一种处理严重不均衡数据的 BERT-BiGRU-WCELoss 短文本警情分类模型

刘 冬[1] 翁海光[1] 陈一民[2*]
[1](上海公安学院 上海 200137)
[2](上海建桥学院 上海 201306)

摘 要 针对 110 报警类警情文本数据存在着文本长度极短且样本类别分布严重不均衡的问题,提出一种 BERT-BiGRU-WCELoss 警情分类模型。该模型通过中文预训练 BERT(Bidirectional Encoder Representations from Transformers)模型抽取文本的语义;使用 BiGRU(Bidirectional Gated Recurrent Unit)综合提炼文本的语义特征;通过优化自适应权重损失函数 WCELoss(Weight Cross Entropy Loss function)给少数类样本赋予更大的损失权重。实验结果表明:该模型在某市 2015 年某一自然月的 110 报警数据集上取得了 95.83% 的分类准确率,精准率、召回率、F1 值和 G_mean 均高于传统深度学习模型和交叉熵损失训练的模型。

关键词 BERT BiGRU 警情分类 非均衡数据 短文本 样本加权

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.09.031

A BERT-BIGRU-WCELOSS CLASSIFICATION MODEL FOR HANDING SEVERELY UNBALANCED SHORT ALERT TEXT DATA

Liu Dong[1] Weng Haiguang[1] Chen Yimin[2*]
[1](*Shanghai Police College*, *Shanghai 200137*, *China*)
[2](*Shanghai Jian Qiao University*, *Shanghai 201306*, *China*)

Abstract In response to the problem of extremely short text length and severely imbalanced distribution of sample categories in 110 alarm text data, this paper proposes a BERT-BiGRU-WCELoss alarm classification model. The model extracted the semantics of the text through the Chinese pre trained BERT (Bidirectional Encoder Representations from Transformers) model. BiGRU (Bidirectional Gated Recurrent Unit) was used to comprehensively extract the semantic features of the text. By optimizing the adaptive weight loss function WCELoss (Weight Cross Entropy Loss function), larger loss weights were assigned to minority class samples. The experimental results show that the model achieved a classification accuracy of 95.83% on the 110 alarm dataset of a certain natural month in 2015 in a certain city, with higher accuracy, recall rate, F1 value, and G_Mean than traditional deep learning models and models trained with cross entropy loss.

Keywords BERT BiGRU Classification of alarm text Unbalance data Short text Sample weighting

0 引 言

随着公安科技信息化的快速发展,各地公安机关汇集了大量的警务工作数据,其中包含数量巨大的警情文本数据。传统的警情文本分类采用人工分类的方法,存在着效率低、准确率差和人力成本高昂等问题。尤其是 110 报警类信息,存在着文本长度极短(约 20 个字)、样本分布不均匀等难题,使用传统的机器学习算法很难取得满意的分类结果[1-2]。现代警务工作是基于大数据分析和预测之上的精准警务,但面对如此海量的非结构化短文本数据,如何高效分析和利用这

收稿日期:2023-12-26。上海公安学院科研项目(23xkx53)。**刘冬**,讲师,主研领域:警务数据分析与建模,自然语言处理。**翁海光**,正高工。**陈一民**,教授。

些数据是一个棘手问题[3]。

为了能够快速有效对警情文本数据进行分类,众多警务科技工作者使用机器学习和深度神经网络开展了相关研究。张齐等[4]研究了 XGBoost、GBDT、KNN、SVM 和朴素贝叶斯5 种机器学习算法在8 种盗窃类犯罪简要案情文本的分类性能(训练集 2 017 条,测试集 514 条,样例数据在去除标点符号和停用词后平均长度为 60.9 个字符),XGBoost 算法取得了 92.3% 的精准率,但其只分类了盗窃类案件,并对数据进行了样本均衡处理,删除了发案量少的类别,与真实警情类别分布存在较为明显的差异。

王孟轩等[1]使用 CNN + Bi-LSTM + RNN + Self-Attention + MLP 组成了一个复合深度学习模型,该模型在 9 类刑事案件接报分类数据集(纵火、绑架、骚乱、盗窃、诈骗、抢劫、斗殴、制假和损坏财物)中准确率达到了 97%,但其使用的数据集也经过样本数量均衡处理,且挑选的 9 类案件文字表述具有明显的差异,与真实警情数据分布差别较大。

李昀轩等[5]使用预训练的 BERT 模型结合 Linear、TextCNN 或 DPCNN 等网络组成分类模型,结果表明这些模型在交通警情数据集(共 12 万条,文本长度在 50 至 100 个字)中分类准确率相比传统深度学习都取得了显著提升,其中 BERT + DPCNN 模型取得了 93.5% 的最高分类准确率。但该论文只研究了警情数据中的“人车事故”“车与车事故”“车与非机动车事故”等 6 种交通事故数据,并没有对其他类别的警情数据开展相关研究。

殷小科等[6]对 110 警情数据开展了全类别的分类研究,发现 110 警情数据存在着类别分布不均的问题,为此他们首先训练了 19 个 BERT 分类器,然后将这 19 个分类器组成了一个三层树状结构模型,结果表明该模型在分类任务中取得了 95.8% 的准确率。但从零开始训练 19 个 BERT 模型进行分类,所需要的时间、计算资源很大,一般机器无法实现相关计算任务,并且所需的数据资源巨大,所以基本上是无法承受的。

