DOI:10.3969/j.issn.1005-202X.2021.05.025

医学人工智能

基于双通道神经网络的疾病文本分类方法

袁野,廖薇
上海工程技术大学电子电气工程学院,上海 201620

【摘要】医疗疾病文本的准确分类对医疗信息化的发展具有重要的推进作用,本研究提出一种基于双通道学习的神经网络模型研究疾病文本分类方法。该模型分别使用卷积神经网络和双向长短期记忆网络对患者输入的疾病症状文本进行局部特征以及时序特征学习。此外,在双向长短期记忆网络上引入自注意力机制区分特征对类别预测的贡献值,增强模型的学习能力和可解释性。为使两个通道提取到的特征能够共同决定分类结果,该模型将两种特征进行拼接融合,最后利用softmax分类器得到最终的分类结果。实验结果表明,在疾病文本分类的性能方面,该模型相比其他分类模型具有较高的精确率、召回率和F_1值,分别可达90.61%、90.48%和90.51%。

【关键词】疾病文本分类;自注意力;卷积神经网络;双向长短期记忆网络
【中图分类号】TP391;R319 【文献标志码】A 【文章编号】1005-202X(2021)05-0655-06

Disease text classification model based on two-channel neural network

YUAN Ye, LIAO Wei
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Abstract: The accurate text classification for diseases plays an important role in promoting the development of medical informatization. A method for studying disease text classification based on a two-channel neural network model is proposed. In the model, convolutional neural network and bidirectional long short-term memory network are used to learn the local features and temporal features of disease symptom text input by patients. In addition, self-attention mechanism is introduced to bidirectional long short-term memory network for distinguishing the contribution value of the feature to classification prediction, which enhances the learning ability and interpretability of the model. The model combining two kinds of features for enabling the features extracted from the two channels to jointly determine classification results, and finally softmax classifier is used to obtain the final classification results. Experimental results show that the accuracy, recall rate and F_1 value of the proposed model are 90.61%, 90.48% and 90.51%, respectively, which were higher than those of other classification models.

Keywords: disease text classification; self-attention; convolutional neural network; bidirectional long short-term memory network

前 言

随着医疗信息化的发展,网络问诊逐渐成为了人们获取疾病症状、用药、治疗方案等信息的主要渠道。到医院就诊前或者没有必要去医院时,通过搜索引擎和网络问诊平台查找和咨询健康问题成为了大多数人的首要选择,互联网医疗已成为重要的补充医疗服务。目前的网络问诊形式大多需要患者首先选择咨询的科室,然后输入询问内容,最后平台匹配医生与患者进行交流。在此过程中,科室如何选择依赖于患者的知识和经验,对于不了解的疾病和症状,会出现科室选择错误的情况。因此,如何自动将健康咨询内容分门别类,自动分析疾病文本并给出对应的科室或者类别是目前研究的重点。

文本分类是自然语言处理中的一个经典问题[1],主要是为了解决句子、段落、文档等文本的标签分配问题。良好的文本分类模型有助于提高信息的提取效率,方便用户迅速检索目的信息,在问答系统[2]、情感分析[3]、新闻分类[4]、用户意图分类等领域都有广泛的应用。

【收稿日期】2020-11-28
【基金项目】国家自然科学基金(62001282);上海高校青年东方学者岗位计划资助项目(QD2017043)
【作者简介】袁野,硕士研究生,研究方向:自然语言处理,E-mail: 313065715@qq.com
【通信作者】廖薇,博士,副教授,研究方向:生物医疗与自然语言处理,E-mail: liaowei54@126.com

在疾病文本分类方面，传统的方法是基于机器学习的方法，通过人工筛选文本特征训练分类器。柏挺等[5]研究了朴素贝叶斯和贝叶斯网络在远程医疗文本分类任务上的性能，在特征词选择正确的情况下，增加其数量可以提高分类性能。文献[6]考虑了多种特征选择方法，考虑将问题转换方法与不同特征结合起来。Campillos等[7]提出设计一个通过词语或者句子匹配实现医疗健康文本分类的系统，对麻醉、心脏病和肺部疾病领域的文本能够进行有效分类。传统机器学习方法主要是特征工程，对于特征选择、规则制定需要大量的专业人员投入其中，且往往只适用于特定的疾病垂直领域，通用性和扩展性较差。

