◆特邀专栏◆

基于 RF-BiLSTM-Attention 音乐分类方法的京剧二分类仿真*

龚谊承[1,2**],刘　青[1,2],肖浩逸[1]

(1.武汉科技大学理学院,湖北武汉　430065;2.冶金工业过程系统科学湖北省重点实验室(武汉科技大学),湖北武汉 430081)

摘要:为了普及国粹京剧,本研究提出一种将随机森林(Random Forest,RF)耦合注意力(Attention)机制和双向长短时记忆(BiLSTM)网络的音乐分类方法 RF-BiLSTM-Attention,使用其进行京剧与其他类型音乐的二分类(以下简称"京剧二分类")。首先,提取音乐所有光谱特征,利用 RF 选择重要特征;然后,在 BiLSTM 网络的隐藏层与输出层之间嵌入注意力层,对数据进行分类训练与预测。用来自大众音乐平台和 GTZAN 数据集的 1 500 首音乐进行京剧二分类实验,对比 RF 对循环神经网络(RNN)、长短时记忆(LSTM)网络、BiLSTM 等 9 种模型的影响,结果表明:RF-BiLSTM-Attention 模型的分类准确率为 89.00%,运行时间为 33.22 s,比简单模型中表现最好的 RF-BiLSTM 模型的分类准确率提高 3.33%,运行时间缩短 40.54%;比原始 BiLSTM-Attention 模型的分类准确率提高 6.33%,运行时间缩短 96.89%。与传统音频分类工作相比,本研究考虑了京剧二分类问题,对京剧起着良好的推广作用。

关键词:京剧;双向长短时记忆网络;注意力机制;随机森林;二分类

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1002-7378(2023)03-0322-10

DOI:10.13657/j.cnki.gxkxyxb.20230829.012

随着国家弘扬传统文化力度的不断加大,京剧逐渐进入大众视野,但其在大众,特别是青年群体中尚未能有效推广[1,2]。在京剧并不流行的时代,人们很难区分京剧与其他类型的音乐。因此,京剧与其他类型音乐的二分类(以下简称"京剧二分类")问题的研究就显得十分必要。

近年来,虽然对京剧音频研究较少,但对各种音乐、语音以及自然声音的研究层出不穷。Saunders[3]和 Scheirer 等[4]使用短时能量、过零率等频谱特征直接对语音与音乐进行二分类;Lim 等[5]在支持向量机(SVM)分类器的基础上通过改进分类器等方式改进原始音乐分类方法,降低了计算成本并提高了准确率。考虑到音乐的频谱信息是音乐本质的体现,复杂的音频或许不能被少数的几种频谱信息完全翻译,

收稿日期:2023-03-30　　修回日期:2023-07-04

*国家自然科学基金项目(12171378),冶金工业过程系统科学湖北省重点实验室项目(Y202105)和武汉科技大学研究生教学研究项目(Yjg202116)资助。

【第一作者简介】

龚谊承(1975-),女,副教授,主要从事博弈学习与统计学习研究,E-mail:gongyicheng@wust.edu.cn。

【**通信作者】

【引用本文】

龚谊承,刘青,肖浩逸.基于 RF-BiLSTM-Attention 音乐分类方法的京剧二分类仿真[J].广西科学院学报,2023,39(3):322-330,339.

GONG Y C,LIU Q,XIAO H Y. Beijing Opera Binary Classification Simulation Based on RF-BiLSTM-Attention Music Classification Method [J]. Journal of Guangxi Academy of Sciences,2023,39(3):322-330,339.

Birajdar 等[6,7]通过音频产生的色谱图、光谱图以及其统计描述符保留语音与音乐的纹理后进行二分类,但利用谱图与音乐纹理进行分类增加了分类难度;姚斯强等[8]和万凌艳[9]提取高维频谱特征后,利用线性判别分析(LDA)、核主成分分析法对特征降维,而后利用 SVM 等分类方法进行训练,证实对原始音频特征降维后再利用分类器训练能有效提高分类准确率,但通过线性降维的方式所得到的特征并不是原始特征的子集,故该方式缺乏挑选重要特征的能力。

除了利用传统的音频特征进行音频分类外,也有许多学者考虑到音频的时间序列特性与情感特性,将各种神经网络用于音频研究中。Tu 等[10]和 Zhang 等[11]利用循环神经网络(RNN)的特殊结构与时间学习能力,弥补短期特征的不足并提高对音频长度处理的灵活性,提高模型对混合音频的分类准确率,但 RNN 模型在处理较长时间序列时容易产生梯度消失与梯度爆炸问题,影响分类准确率;郭毓博等[12]将长短时记忆(LSTM)网络应用于笛音分类中,通过多种模型的混合比较得出最佳分类模型;Yu 等[13]在音频特征提取的基础上引入注意力(Attention)机制增强特定频段,以此提高音频分类准确率,但该方法仅考虑特定频段,缺乏对局部时间信息与全局时间信息的联系。Noumida 等[14]将注意力机制与双向门控循环单元模型结合来对鸟类声音分类,该方法能够提高分类性能,但在一定程度上缺乏对音频信号为强时间序列的考虑。

综上所述,本研究提出一种基于 RF-BiLSTM-Attention 的音乐分类方法,将具有特征筛选能力的随机森林(Random Forest,RF)和能够双向学习时间序列的双向长短时记忆(BiLSTM)模型结合,再利用注意力机制抓取音频的全局信息。通过实验对比分析,验证所提出的方法在京剧二分类问题上的可行性,为后续国粹京剧推广相关工作提供研究基础。

1 BiLSTM 网络及其基础网络

BiLSTM 网络适用于具有前后时间联系的音频数据,由一个前向 LSTM 网络与一个后向 LSTM 网络构成[15],改善了 LSTM 网络仅能单向学习时间信息的问题。LSTM 网络由 RNN 改变隐藏层结构所得,能够有效缓解 RNN 在梯度更新过程中造成的梯度消失与梯度爆炸问题。