目前关于警情分类的研究大多集中在“案件接报”“简要案情”等文本长度相对较长的警情,有关 110 报警的警情分类研究很少[7]。针对 110 报警文本数据存在着文本长度极短、类别分布严重不均衡的问题,本文提出一种处理严重不均衡数据的 BERT-BiGRU-WCELoss 110 短文本警情分类模型。由于短文本自身包含的信息量太少,传统的循环神经类网络无法从短文本中学习到足够的特征,分类效果欠佳。但是,预训练 BERT 模型由于在预训练时学习了海量文本信息,因此借助预训练 BERT 模型 + 微调的方法,我们能够更好地从短文本中提取数据特征[8-11]。在处理严重不均衡数据样本时,我们可以从采样方法和损失函数两个方面对模型进行优化改进。同时,考虑到 110 警情类别样本数量在自然时间内就是严重失衡的,因此本文将重点放在了改进损失函数上,通过自适应优化权重技术,为少数类样本赋予了更高的权重,以此达到提升严重不均衡数据分类的准确性[12,15]。

1 模型介绍

本文提出的处理严重不均衡数据的 BERT-BiGRU-WCELoss 短文本警情分类模型先使用预训练 BERT 提取警情文本数据的特征,接着使用 BiGRU 将上述特征进一步融合后做分类,然后采用 MLP(Multilayer Perceptron) + Softmax 进行进一步的细分,最后通过本文所提的自适应优化权重的 WCELoss 损失函数计算模型的损失,误差逆传播后更新网络的权重参数。处理严重不均衡数据的 BERT-BiGRU-WCELoss 短文本警情分类模型的总体结构如图 1 所示。

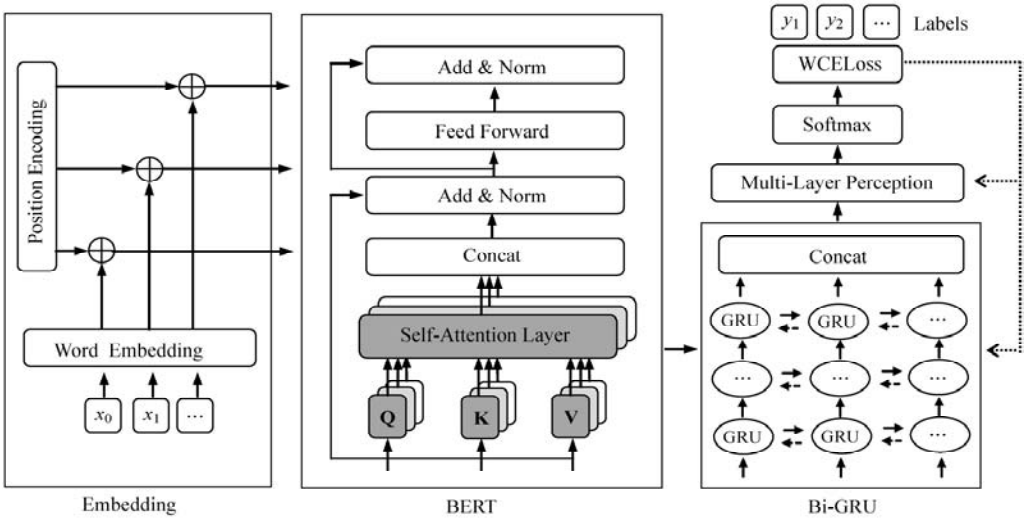


图 1 处理严重不均衡数据的 BERT-BiGRU-WCELoss 的短文本警情分类模型的总体结构

1.1 BERT

BERT 模型结构如图 1 中的 BERT 模块所示,它使用了 Transformer 架构中的编码器(Encoder)部分作为自己的核心模块。Transformer 架构最初的设计用途为机器翻译,它是一种典型的 Seq2Seq 模型,输入的文本序列通过编码器和解码器后可以转换成另一个序列[16]。与传统基于 RNN(Recurrent Neural Network)或 LSTM(Long Short-Term Memory)的 Seq2Seq 模型不同,Transformer 架构只使用了多头注意力机制(Multi-Head Attention)来提取文本中字与字之间的关联性,并不涉及任何循环或卷积。

Transformer 编码器中使用的多头注意力机制是由多个自注意力(Self-Attention)模块组成,其计算过程如图 2 所示。自注意力机制为输入的每个词向量创建了三个可学习参数矩阵 $\boldsymbol{W}_{\mathrm{Q}}$、$\boldsymbol{W}_{\mathrm{K}}$、$\boldsymbol{W}_{\mathrm{V}}$,线性映射的计算方法如式(1)-式(3)所示。

