基于注意力 LSTM 的音乐主题推荐模型

贾 宁 郑纯军
(大连东软信息学院 辽宁 大连 116023)

摘 要 针对传统音乐推荐过程中存在的分类准确率较低、周期较长、难以满足人们在生活中对主题音乐的需求等问题,设计了一种注意力机制与长短期记忆(Long Short-Term Memory,LSTM)相结合的神经网络模型,它由音乐主题模型和音乐推荐模型构成,在使用注意力机制和 LSTM 网络实现音乐情感分类的基础上,音乐主题模型有效地组合了音频码本和主题模型,实现了对某个情感下的音乐主题子类的判别。音乐推荐模型则利用低级描述符(Low-Level Descriptor,LLD)和频谱图,构建手工特征与卷积循环神经网络(Convolutional Recurrent Neural Network,CRNN)特征的联合表示形式,从而获得用户语音表达的情感,并对其进行精准的音乐主题推荐。实验中,针对两个模型分别进行设计,采用两种不同的传统模型作为基线,实验结果表明,与传统的单一模型相比,此模型不仅可以提升主题分类精度,而且可以精准地判断用户语音数据的情感,从而定向地完成主题音乐的推荐。

关键词 音乐主题推荐,长短期记忆网络,注意力机制,卷积循环神经网络,低级描述符,主题模型
中图法分类号 TP183 **文献标识码** A

Model of Music Theme Recommendation Based on Attention LSTM

JIA Ning ZHENG Chun-jun
(Dalian Neusoft University of Information,Dalian,Liaoning 116023,China)

Abstract Aiming at the problems of low classification accuracy,long period,and difficulty in meeting the demand for theme music in people's life,an attention mechanism and LSTM (Long Short-Term Memory) were designed. Based on the neural network model,it consists of a music theme model and a music recommendation model. On the basis of using the attention mechanism and the LSTM network to realize music emotion classification,the music theme model effectively combines the audio codebook and the topic model to achieve Discrimination of a subcategory of music topics under an emotion. In the music recommendation model,a low-level descriptor and a spectrogram are used to construct a joint representation of manual features and Convolutional Recurrent Neural Network (CRNN) features. The emotions expressed by the user's voice are obtained,and the user is given a precise music theme recommendation by using this model. In the experiment,two models were designed separately,and two different traditional models were used as the baseline. The experimental results show that this model not only can improve the classification accuracy of the subject,but also can accurately judge the emotion of the user's voice data,so as to achieve the recommendation of the theme music compared with the traditional single model.

Keywords Music theme recommendation,Long short-term memory network,Attention mechanism,Convolutional recurrent neural network,Low-Level descriptor,Topic model

1 引言

随着音乐流媒体和自媒体的发展,在网络上流行各种形式的海量音频数据。面对检索范围巨大的音乐数据,如何快速地搜寻到用户喜爱的音乐是目前的热点之一,还催生了音乐检索与音乐推荐等相关领域[1]。在音乐推荐领域中,按照相关主题进行音乐分类是一切研究的基础。由于音乐的种类、形式等信息繁多且存在差异性[2],因此对于音乐的分类是一项极其复杂的工作,而且常出现分类准确度低、周期长等问题[3]。基于此,本文选择语音信号的基础特征(情感)为依据进行音乐的分类与推荐。

针对语音信号的情感主要包含两个领域的研究:音乐特征提取和分类模型搭建[4]。许多学者在特征提取方向进行了大量的研究,其目标是寻找标准的特征来表征音乐自身的情感。例如 Trabelsi 等[5]在音频信号提取过程中常采用梅尔频率倒谱系数(MFCC),Li 等[6]采用多贝西小波系数直方图进行声学信号特征的提取,并验证其有效性。Lee 等[7]提出了一种结合 SFS 特征选择的鲁棒特征提取方法:多特征聚类(MFC),以体现更稳定、更高效的分类性能。Du 等[8]从音频信号中提取频谱图特征进行分层,以提高分类的准确性。上述特征不同程度地体现了音乐本身表达的情感,但是由于特征过于单薄,在实际应用过程中存在精度较低、噪声影响较大、鲁棒性较差等问题[9]。反之,若采用多维的特征组合,则会导致超大规模的计算量,此时一般选择成熟的深度学习框

本文受辽宁省自然科学基金项目(20180551068)资助。
贾 宁(1985—),女,副教授,主要研究方向为大数据分析、深度学习等,E-mail:jianing@neusoft.edu.cn(通信作者)。

