文章编号：2095-2163(2020)10-0074-06 中图分类号：TP181;R743.3 文献标志码：A

基于 LSTM 多特征联合的缺血性脑卒中诊断模型

骆轶姝，邵圆圆，陈德华
（东华大学 计算机科学与技术学院，上海 201620）

摘 要：本文以缺血性脑卒中疾病为研究对象，充分考虑疾病发病机制，选取患者当前超声、生化及基本信息 3 种特征检查指标，提出一种基于 LSTM 多特征联合的诊断模型。3 个基于 LSTM(Long short-term memory)模型搭建的双向 LSTM 特征提取子模块，联合训练学习各类型数据的前向和后向信息；增加自注意力机制学习特征间的关联性，并分配权重。实验结果表明，融合自注意力机制的多特征模型在不同分类评估标准下总体性能达 84%，为缺血性脑卒中的临床辅助诊断提供一种方法，为医生对该疾病的鉴别诊断提供参考。

关键词：缺血性脑卒中；LSTM；多特征联合；辅助诊断

Diagnosis Model of Ischemic Stroke based on LSTM with Multi-Feature Combination

LUO Yishu, SHAO Yuanyuan, CHEN Dehua
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 Taking ischemic stroke as the research object, a diagnosis model based on LSTM with multi-feature combination was proposed. Considering the pathogenesis of the disease, three characteristics of the patients were selected: ultrasound, biochemistry and basic information; based on the LSTM(long short-term memory) model, three bidirectional LSTM feature extraction sub models were constructed to jointly train and learn the forward and backward information of various types of data; the relevance between the learning characteristics of self-attention mechanism was increased and the weight was allocated. The experimental results show that the multi-feature model based on self-attention mechanism has the advantages in different classification evaluation criteria, which can reach 84%. The method for the clinical diagnosis of ischemic stroke was proved effective and could be used as a reference for doctors in differential diagnosis of the disease.

【Key words】 ischemic stroke; LSTM; multi-feature combination; auxiliary diagnosis

0 引 言

近年来，人工智能在疾病诊断中的应用不断延伸，缺血性脑卒中疾病的临床辅助诊断也得到越来越多关注。缺血性脑卒中作为一种急性脑血管疾病，占中国脑卒中约 70%左右[1]；且随着人们工作压力及生活方式的改变，呈现发病率高，发病原因复杂的发展趋势，为临床医生带来诊断压力[2]。因此，基于人工智能的缺血性脑卒中辅助诊断问题的研究，对医生和患者来说，均具有重要意义。

本文以上海市某医院的真实患者电子病历为基础，考虑缺血性脑卒中疾病的发病原因，选取当前病历数据中的超声、生化以及个人基本信息作为源数据，在 LSTM 模型基础上搭建双向 LSTM 多特征提取子模型，实现了多特征联合的缺血性脑卒中的辅助诊断。相对传统诊疗模式强化了客观因素，为医生对该疾病诊断提供有效辅助。

1 相关工作

国内外学者关于疾病智慧医疗辅助诊断开展了大量研究。有些学者在支持向量机、决策树等机器学习模型下，实现对疾病数据的线性学习，但该类方法难以捕获复杂特征学习问题。近年来，以 LSTM 模型为基础的疾病诊断方法受到广泛关注，可以建立增加了特征序列输入的学习模型，例如实现基于 LSTM 模型的心脏病诊断[3]、脑血管疾病诊断对疾病时序检查特征的学习等[4]。也有学者在此基础上综合后向特征计算，提出双向 LSTM 模型[5]，该方法在文本分类问题中表现较好，例如融合前向和后向特征的双向 LSTM 模型实现对心血管疾病病历数据挖掘的辅助诊断[6]。