$$\boldsymbol{Q}=\boldsymbol{W}_{\mathrm{Q}}\times\boldsymbol{A} \tag{1}$$

$$\boldsymbol{K}=\boldsymbol{W}_{\mathrm{K}}\times\boldsymbol{A} \tag{2}$$

$$\boldsymbol{V}=\boldsymbol{W}_{\mathrm{V}}\times\boldsymbol{A} \tag{3}$$

式中:$\boldsymbol{W}_{\mathrm{Q}}$、$\boldsymbol{W}_{\mathrm{K}}$和 $\boldsymbol{W}_{\mathrm{V}}$为三个线性变换矩阵,$\boldsymbol{A}$ 为输入词向量矩阵,输入 $\boldsymbol{A}$ 与三个线性变换矩阵相乘后得到 $\boldsymbol{Q}$(query)、$\boldsymbol{K}$(key)、$\boldsymbol{V}$(value)三个矩阵。矩阵 $\boldsymbol{Q}$ 与矩阵 $\boldsymbol{K}$ 的转置点乘计算,接着 Softmax 归一化获得权重矩阵,最后再与矩阵 $\boldsymbol{V}$ 计算就获得融合了当前字与上下文信息的增强语义向量矩阵。计算如式(4)所示。

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\times\boldsymbol{V} \tag{4}$$

式中:d_{k} 是输入向量的长度。

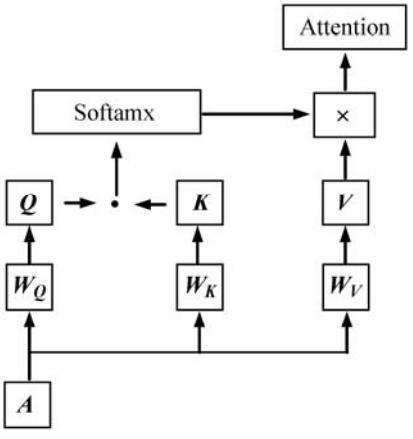


图 2　自注意力机制计算流程

多头注意力的计算公式如式(5)、(6)所示。

$$h_{\mathrm{ead}_i}=\mathrm{Atttention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \tag{5}$$

$$M_{\mathrm{ultiHead}}=\mathrm{Concat}(h_{\mathrm{ead}_1},h_{\mathrm{ead}_2},\cdots,h_{\mathrm{ead}_h})\boldsymbol{W}_o \tag{6}$$

式中:h 为多头注意力的头数;$\boldsymbol{W}_o$ 为可学习的参数矩阵。

BERT 模型的训练方法有两种:

(1) Masked LM:随机抹掉输入文本中的部分词,然后模型需根据剩余的上下文去预测被抹掉的词。在实际训练过程中不是简单地将所有被抹掉的词都用[MASK]替换,有 10% 的概率会使用任意其他词替换,还有 10% 的概率保持原有的词汇不变。这样模型就不知道哪些词被替换过,从而必须记住每个输入词的上下文分布信息。

(2) Next Sentence Prediction:是指从某篇文章中随机地抽取两个句子 A 和 B,然后让模型判断句子 B 是否紧挨在句子 A 之后[8]。

Masked LM 和 Next Sentence Prediction 都是自监督学习方法,因此 BERT 可以使用海量的无标签语料作为训练数据,无须事先人工打标。预训练的 BERT 能够很好地提取出文本的特征,然后使用这些特征应用到下游任务中(如文本分类等),可以取得比传统深度学习算法更好的效果。

本文中预训练 BERT 模型使用经我们改写和适配后的开源"bert-base-chinese"模型,使用海量中文文本进行预训练(共约 2.5 亿个字),字典大小为 21 128 个字,层数为 12,词嵌入维度为 768,参数总量为 1.03 亿。通过高维度以及巨量数据的训练,我们可以更好地从短文本中提取出语义特征,提高模型分类性能。

1.2 BiGRU

为了充分获取文本数据中的信息,在 GRU 的基础上,我们提出采用双向计算的 GRU 网络,即 BiGRU。GRU 是在 LSTM 基础上改进后的一种循环神经网络[17]。GRU 单元结构如图 3 所示。GRU 架构将 LSTM 中的遗忘门和输入门整合成了更新门 $\boldsymbol{z}_t$,更新门控制着上一步的信息输入量以及当前步骤的状态信息。重置门 $\boldsymbol{r}_t$前一步骤的多少信息被写入当前步骤的状态信息。更新门和重置门都使用 sigmod 函数,其输出值范围在 0~1,通过输出值的大小来实现控制信息的通过量。GRU 单元状态计算过程如式(7)-式(10)所示。

$$\boldsymbol{z}_t=\sigma(\boldsymbol{W}_{\mathrm{z}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{z}}) \tag{7}$$

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_{\mathrm{r}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{r}}) \tag{8}$$

$$\boldsymbol{h}'_t=\tanh(\boldsymbol{W}_{\mathrm{h}}[\boldsymbol{r}_t\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{h}}) \tag{9}$$

$$\boldsymbol{h}_t=\boldsymbol{z}_t\boldsymbol{h}_{t-1}+(\boldsymbol{1}-\boldsymbol{z}_t)\boldsymbol{h}'_t \tag{10}$$

式中:$\boldsymbol{W}_{\mathrm{z}}$、$\boldsymbol{W}_{\mathrm{r}}$和 $\boldsymbol{W}_{\mathrm{h}}$为权重矩阵;$\boldsymbol{b}_{\mathrm{z}}$、$\boldsymbol{b}_{\mathrm{r}}$和 $\boldsymbol{b}_{\mathrm{h}}$为偏置项;$\boldsymbol{h}'_t$为当前步骤计算获得的信息更新量;$\boldsymbol{h}_t$为当前步骤的输出,它是前一步骤状态信息与当前步骤信息更新量的综合。

