基于不平衡社交媒体文本的抑郁症检测方法

郭耀木,刘 鹏,孙源乐,白其炜,张少华,刘 建*
(合肥工业大学 计算机与信息学院,安徽 宣城 242000)

摘 要:针对目前基于社交媒体数据的抑郁症检测模型难以适应不平衡数据和评估指标不全面的问题,提出一种基于文档自适应增强 Bagging-τSS3(Document Adaptive Enhanced Bagging-τSS3,DAEB-τSS3)模型的社交媒体文本数据抑郁检测方法和一种新的机器学习评价指标 GF(α,β)-Score。在 τ-SS3 模型基础上引入置信度加权处理,增强少数类数据影响;同时,采用文档自适应增强 Bagging 方法进行集成学习,改进 Bagging 的随机采样为分层采样并对少数类数据文档进行自适应增强以提升模型适应不平衡数据的能力;最后在模型评价阶段,使用 GF-Score 进行自动参数选择,丢弃表现不佳的基学习器,提升模型的可信度和稳定性。在 E-Risk2017 抑郁症检测数据集上的实验结果表明,DAEB-τSS3 有更强的适应不平衡数据集的能力,相较于 τSS3、双向长短时记忆网络和 ERNIE 3.0 等模型有显著性能提升,GF-Score、F1-Score 和 G-Mean Score 平均提升 13%,0.7% 和 26.9%,可以更加有效地实现基于不平衡社交媒体文本的抑郁症检测。

关键词:不平衡数据集;抑郁检测;集成学习;文本分类;社交媒体文本数据
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2024)04-0153-09
doi:10.20165/j.cnki.ISSN1673-629X.2024.0023

A Detection Method for Depression Based on Imbalanced Social Media Text

GUO Yao-mu,LIU Peng,SUN Yuan-le,BAI Qi-wei,ZHANG Shao-hua,LIU Jian*
(School of Computer and Information,Hefei University of Technology,Xuancheng 242000,China)

Abstract:To address the challenges faced by the current depression detection model based on social media data,such as difficulties in handling imbalanced data and incomplete evaluation indicators,we propose a new approach called Document Adaptive Enhanced Bagging-τSS3 (DAEB-τSS3). This method utilizes social media text data for depression detection and introduces a novel machine learning evaluation metric called GF(α,β)-Score. Building upon the τ-SS3 model,we incorporate confidence weighting to amplify the influence of certain data types. Additionally,we employ the Bagging method to enhance integrated learning,improving the sampling process from random sampling to layered sampling. This adaptive enhancement focuses on a select number of data documents,thereby improving the model's ability to handle imbalanced data. In the model evaluation stage,-we utilize GF-Score for automatic parameter selection and discard underperforming base learners,thereby enhancing the model's reliability and stability. Experimental results on the E-Risk2017 depression detection dataset demonstrate that DAEB-τSS3 exhibits superior adaptability to imbalanced datasets and outperforms τSS3,bi-directional long-term memory networks,and ERNIE 3.0 models. The average improvements in GF-Score,F1-Score,and G-Mean Score are 13%,0.7%,and 26.9%,respectively,enabling more effective depression detection based on imbalanced social media texts.

Key words:imbalanced dataset;depression detection;ensemble learning;text classification;social media text data

0 引 言

抑郁症是全球常见的心理疾病,据世界卫生组织数据统计,全球超过 3.5 亿人受抑郁症困扰,它已成为第四大疾病。然而,由于患者对精神疾病了解不足,很多人未及时就医从而错失最佳治疗时机[1]。随着新社交网络发展,人们倾向在社交媒体上分享生活和观点,这为抑郁症检测提供了有用线索。

社交媒体文本易收集且数据多,因此,许多机器学习模型被应用于基于社交媒体文本的抑郁症检测中。

王垚等人采用基于词向量的多维度正则化 SVM

收稿日期:2023-07-17 **修回日期**:2023-11-21
基金项目:国家自然科学基金青年基金(JZ2019GJQN0385);安徽省大学生创新训练项目(S202210359346);合肥工业大学大学生创新训练项目(X202310359868)
作者简介:郭耀木(2001-),男,研究方向为自然语言处理、机器学习;通讯作者:刘 建(1986-),男,博士,副教授,硕导,研究方向为机器学习算法研究与应用、自然语言处理、计算机视觉、目标跟踪等。

模型[2]进行社交网络抑郁检测;张梦娜等人利用时间序列特征和多示例学习[3]进行抑郁检测,但上述模型检测能力不理想;陈妍等人采用基于情感信息融合注意力机制方法[4]进行抑郁检测;张慧使用 Bert + BiLSTM 的方式[5]进行基于文本的抑郁检测;张宗佳[6]则提出一种基于时间感知的社交媒体文本抑郁检测方法。上述方法准确率高,但模型庞大,需要大量数据和硬件支持,且由于采用深度学习方法,可解释性差。Sergio G. Burdisso 等人提出一种基于置信度的文本分类模型 τ -SS3[7],将文本的局部、全局置信度的概念融入文本分类方法中并应用于抑郁症检测,模型足够轻量化、可解释性和分类能力较强。

传统分类问题假定数据集类别平衡,然而在实际应用中,如社交媒体抑郁症数据,某类样本明显多于另一类,形成不平衡数据集[8]。这会导致机器学习模型在处理此类任务时产生错误。