1.1 RNN

RNN 广泛应用于序列数据,独特的链式结构使其增加了对信息的记忆功能。其结构如图 1 所示。

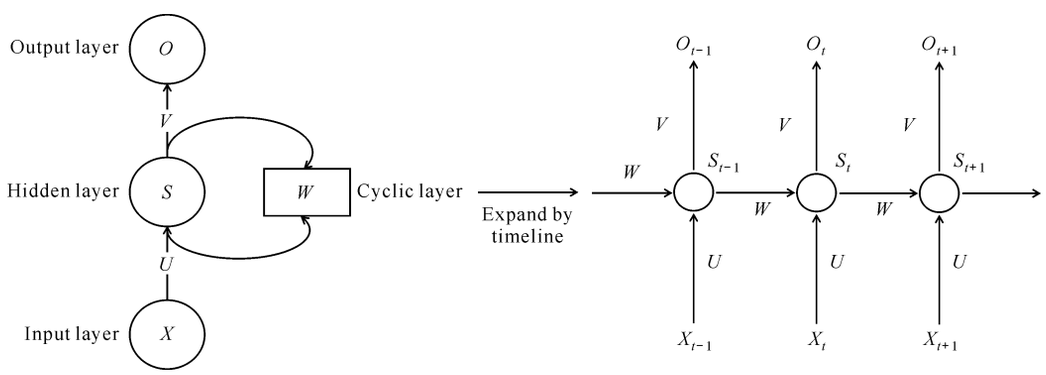

图 1 RNN 基本结构

Fig. 1 Basic structure of RNN

由图 1 可见,RNN 包含输入层 X、隐藏层 S 以及输出层 O,在隐藏层实现循环。隐藏层状态与输出计算公式如式(1)、(2)所示。

$$S_t=\phi(U\cdot X_t+W\cdot S_{t-1}+b_s), \tag{1}$$

$$O_t=V\cdot S_t+b_o, \tag{2}$$

式(1)、(2)中,S_t、O_t 分别表示 t 时刻的隐藏层与输出层状态,ϕ 为激活函数,b_s、b_o 分别为隐藏层与输出层偏置,W、U、V 分别为隐藏层、输入层、输出层的权值矩阵。通常运用梯度下降对权值矩阵参数 W、U、V 更新,更新公式如式(3)-(5)所示。

$$W=W-\mathrm{lr}\sum_{i=0}^{t}\frac{\partial e_t}{\partial W}=W-\mathrm{lr}\sum_{i=0}^{t}\frac{\partial e_t}{\partial O_t}\frac{\partial O_t}{\partial S_t}\frac{\partial S_t}{\partial S_i}\frac{\partial S_i}{\partial W}, \tag{3}$$

$$U=U-\mathrm{lr}\sum_{i=0}^{t}\frac{\partial e_t}{\partial U}=U-\mathrm{lr}\sum_{i=0}^{t}\frac{\partial e_t}{\partial O_t}\frac{\partial O_t}{\partial S_t}\frac{\partial S_t}{\partial S_i}\frac{\partial S_i}{\partial U}, \tag{4}$$

$$V=V-\mathrm{lr}\sum_{i=0}^{t}\frac{\partial e_t}{\partial V}=V-\mathrm{lr}\sum_{i=0}^{t}\frac{\partial e_t}{\partial O_t}\frac{\partial O_t}{\partial S_t}\frac{\partial S_t}{\partial S_i}\frac{\partial S_i}{\partial V}, \tag{5}$$

式(3)-(5)中,e_t 为每次输出值 O_t 与真实值之间的

误差,lr 为学习率 Learn rate 的缩写,S_i 为 i 时刻隐藏层状态。

由式(3)-(5)可看出,在递归计算梯度过程中,如果权值矩阵的值小于 1,那么梯度则会逐渐趋于 0,造成梯度消失;如果权值矩阵的值大于 1,梯度则会趋于无穷,造成梯度爆炸。

1.2 LSTM 网络

LSTM 网络是一种特殊的 RNN,其在原本的 RNN 隐藏层结构基础上增加了遗忘门、记忆门、输出门 3 个门控装置以及细胞状态,有效解决了 RNN 可能带来的梯度消失与梯度爆炸问题[16]。RNN、LSTM 网络隐藏层结构分别如图 2、图 3 所示。

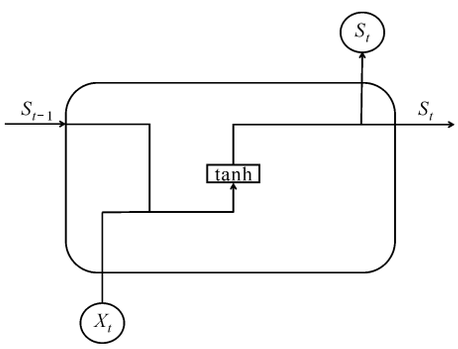

图 2 RNN 隐藏层基本结构

Fig. 2 Basic structure of RNN hidden layer

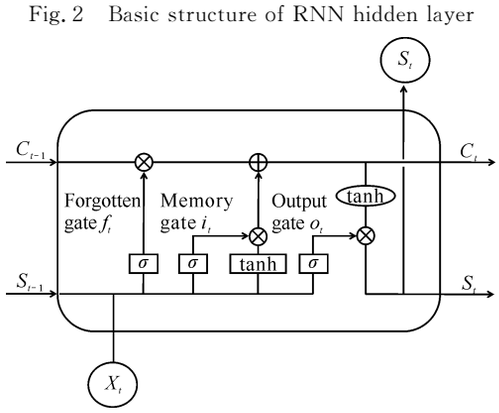

图 3 LSTM 网络隐藏层基本结构

Fig. 3 Basic structure of LSTM network hidden layer

由图 3 可见,t 时刻的输入为 X_t,细胞状态为 C_t,隐层状态为 S_t,遗忘门为 f_t,记忆门为 i_t,输出门为 o_t。其中细胞状态用于保存先前节点信息,隐层状态为当前输入提供门控信号,3 个门用于遗忘细胞中的原始信息与记忆新信息。具体计算公式如式(6)-(11)所示。