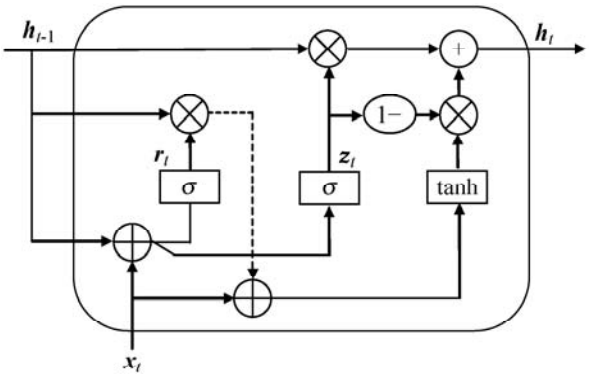


图 3 GRU 单元结构示意图

与 RNN、LSTM 类似,GRU 网络中的文本数据计算方向除了传统的正向计算也可以采用正反同时计算[18]。本文为了充分获取文本数据中的信息,因此我们提出采用双向计算的 GRU 网络,即 BiGRU。BiGRU 的结构如图 1 中的 BiGRU 模块所示。反向 GRU 的计算公式与正向一致,输入顺序由正向变为反向输入,从文本的最后一个字开始直至第一个字。BiGRU 的最终输出结果计算过程如式(11)所示。

$$\boldsymbol{h}_T = \text{Concat}(\overrightarrow{\boldsymbol{h}}_T, \overleftarrow{\boldsymbol{h}}_T) \tag{11}$$

式中:$\overrightarrow{\boldsymbol{h}}_T$为最后时刻 T 的正向 GRU 输出结果,$\overleftarrow{\boldsymbol{h}}_T$为反向 GRU 最后时刻的输出结果,将两者拼接在一起即为 BiGRU 输出结果。

1.3 WCELoss

传统多分类任务的损失多采用交叉熵损失函数进行计算,计算方法如式(12)所示。

$$L_{ce} = -\frac{1}{N}\sum_{i}^{N} y_i \log(P_i) \tag{12}$$

式中:N 为样本总数;y_i 为第 i 个样本的标签值;P_i 为模型预测的分类结果。由于在计算损失时传统的交叉熵损失函数给每一个样本的损失权重都是相等的,因此在处理严重不均衡样本数据时模型只需将结果尽可能地预测为多数类就能获得较小的损失,模型将无法预测出少数类样本[19-20]。我们在 Focal Loss 的基础上,提出通过自适应优化权重技术,为少数类样本赋予更高的权重,以此达到提升严重不均衡数据分类的准确性。它能够较好处理非均衡样本集的损失函数。首先我们为每一类的样本赋予不同的损失权重,并为困难样本赋予更高的损失权重。其计算方法如式(13)所示。

$$L_{wce} = -\alpha_i (1-p_i^t)^{\gamma} \log(p_i^t) \tag{13}$$

式中:p_i^t为第 i 个样本在第 t 训练轮次中模型预测出来对应真实标签上的概率,$-\log(p_i^t)$ 对应交叉熵,而 $(1-p_i^t)^{\gamma}$为样本预测困难程度权重,γ 值为权重指数,我们令其为 2。在模型训练刚开始时,模型的学习能力较弱,困难样本对应的 p_i^t越小,则其交叉熵损失 $-\log(p_i^t)$和样本困难权重$(1-p_i^t)^{\gamma}$越大,也就意味着困难样本比普通样本拥有更大的损失。随着模型训练轮次的增加,模型的学习能力逐渐加强,模型在困难样本上的预测值逐渐接近于真实标签(p_i^t逐渐接近于 1),则困难样本的交叉熵损失和样本困难权重都将快速减小,训练的总损失也随之减小。

α_i为第 i 个样本的类别权重系数,对于非均衡数量样本集,可以给数量少的样本类别赋予更高的权重。在样本集$\{x_i, y_i\}_{i=1}^{N}$,$y_i \in \mathbf{R}^J$,其中:x_i表示第 i 个样本的文本数据,y_i表示第 i 个样本的标签值,类别共有 J 种。类别权重 α_i最简单的计算方法是通过类样本占总数的百分比来计算,其计算公式为:

$$\alpha_i = 1 - \frac{\sum_{i=1}^{N}\prod(y_i = j)}{N} \tag{14}$$

式(14)中样本类别权重与类别样本数量占比成简单反向关系,占比越少的类别拥有更大的损失权重[21]。但对于数据分布极度不平衡的数据集,上述计算式对于少数样本类的权重变化不够灵敏,无法合理体现出少数类样本的权重,模型仍将大部分样本预测为多数类。如本文所使用的 110 报警数据集中,第 0 类的样本数量是第 3 类的 197 倍,但通过式(14)计算第 3 类的权重仅为第 0 类的 2 倍,类别权重的增大幅度仍无法弥补数量上的巨大差异。与之相反,如果简单使用类别样本占比数的倒数如式(15)作为类别权重 α_i,那么少数类的权重 α_i将远大于其他类别,从而导致模型简单地将更多的样本预测为少数类,大幅度降低模型预测的准确率。

$$\alpha_i = \frac{N}{\sum_{i=1}^{N}\prod(y_i = j)} \tag{15}$$

为了能够在增大少数类权重的同时防止少数类的类别权重过大,我们使用对数函数对类别权重进行调整。对数函数具有单调递增性,且在(0,1)空间范围内随着输入值的减小而较快速度减小,同时对数函数输出值减小的速度慢于式(15),可以有效防止少数类样本权重过大。计算式如下:

$$\alpha_i = -\log_{10}\left(\frac{\sum_{i=1}^{N}\prod(y_i = j)}{N}\right) \tag{16}$$

