Computer Science and Application 计算机科学与应用, 2020, 10(12), 2280-2290
Published Online December 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1012240

Hans汉斯

融合LSTM和注意力机制的音乐分类推荐方法

冯鹏宇，陈平华，申建芳

广东工业大学，计算机学院，广东 广州
Email: wwwfffwangyi163@163.com, phchen@gdut.edu.cn, tysjf@gdut.edu.cn

收稿日期：2020年11月23日；录用日期：2020年12月16日；发布日期：2020年12月23日

摘　要

针对音乐资源过于庞大，现有的音乐推荐方法分类准确度不高，对用户情感的识别较模糊导致人们在生活中难以寻找到偏好音乐的问题，本文提出一种将长短期记忆神经网络(Long Short-Term Memory, LSTM)与注意力机制(Attention Model, AM)相融合的音乐分类及推荐方法，该方法由音乐分类模型和音乐推荐模型两部分组成。首先对音频数据的声学特征进行捕获，构成含有多维特征的序列后，通过LSTM神经网络和注意力机制对音乐进行情感分类，接下来采集用户的历史收听记录，选取最近的十首歌曲并生成频谱图，结合CNN (Convolutional Neural Networks, CNN)对用户当前情感进行识别，提升推荐的高效性。实验部分将新提出的模型与其他传统音乐分类模型进行多组对比测试，结果显示与近年来现存的模型相比，新提出的模型明显提升了情感判断及用户情感识别的准确度，音乐推荐的准确度有所增强。

关键词

音乐推荐，音乐分类，长短期记忆网络，注意力机制，卷积神经网络

Music Classification and Recommendation Method Combining LSTM and AM

Pengyu Feng, Pinghua Chen, Jianfang Shen

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong
Email: wwwfffwangyi163@163.com, phchen@gdut.edu.cn, tysjf@gdut.edu.cn

Received: Nov. 23rd, 2020; accepted: Dec. 16th, 2020; published: Dec. 23rd, 2020

Abstract

In view of the huge amount of music resources, the existing music recommendation methods have low classification accuracy, fuzzy recognition of user emotions, and low concentration of target data analysis, which makes it difficult to satisfy people's preference for music in daily life. Due to demand and other issues, a music classification and recommendation method combining Long

文章引用: 冯鹏宇, 陈平华, 申建芳. 融合 LSTM 和注意力机制的音乐分类推荐方法[J]. 计算机科学与应用, 2020, 10(12): 2280-2290. DOI: 10.12677/csa.2020.1012240

Short-Term Memory and Attention Model is proposed. The method consists of a music classification model and a music classification model. The recommended model consists of two parts. First to capture audio data of various acoustic characteristics, constitute a sequence containing multi-dimensional characteristics, through the LSTM Neural network classification of music emotion and attention mechanism; the next, gathering user history to record, select its most recent ten songs and generate the spectrum diagram, combined with CNN (Convolutional Neural Networks, CNN) to accurately identify the user's current emotion, recommend the efficiency of ascension. The experimental part compares the new model with other traditional music classification models, and the results show that compared with the existing models in recent years, the new model significantly improves the accuracy of emotion judgment and user emotion recognition, and the accuracy of music recommendation is enhanced to some extent.

Keywords

Music Recommendation, Music Classification, LSTM, AM, CNN

Copyright © 2020 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着移动技术的不断发展，音乐自媒体的高速发展使得众多的音乐网站都存在大量的用户。目前听音乐已经成为人们在最大程度上在碎片化时间中寻求放松的一种有效方法，难以找到自己喜爱的音乐无疑会降低音乐库的使用效率，体验感不高，影响用户友好度[1]。在这些庞大的音乐库中，怎样快速搜索到用户在最近一段时间内心仪的音乐并提高音乐库的使用效率，提升用户体验是目前大热的研究方向。繁重的音乐信息致使音乐的分类准确度不高，对用户情感判别的实时性较差。对于这个问题，本文选用语音信号的音频特征以及用户收听的历史记录作为依据，来对音乐进行情感分类以及推荐。