$$\tilde{C}_t=\tanh(W_{XC}X_t+W_{Sc}S_{t-1}+b_c), \tag{6}$$

$$C_t=f_tC_{t-1}+i_t\tilde{C}_t, \tag{7}$$

$$f_t=\sigma(W_{Xf}X_t+W_{Sf}S_{t-1}+W_{Cf}C_{t-1}+b_f), \tag{8}$$

$$i_t=\sigma(W_{Xi}X_t+W_{Si}S_{t-1}+W_{Ci}C_{t-1}+b_i), \tag{9}$$

$$o_t=\sigma(W_{Xo}X_t+W_{So}S_{t-1}+W_{Co}C_{t-1}+b_o), \tag{10}$$

$$S_t=o_t\tanh(C_t), \tag{11}$$

式(6)-(11)中,$\tilde{C}_t$ 为临时细胞状态,W_{XC}、W_{SC} 分别表示神经元从输入 X、隐藏层 S 到细胞状态 C 之间的连接权值,W_{Xy}、W_{Sy}、W_{Cy} 分别表示神经元从输入 X、隐藏层 S、细胞状态 C 到 y(y=遗忘 f,记忆 i,输出 o)之间的连接权值,b_c、b_f、b_i、b_o 分别为每个层的偏置量,σ 与 tanh 为激活函数。LSTM 网络权值在更新时将 RNN 权值更新时的 $\frac{\partial S_t}{\partial S_i}$ 替换为 $\frac{\partial C_t}{\partial C_i}$,将其展开如式(12)所示:

$$\frac{\partial C_t}{\partial C_{t-1}}=\frac{\partial C_t}{\partial f_t}\frac{\partial f_t}{\partial S_{t-1}}\frac{\partial S_{t-1}}{\partial C_{t-1}}+\frac{\partial C_t}{\partial i_t}\frac{\partial i_t}{\partial S_{t-1}}\frac{\partial S_{t-1}}{\partial C_{t-1}}+\frac{\partial C_t}{\partial \tilde{C}_{t-1}}\frac{\partial \tilde{C}_{t-1}}{\partial S_{t-1}}\frac{\partial S_{t-1}}{\partial C_{t-1}}+\frac{\partial C_t}{\partial C_{t-1}}。 \tag{12}$$

在 RNN 的梯度更新过程中,$\frac{\partial S_t}{\partial S_{t-1}}$ 会一直处于[0,1]区间内或者一直大于 1,连乘时梯度值趋于 0 或者趋于无穷,造成梯度消失或梯度爆炸。但是不论 $\frac{\partial C_t}{\partial C_{t-1}}$ 取值大于 1 还是处于[0,1],都可以通过调整遗忘门 f_t 的值,将 $\frac{\partial C_t}{\partial C_{t-1}}$ 的值控制在 1,缓解梯度消失与梯度爆炸问题。然而,LSTM 网络单向的传递过程缺乏对音频等具有前后联系的序列同时学习前后时间信息的能力。

1.3 BiLSTM 网络

BiLSTM 网络由两个反向传播的 LSTM 网络构成,结构如图 4 所示。由图 4 可见,t 时刻的输入 X_t 同时进入前向传播的 LSTM 网络与后向传播的 LSTM 网络,因此其可以同时学习 $t-1$、$t+1$ 两个时刻的信息,并综合输出 t 时刻隐层状态。

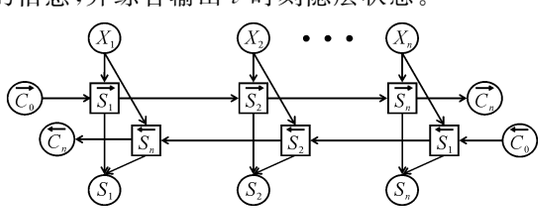

图 4 双向长短时记忆网络传播过程

Fig. 4 Propagation process of bidirectional long short-term memory network

对于具有强前后联系的音频时间序列,选用

BiLSTM 网络可以同时学习某时刻前、后两方向的信息,加强信息学习能力以提高分类准确率。但 BiLSTM 网络本质还是一个长短时记忆网络,仍缺乏对时间序列全局信息的学习能力,并且在学习某时步信息时,太过于依赖前一时刻的计算结果。

2 基于 RF-BiLSTM-Attention 的音乐分类模型

音乐频谱特征是音乐本质的体现,但音频特征维度过大不利于后续模型训练,因此提取全部光谱特征后,采用能够保留原始音频信息的非线性降维方式 RF 进行特征挑选。使用 BiLSTM 网络对具有强时间序列的音频文件进行训练,能够获取局部时间信息,但其在长时间序列上仍会产生信息损失,并且计算过程中过于依赖序列前后关系。因此本研究提出在 BiLSTM 网络上融入注意力机制,以期增强模型长时间序列学习能力并减少计算时间。

2.1 基于 RF-BiLSTM-Attention 的模型框架

基于 RF-BiLSTM-Attention 的模型包括基于 RF 的数据预处理和 BiLSTM-Attention 模型构建两个部分,模型框架如图 5 所示。为在不增加训练难度的基础上获取最重要的频谱特征,首先提取所有光谱特征;然后利用 RF 对特征重要性打分,挑选重要特征进行分类训练。为增强模型对于音频时间序列前后逻辑以及全局关联的学习,先以适用于时间序列的 BiLSTM 网络为基础模型学习音频局部信息;之后通过在其隐藏层与输出层之间嵌入注意力层,给每个时间帧赋予不同权重,同时学习音频全局信息并使其聚焦于更能反应音乐类型的时间帧,提高分类准确率。其中在 BiLSTM 网络中嵌入注意力机制是核心内容,其融合过程也是研究难点。