2 实验及结果分析

2.1 模型描述

为了进行对比参照,本文同时训练了九个模型:

(1) RNN 模型;(2) TextCNN 模型;(3) GRU 模型;(4) 3 个基于预训练 BERT + MLP 的模型,分别使用交叉熵损失函数(BERT-MLP-CE)、Fcolloss 损失函数(BERT-MLP-FL)、优化后的自适应权重损失函数 WCELoss(BERT-MLP-WCE);(5) 3 个基于 BERT + BiGRU 的模型,同样分别使用上述三个损失函数进行训练,记作:BERT-BiGRU-CE、BERT-BiGRU-FL 和 BERT-BiGRU-WCELoss。RNN、TextCNN 和 GRU 模型均使用不带权重交叉熵损失函数训练。GRU 模型使用 4 层双向 GRU 模块,每层 128 个 GRU 单元。上述 9 个模型的 batch_size 和学习率保持一致。

2.2　数据描述

本文选取某地区 2015 年某一自然月 110 报警文本数据共 43 000 条。数据集使用分层采样法,按照训练集(80%)、验证集(10%)和测试集(10%)进行分层划分,划分结果如表 1 所示。

表 1　数据集信息

类别	训练集	验证集	测试集
数量	34 400	4 300	4 300

表 2 是数据集中每条文本数据包含的字符个数统计(包含标点符号):数据的字符长度平均值为 23.7,最短的只有 2 字符,75% 的文本长度少于 29 个字,属于典型的短文本。

表 2　每条数据包含的字符个数统计

类别	数值
平均值	23.7
标准差	15.8
最小长度	2.0
最大长度	355.0
25% 分位	14.0
50% 分位	20.0
75% 分位	29.0

本实验数据涉及报警案由共 11 类,数据按照自然月的占比分层划分,不进行样本数量均衡处理。具体每一类的数量占比如表 3 所示。可以看出数量最多的是类别 0,占比超过了总数的 50%,类别 6、7、9 占比在 10% ~15% 之间,而类别 2、3、4、5、8 和 10 共 6 种类别占比均少于 2%,是少样本类别。在面对样本数量不均衡数据集时,模型通常倾向将样本预测为数量较多的类别而忽视少样本类别,因此少样本类别的召回率也是评价模型能力非常重要的指标之一。

表 3　样本类别占比数

类别	数量占比/%
0	51.24
1	4.52
2	1.89
3	0.26
4	1.43
5	0.48
6	15.48
7	12.99
8	0.98
9	10.26
10	0.47

2.3　实验评价指标

文本分类实验评价指标有 5 个,分别为准确率(Accuracy)、精准率(Precision)、召回率(Recall)、F1 值和 G_mean。

准确率即为预测类别正确的样本数占总样本数的百分比。

$$A_{\text{ccuracy}} = \frac{\text{分类正确的样本数}}{\text{总的样本数目}} \tag{17}$$

精准率 P_{recision} 表示预测为正例的样本中真实为正例的比例,T_{P} 表示预测为正例且标签也是正例的样本个数,F_{P} 表示预测为正例但标签为反例的样本个数。

$$P_{\text{recision}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{P}}} \tag{18}$$

召回率 R_{ecall} 表示正例样本中有多少比例的样本被正确预测为正例。F_{N} 为预测为反例但标签为正例的样本个数。

$$R_{\text{ecall}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{N}}} \tag{19}$$

F_1 值是对精准率和召回率的综合兼顾。

$$F_1 = \frac{2 \times P_{\text{recision}} \times R_{\text{ecall}}}{P_{\text{recision}} + R_{\text{ecall}}} \tag{20}$$

对于多分类任务,模型总的 Precision、Recall、F1 值是所有类别 Precision、Recall、F1 值的算术平均。

对于非平衡数据分类,G_mean 一种更有效的评价指标,其计算公式为:

$$G_{\text{\_mean}} = \sqrt{P_{\text{recision}} \times R_{\text{ecall}}} \tag{21}$$

2.4　实验环境和参数配置

本文实验使用编程语言为 Python 3.9,深度学习框架为 PyTorch2.0.1,操作系统为 Windows 10,处理器为 Intel i9,主频 2.80 GHz,内存 32 GB,GPU 为 NVIDIA

3080Ti。

神经网络模型训练参数如表 4 所示。

表 4 神经网络模型训练参数

参数名称	参数值	参数说明
Epoch	10	训练集通过神经网络训练的次数
Learning rate	0.001	学习率
Batch size	32	每批次同时处理的样本数量