就目前的研究现状而言，音频特征和用户的音乐理解存在巨大的语义鸿沟，推荐效果不好[2]。传统音乐推荐方法，例如基于文本标注的推荐算法[3]，需要大量的人工标注，存在耗时长且低效的问题。为此 Huang 等结合了卷积神经网络(Convolutional Neural Networks, CNN)和注意力机制，对音乐中情感表达较明显的部分实现了高亮[4]；Mirsamadi 等融合了注意力机制与循环神经网络(Recurrent Neural Network, RNN)，集中提取了一些与情感相关联的短时帧级声学特征来自动识别说话者的情感[5]；Picza KJ 将简单的卷积神经网络结构用于 Log 梅尔频谱图来分类处理声音[6]；Zhang 等融合卷积网络层结构与混合样本生成的新样本训练网络，极大的提升了频谱识别模型的效果[7]。

近期随着深度学习技术的不断革新，将长短期记忆网络 LSTM 应用于音乐信息检索(Music Information Retrieval, MIR)取得了巨大的成功[8]。同时也总结出基于原始声学信号分解对情感识别问题进行特征提取的两种方法：一是从原始音频文件中捕捉最本质的声学特征，提取出信号特征[9]；二是直接在原始音频上运行深度学习框架。然而仅利用这些数据并不足以跟进用户近期的喜好变化和情绪转换，因此本文提出一种融合 LSTM 和注意力机制进行音乐分类及推荐的方法，其中音乐分类模型高效的结合了音乐本身的多项音频特征，融合 LSTM、注意力机制对音乐进行了情感分类，训练音乐的情感分类模型，接下来对用户的历史收听记录进行筛选，选取最有代表性的数据与 CNN 相结合来对音乐进行推荐。

本文与其它常见的模型作对比，采用了多组对比测试，实验结果可验证音乐情感分类以及用户情感分类的准确性，对用户进行更精准的音乐推荐，实时性较高。

2. 音乐分类模型设计

融合 LSTM 和注意力机制的音乐分类模型设计流程如图 1 所示。从训练集数据中提取出多维突出音乐情感的代表性语音特征，将其传送至 LSTM 注意力机制神经网络中来搭建分类模型，并利用这个模型对音乐进行情感分类。

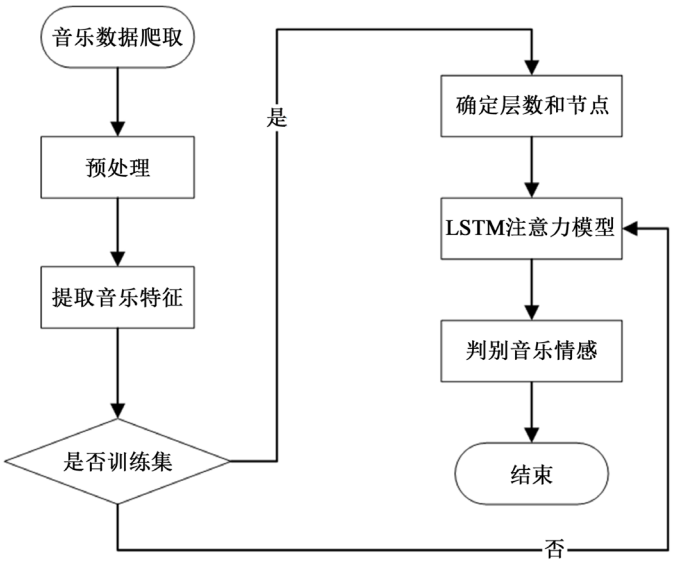

Figure 1. Music emotion classification process
图 1. 音乐情感分类流程

2.1. 提取音频的情感特征

音乐在视觉表现上是以频谱的形式存在的，这里需要选用在频谱信号上最有情感代表性且最合适的特征来提取出情感特征，依此分析出音乐所属的情感类别。本文计划在此处捕捉最原始的声学特征，完成从信号到特征向量的转换。此处需要结合低级描述符 LLD，LLD 作为能够用于描述音频信号本质特征的变量，主要分为韵律特征、谱特征和音质特征三大类[10]。韵律特征是语言的一种音系结构，是情绪表达的重要形式之一，主要包含语调、时域分布和重音三个方面；谱特征是声音及其他信号的视觉表示，最典型的特征提取方法即为梅尔频谱倒谱系数(Mel-scale Frequency Cepstral Coefficients, MFCC)；音质特征主要包含响度，音域、音高、音素和纯度等[11]，本文在这里选用的低级描述符 LLD 包括 MFCC、共振峰、短时能量、基音频率和短时过零率，将这些作为情感分类的特征参数，其主要思想是对音乐音高(频率)、音长(节奏)、响度(重音)的刻画等[12]。