随着Mikolov等[8-9]引入了词向量模型Word2vec，深度学习在疾病文本分类任务中开始快速地发展。Word2vec使用神经网络将词语映射到维度较低的向量空间中，使得向量能够表达语义信息[10]。文献[11]使用卷积神经网络(Convolutional Neural Network, CNN)在句子粒度对临床文本进行分类，多层CNN可以学习到更多的语义特征。文献[12]对多个机器学习方法在中医病历分类中的应用进行了实验，并且提出一种结合深度学习的中医病历文本的表示方法。Chen等[13]提出一种基于注意力机制的双向长短期记忆网络(Bidirectional Long Short-Term Memory Network, BiLSTM)模型，实现根据文本内容进行门诊类别分类的功能。在现有的基于深度学习的疾病文本分类方法中，使用CNN网络缺乏对于文本序列特征的学习能力，使用LSTM网络只能对序列的单个方向进行特征提取，单一模型所考虑的特征存在一定的局限性，难以覆盖疾病文本所有重要的特征层面。

针对上述问题，为了探索疾病文本与类别的潜在关联特征，本文提出一种基于双通道神经网络的疾病文本分类模型(Text Classification Model for Disease, TCMD)，使用词向量进行文本表示，解决短文本的特征稀疏性；将词嵌入后的文本并行输入结合自注意力机制的双向长短期记忆网络和CNN中，进行不同层面的特征提取，增强了句子的整体序列特征以及局部词序特征。实验结果表明，TCMD比现有方法具有更好的分类性能。

1 TCMD模型构建

本文提出的TCMD模型结构图如图1所示，将疾病文本并行输入到两种学习网络中学习不同的特征，最后将两个通道的特征进行拼接融合，共同决定分类结果。TCMD模型主要由以下几部分组成：(1)词嵌入层将文本词向量表示；(2)CNN通道使用3个窗口大小不同的卷积核提取文本局部特征，通过最大池化获取其中最显著的特征；(3)BiLSTM_Attention通道提取疾病文本上下文语义信息，引入自注意力机制对重要词语赋予更高的权重，加强局部关注度；(4)拼接以上两部分的输出作为最终的疾病文本特征；(5)最后通过Softmax预测分类结果。

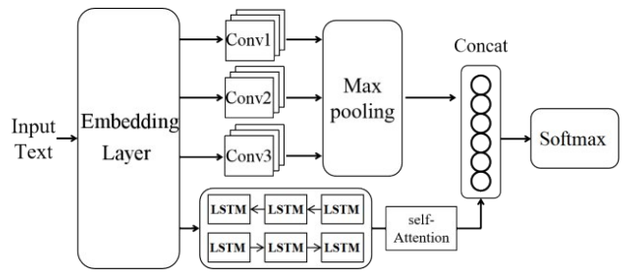

图1 TCMD模型

Fig.1 Text classification model for disease (TCMD)

1.1 文本预处理

对于原始文本需要进行预处理操作，过程如下：(1)首先对文本进行分词处理，去除对文本中出现的标点符号、停用词以及特殊字符等。(2)建立字典，将词语进行统计和编码。(3)统一文本的最大长度L，若文本词语数大于L则进行截断，若文本词语数小于L或者出现了未登录词语时使用0进行向量填充，使长度达到L。由于本文使用文本数据长度平均值为78，故本文L设为80。(4)最后进行文本向量化，将文本序列$\boldsymbol{s}=(w_1, w_2, \cdots, w_L)$中每一个词语$w_i$转化为预先使用Word2vec训练好的$N$维词向量$\boldsymbol{v}_i$，得到维度为$L\times N$的文本矩阵表示，如式(1)所示：

$$\boldsymbol{s}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_L),\ \boldsymbol{v}_i\in R^N \tag{1}$$