不平衡数据导致分类不准确的原因在于绝大多数模型基于平衡数据推导,导致模型更多学习多数类特征,影响分类可信性;少数类样本难以泛化评估;数据不平衡使传统评价指标失去可信度。此类任务的难点在于:如何进行文本抑郁检测模型的合理设计;如何实现数据集特征的平衡化;如何合理选择评价指标并基于评价指标对模型进行参数优化。

为减弱不平衡数据集对分类造成的影响,研究者们提出的处理方案大致可分成三个层面[9]:

(1)数据处理层面,对不平衡数据集进行适当处理以减轻数据集不平衡所带来的影响,典型的方法包含欠采样、过采样和混合采样。

欠采样是删除多数类示例以平衡类别分布。如基于 K 近邻的欠采样[10]、基于聚类的欠采样[11]等。

过采样通过多次对样本的随机抽取来平衡正负样本。典型方法包括 SMOTE[12]及相关变体[13-14],ADASYN 算法[15]。

混合采样综合了欠采样和过采样。林舒杨等结合了基于 SMOTE 的过采样和基于聚类的降采样,平衡了引入噪声和样本丢失的矛盾。张明等[16]通过“变异系数”区分稀疏和密集域,用 BSMOTE 处理稀疏少数类,用 IS 算法处理密集多数类。

(2)算法设计层面,通过改进分类模型来适应不平衡数据。如代价敏感学习[17],集成学习如 Bagging[18]、梯度提升树(Gradient Boosting Decision Tree,GBDT)[19]、随机森林(Random forest,RF)[20],单类学习[21],主动学习[22]等。深度学习[23]通过学习微小差异和模式来区分类别,捕捉少数类特征;预训练模型[24]用大规模数据训练,通用性和泛化能力更强。

(3)评价标准层面,考虑使用 F1 - Score[25],G - Mean Score 等更全面的指标,并且可以通过混淆矩阵[26]将预测分类结果和实际分类结果以矩阵的形式直观地展示出来。

综上,为提升训练数据集不平衡条件下的模型分类性能,该文提出一种基于文档自适应增强 Bagging - τSS3(Document Adaptation Enhanced Bagging - τSS3,DAEB - τ -SS3)模型的抑郁检测方法。首先,提出了置信度增强 τ -SS3 模型。鉴于 τ -SS3 模型是目前在 E-Risk2017 公开数据集上表现最好的方法,该文选择 τ -SS3 模型作为基准模型。但该模型在不平衡数据集上表现不佳,分类结果偏向于多数类数据。因此,该文在 τ -SS3 模型的分类阶段进行置信度加权,增强少数类数据对模型的影响。然后将置信度增强 τ -SS3 模型作为基学习器进行文档自适应增强 Bagging 集成学习。在 Bagging 的数据输入阶段,改进 Bagging 的随机采样为分层采样,对文档内文本进行自适应增强选择并丢弃表现过差的基学习器,同时采用动态 N - grams 进行学习,有效缓解了文献[7]中 N-grams 固定导致的适应性减弱的问题。最后,在模型评价阶段,用 GF-Score 进行自动参数选择。最终实验结果显示,改进模型在检测能力上相对于基准模型有了显著提升,表明提出的方法可以有效改善社交媒体文本数据集不平衡条件下的抑郁症检测模型性能。

1 文中方法

DAEB - τSS3 抑郁检测模型结构如图 1 所示。

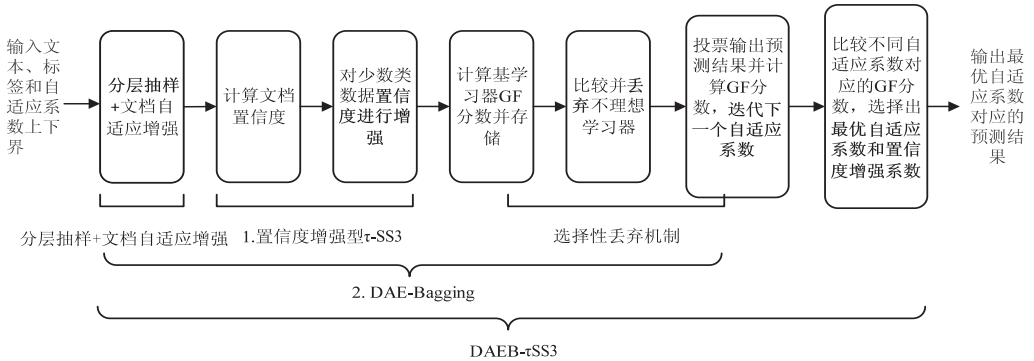

图 1 DAEB - τ SS3 结构框图

该模型包含以下三个部分:(1)置信度增强型 τ -SS3;(2)由分层抽样、文档自适应增强机制和选择性丢弃机制组成的文档自适应增强 Bagging(Document Adaptation Enhanced-Bagging,DAE-Bagging);(3)基于 GF-Score 的参数调整。

