文章编号：2096-1618(2024)02-0170-08

基于深度学习的中文临床实验筛选标准的分类

刘子琦， 胡建成， 牟谷芳
(成都信息工程大学应用数学学院，四川 成都 610255)

摘要：针对大多数临床实验筛选标准的分类研究都集中在英文资格标准上，研究适合中文资格标准的分类模型，利用第五届中国健康信息处理会议开发的中文临床实验短文本数据集，结合神经网络和预训练语言模型对分类任务进行构建和微调，比较分析 Word2vec-BiLSTM 模型、CNN 模型、RNN 模型、预训练语言模型在此应用上的效果差异，并通过实验得到预训练模型 ERNIE 的分类效果优于其他模型。针对数据不平衡这一特征，对数量较少的类别语料进行数据增强后可有效提升模型的性能和效果，结果显示 ERNIE 模型的宏观平均 *F*1 值和微观平均 *F*1 值分别可达到 0.8281 和 0.8537。

关 键 词：临床实验；医学短文本分类；深度学习；预训练模型
中图分类号：TP391.1 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2024.02.007

0 引言

近年来随着大数据的发展，从科学研究到与生活息息相关的各个不同领域，大数据都在创造爆发式增长的数据量，如何高效整理和分析数据是一个巨大的挑战，而挖掘和提取数据则是利用数据的基础，其中文本分类是数据提取中的重要组成部分。文本分类是按照一定的分类规则对文本进行自动划分类别的过程，在很多领域有着非常广泛的应用场景，特别是在搜索、推荐和对话等场景随处可见[1]。

医学数据是一个巨大的知识库，如何将医学文本中的不规则数据转化成结构化数据不仅具有巨大的商业价值，也具有很高的社会价值[2]。临床实验是指通过人体受试者进行的科学研究，筛选标准是临床实验负责人拟定的鉴定受试者是否满足某项临床实验的主要指标，一般为无规则的自由文本形式[3]。临床实验的招募工作一般是通过人工筛选相应指标完成，而这种针对人体展开系统性药物研究的筛选程序严谨且繁重，自动比较和筛选受试者的病历记录和相应指标具有很大的应用前景和临床医学价值。目前，这类研究大多应用在英文临床实验筛选标准及英文电子健康记录数据[4]。尽管针对中文电子健康数据的研究也取得了很多进展，但与中文临床实验筛选标准的自然语言处理的研究则很少。本文则针对中文临床实验筛选标准进行分类。

早期的文本分类模型主要是通过基于规则匹配模板，Regev 等[5]利用动名词短语设计一些规则后利用模型匹配的方式得到分类结果。基于规则的分类模型在简单领域内能取得较好的效果，却特别依赖大量的关联规则，计算速度慢、通用性较差。

Luo 等[6-7]下载了 27278 条来自美国临床实验注册中心网站中真实世界英文临床实验筛选标准语句，使用一体化医学语言系统语义类型构建句子特征，通过由底向上的层次聚类算法和人工归纳总结，设计了不同的机器学习分类器且取得不错的效果。Marafino 等[8]利用支持向量机来分类临床诊断文本，采用向量化、正规化和词语切分等处理特征，构建的支持向量机分类器能够识别临床记录中的一系列诊断，用于风险调整。Yi 等[9]引入基于医学先验知识的马尔可夫模型来提高对医学文献分类的准确率。基于传统机器学习方法的分类器一定程度上实现了分类的自动化，但构建分类器的前提往往需要繁琐的人工特征工程。

王培等[10]基于双向语言 BERT 模型实现中医深层全局语义的特征表示，并进行中医临床文本的分类研究。李启行等[11]提出一种基于注意力机制的双层次文本分类模型用于对生物医学文本进行有效分类。钟桂凤等[12]为提高文本分类的准确性和运行效率，提出一种 Word2vec 文本表征和改进注意力机制 AlexNet-2 的文本分类方法。这些神经网络方法的核心是通过嵌入模型将文本映射到低维连续的特征向量，再通过神经网络分类器进行分类，因此不需要复杂的特征工程。

本文通过预训练语言模型和深度学习模型，针对中文临床实验筛选标准的分类任务进行微调，对比分

收稿日期：2022-12-28

析不同网络对中文短文本分类的结果。实验显示,不同神经网络模型差别较大,主要原因在于每个模型的语义特征构造不同、参数不同、对数据的自适应性不同。预训练模型优于其他模型的主要原因在于预训练模型是在海量数据上进行训练并保留下来的通用语言表示,拥有强大的语言表征和特征提取能力。