Word2vec是词嵌入的一种表示，从大量文本语料中学习词语语义信息，通过一个低维的嵌入空间使得语义上相似的单词在该空间内距离很近，拥有很好的计算特性，避免了使用词袋模型表达文本时的维度灾难和语义信息缺失的缺点。Word2vec提出了两个神经网络语言模型：连续词袋模型Continuous Bag of Words(CBOW)和Skip-gram模型。CBOW模型和Skip-gram模型都属于浅层神经网络，包括输入层、隐藏层和输出层，在对语言模型进行建模的同时获得词在向量空间上的词向量表示。

对于文本$s=(w_1,w_2,\cdots,w_L)$，Skip-gram模型使得式(2)取到最大值：

$$F=\frac{1}{N}\sum_{t=1}^{N}\sum_{-c\leqslant i\leqslant c,i\neq 0}\log p(w_{t+i}|w) \tag{2}$$

其中，c表示训练窗口的大小，即当前词w_t的前面c个词和后面的c个词。

1.2 CNN多尺度特征提取通道

CNN通道主要由输入层、卷积层、池化层组成，整体框架如图2所示。

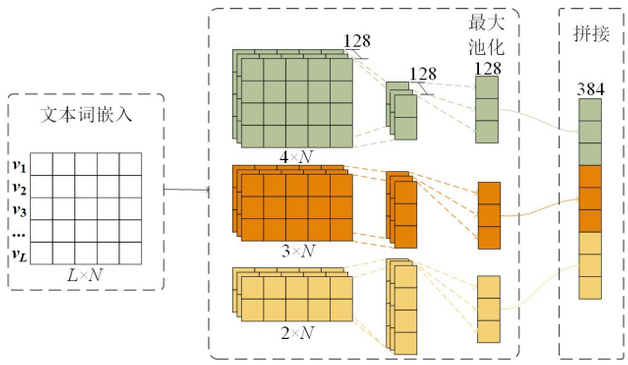

图2 多尺度卷积CNN通道

Fig.2 Multi-scale convolutional neural network channels

将向量化后的文本序列(式(1))作为CNN通道的输入层，通过设计多个尺寸不同的卷积核增加特征多样性，得到各尺寸下的特征信息，每个窗口对输入进行卷积计算的公式为：

$$y_m^h = g\left(\sum_{m\in L} \boldsymbol{W}_h \cdot \boldsymbol{v}_{m:m+h-1} + b_h\right) \tag{3}$$

其中，g为激活函数，本文采用ReLU激活函数，$\boldsymbol{W}_h \in R^{h\times N}$表示卷积核的权重矩阵，$\boldsymbol{v}_{m:m+h-1}$表示$m$至$m+h$-1窗口内的词向量矩阵，$b_h$为偏置，$m$代表卷积核滑动窗口的位置，将上述所有输出特征连接起来就得到了卷积层的输出Y^h，如式(4)所示：

$$Y^h = [y_1^h, y_2^h, \cdots, y_{L-h+1}^h] \tag{4}$$

其中，y_i^h表示第i个大小为h的卷积核提取的文本特征。

池化层对卷积的输出进行冗余特征过滤，将高维特征进行降维，防止模型过拟合。本文使用最大池化(Max-pooling)，对于每一个特征向量，保留Y^h中的最大值作为对应的文本特征$\max(y^h)$。

TCMD模型使用$2\times N$、$3\times N$、$4\times N$大小的卷积核对文本矩阵$\boldsymbol{s}$进行卷积操作，每个尺寸的卷积核数目为128个，步长为1从上往下滑动，则3种尺寸的卷积核卷积后得到的特征输出分别为：