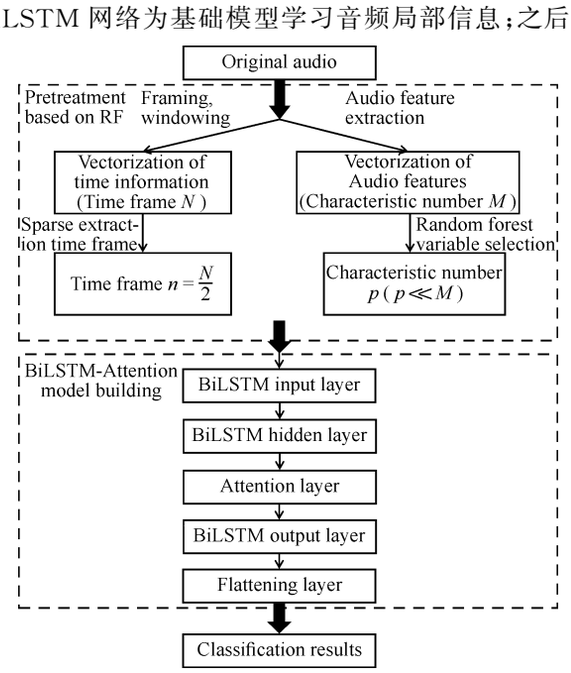

图 5 RF-BiLSTM-Attention 模型框架

Fig. 5 Model framework of RF-BiLSTM-Attention

2.2 基于 RF 的数据预处理

数据预处理是将原本的音频文件转换为包含时间信息与光谱特征信息的向量,并尽量压缩其维度为分类模型训练降低难度。该过程包括音频矩阵生成和 RF 变量选择两个步骤。

2.2.1 音频矩阵生成

音乐信号属于人体语音的一种,要将其转化为分类训练需要的特征,首先对原始音频进行分帧、加窗、傅里叶变换、反傅里叶变换等操作才能获得对应的信号特征。对于给定音乐 S,经过分帧、加窗等,将其划分为 N 个时间帧,对应的时间向量记为 $S(t_1,t_2,\cdots,t_i,\cdots,t_N)$;再通过傅里叶变换、反傅里叶变换等操作获得所有光谱特征 M 个,特征向量记为 $S(a_1,a_2,\cdots,a_i,\cdots,a_M)$。将特征向量与时间向量组合生成一个 M 行 N 列的初始音频矩阵。

一般音频对应的时间与特征都具有较高维度,但高维数据会使得训练难度加大,时间加长。因此,利用稀疏抽取时间帧的方式,即抽取原始时间帧奇数或偶数的时间帧,将原始时间帧 N 降至 $n=N/2$ 帧,使得时间帧维度降为原来的一半,再利用 RF 对所有特征重要性打分并选择重要特征,实现特征降维。

2.2.2 RF 变量选择

Breiman[17] 在 2001 年首次提出 RF 模型,其是一种非线性的降维方式,可以保留原始特征信息。该模型用于特征选择时可以高效地给出特征的重要性得分[18]。通常使用袋外误差,即特征值改变前后的测试集误差率衡量特征的重要程度。特征重要性得分计算过程如图 6 所示。其中 a、c、d 为特征变量。

根据图 6,特征重要性得分计算有以下流程:①在 M 个光谱特征中随机抽取 H 个特征构建 G 棵决策树,再随机选择决策树 g,其中 g 由 D 个特征构成,随机选择特征 G_d,计算测试集(为 M 中未被抽取的部分)误差率 $errOOB_{t1}$;②改变特征 d 的特征值后再次计算测试集误差率 $\mathrm{err}OOB_{t2}$,测试集前、后误差率差值为特征 G_d 在决策树 g 中的重要程度;③遍历 G 棵树中的特征值 d,计算其重要程度的平均值(MDA),用该平均值来衡量该特征在森林中的重要

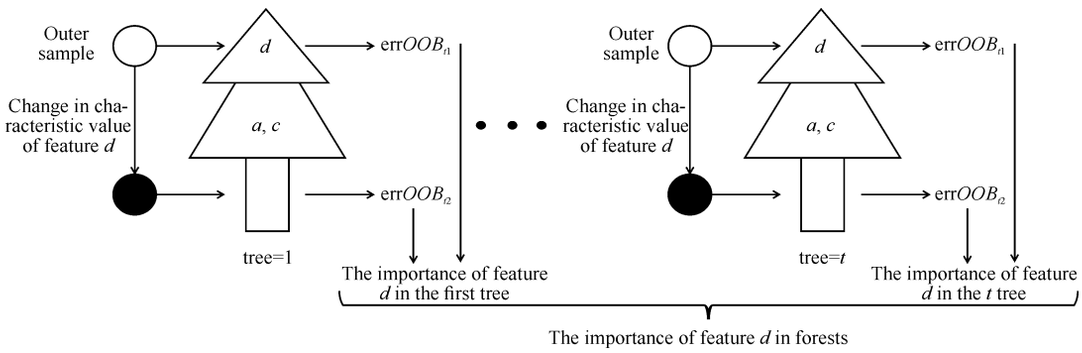

图 6　特征 d 的重要性得分计算流程

Fig. 6　Importance score calculation process for feature d

程度,计算公式如式(13)所示:

$$\mathrm{MDA}(d)=\frac{1}{G_d\ \mathrm{tree}}\sum_{t=1}^{G_d\ \mathrm{tree}}(\mathrm{err}OOB_{t1}-\mathrm{err}OOB_{t2}),\tag{13}$$

其中,G_d tree 为特征 d 在森林中出现的次数,$\mathrm{err}OOB_{t1}$ 为特征 d 的值改变前的测试集误差率,$\mathrm{err}OOB_{t2}$ 为特征 d 的值改变后的测试集误差率。

根据上述步骤计算出所有特征的重要性得分并排序,再选取合适阈值或者所需要的特征个数,从 M 个特征中选取 m 个重要特征。基于 RF 预处理后得到一个 m 行 n 列的音频矩阵,将其作为后续所有模型的输入矩阵进行训练与预测。

2.3　BiLSTM-Attention 模型构建

使用 BiLSTM 对音频进行分类,可以加强模型对音频序列前后关系的学习能力。在此基础上融入注意力机制可加强全局时间的联系,并突出特定时间帧对于音乐分类的重要程度。

2.3.1　BiLSTM-Attention 模型流程

音乐 S 经过数据预处理后得到一个 $m\times n$ 的二维矩阵,将其作为 BiLSTM-Attention 模型的输入进行训练,BiLSTM-Attention 模型训练具体流程如图 7 所示。音乐 S 产生的 $m\times n$ 的二维矩阵进入 BiLSTM 网络输入层后按时间展开为 n 个时间步,其中每个时间步包含 m 个特征,音乐 S 第 i 个时间步的向量记为 $X_i=(a_{S1},a_{S2},\cdots,a_{Sm})_i^{\mathrm{T}}$。输入数据进入 BiLSTM 隐藏层进行训练,学习局部时间信息后从隐藏层输出新向量 $X_i'=(a'_{S1},a'_{S2},\cdots,a'_{Sm})_i^{\mathrm{T}}$,将其直接输入注意力机制层,计算每个时间帧的注意力得分 b^i,为每个时间帧赋予权值,将每帧的权值与经过 BiLSTM 网络学习所得到的信息相结合,得到最终学习结果 $X_i''=(a''_{S1},a''_{S2},\cdots,a''_{Sm})_i^{\mathrm{T}}$。再组合 n 个时间帧的信息,得到 BiLSTM-Attention 模型的输出,进入展平层以及分类器输出最终分类结果。