1　研究理论与技术

1.1　RNN 模型

循环神经网络(RNN)通过使用带自反馈的神经元,能够对任意时间步长的时序数据进行处理。给定一个输入序列 $X_{1:T}=(X_1,X_2,\cdots,X_t,\cdots,X_T)$,循环神经网络通过下面公式更新带反馈边的隐藏层的活性值 h_t:

$$h_t=f(h_{t-1},x_t)$$

$h_0=0$,$f(\cdot)$为一个非线性函数,可以是一个前馈神经网络。

图 1 给出了循环神经网络的实例,其中延迟器是一个虚拟单位,记录神经元最近一次(或几次)的活性值。

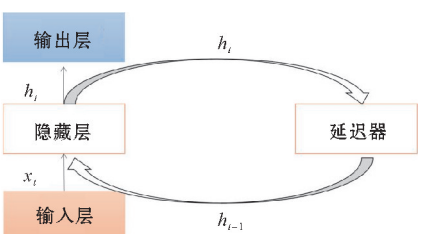


图 1　循环神经网络示例图

1.2　textCNN

textCNN[13]是一种具有少量超参数调优和静态向量的简单卷积神经网络(CNN),CNN 的优势在于无需手动捕捉图像的局部特征,可以自动地对若干单词组成的滑动窗口进行组合和筛选,以获得不同抽象层次的语义信息。CNN 主要用于图像处理的二维数据,其卷积核的宽度和高度一样,卷积核滑动的方向先从左到右、后从上到下。textCNN 一般用作文本分类的一维数据,将词作为文本的最小颗粒,使用的卷积核的宽度与词嵌入向量的维度是一样的,卷积核只在一个方向上滑动。textCNN 在池化层使用 Max-Pooling 对每个特征向量池化成一个值表示当前特征,可以极大降低参数的数目加快收敛,当所有池化层的最终向量拼接送入 Softmax 层就可以得到预测结果。

1.3　LSTM

长短期记忆网络(long short-term memory network,LSTM)是一种为解决梯度爆炸问题而可以有效缓解长程依赖问题的循环神经网络[14]。LSTM 的特点是引入一个新的内部状态 $c\in R^T$ 和门控机制。不同时刻的内部状态以近似线性的方式进行传递,从而缓解梯度消失或梯度爆炸问题,同时门控机制进行信息筛选,可以有效地增加记忆能力。

双向长短期记忆网络(bidirection-long short-term memory network,BiLSTM)[15]由 2 层 LSTM 网络组成,可以共享权重网络,输入信息相同而传递的方向不同。LSTM 网络引入门控机制来控制信息传递的路径,包含输入门 i_t、遗忘门和输出门,分别表示控制上个时刻的内部状态需要遗忘多少信息、控制当前时刻的候选状态需要保存多少信息,以及需要输出给外部的状态信息。这 3 种门的取值区间为(0,1),用一定的比例表示是否运行信息通过。3 种门的计算方式:

$$i_t=\sigma(w_ix_t+U_ih_{t-1}+b_i) \quad (1)$$

$$f_t=\sigma(w_fx_t+U_fh_{t-1}+b_f) \quad (2)$$

$$o_t=\sigma(w_ox_t+U_oh_{t-1}+b_o) \quad (3)$$

式中,$\sigma(\cdot)$为 Logistic 函数,x_t 为当前时刻的输入,h_{t-1}为上一时刻的外部状态,W_*,U_*,b_*为可学习的网络参数。

栈式双向长短期记忆网络(stacked-bidirection-long short-term memory Network,Stacked-BiLSTM)[16]是将多个循环网络堆叠而成的循环神经网络,正向 LSTM 和反向 LSTM 交叠而成,每一层网络的输入是上一层网络的输出,$h_t^{(l)}$ 在 t 时刻第 l 层的隐状态表示为

$$h_t^{(l)}=f(\boldsymbol{U}^{(l)}h_{t-1}^{(l)}+\boldsymbol{W}^{(l)}h_t^{l-1}+\boldsymbol{b}^{(l)}) \quad (4)$$

式中 $\boldsymbol{U}^{(l)}$、$\boldsymbol{W}^{(l)}$、$\boldsymbol{b}^{(l)}$ 为权重矩阵和偏置向量,$h_t^{(0)}=x_t$,即初始的隐状态为初始输入。

1.4　Self-Attention 机制

注意力机制[17]是一种突出对象的某些重要特征的分配机制,即通过句子中词向量的权重表示反映词与句子的关联程度的重要性大小。Self-Attention 是 Transformer 最核心的内容,它不需要将所有信息输入到神经网络中,而是从输入信息内选择一些和任务相关的信息量,通过计算输入信息的注意力分布得到的加权平均值获取输出值。前者是通过打分函数来计算每个输入向量和查询向量的相关性,后者需要再对打分函数得到的结果经过 Softmax 激活函数点乘计算得到加权的输入向量的评分。Self-Attention 实则是查询