$$Y^2 = [\max(y_1^2), \max(y_2^2), \cdots, \max(y_{128}^2)] \tag{5}$$

$$Y^3 = [\max(y_1^3), \max(y_2^3), \cdots, \max(y_{128}^3)] \tag{6}$$

$$Y^4 = [\max(y_1^4), \max(y_2^4), \cdots, \max(y_{128}^4)] \tag{7}$$

3个卷积核的输出直接进行拼接操作，得到CNN多尺度特征提取通道的向量输出为$C = Y^2 \oplus Y^3 \oplus Y^4$，特征维度为384。

1.3 结合自注意力机制的BiLSTM通道

循环神经网络(Recurrent Neural Network, RNN)擅长处理序列数据，但随着序列长度的增加，会产生训练时梯度消失、梯度爆炸以及长期依赖的问题。LSTM作为RNN的一种变体，通过增加输入门i_t、遗忘门o_t、输出门f_t以及记忆状态细胞c_t来解决上述问题，使用门机制控制信息的保留、遗忘以及状态更新，其计算公式如下：

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, \boldsymbol{x}_t] + b_f) \tag{8}$$

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, \boldsymbol{x}_t] + b_i) \tag{9}$$

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, \boldsymbol{x}_t] + b_o) \tag{10}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(\boldsymbol{W}_c \cdot [h_{t-1}, \boldsymbol{x}_t] + b_c) \tag{11}$$

$$h_t = o_t \odot \tanh(c_t) \tag{12}$$

其中，σ表示非线性激活函数，$\boldsymbol{W}$为权重矩阵，b为偏置，$\boldsymbol{x}_t$为t时刻的输入向量，h_{t-1}为前一时刻的输出，c_{t-1}为前一时刻的隐藏状态，c_t和h_t分别为当前时刻的状态和输出。

LSTM只能学习文本的下文信息，而不能学习文本的上文信息。决定疾病文本类别的词语可能分布在句子的任意位置，其语义同时受到上下文信息的影响，因此TCMD通过BiLSTM结构使用两个方向相反的LSTM来捕捉过去和未来的语义信息，并引入自注意力机制对语义信息赋予不同的权重，整体结构如图3所示。

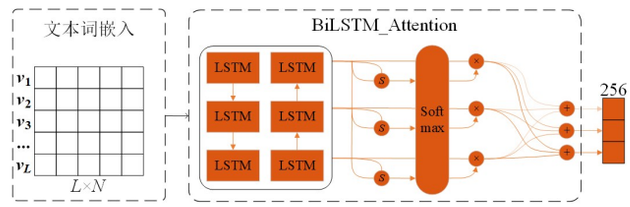

图3 结合自注意力机制的BiLSTM通道

Fig.3 BiLSTM channel combined with self-attention mechanism

本文通过BiLSTM来表示疾病文本的深层信息。BiLSTM对每个时刻输入的句子嵌入进行编码，得到相应的隐层向量，具体过程如下：

$$\overrightarrow{\boldsymbol{h}_t} = \overrightarrow{\mathrm{LSTM}}(\overrightarrow{v_t}) \tag{13}$$

$$\overleftarrow{\boldsymbol{h}_t} = \overleftarrow{\mathrm{LSTM}}(\overleftarrow{v_t}) \tag{14}$$

$$\boldsymbol{H}_t = [\overrightarrow{\boldsymbol{h}_t} \oplus \overleftarrow{\boldsymbol{h}_t}] \tag{15}$$

其中，$\overrightarrow{v_t}$和$\overleftarrow{v_t}$分别表示LSTM从左往右和从右往左读取句子。$\overrightarrow{\boldsymbol{h}_t}$以及$\overleftarrow{\boldsymbol{h}_t}$分别表示前向与后向两个隐层输出，均为128维。$\boldsymbol{H}_t$为最终隐藏层的输出，维度为256。

TCMD模型使用BiLSTM对文本原始的词向量进行编码，分析词语之间的相关性，保留完整的上文和下文信息，同等地考虑两种特征，弥补了CNN只能获得局部信息的不足。

疾病文本中不是每个词都对句子有重要意义，往往包含了大量口语化词语，更需要捕捉哪些词语对分类结果的影响较大。为了区分输入疾病文本中每个词语的重要程度，本文采用自注意力（self-attention）对BiLSTM的输出进行全局性的学习，更加关注重点词语，并且将学习结果与输出序列融合，这样能突出文本的重要信息，建立句子中局部与全局之间的关系，从而更好地表征文本信息。