上述模块分别对应机器学习模型改进、数据特征平衡化、评价指标选择和基于评价指标的模型参数优化问题。

1.1　置信度增强型 τ-SS3

2020 年,Burdisso 等人提出的 τ -SS3 模型是一个优秀的文本分类器,它在 SS3 模型基础上对动态 n 元语法等模块进行了修改。

在 SS3 模型中,文档被划分为段落,段落进一步划分为语句,语句被划分为单词。通过单词的置信度计算公式[1]来计算词级置信度并合成词级置信向量,再将语句中的词级置信向量相加得到语句级置信向量,以此类推得到段落和文档级的置信向量。文档置信向量中的每个元素表示该文档对每个分类标签的置信度。通过比较置信度,实现文本分类。

SS3 模型使用函数 $gv(a,b)$ 计算每个单词的置信度以评估与每个类别相关的单词(gv: global value,全局置信度)。函数 $gv(a,b)$ 取一个单词 a 和一个类别 b,函数在区间[0,1]中输出一个数值,该数值表示单词 a 被认为属于类别 b 的置信度。例如,假设类别是 negative 和 positive。SS3 模型在评估“sad”和“happy”等单词的单一类别置信度时,函数 $gv(a,b)$ 计算结果如下所示:

$$gv(\text{sad},\text{negative}) = 0.2 \tag{1}$$

$$gv(\text{sad},\text{positive}) = 0.8 \tag{2}$$

$$gv(\text{happy},\text{negative}) = 0.7 \tag{3}$$

$$gv(\text{happy},\text{positive}) = 0.1 \tag{4}$$

合并上述置信度,可以得到“sad”和“happy”的置信向量,分别为:

$$gv(\text{sad}) = (0.2,0.8) \tag{5}$$

$$gv(\text{happy}) = (0.7,0.1) \tag{6}$$

但是,SS3 模型使用缺乏识别重要单词序列能力的词袋模型处理输入,这对模型的分类性能会产生负面影响。因此,Sergio G. Burdisso 等人在 SS3 模型基础上加入了动态 n 元语法模块,使用前缀树对每个单词的 k-grams 进行存储。并且,τ -SS3 在标准性能指标和两个 ERDE[27]指标方面表现优于 SS3。

如图 2 所示,观察测试集高置信度词汇云图发现,少数类数据的高置信度词汇明显少于多数类数据的高置信度词汇。τ -SS3 模型是根据文档的全局置信度进行分类,文档全局置信度由各个词汇置信度累加得到,因为少数类数据的高置信度词汇明显少于多数类数据的高置信度词汇,因此在文档分类中,少数类文档的置信度可能低于多数类文档的置信度,造成置信度欺骗,导致文档的错误分类。

(a)多数类词云　　(b)少数类词云

图 2　置信度云图(词汇置信度越高,在本类别词云中体积越大,置信度过小则不显示)

为了解决上述问题,该文提出对少数类别数据置信度进行增强的置信度增强型 τ -SS3 模型。基准的 τ - SS3 模型在进行文档类别判断时会对文档的置信向量进行比较,取较大的置信向量对应的类别作为文档类别。为了减轻不平衡数据集的影响,在置信向量(置信向量包含少数类置信度(Minority Class Confidence,MICC)和多数类置信度(Majority Class Confidence,MACC))判断阶段,对少数类数据标签所在位置元素进行加权处理,加权系数称为置信度增强系数(Confidence Enhancement Coefficient,CEC),如公式 7 所示。在增强文档的少数类置信度后,弥补了因少数类高置信度词汇量少而导致的置信度欺骗问题。

$$gv(\text{MICC},\text{MACC}) = gv(\text{MICC} + \text{CEC} \times gv(\text{MICC}),\text{MACC}) \tag{7}$$

由于手动设定加权值需要进行多次调参影响训练效率,因此将置信度增强系数作为一个可学习参数加入模型训练之中,以此找到最佳的置信度加权系数。图 3 展示了置信度增强型 τ -SS3 的训练流程。

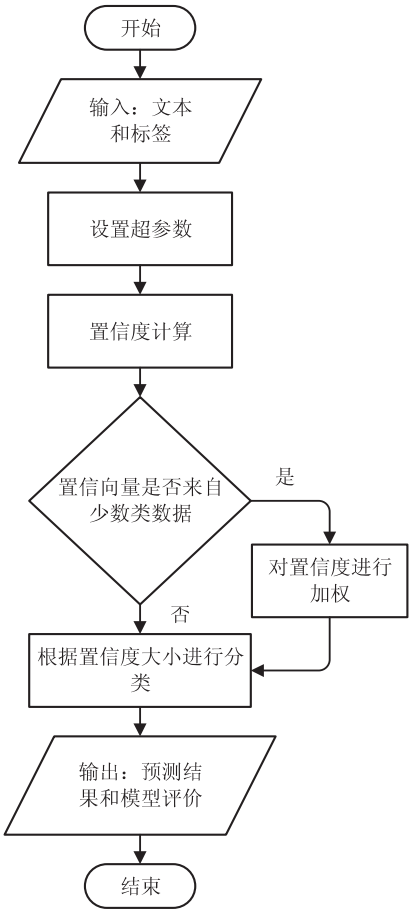

图 3 置信度增强型 τ -SS3(单个学习器)算法流程

1.2 文档自适应增强 Bagging 集成学习模型

传统分类算法在对不平衡数据分类时结果会向多数类倾斜,而集成学习通过训练多个不同的基模型,综合这些基模型的预测结果做出最终预测,可以有效减轻不平衡数据带来的不利影响。