语句到键值对的映射,计算公式分别为

$$\text{Attention}(Q,K,V)=\text{soft max}(\frac{QK^T}{\sqrt{d_k}})V \quad (5)$$

$$Q=\boldsymbol{W}^Q\boldsymbol{X} \quad K=\boldsymbol{W}^k\boldsymbol{X} \quad V=\boldsymbol{W}^V\boldsymbol{X} \quad (6)$$

式中,$\boldsymbol{X}$ 表示输入数据的嵌入向量,Q 表示 Query 查询语句,K 表示 Key 索引,V 表示 Value 内容,$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$ 则是需要训练的参数矩阵。

1.5 预训练模型

BERT 是 2018 年 10 月由 Google AI 研究院提出的一个基于上下文的单词表示模型[18],基于标注语言模型并使用双向 Transformer 进行预训练,采用新的 Masked Language Model 和 Next Sentence Prediction 生成深度的双向语言表征。前者是在输入文本得到嵌入向量后,在句子中随机抽取 15% 的单词作为[MASK]的操作对象,[MASK]有随机替换、标记、保持不变 3 种遮掩方式。这种独特的标记方式可以迫使模型快速根据上下文学习分布式语义,可以更好地预测句子中的单词。后者是让模型学习 2 个句子之间的关联性,通过一个句子判断另一个句子是否跟随其后,可以提高模型在服务问答、推理等语义理解任务上的准确率。二者作为预训练的两大任务,分别捕捉文本的词语和句子级别的特征表示,使得该模型拥有强大的语言表征和特征提取能力,成为 NLP 发展史上的里程碑式的模型。

而对中文文本而言,该建模对象主要聚焦在原始语言信号上,较少利用语义知识。为此,百度提出基于知识增强的 ERNIE 模型。该模型可以通过对词、实体等语义单元的掩码,使模型学习完整概念的语义表示[19]。相较于 BERT 学习原始语言信号,ERNIE 模型能直接对先验语义知识单元进行建模,增强了模型语义表示能力。

预训练模型相较于 RNN、textCNN 和 LSTM 模型可以做到并发执行,能在字、词、句多个层次提取关系特征。相较于 Word2vec 模型,能根据句子上下文获取词义的表征解释,以避免歧义的出现,进而更全面表示文本语义。

2 模型建立

本文探究的是对比不同算法在中文临床实验筛选标准下分类的效果和性能,特别是神经网络模型和语言模型在文本分类上的异同点。

中文临床实验筛选标准虽然文本较短,但所包含的信息丰富且每句话中都有重要的语义和局部特征,它所代表的文本向量可以很大程度上提高模型的分类效果。因此,本文神经网络模型构建的思路主要是针对待分类文本的语义向量的构造,模型构造见图 2 所示。

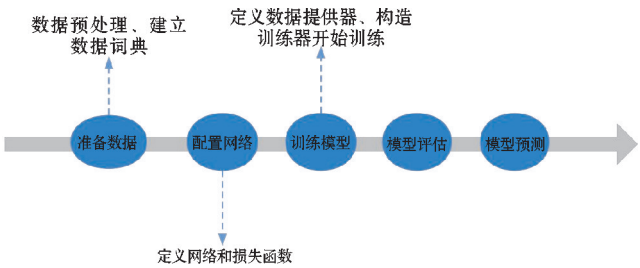


图 2 神经网络模型构造示意图

为更好地解决卷积模型和序列模型中文本长距离的依赖关系,可利用大量未标注的文本数据学习神经模型的参数,故预训练学习通用语言表示有更好的泛化性能和更快的收敛速度。因此,文本预训练模型的构建思路主要是利用模型对输入自我增强以及参数调整,让模型学会高效的自动分类。

2.1 基于 RNN 的分类模型

本文建立基于 Stacked-BiLSTM 的深层循环神经网络模型应用于文本分类,即使用正反方向的 LSTM 循环神经网络。RNN 模型构造如图 3 所示,由训练数据和数据词典得到的 256 维词嵌入向量作为低层 Stacked-BiLSTM 的输入向量,通过对输入向量的正逆序双重处理,后一层的 LSTM 使用之前所有层的信息作为输入,对最后一层 LSTM 通过最大池化得到该输入文本的特定向量表示,抽象出的高级特征映射到和分类类别数同样大小的向量上,最后通过 Softmax 输出类别的预测概率。