架并与之结合使用。例如,Huang 等[10]将卷积神经网络(Convolutional Neural Networks,CNN)与注意力机制相结合,实现了对音乐中富有情绪化部分内容的高亮。Mirsamadi 等[11]将本地注意力机制加入循环神经网络(Recurrent Neural Network,RNN)中,通过集中提取与情感相关的短时帧级声学特征,来实现对说话者情绪的自动识别。

在原始声学信号分解的基础上,总结与情感识别问题相关的特征提取的 3 类方法:1)从原始音频文件中提取信号特征[12],捕获最原始的声学特征;2)直接在原始音频波形上运行的深度学习模型[13];3)利用自动语音识别(Automatic Speech Recognition,ASR)技术将语音转换为文本,然后基于文本完成分析[14],此种方法的成功率较低。目前,以上 3 种方法处于平行探索中,鲜有交叉领域研究存在。

基于此,本文提出了一种基于注意力机制和 LSTM 的神经网络模型进行音乐主题推荐,此网络模型由音乐主题模型和音乐推荐模型构成,前者将注意力机制、LSTM、音频码本和主题模型有效地结合在一起,实现了音乐情感分类与音乐主题分类,后者融合特征提取的前两类方法进行用户的情感识别,形成了低级描述符特征与 CRNN 特征的联合表示形式,在此基础上进行音乐的推荐。

针对多个模型,本文使用了多组对比测试,通过实验结果验证了音乐主题模型和用户情感识别的有效性,同时发现用户情感识别的特征提取方法的效果更佳。

2　基于注意力 LSTM 的音乐主题模型设计

基于注意力 LSTM 的音乐主题模型设计流程如图 1 所示。首先针对已有训练集,提取多维语音相关特征,将其馈送至注意力 LSTM 网络中,完成此模型的构建。此外,系统将自动从指定网站中爬取新增的音乐文件及其相关信息,利用此模型完成情感相关的分类。由于此次分类的颗粒度较粗,因此需要利用主题模型来进一步细化音乐类型,从而得到该音乐所属的子类,即主题类别。

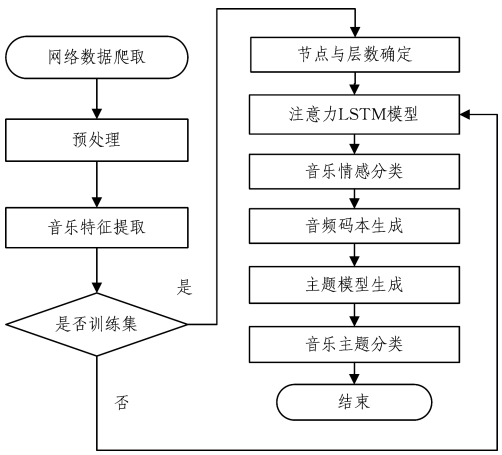

图 1　音乐主题模型设计流程

2.1　音乐数据爬取

针对互联网中的海量音乐文件,本文采用的策略是在规定范围内的音乐资源网站中自动爬取新增的音乐资源,在对其进行数据检查、解析处理的基础上完成数据存储。数据爬取过程是基于 Hadoop 框架[15]实施的,在已知 URL 路径地址的情况下,判断网站对于音乐资源的组织方式,在此基础上爬取海量音频,值得注意的是,必须在爬取过程中确保数据的完备性与正确性,从而避免因网站服务器的不稳定因素而导致的数据获取失败。基于此,每次向服务器发起请求时,应确保获取成功的应答信息,以确保请求被识别。

在数据准备环节中,如何设计爬取策略显得尤为关键。本文采用深度优先的形式获取数据,依照此种策略,在爬取时首先需要总结网站 URL 地址的规律,从中提取部分差异性的路径,将其页面拆分为多组子页面,并放入不同的线程中,以网易云歌曲网站为例,该平台中可利用的音乐文件达到数万个,利用 URL 地址中的参数,在提出请求的基础上获取音乐的基本信息(ID),在解析过程中,首先要完成 4 个相关加密方法的解码操作,然后让任务自动搜索数据库中新添加音乐的 ID,从而实现音乐的自动下载。当任务量较大时,可设计任务列表,将其放入不同的线程中执行。

在解析音乐网站数据时,可以使用差异化的网址规律进行目标音乐的提取,由于不同网站之间的架构存在一定的区分度,针对不同的网站,设计具有特色的解析规则,即提供不同的正则表达式模板,在锁定标签对所包含的数据范围后,解析目标信息。最终,将数据存入 HDFS 中,并自动完成一次备份操作,以便后续数据的存储与查询。