该文以上一节中的置信度增强型 τ -SS3 为基模型,采用 Bagging 集成学习模型进行组合得到最终的分类模型。传统 Bagging 的集成策略是在抽样阶段先进行有放回的简单随机抽样,每次抽样后使用抽样得到的样本训练一个基学习器,最后在集成阶段采用投票原则,得票多的作为预测结果。但是有放回的随机抽样方法存在两个问题:(1)有放回的简单随机抽样得到的训练数据集中少数类别数据的占比远小于多数类别数据,基学习器得到的数据依然是不平衡的;(2)效果差的基学习器在集成阶段会起到负面影响。针对上述两个问题,对 Bagging 的抽样阶段和集成阶段进行改进,采用分层随机抽样替代简单随机抽样,并提出了文档自适应增强策略和选择性丢弃机制。

1.2.1 文档自适应增强策略

针对有放回随机抽样中抽样数据集少数类别数据占比远小于多数类别数据占比的问题,该文将传统 Bagging 的有放回随机抽样的方法调整为有放回的分层随机抽样以降低抽样误差率[28-29]。

通过以上抽样方法,Bagging 基学习器采样得到的样本遵循样本本身的规模比例,相对之前有更好的代表性,但是正负样本数据依然是不平衡的。文献[1]中指出 τ -SS3 模型词汇置信度计算是与文档内词汇的词频相挂钩的,因此增加或者削减文档的数量对模型影响并不显著。

针对数据集的不平衡问题,该文将文档内的文本在原文档内进行随机片段重复以实现增强,这样一定程度上增加了少数类数据关键词汇的词频,从而增加少数类词汇关键词的置信度,达到削弱数据不平衡影响的效果。

该模型引入了一个自适应系数(Adaptive Coefficient,AC),在 Bagging 的分层采样阶段,各文本数据会被读取到文档列表的不同单元中,使用自适应系数将某一文档单元的数据增多从而达到增加词频的效果。计算公式如下(New_Text:文档自适应增强后的新文本,Text:原文本,random$i$:小于等于文本长度的随机数且满足:random$i$ < random(i + 1)):

$$\text{New\_Text} = \text{Text} + \text{AC} \times \text{Text}[\text{random1}:\text{random2}] \tag{8}$$

由于各文档内文本数据量的不一致性,在设置了自适应系数上下界后,模型使用 2.2 节提出的 GF 分数在设定范围内搜索合适的自适应系数。

此外,文献[7]中模型每次训练 n-gram 是固定的,导致模型兼顾词缀范围不足。该文提出的集成学习方法针对 τ -SS3 模型可设置 n-gram[30] 的特点,在每次迭代训练学习器时对 gram 进行调整,兼顾了更多的词缀情况。

1.2.2 基学习器的选择性丢弃机制

在 Bagging 的集成阶段,性能较差的基学习器较多时,通过集成学习得到的最终模型性能也会变差。因此,该文在集成阶段剔除效果过差的基学习器。对于学习器效果的判断标准可以根据任务类型进行选取。丢弃效果过差学习器之后,利用达到评价标准的学习器再进行集成,提升了集成的可信度。

结合基学习器的选择性丢弃机制的文档自适应增强 Bagging 算法流程如图 4 所示。

2 实验分析

2.1 数据集及实验配置

该文先在 E-Risk2017 抑郁检测数据集上进行实

验。为了验证模型的泛化能力,还在 E-Risk2017、DAIC[31]和 E-DAIC[32]组成的混合数据集上进行实验,采用 GF-Score、F1-Score 和 G-Mean Score 评估模型。