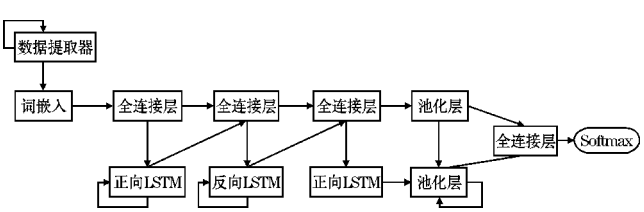


图 3 RNN 分类模型

2.2 基于 Word2vec-BiLSTM 的中文临床实验筛选标准模型

本文采用 Word2vec 模型完成短文本分布式词向量的构造,使用 2 层 BiLSTM 获取上下文双向语义信息,再利用 Attention 机制加强对文本中关键特征的关注以学习不同词的权重分布,最后通过 Softmax 得到分类结果。

词嵌入指的是将单个词在预定义的向量空间中表示为实数向量,即每个单词映射成一个向量,Word2vec 是一种数据中的词转换成蕴含语义信息的低维稠密向量的技术。本文利用 Word2vec 的 Skip-Gram 模型通过当前词预测上下文,得到包含相似语义的嵌入向量,序列长度设定为分词后文本的平均长度 34,向量维度设定为 300。图 4 中,预处理后得到的文本是“研究”“前”“30”“天内”“接受”“临床”“方案”“治疗”,通过 Word2vec 拟合得到的每个词的 id 是嵌入矩阵的索引,如“研究”代表的 id 是 10,这个词的词向量大小为序列长度×向量,即 34×300。

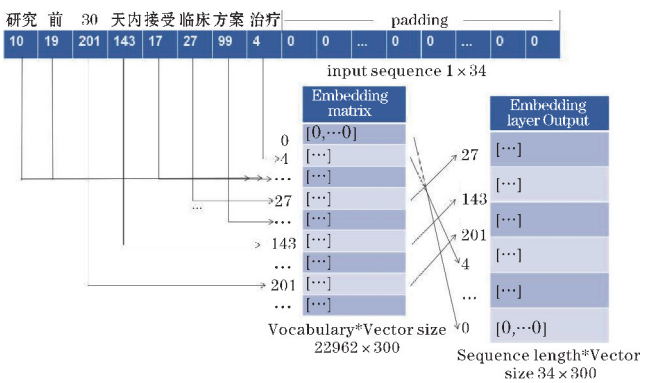


图 4　嵌入向量示意图

将嵌入向量接入 2 层 BiLSTM 实现长距离语义捕捉,提取双向语义依赖关系,再利用 Attention 机制加强对重点词的关注并调整其向量权重,最后通过 2 个全连接层得到每个文本的预测结果。由于实验中医学临床实验筛选标准短文本包含的词长度不一,预处理后分词得到的文本数据最小包含一个词组,最长包含 101 个词组,平均词长度达到 7.5 个,因此要通过分析文本中前后语义的联系和重要特征加强分类效果。从整体模型架构来看,主要是在 BiLSTM 层捕捉双向语义信息后接入 Attention 层,Attention 层先计算 BiLSTM 输出中每个位置词语的向量权重,然后将所有位置词语的向量进行加权和当作文本的表示向量,最后进行 Softmax 分类,输出结果,结构见图 5 所示。

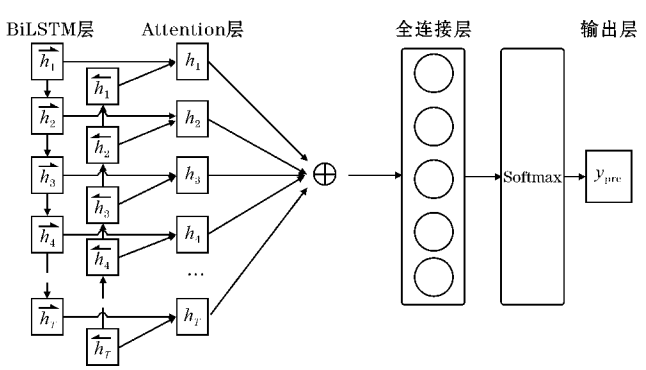


图 5　BiLSTM-Attention 结构示意图

2.3　基于 textCNN 的中文临床试验筛选标准模型

建立基于卷积神经网络的 textCNN 模型应用于文本分类。textCNN 网络结构简单,通过分词构建词向量后设计卷积层,使用卷积提取不同的 n-gram 特征,通过最大池化层后得到卷积后的若干一维向量,拼接的最大值为本层的输出值,该输出值通过全连接的方式连接一个 Softmax 层。