2.2　音乐情感特征提取

在获得音乐数据的基础上,首先需要对音乐的情感特征进行提取,然后进行音乐情感类别的判定,可见选择合适的特征描述符对分类结果至关重要。

本节采用捕获最原始的声学特征的方法,将语音信号转换为语音特征向量,即结合 LLD 和高级统计功能(High-level Statistic Functional,HSF)。常用的 LLD 大致可归纳为 3 种类型:韵律特征、谱特征和音质特征[16]。韵律特征与句法、语篇、信息等结构密切相关,主要包括音高、共振峰、时长、基频、能量等。谱特征是原始信号在声道中激发后产生的测量特征,常见的提取方法有:线性预测系数(LPC)、MFCC 和线性预测倒谱系数(LPCC)等[17]。音质特征主要包括呼吸声音、喉化音、音素、词边界、明亮度等[18]。

许多 LLD 可对音乐类型的判别产生积极的影响,本文选择 MFCC、基音频率、共振峰、频带能量和音乐特有的节奏等特征,在对其进行 HSF 统计的基础上完成特定组合,以此作为音乐情感分类的基本依据。

2.2.1　MFCC 特征

MFCC 从人耳感受响度和幅度的角度出发,体现声道形状在语音短时功率谱的包络特征。它的准确度较高而且抗噪性能较强,可应用于音频分类特征的提取等语音处理领域。

MFCC 特征的计算过程如图 2 所示,在分帧(长度为 512)、加窗的基础上,对每一帧信号做离散傅里叶变换(DFT),计算对数幅度频谱,然后将其经过等带宽的 Mel 滤波器组进行滤波,通过离散余弦变换(DCT)最终得到 MFCC 特征。

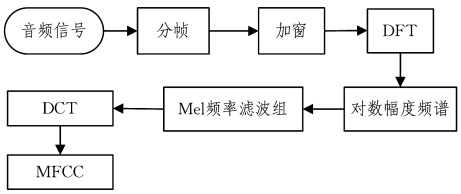

图 2　MFCC 提取方法

MFCC 的计算方法如式(1)所示：

$$MFCC(t,i)=\sqrt{\frac{2}{N}}\sum_{j=1}^{N}\lg[E_{mei}(t,j)]\cos[i(j-0.5)\frac{\pi}{N}] \tag{1}$$

其中，N 为滤波器的数量，$E_{mei}(t,j)$是第 t 时刻、第 j 个滤波器的输出，式(1)的结果为第 t 时刻的 $MFCC$ 参数。

针对音乐信号，MFCC 主要体现为静态特征，其动态特征可以通过 MFCC 特征的一阶差分来体现，基于此，本文提取了 14 维 MFCC 及其一阶差分特征，通过两者融合实现了信息的互补，从而有效地提升了准确率。

2.2.2 基音频率

基音频率一般为声音振动最低频率的表示，其变化形式可以理解为音调，对于音乐而言，音调的差异化非常明显，因此基音频率一般包含音乐的判别信息。本文主要提取当前帧的基音频率，同时记录其最大(小)值、均值、变化范围、标准差、中位数、四分卫数(上、下)等 8 维特征。

2.2.3 共振峰

共振峰(频率)可以理解为声门脉冲通过声道时，在其内部产生共振的频率。该信息包含在频率包络之中，一般认为频谱包络的最大值是共振峰，该信息是音色和音质的重要参数之一。

基于此，本文利用倒谱法计算第一至第三共振峰的平均值、标准差、中位数及其带宽，共获得 12 维特征。

2.2.4 频带能量

由于频带能量的分布体现了音乐强弱的变化程度，而音乐的情感特征往往与能量强弱有关，因此频带能量的分布对音乐类型的判断具有指导意义。本文以 500 Hz 为频带能量区间差值，分别提取 0～4000 Hz 区间内的 8 个频带能量区间的均值，从而获得 8 个维度的特征。

2.2.5 节奏

节奏特征通常包括节拍、拍速和节奏等，可以使用节拍直方图(Beat Histogram)[19]来提取特征，该直方图是对时域信号的若干个滤波的结果，可以从中分别提取两个峰的数值、周期、比值和幅度总和等 6 个特征。节拍直方图的计算流程如图 3 所示。

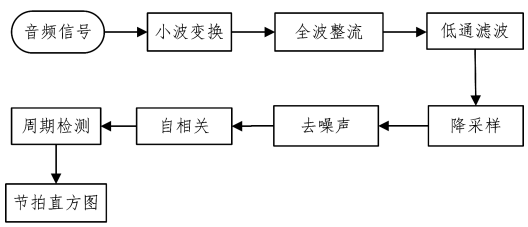

图 3 节拍直方图计算方法

鉴于以上特征均与音乐数据密切相关，本文综合上述 62 维特征，对其进行标签化处理，考虑到 LSTM 的时序性特点，本文将音乐分割成若干个时长为 3s 的段，在每段上提取一个 62 维的特征，从而形成一个特征的序列作为 LSTM 模型的输入，并放入该模型中进行训练。