2.1.1. 梅尔频谱倒谱系数

人是通过声带的振动产生声音的，而人的声道决定了发出什么样的声音，其可以在短时的频谱包络中显示出来。MFCC 是基于人耳听觉特性提出的，广泛应用于语音识别(Speech Recognition)和话者识别(Speaker Recognition)方面，根据人类听觉进行感知[13]，相比其它参数具有更好的鲁棒性，当信噪比降低时仍然具有较好的识别性能。

首先对音频数据进行预处理以提取 MFCC 特征，主要包含预加重、分帧、加窗、快速傅里叶变换(Fast Fourier Transform, FFT)、Mel 滤波器组、对数运算和离散余弦变换(Discrete Cosine Transform, DCT)等操作。本文中将对音频数据进行预加重处理，即增强信号上升沿和下降沿处的幅度。另外，由于音乐信号

具有不平稳性，因此为了最大程度上保证信号的完整性，需要对信号进行分帧处理，将每一帧乘以汉明窗来增加左端和右端的连续性，经过 FFT 得到每一帧在频谱上的能量分布，对对数频谱进行计算。接下来对经过等带宽的 Mel 滤波器器组进行滤波处理来对频谱进行平滑化，将上述对数能量带入 DCT 来得到 MFCC 特征。MFCC 特征的提取过程如图 2 所示。

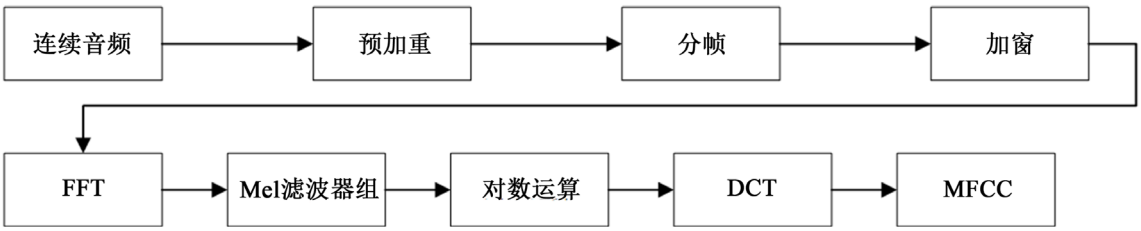

Figure 2. MFCC extraction process
图 2. MFCC 的提取过程

MFCC 是在 Mel 标度频率域提取出来的倒谱参数，Mel 标度描述了人耳频率的非线性特性，它与频率的关系如式(1)所示：

$$\mathrm{Mel}(f)=2595*\lg\left(1+\frac{f}{700}\right) \tag{1}$$

其中，f 为频率，其单位是 Hz。

MFCC 的计算方法如式(2)所示：

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

其中，$\mathrm{MFCC}(t,i)$ 为 t 时刻的 MFCC 参数，N 为滤波器数量，$E(t,j)$ 是 t 时刻第 j 个滤波器的输出。本文在此处共提取了 18 维 MFCC 特征，采用了更符合人耳听觉特征的参数，信号的完整性得到了大幅度提高。

2.1.2. 共振峰

共振峰是在声音的频谱中能量相对集中的一些区域，共振峰不但是音质的决定因素，而且反映了声道(共振腔)的物理特征。它是反映声道谐振特性的重要特征，是语音信号处理中非常重要的特征参数，代表了发音信息最直接的来源，本文捕捉了共振峰的 8 维特征，包含共振峰频率、频带宽度和幅值。