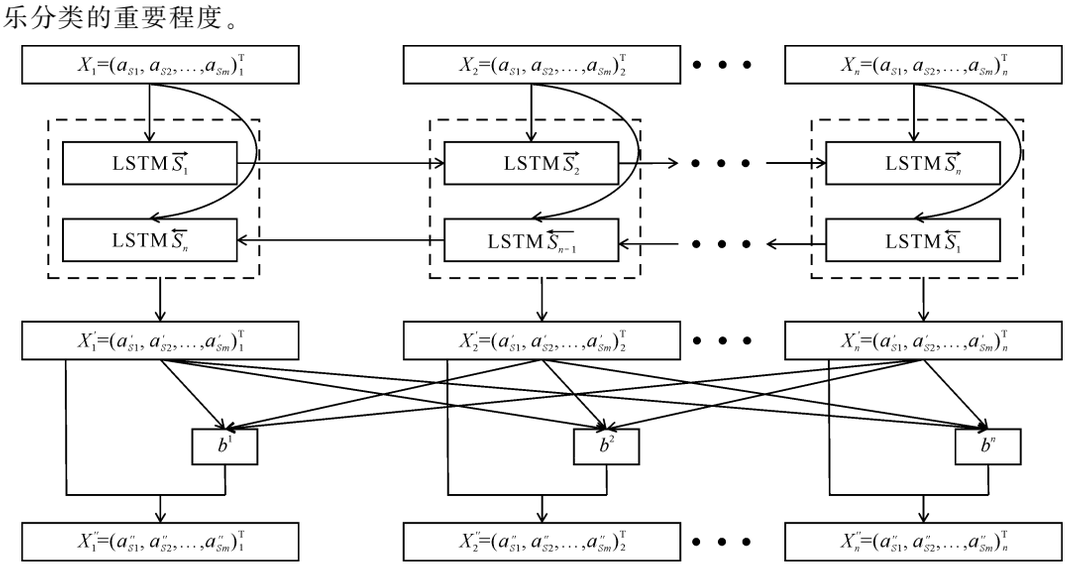

图 7　BiLSTM-Attention 模型流程框架

Fig. 7　Process framework of BiLSTM-Attention model

2.3.2 注意力机制计算原理

注意力机制在学习某时间步信息时,同时考虑所有时间步产生的影响,并且不依赖前后时间步计算出的信息[19],弥补了原始 BiLSTM 网络在学习过程中无法考虑全局带来的信息损失,解决了依赖前个时间步计算结果造成的运算速度慢的问题。

注意力机制中每个输入都会引申出 3 个向量:q、k、v,其中 q 为查询向量 query,k 为键向量 keys,v 为值向量 value。通过 3 个向量与输入的计算,得出最终的注意力得分。第 i 个时间步的注意力得分 b^i 计算过程如图 8 所示。

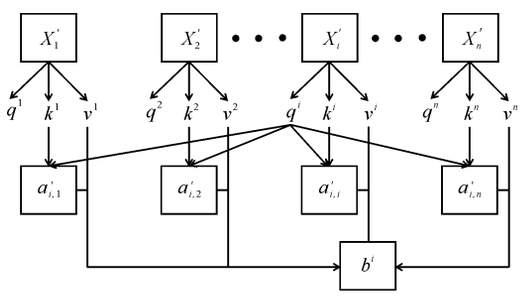

图 8 时间步 i 的注意力得分计算流程

Fig. 8 Attention score calculation process of time step i

图 8 中,q^i、k^i、v^i 和 $\alpha'_{i,j}$、b^i 分别为对应时间步的 q、k、v 向量和注意力权值、注意力得分。计算公式如式(14)所示:

$$q^i = W^q \cdot X_i;k^i = W^k \cdot X_i;v^i = W^v \cdot X_i;\alpha'_{i,j} = \text{softmax}(q^i k^j);b^i = \sum_j \alpha'_{i,j} v^i, \tag{14}$$

其中,W^q、W^k、W^v 为对应的权值矩阵。

经过 BiLSTM 网络隐藏层学习后得到 X'_1,X'_2,…,X'_n,计算得到每个时间步所对应的 q、k、v 向量,将第 i 个时间步对应的 q 向量分别与自身以及其他时间步对应的 k 向量相乘,得到每个时间步对应的权值 $\alpha_{i,j}$,再通过软最大化函数 softmax 进行转化得到权值 $\alpha'_{i,j}$,最后将其与对应时间步的 v 向量相乘再相加得到第 i 个时间步的注意力得分 b^i。通过以上步骤,计算出每个时间步对应的注意力得分,实现 BiLSTM-Attention 模型中注意力机制的嵌入。

3 实验结果与分析

基于上述模型框架,将 RF-BiLSTM-Attention 模型、BiLSTM-Attention 模型分别与 RNN、LSTM、BiLSTM 等 9 种模型进行京剧二分类实验及对比实验。

3.1 实验数据集

选取来源于大众音乐平台的京剧和 5 种戏曲,以及 GTZAN 数据集上提供的 10 种音乐流派歌曲,共计 1 500 首歌曲进行实验。采样歌曲类型及数量见表 1。

表 1 样本类型与数量统计

Table 1 Sample type and quantity statistics

音乐流派 Music genre		数量 Number	合计 Total
Beijing Opera		750	750
Other	Ping Opera	50	750
	Huangmei Opera	50	
	Henan Opera	50	
	Shaanxi Opera	50	
	Shaoxing Opera	50	
	GTZAN dataset 10 music genres (50 songs selected for each genre)	500	