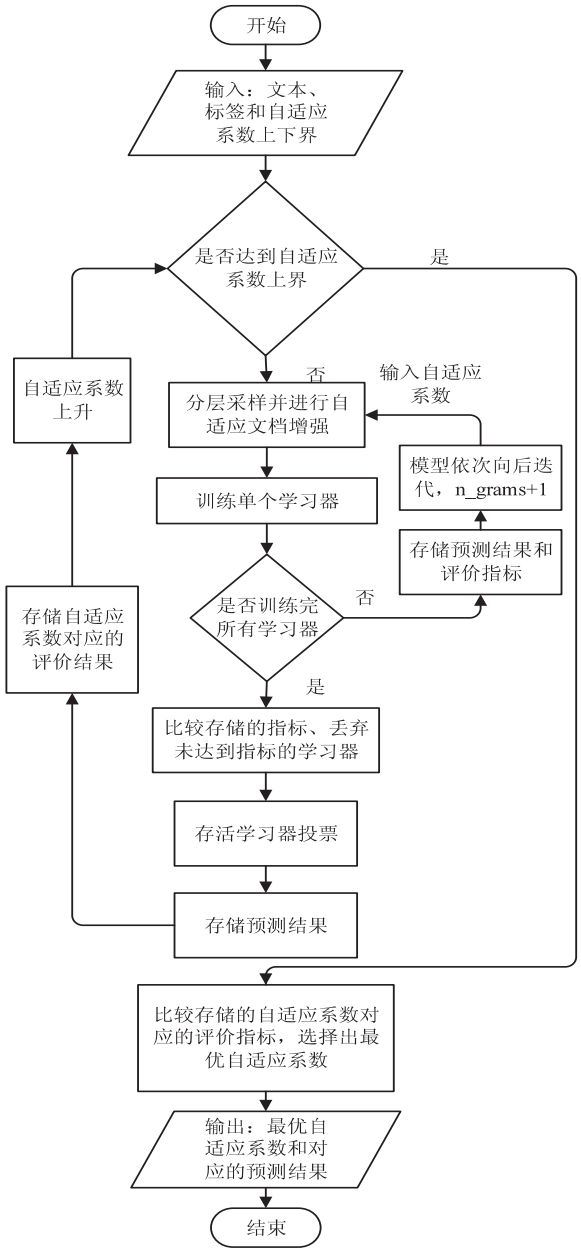

图 4　文档自适应增强 Bagging 算法流程

E-Risk2017 抑郁检测数据集中的数据均来源于社交媒体,包括来自 887 名 Reddit 用户总共 8 200 份文本数据,用户的文本按时间顺序排列,分为 positive(抑郁)和 negative(非抑郁)。数据集中总计有 7 410 份 positive(抑郁)数据,总计 160 488 968 字节,790 份 negative(非抑郁)数据,总计 13 758 691 字节,数据规模比例大约为 12∶1,是典型的不平衡数据集。

DAIC 抑郁检测数据集是疾病分析访谈语料库(Distress Analysis Interview Corpus)的一部分,包括 189 段会话,共 1 609 144 字节。

E-DAIC(Extended Distress Analysis Interview Corpus)抑郁检测数据集是 DAIC 语料库的拓展,共 2 640 573字节。

混合数据集由 E-Risk2017、DAIC 和 E-DAIC 按照各自数据量在总数据中所占比例随机组合而成,分为 positive 和 negative,正负类比例为 10∶1,依然是不平衡数据集,混合数据集也被分为训练集和测试集。

实验的操作系统环境为 Windows11,编程语言 Python3.9,模型训练主要在 CPU(AMD Ryzen 7 6800H)和 GPU(GeForce RTX 3060)上完成。

文中方法代码已在 GitHub 开源:

https://github.com/beizhi23/DAEB--SS3.git

实验参数包含平滑度、显著性、制裁度等,具体参数设置见表 1。由于数据本身不平衡,所以分别在 E-Risk 2017 数据集和混合数据集上采用 Stratified K-Fold($k=5$)进行实验。

表 1　实验参数设置

参数	参数值
平滑度	0.32
显著性	0.8
制裁度	1.7
自适应系数范围	1~4
置信度加权系数	1.15
GF_α	1.0
GF_β	1.0

2.2　评价指标选择

分类任务中,常用的机器学习评价指标及计算公式如下(True Positive, TP; False Negative, FN; False Positive,FP;True Negative,TN):

准确率:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{9}$$

精确率:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

召回率,也称真阳性率:

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

特异度,也称真阴性率:

$$\text{Specificity}=\frac{\text{TN}}{\text{TN}+\text{FP}} \tag{12}$$

综合评价指标 F1-Score、G-Mean Score:

$$\text{F1}-\text{Score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{13}$$

$$\text{G}-\text{Mean Score}=\sqrt{\text{Specificity}\times\text{Recall}} \tag{14}$$

准确率、精确率等单一指标不足以全面评估模型

性能。尤其在样本不平衡的情况下,单一评价指标均存在一定程度的失真。样本不平衡时,对于精确率,TP+FP 或 TN+FN 在数量上优势明显,即使少数类扩大十倍,其对精确率影响仍不超 1%,故精确率不足以评估模型性能。Recall 也是如此,当正类(po)远多于负类(ne),FN 对 Recall 影响微小,但反之,FN 变化则会对 Recall 产生更大影响,导致过度关注 ne 数据,评价失真。Specificity 与 Recall 情况类似。

F1-Score 和 G-Mean Score 综合了精确率、召回率等指标,但在样本不平衡情况下,它们评估模型性能的可信度也会降低。对于 F1-Score,当 po 多于 ne 时,Recall 削弱了 ne 对 F1-Score 的影响,Precision 不会失真,故 po 多于 ne 时 F1-Score 会更加关注 po 类数据。而当 ne 多于 po 时,Recall 会更关注 ne 类,Precision 依然不会失真,继而导致 F1-Score 更加关注 ne 数据。综上所述,F1-Score 更加关注不平衡类别中的多数类数据。同理,G-Mean Score 则更加关注不平衡类别中的少数类数据。

为了可以全面地评价模型,并基于评价指标进行参数调整,该文提出一种新的评价指标 $GF_{(\alpha,\beta)}$-Score。

根据上文所述,在数据不平衡时 F1-Score 和 G-Mean Score 关注的类别不同,为了平衡两者对分类类别的关注程度,对 F1-Score 和 G-Mean Score 求调和平均值,公式如下:

$$GF-Score=\frac{2\times F1\times G-Mean}{F1+G-Mean} \tag{15}$$

但是,在某些分类任务中,可能希望更关注其中一类数据,所以可以对 F1-Score 和 G-Mean Score 求加权调和平均值,可得公式 16:

$$GF_{(\alpha,\beta)}-Score=\frac{(\alpha+\beta)\times F1\times G-Mean}{\alpha F1+\beta G-Mean} \tag{16}$$