2.5 实验结果分析

表5 和表6 是9 个模型的 Accuracy、Precision、Recall、F1 值和 G_mean 值,图 4 是 6 个模型的混淆矩阵。我们可以看到 RNN 模型的各项评价指标都是最低的,其分类准确率仅为 50.58%。这是由于 RNN 模型几乎将所有的数据都预测为了样本数最多的 0 类,无法预测出其他类别的样本。虽然 TextCNN 取得 85% 以上的较好分类准确率,但其 G_mean 值只有 0.387 4,这是由于 TextCNN 模型将所有的样本都预测成了数量前 5 的类别,无法预测出剩下的 6 个少样本类别。GRU 模型在多数类和少样本类表上的表现均优于 RNN 和 TextCNN,但其在少样本类别 3、4 和 5 上表现仍比较糟糕,准确率都没有超过 50%。特别是类别 3,全部预测错误。上述实验结果表明传统深度学习模型需要大量的训练样本才能较好获取文本语义信息,在训练样本较少时无法准确提取文本的语义特征,此时分类性能较差。

表 5 不同模型的 Accuracy、Precision 和 Recall 值

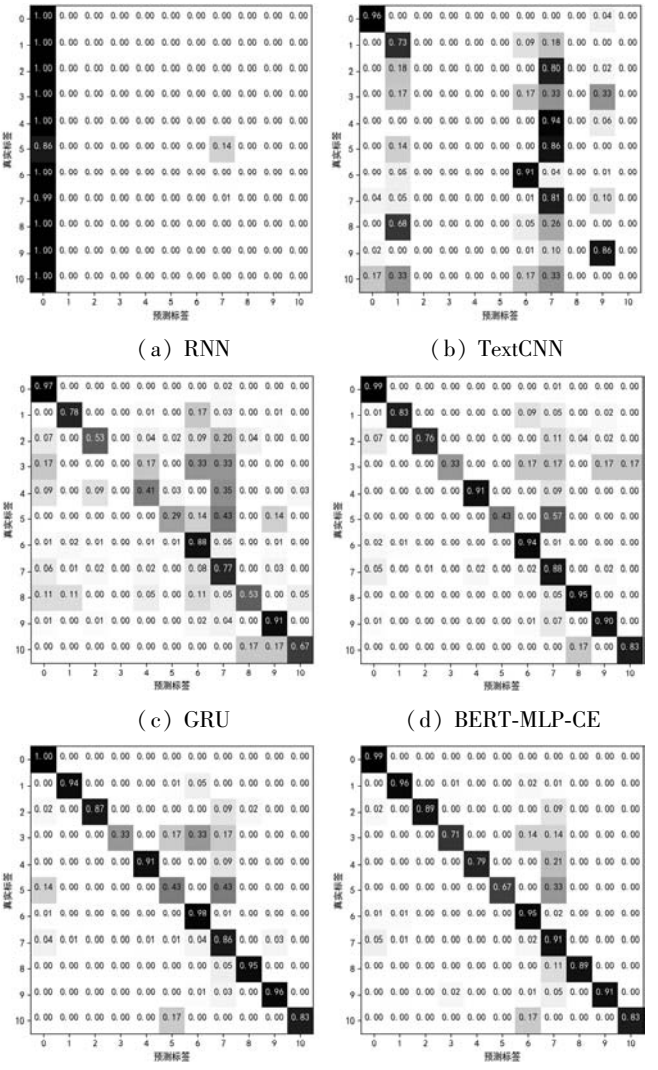
模型	Accuracy/%	Precision	Recall
RNN	50.58	0.106 6	0.091 5
TextCNN	85.77	0.347 1	0.387 4
GRU	88.69	0.625 6	0.612 2
BERT-MLP-CE	94.07	0.866 7	0.795 2
BERT-MLP-FL	93.46	0.787 4	0.777 7
BERT-MLP-WCE	94.02	0.808 1	0.849 5
BERT-BiGRU-CE	**95.92**	0.874 2	0.822 8
BERT-BiGRU-FL	95.64	0.818 3	0.821 1
BERT-BiGRU-WCELoss	95.83	**0.889 6**	**0.865 8**

表 6 不同模型的 F1 和 G_mean 值

模型	F1	G_mean
RNN	0.062 3	0.098 8
TextCNN	0.364 4	0.366 7
GRU	0.616 1	0.618 9
BERT-MLP-CE	0.813 5	0.830 2

续表 6

模型	F1	G_mean
BERT-MLP-FL	0.774 9	0.782 5
BERT-MLP-WCE	0.823 9	0.828 5
BERT-BiGRU-CE	0.841 8	0.848 1
BERT-BiGRU-FL	0.818 3	0.821 1
BERT-BiGRU-WCELoss	**0.871 9**	**0.877 6**



(a) RNN (b) TextCNN (c) GRU (d) BERT-MLP-CE (e) BERT-BiGRU-CE (f) BERT-BiGRU-WCE

图 4 不同模型的混淆矩阵

借助于预训练 BERT 所学习的海量预训练文本信息,基于 BERT 的警情分类模型分类性能比传统深度学习模型提升明显,尤其是少样本类别上。相比 GRU 模型,使用最简单的 BERT-MLP-CE 模型,分类准确率大幅提升了 5.33%,精准率、召回率和 F1 值均有 20% 以上的提升。3 个 BERT-BiGRU 组合模型相比 BERT-MLP 模型在分类准确率、精准率、召回率、F1 值、G_mean 值 5 个指标数又获得进一步提升,其中 BERT-BiGRU-CE 取得了最高 95.92% 的分类准确率。实验结果表明 BiGRU 相比多层感知机能够更有效地综合 BERT