如表 1 所示,本研究选取京剧与其他类型音乐各 750 首,保证了数据集的平衡性;且其他音乐类型中包含戏曲和流行音乐,增加了数据集的丰富度,加大了分类可信度。将上述两类音频文件对应进行独立热编码,“京剧”标签为 0,“其他”标签为 1。

3.2 模型评价指标

在分类问题中,准确率是衡量模型分类好坏标准最常见的指标,它能够直接反映正确分类的比例。由于本研究为二分类问题,因此采取混淆矩阵定义的准确率对京剧二分类模型分类好坏进行评判。混淆矩阵由真实值 Positive (TP)、Negative (TN)以及模型认为的 Positive (FP)、Negative (FN) 4 个值所组成,可计算准确率(Accuracy),计算公式如式(15)所示:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}。 \tag{15}$$

3.3 基于 RF 的音频数据预处理

截取原始音频文件中每首歌曲的前 10 s,并划分为 431 个时间帧。每个时间帧抽取 14 大类共计 205 维光谱特征。经过稀疏抽取时间帧与 RF 降维后,得到的后续模型输入矩阵为 216 行、17 列。其中通过 RF 所给出的特征重要性得分如图 9 所示,降维后的特征如图 10 所示。

从图 9 可以看出,特征重要性得分小于 0.01 的特征占总特征的 85%,且该部分特征的重要性得分差别不大,因此选取 0.01 作为特征筛选的阈值。由于梅尔频谱特征共 128 维,能够提供丰富的频域信息与时域变化,且涵盖声音特性信息,因此图 10 中筛选

出的 17 维特征中 71%为梅尔频谱特征。

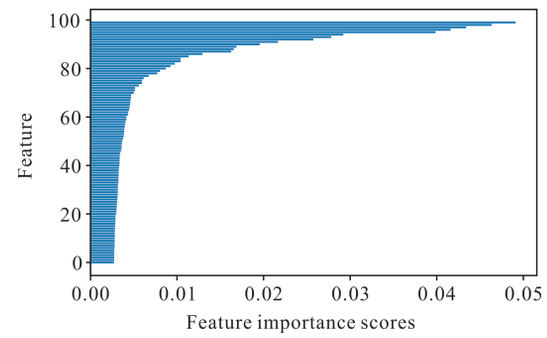

图 9　前 100 个特征重要性得分

Fig. 9　Top 100 feature importance scores

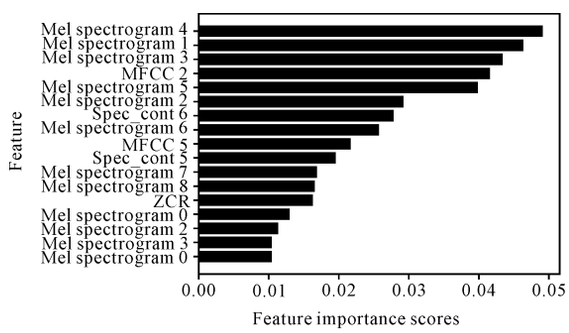

图 10　重要性得分大于 0.01 的特征变量

Fig. 10　Feature variables with importance scores greater than 0.01

3.4　实验结果与分析

将数据划分为训练验证集和测试集,其中训练验证集包含 1 200 首歌曲,测试集包含 300 首。将训练验证集作为模型的输入进行训练,再使用测试集对训练后的模型进行测试和评估。为更好地说明本研究所提出的 RF-BiLSTM-Attention 模型对于分类效果的准确率以及训练时间上的优势,将该模型与 RNN、LSTM、BiLSTM、BiLSTM-Attention 等 9 种模型进行比较,并分别使用未经过 RF 处理的数据与经过 RF 处理的数据进行对比实验,实验结果如表 2、表 3 所示。

当原始特征数据未通过 RF 对音频特征进行筛选,由表 2 可以看出:

①简单模型的平均分类准确率高于复杂模型,且平均运行时间低于复杂模型。说明当数据信息量较大时,相比于复杂模型,简单模型不容易产生过拟合现象,对新数据的泛化能力较好,同时,其更不容易受到数据噪声和不确定性的影响而导致预测错误,更具有鲁棒性。

②在简单模型中,BiLSTM 模型的分类准确率最高,为86.33%,说明BiLSTM模型相较于其他模型具有较好的时间信息学习能力,更适合于音频分类任务。

③在复杂模型中,BiLSTM-Attention 模型分类准确率最高,为 82.67%,说明将 BiLSTM 模型与注意力机制相融合确实能在一定程度上提高音频分类的准确率。

当原始特征数据通过 RF 对音频特征进行筛选后,由表 3 可以看出:

①BiLSTM-Attention 模型的准确率是所有模型中最高的。LSTM 网络改进了 RNN 隐藏层结构,缓解了梯度问题,因此对比 RNN 其分类准确率有明显提升;BiLSTM 模型通过一个前向和一个后向的 LSTM 网络加强了同时学习时间序列前后信息的能力,因此强于 LSTM 模型;而 BiLSTM-Attention 模

表 2　未经过 RF 处理的模型实验结果

Table 2　Model experiment results without RF treatment

神经网络模型 Neural network model		准确率/% Accuracy/%	运行时间/s Run time/s
Simple model	CNN	83.67	185.67
	Attention	71.00	170.00
	GRU	85.00	505.05
	BiGRU	86.00	788.45
	RNN	78.00	243.10
	LSTM	83.30	534.04
	BiLSTM	86.33	807.13
Complex model	CNN + LSTM	74.00	158.52
	BiGRU + Attention	77.00	570.44
	BiLSTM-Attention	82.67	1 035.88

表 3　经过 RF 处理的模型实验结果

Table 3　Model experiment results after RF treatment

神经网络模型 Neural network model		准确率/% Accuracy/%	运行时间/s Run time/s
Simple model	CNN	84.33	3.25
	Attention	79.67	6.06
	GRU	83.00	47.45
	BiGRU	85.33	70.15
	RNN	80.67	20.80
	LSTM	84.33	41.45
	BiLSTM	85.67	54.19
Complex model	CNN + LSTM	85.33	14.17
	BiGRU + Attention	86.00	30.90
	BiLSTM-Attention	89.00	32.22