图 6 所示的卷积示意图中,原文本分词后得到的词是“研究”“前”“30”“天内”“接受”“临床”“方案”“治疗”,假设词向量大小为 8×5,本文中的卷积核大小为 3、4、5,每个过滤器的个数是 100,即每个 kerner_size 有 100 个输出通道。通过 2 组 3 种(3×5,4×5,5×5)卷积核提取特征后得到 3 个维度分别为 6、5、4 的卷积向量,再通过激活函数生成 3 个卷积结果,最后通过最大池化层得到最大值作为输出结果。

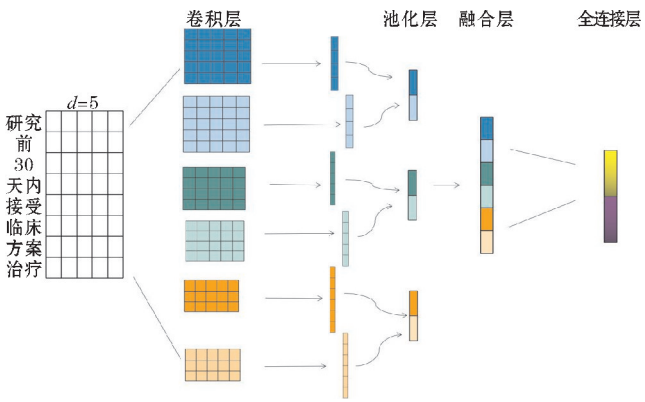


图 6　卷积示意图

2.4　预训练模型

本文使用的 2 个预训练语言模型均基于 BERT 模型方法,BERT 模型的输入包括词向量、段落向量和位置向量。将输入序列经过上述 3 层处理后相加就得到输入数据的最终表示向量,输出则是文本中各个字符融合全文语义信息后的向量表示,如图 7 所示。

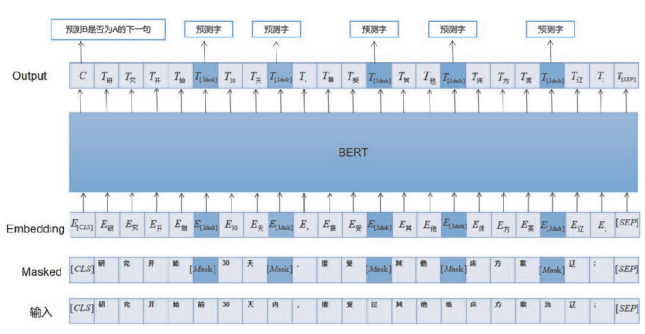


图 7　BERT 分类示意图

本文实验使用的 2 个预训练语言模型分别为 Google 发布的中文版 BERT-base-Chinese[18] 和百度 2019 年发布的的 ERNIE 1.0[19]。二者的主要区别在于初始 BERT 模型以字符为单位进行掩码,而 ERNIE 模型还可以对短语和实体为单位进行掩码。显而易见,中文词富含的语义信息比单个字符更丰富,且可以通过前后语义学到更多的知识。

3 实验结果与分析

3.1 数据来源

第五届中国健康信息处理会议(CHIP2019)[20]共享评测任务 3 开放了中文临床试验筛选短文本数据集,包括 44 种语义类别定义和 38341 条筛选标准,本文选择其中的训练集(22962 条数据)和验证集(7682 条数据)分别作为实验训练集和测试集。训练集具体信息如表 1 所示共有 3 列信息,其中 id 表示样本编号,Label 表示所属类别,text 表示相应的短文本。训练集标签与文本关系分布如图 8 所示。由图 8 可以看出数据集不平衡,14 类标签所支持的语句不及 50 句,Disease 这一标签的语料多达 5000 多条,而 Ethical Audit 这一标签只有 12 条语料,故数据增强理论上可以提升模型效果。

表 1 训练集示例

id	Label	text
s1	Therapy or Surgery	研究开始前30 天内,接受过其他临床方案治疗;
s2	Sign	(9)严重的听力或者视力损害
s3	Addictive Behavior	(10)现在或曾经滥用药物或酗酒,或者每天饮用相当于 30 毫升酒精的酒精饮料。