2.1.3. 短时平均能量

短时能量意为较短时间内的语音能量，是音频信号的时域特征。这里的较短时间通常指的是一帧，即一帧时间内的语音能量就是短时能量。语音的短时能量就是将语音中每一帧的短时能量都计算出来。它主要用于区分浊音段和清音段，对声母和韵母进行分界，每一帧中所有语音信号的平方和都可以作为辅助的特征参数用于信号识别，其计算公式如式(3)所示。

$$E_n=\sum_{m=n-(N-1)}^{n}\left[x(m)\omega(n-m)\right]^2 \tag{3}$$

对于每一帧 m 而言，E_n 为 n 时刻的短时平均短时间内的语音能量，$\omega(m)$ 是窗口函数，在这里它是一种方窗，N 为窗长且其选择与语音周期相关，这里拟将帧长设定为 10~30 ms，提取出 10 维特征。

2.1.4. 基音频率

在声音的振动中，由频率最低的振动所发出的声音叫做基音，其余为泛音。基音频率的长短、薄厚、韧性、劲度和发音习惯有关系，在很大程度上反应了声音的特征。针对此特征，本文主要提取了当前帧的 6 维特征，包含最值、均值、方差、中位数和振幅。

2.1.5. 短时平均过零率

过零率(Zero-Crossing Rate, ZCR)是指一个信号的符号变化的比率，是对敲击声音的进行分类的主要特征，一帧音频中信号波形穿过横轴，即改变符号的次数即为过零率。由于信号具有短时性，在本文将信号以 30 ms 为一段分为若干帧进行分析，两帧起始点相隔 5 ms，提取出 10 维特征。

上文提取出的 50 维音频特征均为接下来音乐情感分类的重要依据。另外，LSTM 是一种具有长记忆特征的循环神经网络(Recurrent Neural Networks, RNN)，因此还需要把音乐的时长进行分割，每 3 秒为一段并同时提取出一个 50 维的特征，得到的特征序列即为接下来 LSTM 模型的输入数据。

2.2. 融合 LSTM 和 AM 的情感分类模型

本文计划在这个模型中实现对音乐的情感类别进行划分，融合了双层 LSTM 结构和注意力机制的神经网络模型由三层 LSTM、注意力层及输出层组成。其模型框架如图 3 所示。

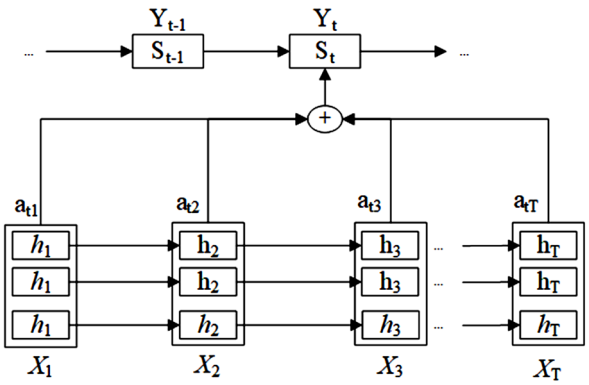

Figure 3. LSTM attentional neural network model
图 3. LSTM 注意力神经网络模型

2.2.1. 三层 LSTM 结构

RNN 会面临梯度消失的问题，梯度是用于更新神经网络的权重值，消失的梯度问题即为在时间的推移传播过程中梯度下降，易导致难以继续学习的问题，而 LSTM 作为 RNN 的一种特殊类型，拥有通过门结构来去除或增加信息到细胞状态的能力。它由 Hochreiter 和 Schmidhuber 于 1997 年提出，其关键在于细胞状态，水平线在上方贯穿运行，非常适合于处理和预测时间序列中间隔和延迟非常长的重要事件。LSTM 的整体架构[14]如式(5)~(10)所示。

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{4}$$

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{5}$$

$$\tilde{C}_t = \tanh\left(W_{\tilde{c}}\left[h_{t-1}, x_t\right] + b_{\tilde{c}}\right) \tag{6}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{8}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{9}$$

其中，权重参数分别是 W_f，W_i，$W_{\tilde{c}}$，W_o，偏置值分别是 b_f，b_i，$b_{\tilde{c}}$，b_o。首先将输入的特征序列 x_t 和前一层的隐藏层 h_{t-1} 状态相结合并通过激活函数 sigmoid 来构成忘记门 f_t，会读取 h_{t-1} 和 x_t，输出一个 0 到 1 之间的数给每个在细胞状态 C_{t-1} 中的字。接下来将前一层隐藏层状态的信息 h_{t-1} 和当前输入的信息 x_t 传递到 tanh 函数中，去创造一个新的候选值变量，将 sigmoid 输出值与 tanh 的输出值相乘，sigmoid 的输出值将决定 tanh 的输出值中哪些数据是重要且需要保留的。输出数据将作为接下来注意力机制结构的输入序列。