型在 BiLSTM 模型的基础上引入注意力机制,加强了模型对于时间序列全局信息同时学习的能力,进一步提高了分类准确率。

②BiLSTM-Attention 模型在保证分类准确的同时,运行时间相对较短。由于该模型在 BiLSTM 模型的基础上加入注意力机制,减缓了 BiLSTM 网络对序列前后关系的依赖,因此运行时间相对较短;而 BiLSTM 模型是两个 LSTM 模型的结合,运算复杂,因此运行时间比 LSTM 更长;RNN 由于内部运算简单,运行时间更短,但准确率较低。

③对比所有模型发现,BiLSTM 是所有简单模型中对京剧二分类准确率最高的模型,证明 BiLSTM 模型对音频局部时间信息学习能力较强。单独利用 Attention 模型进行分类训练,准确率较低,但将注意力机制与其他网络模型结合后,均能提高原始模型分类准确率且缩短运行时间,证明注意力机制能学习音频的全局时间联系,加强模型的学习能力并降低计算难度。因此,结合 BiLSTM 与注意力机制后的 BiLSTM-Attention 模型涵盖两个基础模型各自的优势,使得分类准确率达到最优。

根据上述分析可知,本研究提出的 RF-BiLSTM-Attention 模型对于京剧与其他类型音乐的分类准确率最高,运行时间居中。相较于简单模型中分类准确率最好的 RF-BiLSTM 模型,RF-BiLSTM-Attention 模型的分类准确率提高 3.33%,运行时间缩短 21.97 s。

从图 11 可以看出 70%的模型在经过 RF 预处理后再进行训练得到的准确率更高,说明使用 RF 对特征进行筛选,能够有效地保留音频本质信息,并减少由于特征过多带来的信息重复所造成的特征冗余,使得模型的分类准确率增加。从图 12 可以看出所有模型经过 RF 预处理后运行时间都比未预处理的运行时间缩短 90%以上,说明 RF 能够有效筛选出重要特征,实现有效降维,使得原始数据复杂度降低,从而减少运行时间。由上述分析可以看出,将 RF 加入 BiLSTM-Attention 模型中,能够有效地提高模型的运行速度并提高模型分类准确率。

将 RF-BiLSTM-Attention 模型与其部分组件进行比较,可由表 4 得出以下结论:

①加入 RF 对音频特征进行筛选后,能有效降低数据的冗余,减少噪声数据对后续分类准确率造成的影响,以此提高模型分类准确率。同时,筛选特征后,数据维度降低,能使模型训练时间缩短。加入 RF 的模型的分类准确率最高提高 8.67%,运行时间最高减少 96.89%。

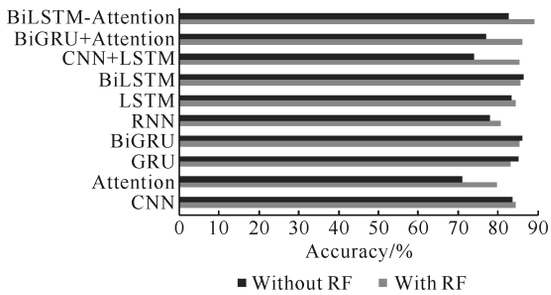

图 11 有无 RF 条件下的准确率对比

Fig. 11 Comparison of RF accuracy with and without RF

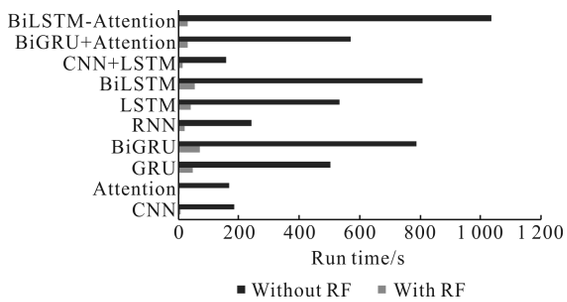

图 12 有无 RF 条件下的运行时间对比

Fig. 12 Comparison of RF operating time with and without RF

表 4 各个模型组件的消融实验结果

Table 4 Results of ablation experiments on various model components

神经网络模型 Neural network model	准确率/% Accuracy/%	运行时间/s Run time/s
BiLSTM	86.33	807.13
RF-BiLSTM	85.67	54.19
Attention	71.00	170.00
RF-Attention	79.67	6.06
BiLSTM-Attention	82.67	1 035.88
RF-BiLSTM-Attention	89.00	32.22

②仅在 BiLSTM 模型的基础上融合 Attention 时,BiLSTM-Attention 模型比 BiLSTM 模型的分类准确率降低 3.66%,说明在数据较为复杂时,由于 BiLSTM-Attention 模型对数据有更好的学习能力,因此更易产生过拟合现象,导致模型的分类准确率下降。相比于 RF-BiLSTM 模型,RF-BiLSTM-Attention 模型的分类准确率提高 3.33%,运行时间缩短 40.54%。这是由于注意力机制能够学习音频的局部信息并且不依赖时间帧之间的关系,因此能够提高模型分类准确率并缩短运行时间。

经过模型的各个组件的对比,表明经过 RF 预处

理后的数据再利用耦合了注意力机制的 BiLSTM 网络,即本研究所提出的 RF-BiLSTM-Attention 模型,在本研究的分类任务中表现出最佳性能,证实了 3 个组件的协同效应。

4 结论

本研究提出一种 RF-BiLSTM-Attention 模型用于京剧与其他类型音乐的二分类研究。该模型通过随机森林选择特征变量得分较高的有效特征,然后利用 BiLSTM-Attention 同时学习音频局部时间信息与全局时间信息。以京剧与其他类型音乐为算例,将 RF-BiLSTM-Attention 模型与 RNN、LSTM、BiLSTM 等 9 种模型进行对比实验,证实 RF-BiLSTM-Attention 模型可以通过随机森林在不影响时序全局建模的前提下,使用袋外误差衡量初始特征重要性,刻画同序列下的特征重要程度,实现特征维度的降低,减少运行时间;并可利用 BiLSTM-Attention 学习音频的时间信息,减少因为位置差异而产生的信息损失,增加特征拟合能力,提高模型分类准确率。但该模型在京剧分类时仅根据音频提取出的特征进行分类训练,京剧统计特点及特定特征对于京剧分类效果的影响还有待进一步研究。

参考文献

[1] 王文照.北京京剧票房运营调研报告:以"北京新荣春青年京剧社"为例[D].北京:中国艺术研究院,2021.