模型提取出的文本语义特征,从而提高模型的分类性能。BERT-BiGRU-WCELoss 模型通过使用自适应优化权重技术,相比 BERT-BiGRU-CE 模型在牺牲 0.09% 微小准确率的情况下,Precision 提升了 1.54%,Recall 提升了 4.3%,F1 值提升了 3.58%,G_mean 值提升了 3.48%,综合性能更出色。

模型在训练过程中的训练损失也是衡量模型性能的一个重要标准。图 5 是本文所使用的部分警情分类模型在训练过程中的平均训练损失。我们可以看到,TextCNN 由于在少样本类别上的糟糕表现,其训练损失在所有训练轮次中都是最大的。BERT-BiGRU-WCELoss 模型的训练损失在所有训练轮次中都是最低的,并且随着训练轮次的增加仍有缓慢减小的趋势。GRU 模型的损失变化曲线与其他 5 个模型均不同,其在前 3 个训练轮次中训练损失迅速减小,从第 4 个训练轮次之后训练损失仍以较快速度下降。在第 7 个训练轮次之后,GRU 模型的训练损失已小于 BERT-MLP-CE 模型,但其在测试集上的准确率要比前者低了 5.33%。

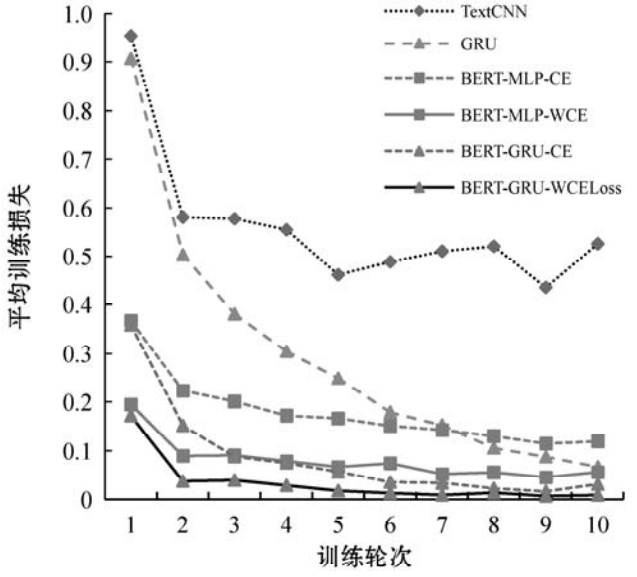


图 5　不同模型的训练损失变化曲线

3　结　语

针对 110 报警类文本数据存在着文本长度极短、类别分布严重不均衡的问题,我们首先使用预训练模型提升模型对文本的语义提取能力,接着通过优化自适应权重的方法给少样本类别赋予更大损失权重,从而提升模型在少样本类别上的表现。本文提出的 BERT-BiGRU-WCELoss 模型,其中 BERT 是基于海量中文语料的预训练模型,它可以借助预训练时所学的语料知识较好地从短文本中抽取出有用信息;BiGRU 是双向门控循环神经网络,相比传统的多层神经网络,它更能够综合提炼 BERT 模型的输出;损失函数 WCELoss 可以根据样本类别的数量占比权重自适应地调整类别和样本的权重,为少类别样本和易错样本赋予更大的损失权重,从而提升模型在少类别样本上的表现,能够很好地分类严重不平衡数据集。本文提出的 BERT-BiGRU-WCELoss 模型在某市某年某一自然月 110 报警类文本数据集上取得了较好的分类结果。同时实验结果表明基于预训练 BERT 模型的分类模型在非均衡短文本数据集上比传统深度学习表现更好。下一步的研究重点是让模型在有限数据集中学习到更多的样本特征,从而进一步提升模型对不平衡数据的分类准确率。

参　考　文　献

[1] 王孟轩,张胜,王月,等. 改进的 CRNN 模型在警情文本分类中的研究与应用[J]. 应用科学学报,2020,38(3):388-400.

[2] 王云,李丛. 基于自适应引力搜索的支持向量机在公安巡防警情分类中的应用研究[J]. 计算机应用与软件,2020,37(7):56-60.

[3] 章磊,王攀,何芬. 自然语言处理在警情智能分析中的应用[J]. 警察技术,2021(5):39-43.

[4] 张齐,李雪琛. 基于机器学习的多标签盗窃犯罪类型识别方法研究[J]. 中国人民公安大学学报(自然科学版),2023,29(1):88-93.

[5] 李昀轩,李萌,陆建,等. 基于多任务迁移学习的交通警情信息自动处理方法[J]. 中国公路学报,2022,35(9):1-12.

[6] 殷小科,王威,王婕,等. 分层文本分类在警情数据中的应用[J]. 现代计算机,2021,27(23):86-90.

[7] 李卫红,童昊昕. 针对非平衡警情数据改进的 K-Means-Boosting-BP 模型[J]. 中国图象图形学报,2017,22(9):1314-1324.