2.3 基于注意力 LSTM 的音乐情感分类

本节拟实现对于音乐的情感分类，基于此，设计了一个基于注意力 LSTM 结构的神经网络，其包含 2 个 LSTM 层、注意力层和输出层。此模型的整体框图如图 4 所示。

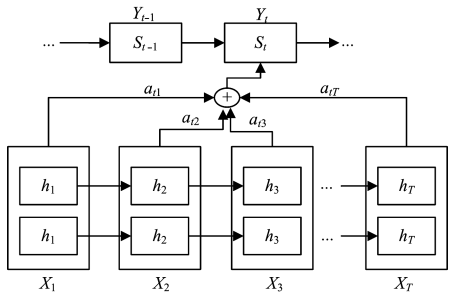

图 4 注意力 LSTM 神经网络整体框图

2.3.1 LSTM 层

传统的 RNN 中添加了与时间结点有关的自我连接形式，以“环”的形式体现，下一个时间步会受到本次的影响，从而彰显了其对于时间序列的建模能力，这种形式与音乐类声学信号随时间变化的特点十分吻合。然而，在自我循环的过程中，传统的 RNN 易产生梯度消失或梯度爆炸问题，因此在其基础上，Hochreiter 和 Schmidhuber 在 1997 年提出了 LSTM 模型，即增加了存储长期有效数据的单元，从而解决了梯度的问题。

在 LSTM 中引入了长时间信息有效性的机制，这些信息有选择性地被控制并保存下来。LSTM 采用的策略是，在每个神经元内部设计输入门、输出门和忘记门。选用误差函数反馈权重，通过忘记门决定记忆单位是否被清除。LSTM 结构式(2)所示：

$$\begin{aligned}
f_t&=\sigma(W_f[h_{t-1},x_t]+b_f)\\
i_t&=\sigma(W_i[h_{t-1},x_t]+b_i)\\
\widetilde{C}_t&=\tanh(W_{\tilde{c}}[h_{t-1},x_t]+b_{\tilde{c}})\\
C_t&=f_t*C_{t-1}+i_t*\widetilde{C}_t\\
o_t&=\sigma(W_o[h_{t-1},x_t]+b_o)\\
h_t&=o_t*\tanh(C_t)
\end{aligned} \tag{2}$$

其中，W_f，W_i，$W_{\tilde{c}}$，W_o 是权重参数，b_f，b_i，$b_{\tilde{c}}$，b_o 是偏置，x_t 作为输入序列，结合上一个隐藏层 h_{t-1}的状态，通过激活函数构成忘记门 f_t。输入门 i_t 和输出门 o_t 也由 x_t 和 h_{t-1}计算。忘记门 f_t 与前单元状态 C_{t-1}联合以确定是否丢弃信息。本节采用了双层 LSTM 模型，并将其输出与注意力机制相结合，突出在音乐序列中的重要特征。

2.3.2 注意力层

考虑到人类大脑中对事物的感知是一个有选择性的集中注意力的过程，这种注意力机制可以被应用于深度学习领域，注意力可以被描述为用于分配有限信息处理能力的“选择机制”，它有助于快速分析目标数据，配合信息筛选和权重设置机制，提升模型的计算能力。

对于输入 x 的序列中的每个向量 x_i，可以根据式(3)计算注意力权重 α_i，其中 $f(x_i)$是评分函数。

$$\alpha_i=\frac{\exp(f(x_i))}{\sum_j\exp(f(x_j))} \tag{3}$$

注意力层的输出为 $attentive\_x$，是输入序列的权重之和，如式(4)所示：

$$attentive\_x=\sum_i\alpha_i x_i \tag{4}$$

本文在 LSTM 之后引入注意力机制，可以实现音乐情感类型评判结果与特征输出每一项的对应关系。

2.3.3 输出层

本文中，输入为 62 维与情感相关的特征，输出为该音乐

所属的情感类别,采用 Softmax 函数把输入映射到[0,1]空间中,并且进行归一化操作,保证其和为 1,将最大的概率值作为判别出的音乐情感。按照效价/激活的分类思想,将可判别的情绪类别分为以下 4 项:悲伤、喜悦、愤怒和平静。

此处并未细分每一种情绪,2.4 节将针对每一种情绪进行主题分类,通过二次分类确定当前歌曲的细节情感及其应用场景,便于向用户推荐符合心境的歌曲。

2.4　基于 LDA 的音乐主题分类

在上文生成情感分类的基础上,本节针对表达每一类情感的音乐,使用 K-means 聚类方法,将每首音乐生成音频码本,然后使用隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)获得主题模型,从而获得在当前情感类别下的音乐主题。