[2] 袁思源.X 大学附中"京剧社"发展现状研讨[D].重庆:西南大学,2018.

[3] SAUNDERS J. Real-time discrimination of broadcast speech/music [C]//1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Atlanta, GA, USA: IEEE, 1996, 2: 993-996.

[4] SCHEIRER E, SLANEY M. Construction and evaluation of a robust multifeature speech/music discriminator [C]//1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. Munich, Germany: IEEE, 1997, 2: 1331-1334.

[5] LIM C, CHANG J H. Efficient implementation techniques of an SVM-based speech/music classifier in SMV [J]. Multimedia Tools and Applications, 2015, 74(15): 5375-5400.

[6] BIRAJDAR G K, PATIL M D. Speech and music classification using spectrogram based statistical descriptors and extreme learning machine [J]. Multimedia Tools and Applications, 2019, 78: 15141-15168.

[7] BIRAJDAR G K, PATIL M D. Speech/music classification using visual and spectral chromagram features [J]. Journal of Ambient Intelligence and Humanized Computing, 2020, 11: 329-347.

[8] 姚斯强,胡剑凌.线性判别分析和支持向量机的音乐分类方法[J].电声技术,2006,12:6-10.

[9] 万凌艳.基于特征提取与 SVM 的多乐器信号快速识别[J].自动化与仪器仪表,2022,26:17-21,26.

[10] TU W P, YANG Y H, DU B, et al. RNN-based signal classification for hybrid audio data compression [J]. Computing, 2020, 102: 813-827.

[11] ZHANG Z X, LIU D, HAN J, et al. Learning audio sequence representations for acoustic event classification [J]. Expert Systems with Applications, 2021, 178: 115007.

[12] 郭毓博,陆军,段鹏启.基于深度学习的竹笛吹奏技巧自动分类[J].应用科学学报,2021,39(4):685-694.

[13] YU W, HUA M, ZHANG Y. Audio classification using attention-augmented convolutional neural network [J]. Knowledge-Based Systems, 2018, 161: 90-100.

[14] NOUMIDA A, RAJEEV R. Multilabel bird species classification from audio recordings using attention framework [J]. Applied Acoustics, 2022, 197: 108901.

[15] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2618.

[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

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

[18] GENUER R, POGGI J M, TULEAU-MALOT C. Variable selection using random forests [J]. Pattern Recognition Letters, 2010, 31(14): 2225-2236.

[19] NIU Z Y, ZHONG G Q, YU H. A review on the attention mechanism of deep learning [J]. Neurocomputing, 2021, 452: 48-62.

(下转第 339 页 Continued on page 339)

A River Ship Target Detection Algorithm Based on YOLOv3 and Transfer Learning

LI Hui[1],JIANG Quan[2**]
(1. School of Electronic Information,Guangxi Minzu University,Nanning,Guangxi,530006,China;2. School of Artificial Intelligence,Guangxi Minzu University,Nanning,Guangxi,530006,China)

Abstract:In order to strengthen the monitoring and tracking of granular material transportation ships in river monitoring video images,so as to assist in the realization of intelligent and efficient river sand mining supervision and granular material scheduling,a river granular material ship target detection algorithm is proposed based on You Only Look Once version 3 (YOLOv3) algorithm and transfer learning. Firstly,the COCO dataset is used to train the initial YOLOv3 algorithm,and the pre-training weight of the model is obtained. Then,the image data of sand mining and sand transportation ships collected from the monitoring equipment around the important rivers in Guangxi are processed to obtain a high-quality ship dataset. Finally,driven by this dataset,the transfer learning pre-training weight is used to train the YOLOv3 detection model for key targets such as river sand mining ships. The model uses Darknet-53 as the backbone network and integrates multi-scale feature maps to achieve the detection of small,medium and large targets. The experimental results show that the average accuracy value and detection speed of the algorithm on the test set reach 98.00% and 17.78 fps,respectively,which is of practical significance for improving the supervision efficiency of river sand mining and realizing the intelligent scheduling of granular materials.
Key words:YOLOv3;transfer learning;granular material ship;target detection;sand mining supervision
责任编辑:陆　雁,陈少凡

(上接第 330 页 Continued from page 330)

Beijing Opera Binary Classification Simulation Based on RF-BiLSTM-Attention Music Classification Method

GONG Yicheng[1,2**],LIU Qing[1,2],XIAO Haoyi[1]
(1. School of Science College,Wuhan University of Science and Technology,Wuhan,Hubei,430065,China;2. Hubei Province Key Laboratory of Systems Science in Metallurgical Process (Wuhan University of Science and Technology),Wuhan,Hubei,430081,China)

Abstract:In order to popularize Beijing Opera,a music classification method of RF-BiLSTM-Attention was proposed based on Random Forest (RF) coupled Attention mechanism and Bidirectional Long Short Term Memory (BiLSTM) network,which was used to classify Beijing Opera and other types of music (hereinafter referred to as 'Beijing Opera binary classification'). Firstly,all spectral features of music were extracted,and important features were selected by RF. Then,the attention layer was embedded the hidden layer and the output layer of the BiLSTM network to classify,train and predict the data. Using 1 500 pieces of music from the popular music platform and GTZAN dataset for Beijing Opera binary classification experiments,the effect of RF on 9 models such as Recurrent Neural Network (RNN),Long Short Term Memory (LSTM) and BiLSTM was compared. The results showed that the classification accuracy of RF-BiLSTM-Attention was 89.00%,with a run time of 33.22 s. Compared with the best performing RF-BiLSTM model in the simple model,the classification accuracy was improved by 3.33%,and the run time was reduced by 40.54%. Compared with the original BiLSTM-Attention model,the classification accuracy was increased by 6.33%,and the run time was shortened by 96.89%. Compared with traditional audio classification work,this article considers the binary classification of Beijing Opera,which plays a good role in promoting Beijing Opera.
Key words:Beijing Opera;BiLSTM network;attention mechanism;random forest;binary classification
责任编辑:梁　晓