基于上述研究，本文提出 LSTM 多特征联合的缺血性脑卒中辅助诊断模型，运用数据预处理方法，设计从不同特征提取子模型中提取信息并进行向量

基金项目：上海市经信委人工智能创新发展专项资金(RX-RJJC-08-16-0483,2017-RGZN-01004)。

作者简介：骆轶姝(1974-)，女，博士，副教授，主要研究方向：数据库、数据仓库与智慧医疗；邵圆圆(1996-)，女，硕士研究生，主要研究方向：智慧医疗。

收稿日期：2020-05-13

融合，降低不同类型检查数据间差异所带来的模型学习能力；另外，模型中多特征层次上自注意力机制的特征加权，弥补不同特征间存在的信息关联性，提升模型分类性能。

2　模型建立

2.1　多特征联合建模

基于 LSTM 多特征联合的缺血性脑卒中诊断模型包括输入层、特征提取层、分类层和输出层。模型总体结构如图 1 所示。

其中输入层由预处理的超声指标、生化检查指标和基本信息组成；特征提取层经 3 个双向 LSTM 搭建的子模型学习特征信息；分类层的各特征向量，是在模型特征融合的基础上，增加自注意力机制分配获得；输出层用于输出疾病诊断结果。

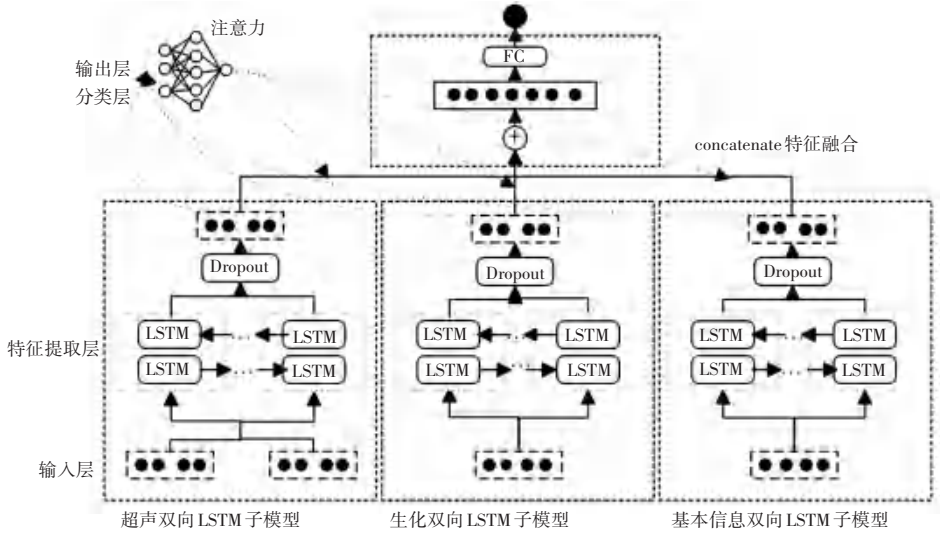

图 1　模型总体结构

Fig. 1　Overall structure of the model

2.2　特征提取

双向 LSTM 建立的 3 个特征提取子模块分别为超声特征提取、生化检查特征提取及基本信息特征提取。

(1)超声特征提取。将患者结构化后的颈动脉超声指标作为该超声特征提取模块的输入，提取超声中有关影响疾病的重要信息。超声特征提取子模块的设计如图 2 所示。

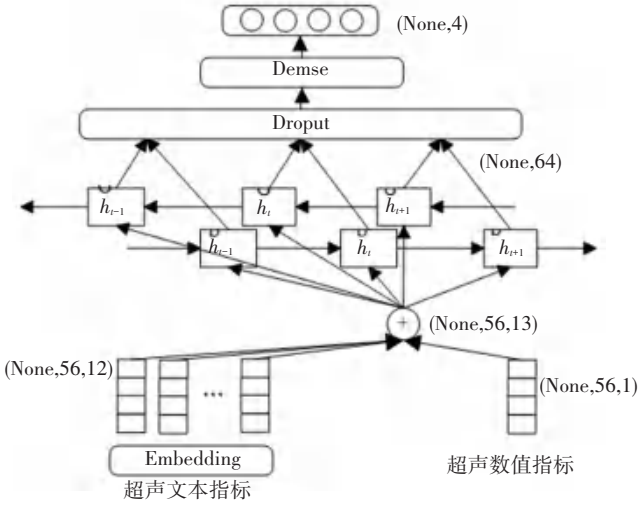

图 2　超声特征提取子模块

Fig. 2　Submodule of feature extraction for ultrasonic

由图 2 可知，针对超声中的文本指标，采用神经网络中 Embedding 层加载 Word2vec 模型实现向量化；并将超声中的数值指标填充为相同形状的 1 维特征；融合后输入双向 LSTM 模型中进行信息提取，其中 t 时刻前向隐藏层特征信息的计算如式(1)~(6)所示。