自注意力模块的输入由$\boldsymbol{Q}$（Query）、$\boldsymbol{K}$（Key）和$\boldsymbol{V}$（value）构成，如式（16）所示。输出是带有权重和的$\boldsymbol{V}$向量，具体算法步骤如下。

$$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{H}_t \tag{16}$$

（1）将$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$进行线性变换：

$$\boldsymbol{Q}'=\boldsymbol{W}^{Q}\boldsymbol{H}_t,\quad \boldsymbol{K}'=\boldsymbol{W}^{K}\boldsymbol{H}_t,\quad \boldsymbol{V}'=\boldsymbol{W}^{V}\boldsymbol{H}_t \tag{17}$$

其中，$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{V}$分别为$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的权重矩阵。

（2）将步骤（1）中的$\boldsymbol{K}'^{\mathrm{T}}$与$\boldsymbol{Q}'$进行点积运算，打分函数采用缩放点积函数，通过除以$K$的维度进行缩放，使内积不会过大。再通过softmax归一化为概率分布，输出自注意力权重向量$\boldsymbol{S}$：

$$\boldsymbol{S}=\mathrm{softmax}\left(\frac{\boldsymbol{K}'^{\mathrm{T}}\boldsymbol{Q}'}{\sqrt{d_K}}\right) \tag{18}$$

（3）将步骤（2）得到的自注意力权重向量$\boldsymbol{S}$与$\boldsymbol{V}$相乘，形成句子自注意力模块的最终加权输出AB，其维度为256：

$$AB=\boldsymbol{S}\boldsymbol{V} \tag{19}$$

自注意力机制的增加改变了BiLSTM输出的隐藏状态，对于编码的结果加入了权重的影响，能够突出重要特征。

1.4 模型优化与分类预测

为使文本序列特征与局部特征建立联系，将双通道输出的特征表示进行拼接得到维度为640的最终特征向量$\boldsymbol{U}=[Y^2\oplus Y^3\oplus Y^4\oplus AB]$，令其作为softmax分类器的输入，共同决定文本的类别结果，计算公式为：

$$\hat{y}=\mathrm{softmax}\left(\boldsymbol{W}_f\cdot\boldsymbol{U}+b_f\right) \tag{20}$$

其中，$\hat{y}$为TCMD模型预测的文本类别概率，$\boldsymbol{W}_f$、b_f分别是全连接层权重矩阵和偏置。

最后，通过最小化交叉熵来优化模型，如式（21）所示：

$$\mathrm{loss}=-\sum_{i=1}^{T}\sum_{j=1}^{C}y_i\log\hat{y}_i^j+\lambda\|\theta\|^2 \tag{21}$$

其中，T表示训练数据集，C为文本类别数，y_i为文本实际类别，λ为正则，θ为设置的参数。

2 实验结果与分析

2.1 实验数据

本文实验数据集来自网络问诊平台，共九大类疾病文本，分别是呼吸科（C1）、内分泌科（C2）、神经科（C3）、内科（C4）、消化科（C5）、心血管科（C6）、耳鼻喉科（C7）、营养保健科（C8）以及神经脑外科（C9），每个类别的数据量为1万条，数据总量9万条，其中70%为训练集，10%为验证集，剩余20%为测试集。

2.2 评估指标

本文评估指标采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F_1值（F-score）、ROC以及ROC曲线下的面积（Area Under Curve, AUC）[14]。精确率用于检验结果的有效性，召回率检查结果的完整性，F_1值调和平均准确率与召回率。ROC曲线的横纵坐标分别为特异性（FPR）和敏感度（TPR），综合衡量模型的有效性和可靠性；AUC反映了模型的分类性能，其值越接近于1，模型分类性能越好。

2.3 实验参数设置

本文实验环境如下：操作系统Win10，CPU型号为Intel Core i5-9400F，GPU为GeForce GTX 1660s，内存大小16 G，深度学习框架TensorFlow1.15.0，编程语言Python3。