2.2.2. 注意力机制

人们在注意某个目标或某个场景时会利用有限的注意力资源从海量信息中筛选出最具价值的信息。而深度学习中注意力机制的核心目标和选择性视觉注意力机制相似，可以极大提高信息处理的效率和准确性的注意力机制，有助于克服 RNN 中计算效率低下等问题，其核心在于注意力权重的学习，通过在体系结构中加入一个额外的前馈神经网络来学习。注意力权重 α_i 的计算方法如式(11)所示。

$$\alpha_i = \frac{\exp\left(f\left(x_i\right)\right)}{\sum_j \exp\left(f\left(x_j\right)\right)} \tag{10}$$

其中，是用来评分的函数，而是针对于输入序列中的每个向量而言的，每一个向量都对应一个权重。

这些注意力权重都被用于构建内容向量，该向量作为输入传递给解码器，而编码器中所有隐藏状态和相应注意力权重的加权和即为注意力层的输出结果。输出结果 $attentive_x$ 的计算方法如式(11)所示。

$$attentive_x = \sum_i \alpha_i x_i \tag{11}$$

本文使用的融合 LSTM 和注意力机制的情感分类模型，将特征输出和音乐所属情感类别更加紧密的联系对应起来，使计算资源分配更合理。

2.2.3. 输出层

在输出层中，使用 softmax 函数将 50 维与情感相关的特征序列映射到[0,1]范围内，为了数据处理方便，把数据定为总和为 1 的 0 到 1 之间的小数。音乐所属类别取决于最大的值。由此音乐的情感被分类为快乐、清新、放松和伤感。

3. 音乐推荐模型

在对音乐进行情感分类后，还不能直接向用户推荐音乐，需要先对用户当前的情绪和情感进行分析之后才能确定拟推荐的音乐种类。音乐推荐模型的设计流程如图 4 所示。

本文在这里设计了一个音乐推荐模型，首先针对用户历史收听记录中的歌曲进行特征获取及一系列处理，然后再经过 LSTM 层、AM 层以及 softmax 层，将特征向量降维处理并进行用户的情感分类。以上处理有利于识别出用户在最近一段时间内的情感类别，对音乐推荐有重大的参考价值。

3.1. 用户的情感分析

本部分将对设计一个利用用户的历史收听记录对用户的情感进行判别，把上文中选用的包含 MFCC、共振峰、短时能量、基音频率和短时过零率的低级描述符 LLD、LSTM 和 AM 相融合，即得到用户情感判断模型，其模型图如图 5 所示。