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{1}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{2}$$

$$\widetilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c), \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t, \tag{4}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tag{5}$$

$$h_t = o_t \times tanh(C_t). \tag{6}$$

式中：W_f、W_i、W_c、W_o 为共享权值参数矩阵，b_f、b_i、b_c、b_o 为偏置值，通常随机初始化。由 $t-1$ 时刻输出的特征向量与当前时刻输入特征的计算，得到 t 时刻隐藏层的 h_t 的特征信息。最后由该时刻的两个隐藏单元的输出向量连接构成该时刻输出。计算如式(7)~式(9)所示。

$$\overrightarrow{h_t} = LSTM(\overrightarrow{h_{t-1}}, x_t), \tag{7}$$

$$\overleftarrow{h_t} = LSTM(\overleftarrow{h_{t+1}}, x_t), \tag{8}$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]. \quad (9)$$

由 Dropout 以一定概率丢弃神经元个数,减少模型复杂带来的过拟合问题。最后经一个 Dense 全连接层将该模块提取的特征向量做非线性映射转化为(None,4)形状的特征向量。

(2)生化检查特征提取。生化检查特征提取子模块的设计如图 3 所示。首先,直接利用预处理的生化指标转化为三维特征,由双向 LSTM 模型中神经元计算生化检查中特征的前向和后向特征,充分提取特征中具有的信息;其次,连接 Dropout 网络丢弃层,由一个 Dense 全连接层将高维的特征压缩为(None,4)形状的特征向量。

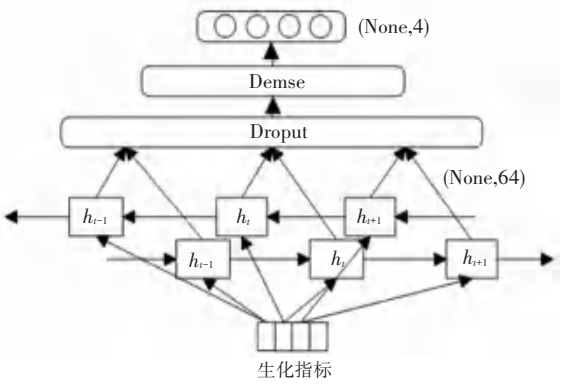

图 3 生化检查特征提取子模块

Fig. 3 Submodule of feature extraction for biochemical examination

(3)基本信息特征。提取基本信息特征子模块的设计原理同生化指标模型设计,仅输入特征形状为(None,12,1),故此处分析省略。

2.3 疾病分类

疾病分类模块将各子模块提取的形状相同的特征向量连接自注意力机制[7],分配特征权重,实现多特征联合的诊断模型训练学习。其中自注意力计算如式(10)所示。

$$A_t = \sum_{t'}^{T} \alpha(x_t, x_{t'}) f(x_t, x_{t'}). \quad (10)$$

式中:$\alpha(x_t, x_{t'})$ 表示特征向量中的每个特征与该特征向量之间加权值,突出各类型特征的重要程度。最后经两层 Dense 全连接层由 Sigmoid 激活函数作为分类器,输出结果。其中 Sigmoid 计算如式(11)所示。

$$\sigma(c_j) = \frac{e_j^c}{1 + e_j^c}. \quad (11)$$

3 多特征数据预处理

3.1 多特征数据

患者病历中 3 种多特征数据通过医疗卡号和住院号实现关联。

(1)超声数据。作为缺血性脑卒中发生常见的原因之一,颈动脉超声一定程度上可以反映缺血性脑卒中发生与否及严重程度。结构化后的指标数据组成的超声数据,见表 1。

表 1 超声数据

Tab. 1 Ultrasound data

医疗卡号	住院号	部位	方位	属性	属性值
11111111	100001	颈总动脉	左侧	内径	正常
				IMT	0.6
				RI	0.72
				PSV	72
				斑块	可见
				回声性质	强回声
				狭窄率	<50%
				血管外形	直
				血流方向	偏窄
				阻力指数	在正常范围
				流速	在正常范围
				内膜回声	毛糙
				边缘	欠规则
				管腔内径	正常
				管径	对称
				血流	连续完整
				内中膜厚度	正常