本文使用Word2vec预训练词向量，维度为64，词典大小为5 000；CNN卷积窗口大小分别为2、3、4，卷积核数量为128个；正向和反向的LSTM单元大小均设置为128，共享词嵌入输入；采用ReLU激活函数；dropout设置为0.5以防止过拟合；训练批次batch_size设置为128；使用交叉熵作为损失函数；优化器使用Adam；初始学习率设置为0.001，网络迭代次数epoch设置为20。

2.4 实验结果分析

2.4.1 模型性能分析 为了验证TCMD模型在疾病文本分类任务上的性能，设置实验参数，使用训练集进行模型训练，使用测试集对模型进行分类性能评估，实验结果如图4所示。

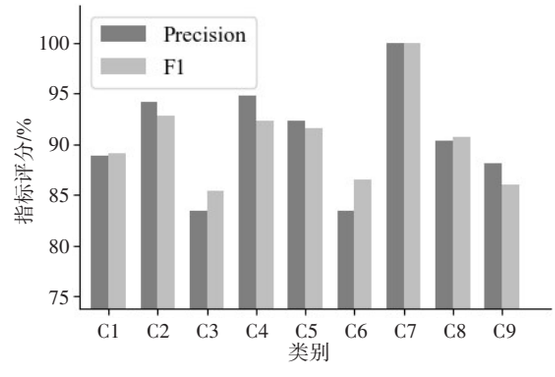

图4 TCMD模型分类测试结果

Fig.4 Classification test results of TCMD

由于数据集中包含许多领域专业词汇以及罕见词，模型对于部分类别的疾病文本不能充分学习；另一方面，数据集中的描述文本包含了大量的非正式语言，一定程度上会导致模型学习到语义混乱的文本特征。由图4可知，整体来看，TCMD模型取得了不错的分类效果，F_1值均超过了85%，其中类别C7的精确率与F_1值最高。类别C3与C6的指标评分较其他类别略低，是因为在本文设置的文本序列长度为80，且其他7个类别的平均文本长度达到78的情况下，C3与C6的文本平均长度都没有超过70，所以这两类文本在预处理时增加了许多空白位，对模型提取语义特征产生影响，从而影响模型的分类性能。

图5展示了TCMD模型在9种疾病文本分类上的ROC曲线。各类别疾病文本的AUC波动较小，平均AUC值为0.989 1，说明模型在各类别上都能达到很好的分类效果。

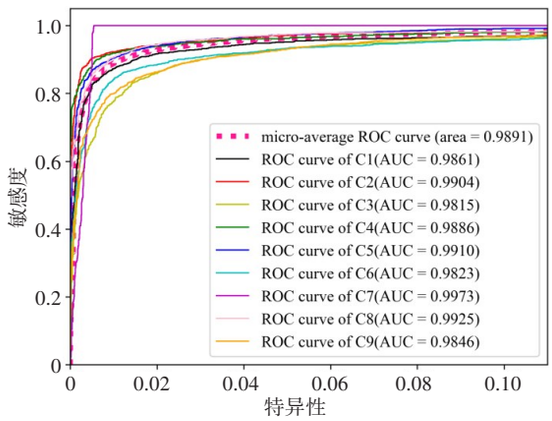

图5 TCMD模型的ROC曲线

Fig.5 ROC curve of TCMD

2.4.2 对比实验分析　为进一步验证TCMD模型的分类性能，本节设置多个分类模型对比实验，包括传统的分类模型支持向量机（SVM）、朴素贝叶斯（Bayes）以及深度学习中的FastText、CNN、LSTM、RCNN[15]分类模型，所有模型在同一数据集上进行训练以及测试，对比传统的机器学习方法和深度学习方法在疾病文本分类任务上的优劣。各模型设置参数如下：（1）传统机器学习方法：SVM和朴素贝叶斯。（2）FastText，此方法中上下文窗口大小设置为5，语言模型为2-gram。（3）CNN，此方法为本文模型中CNN通道采用的方法，超参数与TCMD中的CNN相同。（4）LSTM，此方法使用词嵌入方式，利用单向LSTM网络提取序列特征，超参数与TCMD中LSTM一致。（5）RCNN，RCNN结合RNN与CNN，将CNN网络中卷积层替换为双向RNN，隐藏层个数设置为128。对比实验的性能评估指标为分类精确率、召回率以及F1值，结果如表1所示。