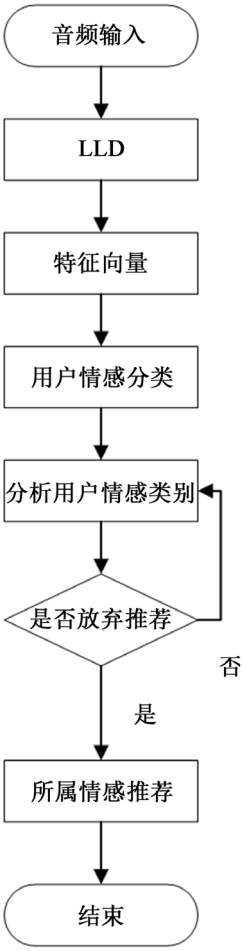

Figure 4. Flow chart of music recommendation model
图 4. 音乐推荐模型流程图

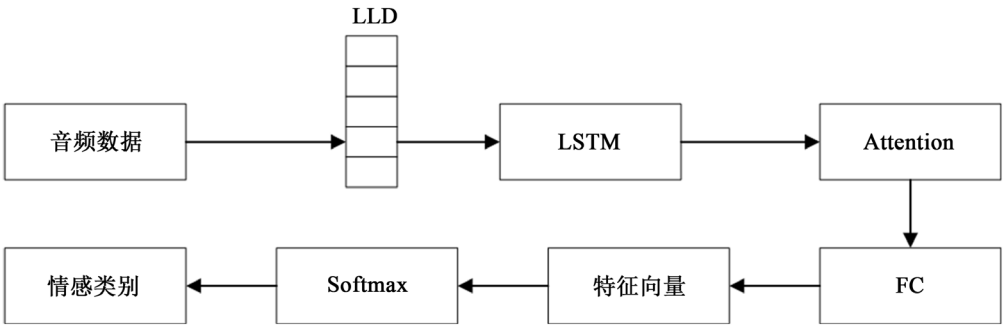

Figure 5. User's emotion discrimination model
图 5. 用户的情感判别模型

3.1.1. 历史收听音乐的特征提取

本部分的特征提取过程与 2.1 节与 2.2 节中的特征提取过程与情感分类过程基本一致，需要获取音乐的 LLD 描述，将音乐继续分段处理，每一段时长为 3 s，且每一段里都需要包含 18 维 MFCC、8 维共振峰、10 维短时平均能量、6 维基音频率以及 8 维短时过零率等总计 50 维的原始音频特征，并将得到的特征序列进行接下来的加权平均处理。

3.1.2. 历史收听记录的选择及情感分类

首先，针对用户历史收听记录的选择问题，可以知道：历史记录大致分为全部时间内的播放次数排名和近一周内的播放次数排名，且距今时间越短、播放次数越多的音乐参考价值越大。此处选取近一周内前 k 首歌曲，分别提取出每一首歌曲的特征序列并进行加权平均处理，新得到的特征序列具有较高的实时性。加权平均的公式如式(13)所示。

$$F(i,k)=\frac{\sum_{i}^{i+k}e_i}{k} \tag{12}$$

其中，k 为选定的歌曲数量，e_i 为第 i 首歌曲的特征序列，$F(i,k)$ 是把从第 i 首到第 $i+k$ 首歌曲的特征序列进行加权平均而得到的新序列。这段新的特征序列将作为接下来三层 LSTM 结构的输入数据，再经过注意力层和隐层来对信息进行筛选，加强输出与情感类别的对应关系，接下来得到的数据悉数映射到可以用于多分类的 softmax 层，以此判别用户当前的情感分类，此处类别依然是：快乐、清新、放松和伤感。

3.2. 音乐推荐模型的具体步骤

由上文中的图 4 可总结出音乐推荐模型的具体步骤。

1) 获取用户的历史收听记录，对记录中的音乐进行筛选；

2) 针对选定的音乐，提取出每首音乐的 50 维原始音频的信号特征，并对这些音乐的特征序列取加权平均，得到新的特征序列；

3) 将新的特征序列依次通过三层 LSTM 和 AM 结构，得到新的特征向量，对用户当前的情感类别进行分析和确定；

4) 用户可以选择是否接受此次推荐，若接受则推荐成功，以后将继续向用户推荐该情感标签下的音乐；

5) 若用户不接受此次推荐，则需要返回步骤 3，重新分析用户当前的情感状态以便接下来推荐其他音乐。

这样处理的好处是在对原始音频信息进行分析的同时更多的考虑到用户的情绪转变，最大限度的利用了用户历史数据等音乐资源，更加全面的利用了音乐的各方面特征信息，周期短，更贴近用户当前的情绪状态，能够大幅度提高音乐推荐的效率。

4. 实验

4.1. 数据集的选择

由于 Last.FM、QQ 音乐、虾米音乐以及酷狗音乐等网站难以获取到其他用户近一周内的播放歌曲的种类及频数，因此本实验选用网易云音乐的数据集，采集到带有歌曲名、歌手名、专辑名和情感标签等信息的音乐，在之后的训练集和对比实验中。以歌曲名、歌手名和专辑名来判断一首歌曲是否被重复推荐，以带有情感标签的音乐来训练数据集。

4.2. 实验参数的设置

本实验从网易云音乐上爬取了 10,000 首音乐，按照前文分类模型所述，最终分为四种情感类型，分别是：快乐、清新、放松、伤感。每一首音乐都带有不止一个情感标签，且时长均在 6 分钟以内。将每一首音乐都以 3 s 为一段进行分割，得到按时间顺序排列的音频段。