2.4.1　音频码本的生成

K-means 是最常见的聚类方法之一,属于无监督的学习方式,其基本思想是,针对给定的样本,根据样本之间距离的大小,将样本划分为 K 个类,保证类内的间距最小,而类之间的距离尽可能的最大。步骤实现如下:

假设 K 个类中心分别是 $a_1,a_2,\cdots,a_k$,每个类的样本数量是 $N_1,N_2,\cdots,N_k$。使用平方误差作为目标函数,得到式(5):

$$J(a_1,a_2,\cdots,a_k)=\sum_{j=1}^{K}\sum_{i=1}^{n}(x_i-a_j)^2/2 \tag{5}$$

若要使平方误差 $J(a_1,a_2,\cdots,a_k)$ 达到尽量小,可对 $J(a_1,a_2,\cdots,a_k)$ 函数求偏导,得到类中点的更新方法,如式(6)所示:

$$\frac{\partial J}{\partial a_j}=\sum_{i=1}^{n}(x_i-a_j)\longrightarrow 0\Rightarrow a_j=\frac{1}{N_j}\sum_{i=1}^{n}x_i \tag{6}$$

在 2.3 节已经获得了 62 维特征向量,使用 K-means 聚类方法获得每个类的中心,本文中的 K 的数值可以采用式(7)获得。

$$K=\sqrt{m/2} \tag{7}$$

其中,m 为训练样本的个数,将训练样本的所有特征进行聚类,生成音频码本,将语音特征转换为音频词。

为了实现 LDA 主题模型的输入,需要将每个训练音频归一化为相同维度的特征向量,本文使用了音频词袋[20](Bag of Audio Words,BoAW)来表示音频中的计数向量。通过计算训练集中每个音频的每帧特征与音频码本中每个音频单词的欧氏距离,来确定每个特征与音频码本中的哪个“音频单词”对应,并计算次数信息,BoAW 针对每一帧音频特征进行计数统计,将其归一化后作为主题模型的输入。

2.4.2　主题模型的生成

LDA 模型是一种非监督主题模型,可以用来识别大规模文档集或语料库中潜藏的主题信息。本文在获取 BoAW 的基础上,使用 LDA 模型生成语音潜藏的主题信息。

本文认为每个音乐文件由一个到多个主题组成,每个主题蕴含多个主题短语,每个主题短语蕴含多个主题单词,而每个主题单词由多个语音单词构成。如果要生成一个音乐文件,其中的每个音频单词出现的概率公式如下:

$$P(\text{语音单词}|\text{音乐})=\sum_{\text{主题单词}}P(\text{语音单词}|\text{主题单词})*P(\text{主题单词}|\text{音乐}) \tag{8}$$

主题模型如图 5 所示,其中 θ 是一个主题向量,在语音中是一个主题单词向量,向量的每一列表示每个主题单词在语音中出现的概率,该向量为非负归一化向量;$p(\theta)$ 是 θ 的分布,具体为 Dirichlet 分布,即分布的分布。

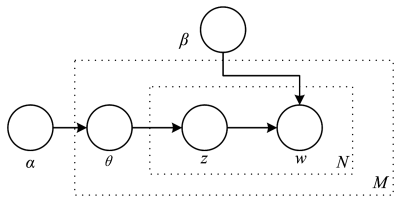

图 5　LDA 主题模型

之前在对音乐情绪进行判断时,获得的分类颗粒度较大,针对每一种情绪的音乐,均需要单独分组来完成对音乐主题的分类,其中分类的主题与场景有关,主题模型的数量是超参数,可通过实验获得最佳方案。

3　音乐推荐模型设计

在获得海量音乐主题的基础上,设计音乐推荐模型,主要由用户情感判断和音乐推荐两个阶段构成。前者主要针对用户提供的语音片段,判断当下用户反馈的具体情感,后者在此基础上,结合音乐主题模型,向用户推荐所属情感分类或主题分类的音乐。具体音乐推荐模型设计流程如图 6 所示。