表1 疾病文本分类模型实验结果（%）

Tab.1 Experimental results of different text classification models for diseases (%)

模型	精确率	召回率	F_1值
SVM	81.92	62.70	63.58
Bayes	83.18	81.91	82.54
FasText	85.35	84.99	85.11
CNN	89.39	89.14	89.18
LSTM	87.95	87.82	87.81
RCNN	89.11	89.02	89.01
TCMD	90.61	90.48	90.51

从表1可以看出，深度学习模型性能均要优于传统的机器学习模型SVM和Bayes，原因是深度学习能够提取到更丰富的分类特征。其中，SVM模型通过组合多个二分类器来构建SVM多分类器，虽然该模型有着较好的分类精准率，但召回率远远低于其他模型，从而导致F_1值的降低，分类性能不佳。

另一方面，由表1中性能数据可知，TCMD的精确率、召回率和F_1值分别为90.61%、90.48%、90.51%，相比FastText各指标提升了5.26%、5.49%以及5.40%，TCMD模型能够提取长期上下文依赖特征，而FastText模型对输入文本进行N-gram处理，只能够获取局部词向量特征以及词序顺序，故评估指标较低于TCMD。相比于RCNN模型，TCMD不仅能够对上下文特征进行学习，通过自注意力机制还加强了重要词语的特征信息，故性能略有提升，精确率和F_1值分别均提高了1.50%。相比于CNN与LSTM单模型，TCMD的F_1值分别提高了1.33%、2.70%，主要原因是CNN单模型、LSTM单模型分别只考虑了文本局部特征、下文信息特征对分类结果的影响，切入面单一。而TCMD能够充分考虑两个特征层面，通过结合CNN与BiLSTM的优点，积极提取对文本分类起到正面作用的特征，发挥出CNN局部特征提取的优势以及BiLSTM对不同距离的双向语义信息的保留和筛选能力。在此基础上，注意力对语义信息进行权重分配，学习了句子中不同词语对于文本分类结果的重要程度，故分类效果有所提高。

综上，TCMD模型各项评估指标优于其他分类模型，说明了基于双通道神经网络的方法能够有效提升疾病文本的分类性能。

在深度学习模型训练方面，训练数据量的改变会对模型的分类性能产生显著的影响。通过改变训练数据集的大小来分析数据量与模型的分类准确率的关系。以数据总量的10%为步长设置训练数据集大小，各模型训练结果如图6所示。由图可知，随着

数据量的增加，5种模型的准确率呈上升趋势，但在数据量最少时（10%），TCMD模型的训练准确率远远高于其他模型，说明对于小数据集TCMD模型仍有着较好的分类性能。CNN和LSTM的准确率需要训练数据量分别达到60%和70%之后才保持在90%以上，而TCMD模型在数据量达到40%之后即可达到相同的性能指标，表现出良好的分类能力。

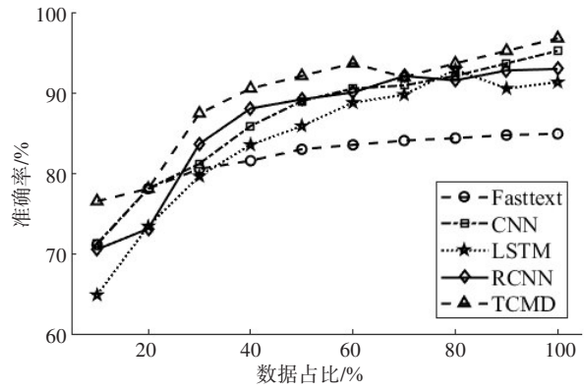

图6 各模型在不同数据量下的准确率

Fig.6 Accuracy of each model under different data volumes

3 结 语

以往的疾病分类研究受到疾病文本数据库的限制，只能处理极少数疾病文本的分类任务；或者分类模型考虑到的特征粒度较为单一，性能尚有较大提升空间。本研究在数据集方面使用了充足的多类别疾病文本数据，在模型上兼顾文本的局部特征和上下文语义特征，同时在BiLSTM层后加入自注意力机制，用以提取句子的全局信息特征，能够区分词语对结果的重要程度，实验结果表明本文模型的分类精度更高、性能更稳定。

