactive social media to protect adolescent online safety[A]. International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing[C]. Washington: USA, IEEE Computer Society, 2013. 71 - 80.

[10] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati. Hate speech detection with comment embeddings[A]. In Proceedings of the 24th International Conference on World Wide Web[C]. Florence: Italy: ACM, 2015. 29 - 30.

[11] Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, Vassilis P Plagianakos. Convolutional neural networks for toxic comment classification[A]. Proceedings of the 10th Hellenic Conference on Artificial Intelligence [C]. Patras: Greece, ACM, 2018. Article No. 35.

[12] Betty vanAken, Julian Risch, Ralf Krestel, Alexander Loser. Challenges for toxic comment classification an in-depth error analysis[A]. Proceedings of the Second Workshop on Abusive Language Online [C]. Brussels: Belgium, ACL, 2018. 33 - 42.

[13] David H Wolpert. Stacked generalization [J]. Neural Network, 1992, 5(2): 241 - 59.

[14] Breiman Leo. Stacked regressions [J]. Machine Learning, 1996, 24: 49 - 64.

[15] Ashley I Naimi, Laura B Balzer. Stacked generalization: an introduction to super learning[J]. European Journal of Epidemiology, 2018, 33(5): 459 - 464.

[16] Jin Chen, Cheng Wang, Runsheng Wang. Using stacked generalization to combine SVMs in magnitude and shape feature spaces for classification of hyperspectral data[J].

IEEE Transactions on Geoscience and Remote Sensing, 2009,47(7):2193 – 2205.

[17] Lei Gao, Ruihong Huang. 2017. Detecting online hate speech using context aware models[A]. Proceedings of the International Conference Recent Advances in Natural Language Processing [C]. Varna: Bulgaria, INCOMA Ltd,2017. 260 – 266.

[18] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy. Hierarchical attention networks for document classification[A]. The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [C]. San Diego California, USA: ACL,2016. 1480 – 1489.

[19] Sida Wang, Christopher D Manning. Baselines and bigrams: simple, good sentiment and topic classification [A]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics[C]. Jeju Island: Korea, ACL,2012. 90 – 94

[20] Mai Ibrahim, MarwanTorki, Nagwa El-Makky. Imbalanced toxic comments classification using data augmentation and deep learning[A]. 17th IEEE International Conference on Machine Learning and Applications[C]. Orlando, USA: IEEE,2018. 875 – 878.

[21] Kai Ming Ting, Ian H Witten. Issues in stacked generalization[J]. Journal of Artificial Intelligence Research, 1999, (10):271 – 289.

作者简介

吕 品 女,1973 年 3 月出生,湖北鄂州人,现为上海电机学院副教授、博士,硕士生导师,研究方向为数据挖掘、情感分析与机器学习.
E-mail:lvp@ sdju. edu. cn

于文兵 男,1972 年 10 月出生,湖北洪湖人,现为上海电机学院高级工程师、硕士,研究方向为智能计算、全光通信、光纤传感.
E-mail:yuwb@ sdju. edu. cn

汪 鑫 男,1978 年 3 月出生,安徽黟县人,现为上海电机学院衣讲师、硕士,研究方向为数据挖掘、云计算.
E-mail:wangx@ sdju. edu. cn

计春雷 男,1964 年 1 月出生,上海人,现为上海电机学院教授、博士,硕士生导师,研究方向为大数据、数据挖掘.
E-mail:jicl@ sdju. edu. cn

周曦民 男,1961 年 2 月出生,上海人,现为上海超级计算机中心主任、教授级高级工程师,上海大数据联盟副理事长,研究方向为信息安全、大数据与人工智能.
E-mail:xmzhou@ ssc. net. cn