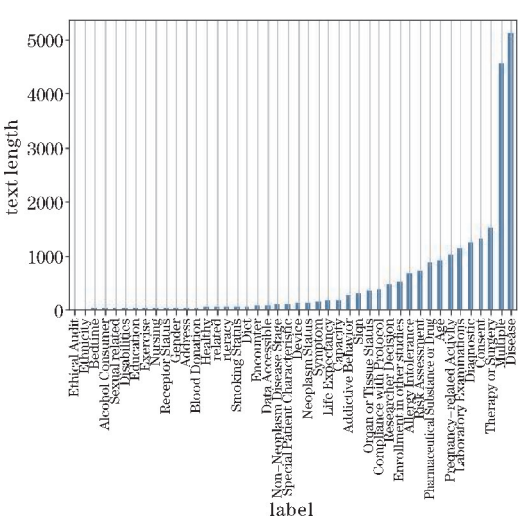


图 8 文本长度分布

通过数据增强,不仅可以缓解样本不平衡问题,还可以通过增加训练数据提升模型的鲁棒性和泛化能力。Wei 等[21]提出用于文本分类任务的简单数据增强技术,即 EDA 技术,包括 4 种方法:同义词替换、随机插入、随机交换、随机删除,可以提高卷积和神经网络的性能。具体实例如表 2 所示,增强方法有随机增加标点符号、随机删除词、随机交换词等。

表 2 增强数据实例

增强前	增强后
2)本地户籍或常住人口;	2)本地户籍
	2);户籍或常住人口本地
	2)本地异地或城镇居民
	2);户籍或常住人口本地
	2)本地农业户口或人口统计

3.2 实验环境及模型参数

实验在 Windows10 操作系统下展开,16 GB 的 NVIDIA Tesla T4 和 Intel(R) Xeon(R) CPU @ 2.00 GHz的处理器,深度学习框架为 1.12.0+cu113 的 torch 版本,编程语言为 Python,实验环境达标且计算能力满足实验要求。BERT 和基于 BERT 改编的 ERNIE 的网络结构相同,都有着 12 层、12 头、768 隐藏单元大小,共包含 1.1 亿个可训练参数。

3.3 评价指标

实验的评价指标包括宏观准确率(macro precision),宏观召回率(macro recall),Average F_1 值,3 个指标的区间为[0,1],结果越接近 1 说明模型效果越好,计算公式如下:

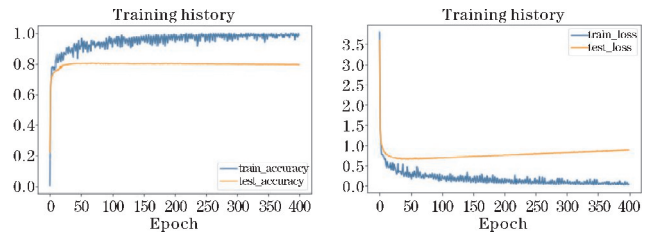
$$准确率\ P_i=\frac{正确预测为类别为\ i\ 的样本个数}{预测为\ i_i\ 类的样本个数}\times 100\% \tag{7}$$

$$召回率\ R_i=\frac{正确预测为类别为\ i\ 的样本个数}{真实为\ i\ 类的样本个数}\times 100\% \tag{8}$$

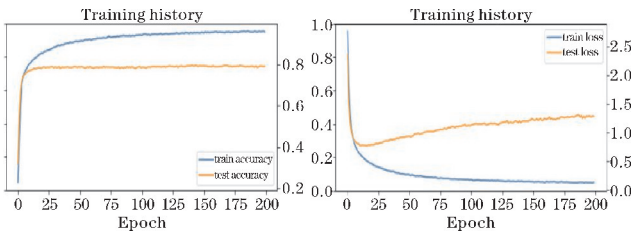
$$平均值\ F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2\times P_i\times R_i}{P_i+R_i} \tag{9}$$

3.4 实验结果与分析

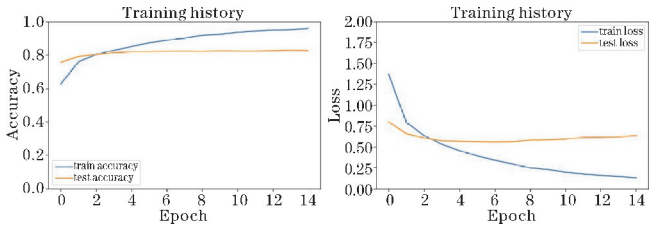
本文通过基于 RNN、Word2vec-BiLSTM、textCNN、BERT 和 ERNIE 的 5 种模型对临床实验筛选标准短文本进行分类实验,实验结果如图 9 和表 3 所示。