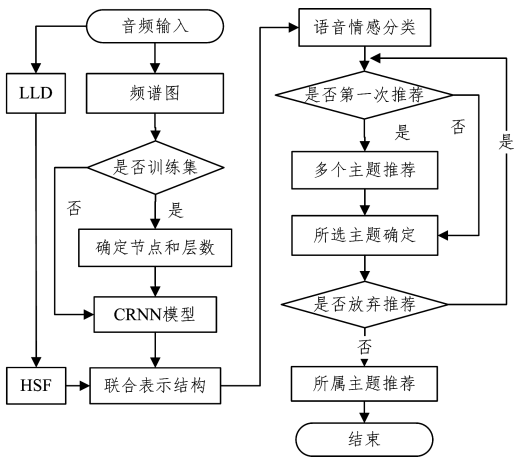

图 6　音乐推荐模型设计流程

3.1　用户情感识别模型

通过 LLD-HSF 和 CRNN 学习的特征,从不同方面描述了情绪的状态,并且它们之间存在互补的特性。本节将以上两种学习特征相结合,提出了一种双通道情感识别模型,两种类型的特征通过隐藏层连接,被投射到同一个空间中。与传统方式相比,增强了情感相关特征的辨别性。

如图 7 所示,作为通道之一,HSF 计算 LLD 的统计数据,从而代表 LLD 的全局动态变化。此部分内容和 2.2 节相似,不再赘述。本节将重点介绍第二个通道 CRNN 及两个通道之间的结合方式。

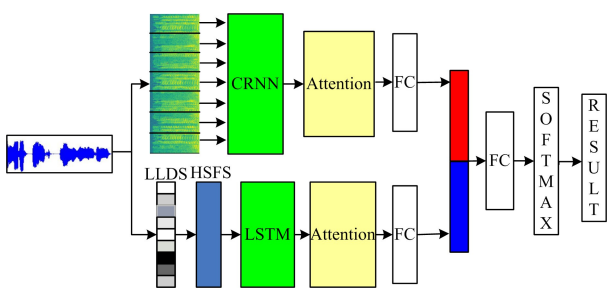

图 7　HSF 与 CRNN 联合模型

3.1.1 频谱图

作为 CRNN 模型的输入，频谱图的横坐标是时间，纵坐标是频率，坐标点值为语音数据能量。由于其采用二维平面表达三维信息，因此能量值的大小是通过颜色来表示的，颜色越深，表示该点的语音能量越强。

本文将以 20 ms 长度为一帧，使用汉明窗加窗，每一帧通过使用快速傅里叶变换计算各个频率的能量值，以 10 ms 为步长进行滑动，分别产生每帧的频谱图，按照时间顺序将所有的频谱图拼接成语音对应的频谱图。

3.1.2 CRNN 模型

CRNN 网络架构由两个部分组成，第一部分是卷积特征提取器，它以频谱图作为输入，对输入图像进行卷积，分为 4 个步骤，即连续的两组卷积和池化，并生成特征图。对于预先分段的语音，可以获得每个片段的 CNN 学习特征。第二部分是 LSTM，其中每个时间步对应于原始音频输入的一段，无需对音频进行削波或填充，而且可以保留段间的长期依赖性。通过最大池化层、最小池化层和平均池化层计算输出的统计数据，并将得到的池化向量连接成一体。

3.1.3 HSF 与 CRNN 联合表示

HSF 与 CRNN 联合表示的设计过程如图 7 所示。整个过程以端到端的方式进行训练，在两个并行通道中同时处理给定的语音。

在一个通道中，在频谱图分段的基础上，把这些段输入到 CRNN 网络。CRNN 输出大小为 384 维向量，然后依次将其加入一个注意力层和一个 128 维的隐藏层。

在另一个通道中，原始波形被分段为帧。从其中提取 LLD 并进一步计算 HSF。随后加入一个 LSTM 层、一个注意力层和一个隐藏层，将高维 HSF 连续映射到低维的特征空间，并最终输出特征向量。

把两个通道输出的两个特征向量连接起来，用作下一个隐藏层的输入，该隐藏层将它们投影到特征空间中再传递给 Softmax 层以进行分类。为了便于后期的情感判断，此处的分类与 2.3.3 节中的分类集合完全吻合。

3.2 音乐主题推荐流程

在获得海量音乐主题分类模型和用户情感分类的基础上，完成对用户的音乐推荐功能。音乐主题推荐具体流程如图 6 所示。具体步骤如下：

(1)根据当前的用户情感，确定其隶属哪种情感分类。

(2)若第一次涉及此情感，则针对当前的情感分类，应用 LDA 主题模型，从每个情感主题中随机挑选一首音乐，组合成歌单为其推荐。然后继续步骤(3)，若非第一次涉及此情感，则直接进入步骤(4)。

(3)用户从当前歌单中挑选音乐，系统确定其喜好的主题。继续步骤(4)。

(4)重复为用户推荐所属情感分类下相关主题的音乐。

(5)如果用户放弃此次推荐，则回到步骤(1)，并清除历史情感信息。

4 实验及其结果分析

4.1 音乐数据集合

本文中涉及的数据均从相关的网站中爬取，经过完备性检查、数据解析与清洗之后，将有效的数据存储至 HDFS 中。获取的数据主要来源于以下几个网站：网易云音乐、QQ 音乐、虾米音乐、酷我音乐等。