其中斑块狭窄率根据美国超声会议中标准转化可输入数据预处理形式[8]。

(2)生化检查数据。生化检查对临床中疾病的筛查验证具有重要意义。本文选取的生化指标共计 8 个,包括 CHOL(总胆固醇)、CRP_1(C 反应蛋白)、GLU1(空腹血糖)、APOA(载脂蛋白 A)、APOE(载脂蛋白 E)、MO#(单核细胞计数)、TG-B(甘油三酯)以及 UHDL(高密度脂蛋白)。生化检查数据见表 2。

表 2 生化检查数据

Tab. 2 Biochemical examination data

医疗卡号	住院号	CHOL	CRP_1	GLU1	APOA	APOE	MO#	TG-B	UHDL
11111111	100001	3.42	0.87	6.58	1.61	3.3	0.40	1.37	1.25

（3）基本信息数据。病历数据中患者基本信息包含性别、出生年月、身高、体重、sbp 收缩压、dbp 舒张压等。出生年月转化为年龄，身高体重转化为 BMI（身体质量指数，衡量人体是否健康及胖瘦的一个指标）。同时高血压、糖尿病及高血脂常伴随缺血性脑卒中患者，因此也作为缺血性脑卒中研究的指标之一加入患者的基本信息中。基本信息数据见表 3。

表 3　基本信息数据

Tab. 3　Basic information data

医疗卡号	住院号	性别	年龄	BMI	sbp	dbp	高血压	糖尿病	高血脂
11111111	100001	男	65	30.07	147.0	104.0	是	否	是

3.2　多特征数据预处理

（1）Word2vec。Word2vec，一种词向量化技术，能够实现语义空间信息到向量空间上的映射。本文使用 Skip-Gram 思想计算词的上下文概率分布，由建立 Word2vec 模型对语料库编码，神经网络中加载实现词向量化。Word2vec 词向量化示意如图 4 所示。

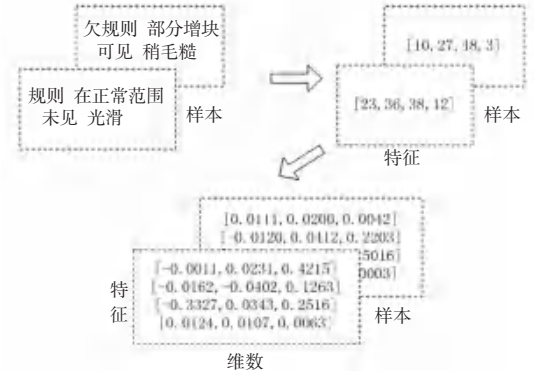

图 4　Word2vec 词向量化示意

Fig. 4　The vectorization of Word2vec

（2）one-hot。one-hot 是一种通过 N 位状态寄存器对 N 个状态编码，实现离散特征映射到欧式空间的独热编码方式，将模型特征中非连续性数值，即离散型数据通过编码的方式进行转换。一方面提高模型的计算特征之间距离的效率，另一方面对数据特征维度上起扩充作用。以性别、高血压为例，经 one-hot 独热编码后，由 0，1 二进制形式表示。性别与高血压向量化见表 4。