在针对音乐的信号特征提取阶段一共提取出 50 维特征，其中 MFCC 18 维、共振峰 8 维、短时平均能量 10 维、基音频率 6 维以及短时过零率 8 维，对特征序列映射到[0,1]范围内时，采用 max-min 进行归一化，缩放数据，并将这种方式应用到训练集和测试集中。在情感分类阶段需要用到三层 LSTM 网络以及 AM，其中学习率定为 0.002，设置 100 个 Epoch 且每个 Epoch 的值均定为 1000，Dropout 的值定为 0.7，选用 tanh 激活函数，Batchsize 定为 128，在优化器上采用 SGD。

在音乐推荐模型中，选用每个用户在最近一周内播放次数最多的前 10 首歌曲，以 2 s 为一段，每段中包含和分类模型中同种类别构成的 50 维音频特征，隐层的激活函数均选用 sigmoid 函数，softmax 的节点数定为 6，sigmoid 之前的隐层节点数定为 64，softmax 损失函数为 CE。

4.3. 实验结果及其对比分析

在本文的实验环境中，CPU 配置为 Intel i7 8th Gen，框架选用 Tensor Flow，编写环境使用 Jupyter Notebook 和 PyCharm。在将数据集分为训练集和测试集时，本文采用五折交叉验证法。

在音乐情感分类模型中，将文中所述的融合了音频 50 维特征、三层 LSTM 以及 AM 的模型与其他模型进行对比，分别为：50 维特征+LSTM；50 维特征+LSTM+AM；50 维特征+三层 LSTM+AM；支持向量机 SVM(Support Vector Machines, SVM)的分类模型；隐含狄利克雷分布 LDA (Latent Dirichlet Allocation, LDA)模型的分类模型。对比的内容即为各模型对情感进行分类的准确度，对比结果如表 1 所示。

Table 1. Emotional classification accuracy of music
表 1. 音乐的情感分类准确度

模型	平均准确率/%
50 维特征+LSTM	65.8
50 维特征+LSTM+AM	68.2
50 维特征+三层 LSTM+AM	71.6
SVM	69.2
LDA	66.3

由表 1 可以看出，仅包含音频特征和传统 LSTM 模型的分类准确度是最低的，说明这时音频的信号特征并没有被最大化的利用，不过在加入注意力机制后，情感类别的准确度和集中度都有所提升。和基于 SVM 的音乐情感分类模型以及基于 LDA 的音乐情感分类模型相比，本文提出的使用音频特征和三层 LSTM 网络、AM 的模型框架是分类准确度最高的。

在对数据集进行训练时，音乐情感的识别率也是一个需要进行对比的衡量点。四种音乐情感的识别率如表 2 所示。

Table 2. Emotional recognition rate of music
表 2. 音乐的情感识别率

识别率/%	快乐	清新	放松	伤感
快乐	74.5	11.2	9.6	4.7
清新	10.1	68.3	16.4	5.2
放松	4.6	11.5	78.2	5.7
伤感	6.1	10.3	10.9	72.7

由表 2 可知，对于快乐、清新、放松和伤感四个情感类别，情绪校价差别较大的情感更容易被区分，而差别较小的，如清新和放松类音乐，二者在识别上更容易被混淆。

最后是根据用户近一周的历史收听记录来对用户的情绪进行分类的模型测试结果，这里将分别选择用户近一周内历史记录中播放次数最多的前 1 首、前 3 首、前 5 首、前 10 首音乐的特征序列作为输入来对用户情绪分类准确度进行对比，对比结果如表 3 所示。

Table 3. Accuracy of User's emotion classification
表 3. 用户的情感分类准确度

模型	平均准确率/%
前 1 首＋LSTM+AM	60.2
前 3 首＋LSTM+AM	64.3
前 5 首＋LSTM+AM	65.4
前 10 首＋LSTM+AM	70.8

由表 3 可以发现，在利用用户近一周内播放次数最多前 10 首音乐的加权平均特征序列后，准确率有明显的提升，相比仅仅利用前 1 首、前 3 首和前 5 首音乐并不足以准确识别用户在最近一段时间内的情感和听歌喜好，因此本文提出的这个音乐推荐模型的可行性是很可观的。