如果 $\alpha>\beta$,G-Mean Score 有更大影响,即更关注少数类,如果 $\alpha<\beta$,F1-Score 有更大影响,即更关注多数类。

当 $\alpha=\beta=1$ 的时候,F1-Score 和 G-Mean Score 影响力相同,即 GF-Score。

基于以上研究,该文选择使用 GF-Score 进行模型的性能评估和参数调整,F1-Score 和 G-Mean Score 作为参考指标。

2.3 实验结果分析

2.3.1 算法比较分析

为了更加直观地展示基于 GF-Score 进行参数调整的优越性,表 2 给出了在不同数据集上分别基于 F1、G-Mean、GF-Score 进行参数调整时 SS3、τSS3 和 DAEB-τSS3 的评价结果。

表 2 不同数据集上基于不同指标参数调整所得模型评价

数据集	模型	评价指标	基于 F1-Score 参数调整	基于 G-Mean Score 参数调整	基于 GF-Score 参数调整
E-Risk2017	SS3	F1-Score	0.92	0.77	0.89
		G-Mean Score	0.46	0.77	0.70
		GF-Score	0.61	0.77	0.78
	τ-SS3	F1-Score	0.97	0.77	0.93
		G-Mean Score	0.53	0.79	0.68
		GF-Score	0.69	0.78	0.81
	DAEB-τSS3	F1-Score	0.95	0.90	0.95
		G-Mean Score	0.66	0.91	0.90
		GF-Score	0.77	0.91	0.92
混合数据集	SS3	F1-Score	0.94	0.81	0.92
		G-Mean Score	0.39	0.71	0.68
		GF-Score	0.55	0.76	0.78
	τ-SS3	F1-Score	0.94	0.77	0.92
		G-Mean Score	0.45	0.76	0.72
		GF-Score	0.61	0.77	0.80
	DAEB-τSS3	F1-Score	0.93	0.82	0.93
		G-Mean Score	0.55	0.80	0.79
		GF-Score	0.69	0.81	0.86

通过表 2 可以看出,基于最优 F1 分数进行模型超参数选择时,模型可以获得较高的 F1-Score,但是 G-

Mean Score,GF-Score 较低。基于最优 G-Mean 分数时,G-Mean Score 和 GF-Score 会出现一定程度提升,但是 F1-Score 会大幅度下降,说明少数类数据分类效果增强,但是多数类数据也大量向少数类标签倾斜从而导致 F1 的降低。而基于 GF 分数进行模型超参数选择时,F1 和 G-Mean 均可以保持在较高水平。

从而可以得出结论:在数据不平衡情况下,F1-Score、G-Mean Score 不宜作为超参数选择的首选指标,应基于 GF-Score 进行参数调整。

基于 GF-Score 进行调参时,将文中模型(DAEB-τSS3)与以下 13 种方法进行了对比:

(1)RF:基于文献[9]中所述,RF 在不平衡数据集上表现良好;

(2)SVM[33]:在多种二分类任务中表现出色;

(3)过采样+SVM:采用随机过采样增加少数类数据集数量;

(4)GBDT:基于文献[9]中所述,GBDT 在不平衡数据集上会取得更加优秀的表现;

(5)SS3:在 E-Risk2017 数据集上表现优秀,获得 E-Risk2017 抑郁症检测挑战赛冠军;

(6)τSS3:文中集成学习基学习器的原始模型;

(7)LightGBM[34](LGBM):一种高效的梯度 Boosting 框架;

(8)Stacking[35]:将 GDBT、SVM、LGBM、AdaBoost[36]、RF 作为基学习器进行 Stacking,融合上述模型优点,具有更强大的分类能力;

(9)TextCNN[37]:TextCNN 通过局部特征学习、多尺度特征融合、参数共享等方法提高了分类性能,优于传统深度学习模型;

(10)过采样+TextCNN:在 TextCNN 基础上采用随机过采样增加少数类数据集数量;

(11)BiLSTM[38]:能有效地捕捉文本序列信息和长距离依赖关系,从而提高分类性能;

(12)过采样+BiLSTM:在 BiLSTM 基础上采用随机过采样增加少数类数据集数量;

(13)ERNIE 3.0[39]:百度推出的大语言模型,有极强的语义理解能力和分类能力。

观察表 3 发现,基于 GF-Score 进行的参数调整显著提升了模型的综合表现。并且本模型算法相对于基模型 τSS3,GF-Score、F1-Score、G-Mean Score 分别提升了 13.6%,2.2%,32.4%,相较于其他算法以及模型,也有着明显的优势。

观察表 4 发现,在混合数据集上,本模型依然优于其他模型,并且相对于基模型 τSS3,GF-Score、F1-Score、G-Mean Score 分别提升了 7.5%,1.1%,9.7%,相较于 TextCNN、BiLSTM 和 ERNIE 3.0 等热门模型及改进算法也有一定优势。

表 3　不同算法在 E-Risk2017 数据集上的表现
(从上向下模型 GF-Score 递增)