表 4　性别与高血压向量化

Tab. 4　Gender and quantitative of hypertension

性别_男	性别_女	高血压_是	高血压_否
1	0	1	0
0	1	0	1

（3）归一化。当实验数据作为同一水平的输入变量输入模型中时，存在纲量不一致问题，不仅影响数据之间的可比性，还会导致分析结果存在偏差。采用离差标准化归一化方法，通过线性变化，将所有指标数值进行压缩，计算如式（12）所示。

$$X' = \frac{X - X.Min}{X.Max - X.Min}. \tag{12}$$

式中，$X.Min$ 为指标 X 数据中的最小值；$X.Max$ 为指标 X 数据中的最大值。以 APOE 数据为例，线性归一化处理如图 5 所示。由图 5 可知，横坐标为 APOE 原数据形式，范围为[2，15.3]，由归一化将其映射到[0，1]之间；其中数据仍保持原特征，提升模型训练收敛速度和精度。

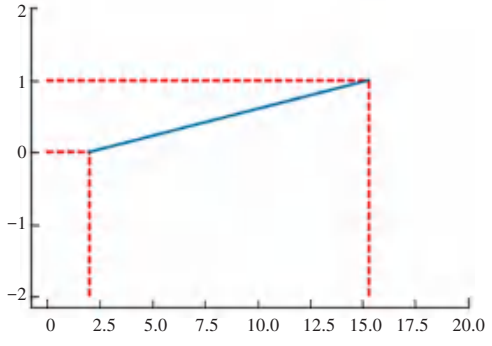

图 5　APOE 线性归一化处理

Fig. 5　Linear normalization of APOE

4　模型训练与结果分析

4.1　实验数据

实验数据来自上海某医院真实病历数据。数据集中筛选处理 797 条缺血性脑卒中患者正样本数据。为进行实验对比，选取 962 条非缺血性脑卒中患者的数据作为实验负样本。模型训练过程中分为训练集和测试集，其中训练集占 80%，测试集占 20%。

4.2　模型训练

实验中采用交叉熵损失函数计算模型预测值与真实值间误差，并设置 Adam 优化器反向优化学习，使得损失最小时，模型训练达最优。多特征联合的缺血性脑卒中辅助诊断模型训练实现如算法 1 所示。

算法 1　缺血性脑卒中辅助诊断模型训练实现

E：迭代次数

B：批大小数据集

$L_{learning\_rate}$：学习率

D_{train}、D_{test}：训练集、测试集

N：神经元个数

n：特征子模型个数

V_n：第 n 个子模型输出的特征向量

V：特征联合向量

A：注意力机层输出向量

${}^{T}\theta_n^i / {}^{C}\theta^i$：第 n 个特征子模型的 i 次迭代模型网络参数 / 分类模型 i 次迭代网络参数 θ

${}^{T}L_n / {}^{C}L_{loss}$：第 n 个特征子模型网络误差 / 分类模型网络误差

Initialize $({}^{T}\theta_1^0, {}^{T}\theta_2^0, \ldots\ldots {}^{T}\theta_n^0, {}^{C}\theta^0)$

For i in E:

${}^{B}D_{train} \leftarrow$ GetMiniBatch (D_{train}, B)

${}^{B}D_{train}^{n} \leftarrow GetNData(D_{train}^{B})$

For j in n:

$V_j \leftarrow$ ModelBLSTM $({}^{B}D_{train}^{j}, {}^{T}\theta_j^i)$

End For

$V \leftarrow$ concatenate (V_n)

$A \leftarrow$ Self-attention (V)

$L_{loss} \leftarrow$ ModelClassify $(A, {}^{C}\theta^i)$

11. $({}^{T}\theta_1^i, {}^{T}\theta_2^i, \ldots\ldots {}^{T}\theta_n^i, {}^{C}\theta^i) \leftarrow$ Adam $(L_{learning\_rate}, L_{loss})$

12. End for

13. Evaluate $(D_{test}, {}^{T}\theta_1, {}^{T}\theta_2, \ldots\ldots {}^{T}\theta_n, {}^{C}\theta)$

14. End

模型在训练中，确定了 learning_rate = 0.001，dropout = 0.5，epoch = 100 时，性能达最优。

4.3 结果分析

实验中采用准确度（*Accuracy*）、灵敏度（*Sensitivity*）、特异度（*Specificity*）、阳性预测率（*PPV*）、阴性预测率（*NPV*）以及 *F1_Score* 的作为评估标准。计算如式（13）、（14）、（15）、（16）、（17）和（18）所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FN + TN}, \tag{13}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{14}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{15}$$

$$PPV = \frac{TP}{TP + FP}, \tag{16}$$

$$NPV = \frac{TN}{TN + FN}, \tag{17}$$

$$F1\_score = \frac{2 \times PPV \times Sensitivity}{PPV + Sensitivity}. \tag{18}$$