5. 总结与展望

对于目前音乐推荐中存在的情感分类不够细致、用户情绪识别实时性不高等问题，本文在音乐分类模型中构造了三层 LSTM 和注意力机制模型，并依此对音乐在频谱特征上表达的情感进行分类；音乐推荐模型创新性的构造了用户历史收听记录中音乐的特征序列进行加权平均处理并再次融合 LSTM 与 AM 模型，利用新构造出的关于音频特征序列来对用户的情绪进行识别并进行音乐推荐。与其它传统的推荐模型及仅仅结合了单一模型的推荐方法相比，本文提出的方法在和用户情感的吻合性方面表现较高，对音乐的推荐精确度更高，但存在情感类别的划分种类较少等问题，下个阶段拟采用网易云音乐中更多的音乐标签来对音乐进行分类，同时对 LSTM 模型进行门单元的改进来提高音频信号的利用率，进一步提高推荐的准确度。

基金项目

国家自然科学基金项目(62073090)；广东省自然科学基金项目(2019A1515010700)；广东省科技计划项目(2020B1010010010、2019B101001021)。

参考文献

[1] 刘杨. 个性化音乐推荐系统的研究与实现[M]. 北京: 北京邮电大学, 2014.

[2] 陈雅茜. 音乐推荐系统及相关技术研究[J]. 计算机工程与应用, 2012, 48(18): 9-16.

[3] Ness, S.R., Theocharis, A., Tzanetakis, G., *et al.* (2009) Improving Automatic Music Tag Annotation Using Stacked Generalization of Probabilistic SVM Outputs. *International Conference on Multimedia*, Vancouver, October 2009, 705-708. https://doi.org/10.1145/1631272.1631393

[4] Huang, Y.S., Chou, S.Y. and Yang, Y.H. (2018) Pop Music Highlighter: Marking the Emotion Keypoints. Audio and Speech Processing.

[5] Mirsamadi, S., Barsoum, E. and Zhang, C. (2017) Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), New Orleans, LA, 19 June 2017, 2227-2231. https://doi.org/10.1109/ICASSP.2017.7952552

[6] Piczak, K.J. (2015) Environmental Sound Classification with Convolutional Neural Networks. *IEEE* 25*th International Workshop on Machine Learning for Signal Processing* (*MLSP*), Boston, MA, 1-6. https://doi.org/10.1109/MLSP.2015.7324337

[7] Zhang, Z., Xu, S., Cao, S., *et al.* (2018) Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. *Chinese Conference on Pattern Recognition and Computer Vision* (*PRCV*), Springer, Cham, 356-367. https://doi.org/10.1007/978-3-030-03335-4_31

[8] Hinto, G., Deng, L., Yu, D., *et al.* (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, **29**, 82-97. https://doi.org/10.1109/MSP.2012.2205597

[9] Van Den Oord, A., Dieleman, S., Zen, H., *et al.* (2016) Wavenet: A Generative Model for Raw Audio. SSW, 125. arXiv:1609.03499.

[10] Palo, H.K., Mohanty, M. and Chandra, M. (2015) Computational Vision and Robotics. *Advances in Intelligent Systems and Computing*, **332**, 63-70.

[11] Roddy, C. (2001) Emotion Recognition in Human-Computer Interaction. *Signal Processing Magazine*, **18**, 32-80. https://doi.org/10.1109/79.911197

[12] 张燕, 唐振民, 李燕萍. 面向推荐系统的音乐特征抽取[J]. 计算机工程与应用, 2011, 47(5): 130-133.

[13] Zhang, L., Wu, D., Han, X., *et al.* (2016) Feature Extraction of Under-water Target Signal Using Mel Frequency Cepstrum Coefficients Based on Acoustic Vector Sensor. *Journal of Sensors*, **4**, 1-11. https://doi.org/10.1155/2016/7864213

[14] Gers, F.A., Schmidhube, J. and Cummins, F. (1999) Learning to Forget: Continual Prediction with LSTM. 9*th International Conference on Artificial Neural Networks*: *ICANN*'99, 850-855. https://doi.org/10.1049/cp:19991218