模型/评价指标	GF	F1	G-Mean
LGBM	0.49	0.95	0.33
RF	0.61	0.95	0.45
SVM	0.63	0.95	0.47
过采样+SVM	0.70	0.87	0.58
Stacking	0.70	0.97	0.55
GBDT	0.72	0.96	0.58
SS3	0.78	0.89	0.70
τ SS3	0.81	0.93	0.68
TextCNN	0.81	0.95	0.70
BiLSTM	0.82	0.96	0.71
过采样+BiLSTM	0.83	0.96	0.73
ERNIE 3.0	0.83	0.95	0.74
过采样+TextCNN	0.84	0.96	0.75
DAEB-τ SS3	0.92	0.95	0.90

表 4　不同算法在混合数据集上的表现
(从上向下模型 GF-Score 递增)

模型/评价指标	GF	F1	G-Mean
LGBM	0.50	0.95	0.34
RF	0.68	0.94	0.53
Stacking	0.67	0.91	0.53
SVM	0.67	0.94	0.52
过采样+SVM	0.73	0.94	0.59
GBDT	0.75	0.95	0.62
SS3	0.78	0.92	0.68
TextCNN	0.78	0.96	0.66
τ SS3	0.80	0.92	0.72
BiLSTM	0.81	0.95	0.70
ERNIE 3.0	0.81	0.96	0.70
过采样+TextCNN	0.83	0.96	0.73
过采样+BiLSTM	0.84	0.94	0.75
DAEB-τ SS3	0.86	0.93	0.79

基于以上实验结果,该文提出的方法可以显著提升 τSS3 模型适应不平衡数据集的能力并使模型的综合表现得到进一步优化,最终在分类效果上优于多种常用算法和热门模型。

2.3.2　消融实验分析

为了测试不同改进模块对模型性能的影响,该文将算法裁剪成 6 组。第一组为原始算法 τSS3,对算法不做任何改进;第二组在 τSS3 算法中使用置信度加权(Confidence weighting,CW);第三组在 τSS3 算法中使用传统 Bagging 进行集成学习;第四组在 τSS3 算法中使用文档自适应增强型 Bagging(Document Adaptive Enhancement Bagging,DAE-Bagging);第五组将置信

度加权型 τSS3 与传统 Bagging 相结合;第六组使用完整的 DAEB-τSS3 算法。使用 GF-Score 作为评价指标,在 E-Risk2017 抑郁检测数据集上进行消融实验,结果如表 5 所示,“√”表示使用该结构,“×”表示不使用该结构。

表 5 消融实验结果对比

模型	CW	Bagging	DAE-Bagging	GF-Score
1	×	×	×	0.81
2	√	×	×	0.85
3	×	√	×	0.84
4	×	×	√	0.90
5	√	√	×	0.88
6	√	×	√	0.92

从表 5 可以看出,通过置信度加权模型 2 比原始算法 τSS3 提升了 4.9% 左右。第三组采用传统 Bagging 的实验结果和第四组 DAE-Bagging 相比效果相差了 6 百分点,说明 DAE-Bagging 优于传统 Bagging,第五组和第六组结果对比也同样体现出了 DAE-Bagging 的优势。

为了进一步验证文中方法的有效性,输出了第一组消融实验的少数类词云与第六组的少数类词云,如图 5 所示。

对比图 5(a)和(b)可以明显看到,使用文中方法之后,少数类词云变得更加饱满,高置信度词汇也更多(增加的主要高置信度词汇用深色表示),进一步验证了文中方法的合理性和有效性。

(a)第一组的少数类词云 (b)第六组的少数类词云

图 5 消融实验词云对比

3 结束语

该文提出一种应用于社交媒体文本不平衡数据下的基于 DAEB-τSS3 模型的抑郁检测方法。将置信度增强 τ-SS3 模型作为基学习器进行文档自适应增强 Bagging 集成学习。在 Bagging 的数据输入阶段,改进 Bagging 的随机采样为分层采样,并对文档内文本进行自适应增强选择并丢弃表现过差的基学习器,同时采用动态 N-grams 进行学习,最后,使用 GF-Score 综合评价指标进行自动参数选择。实验表明,该方法相对于基准模型有更强的适应数据集能力和抑郁症检测能力。未来会将掩码语言模型与该方法相结合,不仅局限于基于词频、词缀和概率分布的分类方法,而是能在兼顾局部和全局语义的基础上,更好地解决不平衡文本分类的问题。

参考文献:

[1] BURDISSO S G,ERRECALDE M,MONTES-Y-GOMEZ M. A text classification framework for simple and effective early depression detection over social media streams[J]. Expert Systems with Applications,2019,133:182-197.

[2] 王 垚,贾宝龙,杜依宁,等. 基于词向量的多维度正则化 SVM 社交网络抑郁倾向检测方法[J]. 计算机应用与软件,2022,39(3):116-120.

[3] 张梦娜,王君岩,龙 洋,等. 基于社交媒体中文本信息的早期抑郁症检测[J]. 中国生物医学工程学报,2022,41(1):21-30.

[4] 陈 妍,罗雪琴,梁 伟,等. 基于情感信息融合注意力机制的抑郁症识别[J]. 智能科学与技术学报,2022,4(4):600-609.

[5] 张 慧. 基于文本预训练模型的抑郁倾向检测方法研究[D]. 济南:山东师范大学,2022.