其中涉及基本概念的混淆矩阵表示，见表 5。

表 5 混淆矩阵表示

Tab. 5 Confusion matrix representation

预测值	真实值	
	缺血性脑卒中	非缺血性脑卒中
缺血性脑卒中	*TP*	*FP*
非缺血性脑卒中	*FN*	*TN*

本文实验首先对比了基于 LSTM 多特征模型（MLSTM）、基于双向 LSTM 和 LSTM 组合的多特征模型（MBLSTM-LSTM）以及基于双向 LSTM 多特征模型（MBLSTM）。不同模型下实验结果对比见表 6。由表 6 可知，MBLSTM 模型优于其他两种模型。从网络模型结构上看，LSTM 实现对输入特征的单向计算，双向 LSTM 综合输入前向和后向的信息，提升了模型分类性能。

表 6 不同模型下实验结果对比

Tab. 6 Comparison of experimental results based on different models %

模型	PPV	Sensitivity	F1_Score	NPV	Specificity	Accuracy
MLSTM	72	78	75	75	68	73
MBLSTM-LSTM	87	69	77	73	89	79
MBLSTM	80	81	81	80	79	80

为进一步验证文中提出多特征模型的有效性，实验对比了单个超声 LSTM（$LSTM_c$）/双向 LSTM（$BLSTM_c$）、生化 LSTM（$LSTM_s$）/双向 LSTM（$BLSTM_s$）、基本信息 LSTM（$LSTM_j$）/双向 LSTM（$BLSTM_j$）诊断模型。各单独特征模型与多特征模型结果见表 7。

表 7 单独特征模型与多特征模型结果对比

Tab. 7 Comparison of the experimental results of single feature models with multi-feature models

模型	PPV	Sensitivity	F1_Score	NPV	Specificity	Accuracy
$LSTM_c$	70	52	59	60	76	64
$LSTM_s$	72	52	60	61	79	65
$LSTM_j$	68	55	61	60	73	63
$BLSTM_c$	67	57	61	61	70	63
$BLSTM_s$	86	66	75	71	89	77
$BLSTM_j$	77	78	77	76	75	76
MBLSTM	80	81	81	80	79	80

由表 7 可知，较单独特征 LSTM 诊断模型、双向 LSTM 诊断模型，多特征诊断模型有效地联合多特

征间的信息，提升模型诊断预测结果，模型整体准确度为 80%左右，发挥了不同类型特征信息对疾病诊断的作用。

考虑到注意力机制对关键特征加权的影响，在多特征模型基础上增加自注意力机制。各模型对比增加注意力机制模型的准确度结果对比如图 6 所示。

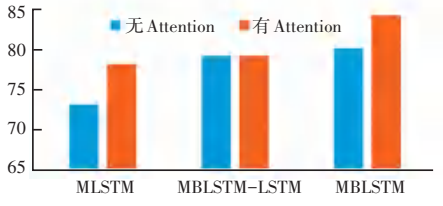

图 6　模型对比增加自注意力机制模型的准确度结果对比

Fig. 6　Comparison of the experimental accuracy results of models with self-attention mechanism models

由图 6 可知，各多特征模型对比有无自注意力机制上，准确度均保持稳定或者有所增加，说明自注意力机制增加了对各特征子模型输出的特征向量权重的计算，并分配了相应的权重值。

5　结束语

本文提出基于 LSTM 多特征联合诊断模型，利用 Word2vec、one-hot 及归一化等数据预处理方法，获取高质量输入数据，加速模型训练的收敛速度；联合患者当前多种检查数据，在建立的双向LSTM子模型下提取特征信息；自注意力机制学习特征间的关联并分配权重，增强模型学习性能，提升分类结果。实验结果表明，该模型诊断效果良好，在准确度、灵敏度、特异性、阳性预测率、阴性预测率以及 F1_score 中性能总体达 84%，且自动辅助诊断降低了主观因素影响，在缺血性脑卒中辅助诊断研究中具有一定的价值，为临床医生缺血性脑卒中疾病诊断提供决策参考。

参考文献

[1] 中华医学会神经病学分会，中华医学会神经病学分会脑血管病学组. 中国急性缺血性脑卒中诊治指南 2018[J]. 中华神经科杂志，2018，51(9)：666-682.