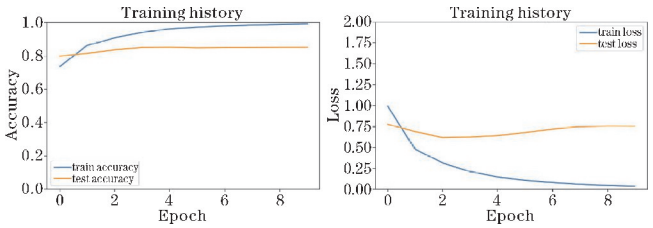
(a) RNN 结果图



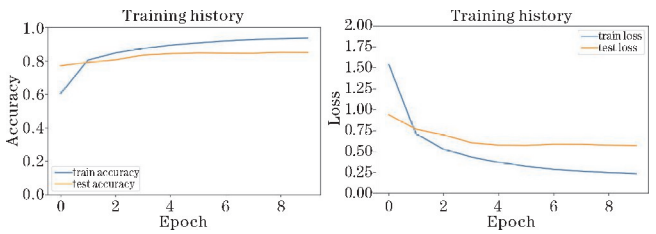
(b) Word2vec-Bilstm 结果图



(c) textCNN 结果图



(d) BERT 结果图



(e) ERNIE 结果图

图 9　实验结果图

表 3　不同模型在临床实验筛选标准分类下的效果对比

Module	增强前后	Macro-average			Micro-average		
		Precision	Recall	F_1-score	Precision	Recall	F_1-score
RNN	增强前	0.7804	0.7014	0.7194	0.6658	0.6658	0.6658
	增强后	0.7741	0.7229	0.7332	0.6647	0.6647	0.6647
Word2vec-BiLSTM	增强前	0.6394	0.6329	0.6075	0.7628	0.7628	0.7628
	增强后	0.6919	0.6179	0.6218	0.7799	0.7799	0.7799
textCNN	增强前	0.8834	0.6380	0.7043	0.8097	0.8097	0.8097
	增强后	0.8808	0.6936	0.7551	0.8172	0.8172	0.8172
BERT	增强前	0.8358	0.8144	0.8206	0.8519	0.8519	0.8519
	增强后	0.8395	0.8296	0.8296	0.8515	0.8515	0.8515
ERNIE	增强前	0.8428	0.8149	0.8246	0.8521	0.8521	0.8521
	增强后	0.8356	0.8274	0.8281	0.8537	0.8537	0.8537

由图 9 和表 3 可知各个模型的准确率和相应的 loss 变化率,可以看出各个模型分类的准确率均不低于 0.6,在测试集中其准确率随着 Epoch 的增大而增大,其 loss 值随着 Epoch 的增大而减小,说明网络尚在学习直至趋于稳定。可以看出,神经网络模型中的 textCNN 模型分类效果最好,其宏观指标的 Precision 值、Recall 值、F_1-score 值分别达到了 0.8834、0.6308 和 0.7043。textCNN 中的非静态词向量特征以预训练中的 Word2vec 初始化词向量,在训练过程中加快收敛的方式构成而成,训练速度快,效果好。预训练模型中的 ERNIE 模型要好于 BERT 模型,其宏观指标的 Precision 值、Recall 值、F_1-score 值分别达到了 0.8428、0.8149、0.8281,主要是由于 ERNIE 模型直接对先验语义知识单元进行建模,增强了模型语义表示能力。实际上,基于预训练语言模型效果好于神经网络模型从理论中也可以得到此结论。因为预训练模型本就是从大规模文本语料库集中训练得到的深度模型,通过基于 Transformer 框架对本地词向量进行训练并微调,使模型有自适应更强的语义信息提取能力。同时可以看出,增强后的数据的确提升了模型分类的性能,tex-

tCNN 增强后宏观指标的 Precision 值、Recall 值、F_1-score 值分别达到了0.8808、0.6936 和0.7551，因为 EDA 方法可以通过同义替换、随机插入和删除等操作加入新词汇。不过，由表 3 可知，EDA 的提升效果不是很高，原因在于本试验数据集极不平衡，对类别较少的数据增强的较少，只扩充了 50 条。因为如果扩充太多数据，可能会改变句子的原意，同时仍保留的是原始类别标签，从而会产生标签错误的句子。

4 结束语

鉴于通用领域语料训练的预训练语言模型应用到医学文本分类会产生由于数据结构分布不同而使分类效果有一定的上升空间，故本文将基于医学大数据的预训练语言模型引入到中文临床试验筛选标准分类研究中，探索了提高医学文本分类效果的方法。实验探究了神经网络和预训练模型应用于中文临床试验筛选标准分类的有效性，得出由百度改进的预训练 ERNIE 模型的分类性能在本实验效果最优。神经网络模型由于数据集的不平衡导致类别较少的语料数据分类结果较差，在今后的改进工作中可以优先扩充数据集。比如，使用数据增强技术和迁移学习，进一步得到语料之间的语义关系以获得更好的结果。

参考文献：

[1] 左圆圆，王媛媛，蒋珊珊，等. 数据可视化分析综述[J]. 微计算信息，2019，(1)：82-83.

[2] 王勇，李帅. 自然语言处理在医学文本挖掘中的应用[J]. 电子技术与软件工程，2019，(7)：190-190.