为确保下载音乐的唯一性，在获取音频文件的同时，还将存储音乐标题、演唱者、作者、时长、专辑等信息。当检索到新音乐与数据库中的此类信息完全一致时，认定此音乐已存在，无需再次获取。由于音乐情感分类属于有监督的，因此训练集合中的音乐均来自于上述网站中的分类歌单。

分析用户情感时，用户的语音数据由相关软件采集，在获得用户语音的基础上在线识别用户情绪。

4.2 特征及模型参数设置

实验采用了从网络中爬取的 2 000 首音乐，这些音乐被平均分为 4 个情感类别：悲伤、喜悦、愤怒和平静。音乐文件的时长均为 3～5 min，每首音乐被分割成长为 3 s 的若干段。

针对音乐情感分类，本文从音频中提取 28 维 MFCC、8 维基音频率、12 维共振峰、8 维频带能量和 6 维节奏特征，每段音乐(3 s)共有 62 维特征。在设计模型时，采用双层单向 LSTM 网络，Batch_size 为 150，最大 Epoch 为 10 000，同时设置学习速率为 0.001，Dropout 为 0.5。神经元激活函数为双曲正切函数，优化器为 Adam，损失函数为均方误差。

针对音乐主题分类，K-means 聚类时选取 $K=150$，LDA 主题模型的数量设置为 5。

针对用户情感识别过程，一个频谱图段的大小为 50 帧，重叠为 25 帧，段长为 515 ms。CRNN 基础结构包含两个卷积层，第一层具有 30 个大小为 40×5 的滤波器，第二层有 30 个大小为 1×3 的滤波器。两个卷积层均采用 Relu 激活函数，两个最大池化层的大小均为 1×3，而且步幅为 2。在 HSF 通道中，Softmax 层有 4 个节点，其上一个隐藏层有 32 个节点。所有隐藏层都采用 Sigmoid 激活函数，使用交叉熵作为目标函数进行训练。

4.3 实验设计及结果分析

由于从网络中获取的音乐信息量较大，通过歌单列表爬取的音乐标记数据较为准确，而且训练集与测试集的数据类别较为平衡，采用 5 折交叉测试的方式，即将测试集分成 5 份，其中 1 份作为测试样本，剩余 4 份作为训练集，按此种形式重复 5 次，从而使每组数据都成为测试集和训练集。

本文使用 Tensorflow 框架进行网络模型结构的搭建，对音乐进行情感分类和主题分类，并实现为用户推荐音乐。

对音乐情感分类模型进行测试，以 MFCC＋传统的 LSTM 模型作为基线，对照模型 1 为 62 维特征＋传统的 LSTM，模型 2 为 62 维特征＋注意力 LSTM，模型 3 为当前系统(62 维特征＋注意力机制＋双层 LSTM)，分别计算上述模型情绪识别的准确性。

表 1 列出了经过实验验证后，不同的音乐情感分类的准确度。

表 1 音乐情感分类的准确度

模型	平均准确度/%
基线：MFCC＋传统的 LSTM 模型	63
模型 1：62 维特征＋传统的 LSTM	67
模型 2：62 维特征＋注意力机制＋ LSTM	69.8
当前系统：62 维特征＋注意力机制＋双层 LSTM	71.5

由表 1 可知，本文提出的策略在所有模型中的分类表现最好，拥有最优的准确率。其次是带有注意力机制的模型，而

基线的准确度最低。由此可以确定,添加注意力机制与多元化特征组合均对模型产生了较大的影响。

表 2 列出了当前系统中具体情感类别的混淆矩阵结果。由结果可知,对于唤醒度较高的情绪,识别准确度较高,例如喜悦、愤怒等类别,反之,针对平静、悲伤等唤醒度较低的类别,识别准确率较低。

表 2　音乐分类详情

(单位:%)

识别率	悲伤	喜悦	愤怒	平静
悲伤	71.5	5	5.5	18
喜悦	6	75.2	6.8	12
愤怒	10.3	8.9	74.3	6.5
平静	12.2	10.9	13.4	63.5

针对用户情绪识别模型进行测试,以 CRNN 模型为基线,对照模型 1 为 HSF+标准 CRNN,模型 2 为当前系统(HSF+ CRNN(含注意力机制)),分别计算其情绪识别的准确性。