[2] 曹新西，徐晨婕，侯亚冰，等. 1990—2025 年我国高发慢性病的流行趋势及预测[J]. 中国慢性病预防与控制，2020，28(1)：14-19.

[3] 李晓坤，郑永亮，刘磊，等. LSTM 与 DeepLearning 技术在疾病诊断中的应用[J]. 黑龙江大学工程学报，2018，9(3)：67-72.

[4] 姚春晓. 基于长短时记忆神经网络的脑血管疾病预测系统研究[D]. 北京交通大学，2019.

[5] GRAVES A，SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks，2005，18(5-6)：602-610.

[6] 郭赛迪. 基于 LSTM 的心血管疾病辅助诊断研究[D]. 郑州大学，2019.

[7] LIN Z，FENG M，SANTOS C N D，et al. A Structured Self-attentive Sentence Embedding[J]. 2017.

[8] 美国心脏病学院. 超声心动图临床应用指南：2003 年版[M]. 科学技术文献出版社，2005.

（上接第 73 页）

新的边界条件如式(18)[5]：

$$\sum_{k=1}^{K}\sum_{n=1}^{N-2}(\bar{E}^{c}_{k,n+1}(z_{n+1};z_{n+1}(v))+\bar{E}^{m}_{O,k,n+2}(z_{n+2};z_{n+2}(v)))+\sum_{n=1}^{N}E^{c}_{F,n}(z_n)\leqslant\varepsilon. \quad (18)$$

3　结束语

(1)因为无人机供能有限，其飞行与计算卸载受到限制。本文的目标是在满足通信质量的前提下，使系统能耗最小化。通过使用 SCA 方法解决了在时延与能耗限制条件下，同时优化无人机飞行路径与上行，下行传输，本地计算数据分配等问题。

(2)本文主要研究了铁路安全监测用无人机云计算算法的优化，给出了无人机云计算算法相关的理论推导。调研了几种不同的迭代算法，包括二分查找，牛顿-拉普森导数方法，最终决定比较朴素的 SCA 算法，SCA 算法需要涉及函数全为凸函数。

(3)本文给出的关于铁路安全监测用无人机计算算法的优化方案是基于正交的信息传输方式，仅进行了初步探讨。考虑到程序的运行速度以及性能，具体的优化结果有待进一步研究与具体实现。不过，铁路安全监测用无人机云计算的研究可继续深化，在未来发展前景广阔。

参考文献

[1] ZENG Y，ZHANG R，Teng Joon LimYang. Wireless Communications with Unmanned Aerial Vehicles：Opportunities and Challenges [J]. IEEE Communications Magazine，2016，54(5)：36-42.

[2] Seongah Jeong，Osvaldo Simeone，Joonhyuk Kang. Mobile Edge Computing via a UAV-Mounted Cloudlet：Optimization of Bit Allocation and Path Planning[J]. IEEE Transactions on Vehicular Technology，2018，67(3)：2049-2062.

[3] JEONG S，SIMEONE O，HAIMOVICH A，et al. Mobile cloud com- puting with a UAV-mounted cloudlet：Optimal bit allocation for communication and computation[J]. IET Commun，2017，11(7)：969-974.

[4] CAO Xiaowen，XU Jie，ZHANG Rui. Mobile Edge Computing for Celluar-Connected UAV：Computation Offloading and Trajectory Optimization [J]. 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)，2018，4(3)：954.

[5] WEI L，HU R Q. Enabling Device-to-device Communication Underlaying Cellular Networks：Challenges and Research Aspects [J].IEEE Commun，2014，52(6)：90-96.