[3] 宗辉，张泽宇，杨金璇，等. 基于人工智能的中文临床试验筛选标准文本分类研究[J]. 生物医学工程杂志，2021，38(1)：105-121.

[4] Hao T, Rusanov A, Boland MR, et al. Clustering clinical trials with similar eligibility criteria features[J]. Journal of biomedical informatics, 2014, 52(C): 112-120.

[5] Regev Y, Finkelstein M, Feldman R, et al. Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002 (task 1) [J]. SIGKDD Explor. Newsl, 2002, 4(2): 90-92.

[6] Luo Z, Stephen B, Weng C, et al. Semi-automatically inducing semantic classes of clinical research eligibility criteria using UMLS and hierarchical clustering[C]. AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2010, 2010: 487.

[7] Luo Z, Yetisgen-Yildiz M, Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering[J]. Journal of biomedical informatics, 2011, 44(6): 927-935.

[8] Marafino B J, Davies J M, Bardach N S, et al. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit[J]. Journal of the American Medical Informatics Association, 2014, 21(5): 871-875.

[9] Yi K, Beheshti J. A hidden Markov model-based text classification of medical documents[J]. Journal of Information Science, 2009, 35(1): 67-81.

[10] 王培，王亚文，卢苗苗. 基于 BERT 模型的中医文本分类研究[J]. 电脑知识与技术，2021，17(27)：13-20.

[11] 李启行，廖薇. 基于注意力机制的生物医学文本分类模型[J]. 中国医学物理学杂志，2022，39(4)：518-523.

[12] 钟桂凤，庞雄文，隋栋. 基于 Word2Vec 和改进注意力机制 AlexNet-2 的文本分类方法[J]. 计算机科学，2022，49(4)：288-293.

[13] Kim Y. Convolutional Neural Networks for Sentence Classification[J/OL]. https://ui. adsabs. harvard. edu/abs/2014arXiv1408. 5882K, 2014, 08, 01/2022, 12, 11, 2022.

[14] Tan M, Santos C, Xiang B, et al. Lstm-based deep learning models for non-factoid answer selection [J/OL]. https://ui. adsabs. harvard. edu/abs/2015arXiv151104108T, 2015, 11, 01/2022, 12, 11, 2022.

[15] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memorynetworks for relation classification[C]. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.

[16] Cai L, Zhou S, Yan X, et al. A stacked BiLSTM neural network based on coattention mechanism for question answering[J]. Computational intelligence and neuroscience, 2019, 2019: 9543490.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J/OL]. https://ui. adsabs. harvard. edu/abs/2017 arXiv170603762V, 2017,

01,01/2022,12,11,2022.

[18] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J/OL]. https://ui. adsabs. harvard. edu/abs/2018arXiv181004805D, 2018, 10,01/2022,12,11,2022.

[19] Sun Y, Wang S, Li Y, et al. ERNIE: Enhanced Representation through Knowledge Integration[J/OL]. https://ui. adsabs. harvard. edu/abs/2019arXiv190409223S, 2019, 05, 01/2022, 12, 11,2022.

[20] 第五届中国健康信息处理会议(CHIP 2019). 评测三:临床试 验筛选标准短文本分类[EB/OL]. http://www. cips-chip. org. cn:8088/evaluation,2020,11,06/2022,12,11,2022.

[21] Wei J,Zou K. Eda:Easy data augmentation techniques for boosting performance on text classification tasks[J/OL]. https://ui. adsabs. harvard. edu/abs/2019arXiv190111196W, 2019, 01, 01/2022,12,11,2022.

Classification of Screening Criteria for Chinese Clinical Trials based on Deep Learning

LIU Ziqi,　HU Jiancheng,　MOU Gufang

(College Applied Mathematics,Chengdu University of Information Technology,Chengdu 610225,China)

Abstract:Classification research for most clinical trial screening criteria focuses on English eligibility criteria. This paper compares the characteristic of classification models suitable for Chinese eligibility criteria, using the Chinese clinical trial short text dataset developed by the 5th China Health Information Processing Conference, combined with neural networks and pre-trained language models to construct classification tasks and fine-tuning, analyzed the differences in the effects of the Word2vec-BiLSTM model, CNN model, RNN model, and pre-trained language model in this application, and obtained through experiments that the classification effect of the pre-trained model ERNIE performsbetter. In view of the characteristic of data imbalance, the performance and effect of the model can be effectively improved after data enhancement of a small number of category corpora. The results show that the macro-average F_1 value and micro-average F_1 value of the ERNIE model can reach 0.8281 and 0.8537, respectively.

Keywords:clinical trials;medical short text classification;deep learning;pre-training model