本文提出的TCMD模型面向的是疾病文本，可以应用于网络问诊、医院智能导诊、医疗文本数据挖掘处理等方面。在下一阶段的工作中，对于本文模型训练时间较长的不足需要加以改进。未来的研究重点是融合疾病文本的其他特征，如将文本长度、医疗文本词典等特征融入模型，构建更好的分类模型。

【参考文献】

[1] KOWSARI K, MEIMANDI K J, HEIDARYSAFA M, et al. Text classification algorithms: a survey[J]. Information, 2019, 10(4): 150.

[2] 岳世峰, 林政, 王伟平, 等. 智能回复系统研究综述[J]. 信息安全学报, 2020, 5(1): 20-34.
YUE S F, LIN Z, WANG W P, et al. Research on intelligent reply system: a survey[J]. Journal of Cyber Security, 2020, 5(1): 20-34.

[3] 洪巍, 李敏. 文本情感分析方法研究综述[J]. 计算机工程与科学, 2019, 41(4): 750-757.
HONG W, LI M. A review: text sentiment analysis methods[J]. Computer Engineering & Science, 2019, 41(4): 750-757.

[4] 薛春香, 张玉芳. 面向新闻领域的中文文本分类研究综述[J]. 图书情报工作, 2013, 57(14): 134-139.
XUE C X, ZHANG Y F. Research review on chinese text classifiction in the news field[J]. Library and Information Service, 2013, 57(14): 134-139.

[5] 柏挺, 朱海云, 龚宏伟, 等. 机器学习在远程医疗中文分类中的应用[J]. 中国数字医学, 2017, 12(3): 79-82.
BO T, ZHU H Y, GONG H W, et al. Application of machine learning in chinese text categorization for telemedicine[J]. China Digital Medicine, 2017, 12(3): 79-82.

[6] GLINKA K, WOŹNIAK R, ZAKRZEWSKA D. Improving multi-label medical text classification by feature selection[C]//International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). Poznan: IEEE, 2017: 176-181.

[7] CAMPILLOS L, BOUAMOR D, BILINSKI E, et al. Description of the patient genesysdialogue system[C]//Proceedings of the 2015 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. New York: Association for Computational Linguistics, 2015: 438-440.

[8] MIKOLOV T, CORRADO G, CHEN K, et al. Efficient estimation of word representations in vector space[C]. International Conference on Learning Representations, 2013: 1-12.

[9] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrasesand their compositionality[J]. Adv Neural Inf Process Syst, 2013, 26: 3111-3119.

[10] LIU M, LANG B, GU Z, et al. Measuring similarity of academic articles with semantic profile and joint word embedding[J]. Tsinghua Science and Technology, 2017, 22(6): 619-632.

[11] HUGHES M, LI I, KOTOULAS S, et al. Medical text classification using convolutional neural networks[J]. Stud Health Technol Inform, 2017, 235: 246-250.

[12] YAO L, ZHANG Y, WEI B, et al. Traditional Chinese medicine clinical records classification using knowledge-powered document embedding[C]//International Conference on Bioinformatics and Biomedicine (BIBM). Shenzhen: IEEE, 2016: 1926-1928.

[13] CHEN C W, TSENG S P, KUAN T W, et al. Outpatient text classification using attention-based bidirectional LSTM for robot-assisted servicing in hospital[J]. Information, 2020, 11(2): 106.

[14] CHENG Y, WANG F, ZHANG P, et al. Risk prediction with electronic health records: a deep learning approach[C]//Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2016: 432-440.

[15] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//AAAI Conference on Artificial Intelligence. Austin: AAAI Press, 2015: 2267-2273.

（编辑：薛泽玲）