表 3 列出了经过实验验证后,不同的用户情感分类的准确度。

表 3　用户情感分类的准确度

模型	平均准确度/%
基线:CRNN	64
模型 1:HSF+标准 CRNN	68.5
当前系统:HSF+ CRNN(含注意力机制)	72.6

由表 1 可知,本文提出的策略在所有模型中的分类表现最好,拥有最优的准确率。而且,针对用户的情感表达,其识别准确率均高于表 2 中的数值,说明 HSF+CRNN 模型在识别过程中存在一定的优势。

结束语　针对音乐主题分类精度不高、推荐准确度较低等问题,本文设计数据采集、清洗、处理、转存、音乐情感分析、音乐主题分析、用户情感识别、音乐推荐等环节,形成了一套完整的分类问题解决方案。其中,音乐主题模型将注意力机制、LSTM、BoAW 和 LDA 主题模型有效地结合在一起,实现了音乐情感分类与音乐主题分类。在音乐推荐模型中设计了两个通道的融合,分别以低级描述符和语谱图作为输入,构建特征的联合表示,并在此基础上完成对用户的音乐推荐。通过实验验证,本文提出的模型分类准确度较高。

在未来的工作中,将继续自动爬取音乐数据并进行情感和主题分类,设立音乐情感分类数据库,并对外提供相应的接口,同时针对现有的神经网络进一步完成模型优化,以提升分类的精度。

参 考 文 献

[1] VELARDE G, CHACÓN C C, MEREDITH D, et al. Convolution-based classification of audio and symbolic representations of music[J]. Journal of New Music Research, 2018: 1-15.

[2] LAKOMKIN E, ZAMANI M A, Weber C, et al. EmoRL: Continuous Acoustic Emotion Classification using Deep Reinforcement Learning[C]// ICRA'18. 2018.

[3] HE H, XIA R. Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification[C]// Natural Language Processing and Chinese Computing. 2018.

[4] RAJANNA A R, ARYAFAR K, SHOKOUFANDEH A, et al. Deep Neural Networks: A Case Study for Music Genre Classification[C]// IEEE International Conference on Machine Learning & Applications. IEEE, 2015.

[5] TRABELSII, AYED D B. On the Use of Different Feature Extraction Methods for Linear and Non Linear kernels[C]// 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT). IEEE, 2012.

[6] LI T, OGIHARA M, LI Q. A Comparative Study on Content-Based Music Genre Classification[C]// International AcmSigir Conference on Research & Development in Informaion Retrieval. ACM, 2003.

[7] LEE K K, PARK K S. Robust Feature Extraction for Automatic Classification of Korean Traditional Music in Digital Library[C]// International Conference on Asian Digital Libraries: Implementing Strategies & Sharing Experiences. Springer-Verlag, 2005.

[8] DU W, LIN H, SUN J, et al. A new hierarchical method for music genre classification[C]// International Congress on Image & Signal Processing. IEEE, 2017.

[9] DEB S, DANDAPAT S. Multiscale Amplitude Feature and Significance of Enhanced Vocal Tract Information for Emotion Classification[J]. IEEE Transactions on Cybernetics, 2018, PP(99): 1-14.

[10] HUANG Y S, CHOU S Y, YANG Y H. Pop Music Highlighter: Marking the Emotion Keypoints[C]// Audio and Speech Processing. 2018.

[11] MIRSAMADI S, BARSOUM E, ZHANG C. Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention[C]// ICASSP. IEEE, 2017.

[12] BERTIN-MAHIEUX T, ELLIS D P. Large-scale cover song recognition using hashed chroma landmarks[C]// 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2011: 117-120.

[13] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[C]// SSW. 2016: 125.

[14] EZZAT S, EL GAYAR N, GHANEM M M. Sentiment analysis of call centre audio conversations using text classification[J]. International Journal of Computer Information Systems and Industrial Management Applications, 2012, 4(1): 619-627.

[15] PALKAR V V, JOEG P. Proposing scalable method for music genre classification[C]// International Conference on Inventive Computation Technologies. 2017.

[16] 韩文静,李海峰,阮华斌. 语音情感识别研究进展综述[J]. 软件学报,2014,25(1):37-50.

[17] PALO H K, MOHANTY M N, CHANDRA M. Computational Vision and Robotics[J]. Advances in Intelligent Systems and Computing, 2015, 332: 63-70.

[18] RODDY C. Emotion recognition in human-computer interaction [J]. Signal Processing Magazine, IEEE, 2001, 18(1): 32-80.

[19] DAVIES M E P, DEGARA N, PLUMBLEY M D. Measuring the Performance of Beat Tracking Algorithms Using a Beat Error Histogram [J]. IEEE Signal Processing Letters, 2011, 18(3): 157-160.

[20] YUAN C, GLASS J. Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embedding from Speech[C]// Interspeech. 2018.