[6] 张宗佳. 基于时间感知的社交媒体文本抑郁检测研究[D]. 兰州:兰州大学,2021.

[7] BURDISSO S G, ERRECALDE M , MONTES-Y-GÓMEZ M. t-SS3:a text classifier with dynamic n-grams for early risk detection over text streams[J]. Pattern Recognition Letters,2020,7(1):1-10.

[8] HE H,GARCIA E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge and Data Engineering,

2008,21(9):1263-1284.

[9] 徐玲玲,迟冬祥. 面向不平衡数据集的机器学习分类策略[J]. 计算机工程与应用,2020,56(24):12-27.

[10] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C]//Artificial intelligence in medicine:8th conference on artificial intelligence in medicine in Europe. Cascais:Springer,2001:63-66.

[11] 林舒杨,李翠华,江　弋,等. 不平衡数据的降采样方法研究[J]. 计算机研究与发展,2011,48(z2):425-431.

[12] CHAWLA N V, BOWYER K W, HALL L O , et al. SMOTE:synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16(1):321-357.

[13] HU S,LIANG Y,MA L,et al. MSMOTE:improving classification performance when training data is imbalanced[C]//Second international workshop on computer science & engineering. Qingdao:IEEE,2010:13-17.

[14] HAN H,WANG W Y,MAO B H. Borderline-SMOTE:a new over-sampling method in imbalanced data sets learning[C]//International conference on intelligent computing. Hefei:Chinese Academy of Sciences,2005:878-887.

[15] HE H,BAI Y,GARCIA E A,et al. ADASYN:adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). Hong Kong,China:IEEE,2008:1322-1328.

[16] 张　明,胡晓辉,吴嘉昕. 基于混合采样的不平衡数据集算法研究[J]. 计算机工程与应用,2019,55(17):68-75.

[17] THAI-NGHE N,GANTNER Z,SCHMIDT-THIEME L,et al. Cost-sensitive learning methods for imbalanced data[C]//International joint conference on neural networks. Barcelona:IEEE,2010:1566-1573.

[18] BREIMAN L. Bagging predictors[J]. Machine Learning,1996,24(2):123-140.

[19] FRIEDMAN J. Greedy function approximation:a gradient boosting machine[J]. The Annals of Statistics,2002,29(5):1189-1232.

[20] BRIEMAN L. Random forests[J]. Machine Learning,2001,45(1):5-32.

[21] AVENUE M,HILL M,COHEN W W,et al. Fast effective rule induction[C]//Machine learning. Tahoe City:IMLS,1995:115-123.

[22] SETTLES B. Active learning literature survey[D]. Madison:University of Wisconsin,2010.

[23] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature,2015,521(7553):436-444.

[24] HAN X,ZHANG Z,DING N,et al. Pre-trained models:past,present and future[J]. AI Open,2021,2:225-250.

[25] CHINCHOR N. MUC-4 evaluation metrics[C]//Conference on message understanding. San Diego:Association for Computational Linguistics,1993:409-450.

[26] DENG X,LIU Q,DENG Y,et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem[J]. Information Sciences,2016,340/341:250-261.

[27] SADEQUE F,XU D,BETHARD S. UArizona at the CLEF eRisk 2017 pilot task:linear and recurrent models for early depression detection[C]//CLEF. Dublin:Information Technology and Nanotechnology,2017.

[28] 苗　杰. 简单随机抽样和分层抽样效率的实证检验[J]. 中国科技投资,2019(33):273.

[29] 周　丹. 分层抽样与简单随机抽样效率比较[J]. 经济视野,2013(1):99-100.

[30] OGADA K,MWANGI W,CHERUIYOT W. N-gram based text categorization method for improved data mining[J]. Journal of Information Engineering and Applications,2015,5(8):35-43.

[31] RINGEVAL F,PANTIC M,SCHULLER B,et al. AVEC 2017:re-al-life depression,and affect recognition workshop and challenge[C]//Audio/visual emotion challenge and workshop. Mountain View:ACM,2017:3-9.

[32] RINGEVAL F,SCHULLER B,VALSTAR M,et al. AVEC 2019 workshop and challenge:state-of-mind,detecting depression with AI,and cross-cultural affect recognition[C]//Audio/visual emotion challenge and workshop. Nice:ACM,2019:3-12.

[33] ZENG Zhiqiang,XIE Yanqi,YU Hong-bin,et al. Fast training support vector machines using parallel sequential minimal optimization[C]//ISKE 2008. Xiamen:[s. n.],2008:997-1001.

[34] QI M. LightGBM:a highly efficient gradient boosting decision tree[C]//Neural information processing systems. Long Beach:Curran Associates Inc. ,2017:3149-3157.

[35] WOLPERT D H. Stacked generalization[J]. Neural Networks,1992,5(2):241-259.

[36] FREUND Y,SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[C]//Conference on learning theory. Berlin:Academic Press,1997:119-139.

[37] KIM Y. Convolutional neural networks for sentence classification[C]//Conference on empirical methods in natural language processing. Doha:ACL,2014:1746-1751.

[38] HUANG Z,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv:1508.01991,2015.

[39] SUN Y,WANG S,FENG S,et al. ERNIE 3.0:large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv:2107.02137,2021.