[8] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB]. arXiv:1810.04805,2018.

[9] 段丹丹,唐加山,温勇,等. 基于 BERT 模型的中文短文本分类算法[J]. 计算机工程,2021,47(1):79-86.

[10] 赵柳杨,杜彦辉,王腾飞. 基于改进 BERT 模型的时政微博评论情感分类[J]. 中国人民公安大学学报(自然科学版),2021,27(1):63-69.

[11] Chen X, Cong P, Lv S. A long-text classification method of Chinese news based on BERT and CNN[J]. IEEE Access, 2022,10:34046-34057.

[12] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV),2017.

(下转第 229 页)

频信号进行动态聚类,从每个聚类中选出一个靠近聚类中心的序列作为关键片段;其次使用 STFT 变换将所选序列转化为一个二维谱图,输入到深度学习模型中;然后使用 ResNet 和 Bi-LSTM 网络的组合模型来学习谱图中的时空信息;最后利用 Softmax 分类器对输出特征进行分类。该方法通过特征聚类和深度学习提高情感识别的准确度,降低整体模型的处理时间。实验结果表明,相较于其他的识别方法,本文方法的性能最优。

参 考 文 献

[1] 王忠民,刘戈,宋辉. 基于多核学习特征融合的语音情感识别方法[J]. 计算机工程,2019,45(8):248 - 254.

[2] Zamil A, Hasan S, Baki S, et al. Emotion detection from speech signals using voting mechanism on classified frames [C]//2019 International Conference on Robotics, Electrical and Signal Processing Techniques,2019:281 - 285.

[3] Badshah A, Rahim N, Ullah N, et al. Deep features-based speech emotion recognition for smart affective services[J]. Multimedia Tools and Applications, 2019, 78(5): 5571 - 5589.

[4] Liu Z, Wu M, Cao W, et al. Speech emotion recognition based on feature selection and extreme learning machine decision tree[J]. Neurocomputing,2018,273:271 - 280.

[5] Hao M, Yan T, Fei Y, et al. Speech emotion recognition from 3D log-mel spectrograms with deep learning network [J]. IEEE Access,2019,7:125868 - 125881.

[6] 高帆,张雪英,黄丽霞,等. 基于 DBM-LSTM 的多特征语音情感识别[J]. 计算机工程与设计,2020,41(2):465 - 470.

[7] Zhao J, Mao X, Chen L, et al. Learning deep features to recognise speech emotion using merged deep CNN[J]. IET Signal Processing,2018,12(6):713 - 721.

[8] Ma X, Wu Z, Jia J, et al. Emotion recognition from variable-length speech segments using deep learning on spectrograms[C]//2018 Conference of the International Speech Communication Association,2018:3683 - 3687.

[9] Chen M, He X, Yang J, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters,2018,25(10): 1440 - 1444.

[10] Zhao J, Mao X, Chen L, et al. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical Signal Processing and Control,2019,47:312 - 323.

[11] Hajarolasvadi N, Demirel H. 3D CNN-based speech emotion recognition using K-means clustering and spectrograms[J]. Entropy,2019,21(5):479.

[12] Wu L, Zhang S, Jian M, et al. Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks[J]. IEEE Access,2019,7:77268 - 77276.

[13] Xu Z, Sun K, Mao J. Research on ResNet101 network chemical reagent label image classification based on transfer learning[C]//2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology,2020:354 - 358.

[14] Guo L, Wang L, Dang J, et al. Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine[J]. IEEE Access,2019,7:75798 - 75809.

[15] Jiang P, Fu H, Tao H, et al. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition[J]. IEEE Access,2019,7:90368 - 90377.

[16] Zhao Z, Bao Z, Zhao Y, et al. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition[J]. IEEE Access,2019,7:97515 - 97525.

(上接第 223 页)

[13] Xu H. Hierarchical cost-sensitive techniques for class imbalance learning[C]//2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD),2021.

[14] Pasupa K, Vatathanavaro S, Tungjitnob S. Convolutional neural networks based focal loss for class imbalance problem: A case study of canine red blood cells morphology classification[J]. Journal of Ambient Intelligence and Humanized Computing,2023,14:15259 - 15275.

[15] Miao L, Liu M, Zhang D. Cost-sensitive feature selection with application in software defect prediction[C]//21st International Conference on Pattern Recognition (ICPR2012), 2012.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB]. arXiv:1706.03762,2017.

[17] Cho K, Merrienboer B V, Gulcehre C, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation[EB]. arXiv:1406.1078,2014.

[18] 林伟. 基于 BiGRU-CNN 的网络舆情情感识别模型[J]. 中国人民公安大学学报(自然科学版),2023,29(2):61 - 66.

[19] 梁越,刘晓峰,李权树,等. 面向司法文本的不均衡小样本数据分类方法[J]. 计算机应用,2022,42(S2):118 - 122.

[20] 宋明,刘彦隆. Bert 在微博短文本情感分类中的应用与优化[J]. 小型微型计算机系统,2021,42(4):714 - 718.

[21] 王雯慧,靳大尉. 基于改进 Focal Loss 和 EDA 技术的 UT 分类算法[J]. 计算机仿真,2023,40(4):346 - 349.