

利用长短期记忆网络进行音乐流派的分类

何 丽,袁 斌

(北方工业大学 计算机学院,北京 100144)

摘 要:针对传统的基于人工标注的文本音乐分类存在人工标注成本高,易于出错,没有涉及到音乐本身的内容的问题,提出了一种基于音乐内容的分类方法,将深度学习中的长短期记忆网络(LSTM)应用到音乐流派分类中。从包含 10 种音乐流派的 1 000 首歌曲中提取出梅尔倒谱系数,频谱质心和频谱对比度三个特征,将提取出来的特征数据输入到 LSTM 网络中进行训练,输出每种音乐类别的概率。对此,进行了三次实验。第一次是将梅尔倒谱系数,频谱质心作为特征数据输入到 LSTM 网络中,第二次是以频谱对比度和频谱质心作为特征数据,第三次是将梅尔倒谱系数,频谱质心和频谱对比度作为特征数据。从实验结果上看,当梅尔倒谱系数,频谱质心和频谱对比度作为特征数据时,模型的分类效果最好,分类准确率最高。实验结果表明,该方法在准确率上比玻尔兹曼机和卷积神经网络等方法都有所提升。

关键词:音乐分类;长短期记忆网络;梅尔倒谱系数;频谱质心;频谱对比度

中图分类号:TP399

文献标识码:A

文章编号:1673-629X(2019)11-0190-05

doi:10.3969/j.issn.1673-629X.2019.11.038

Classification of Music Genres Using Long and Short-term Memory Network

HE Li, YUAN Bin

(School of Computer, North China University of Technology, Beijing 100144, China)

Abstract: In view of the problems of traditional music classification of text based on manual labeling, such as high manual labeling cost, error-prone and not involving music content, we propose a classification method based on music content, which applies long and short-term memory network (LSTM) in deep learning to music genre classification. Three features including Mel cepstrum coefficient, spectral centroid and spectral contrast are extracted from 1 000 songs of 10 music genres. The extracted feature data is input into LSTM network for training, and the probability of each music category is output. Three experiments are carried out on this. The first is to input the Mel frequency cepstral coefficient and spectral centroid as feature data into the LSTM network. The second is to use spectral contrast and spectral centroid as feature data, and the third to use Mel frequency cepstral coefficient, spectral centroids, and spectral contrast as feature data. From the experimental results, when Mel frequency cepstral coefficient, spectral centroids and spectral contrast are used as feature data, the model has the best classification results, the highest classification accuracy. Experiment shows that the method proposed is more accurate than Boltzmann and convolutional neural networks.

Key words: music classification; long short-term memory network; Mel frequency cepstral coefficient; spectral centroid; spectral contrast

0 引 言

随着计算机与手机的普及,越来越多的人选择在互联网上听音乐。面对互联网上海量的音乐,快速、精准地检索出用户想要的音乐已经变得越来越重要。近年来,越来越多的研究者开始进入音乐信息检索领域。目前音乐信息检索^[1]主要的研究领域包括音乐流派的识别分类、作曲家识别分类、乐器识别分类、歌手识别分类、情感识别分类、人声分离、音乐自动生成等。

音乐分类是音乐信息检索的一个重要分支,正确的音乐分类对于提高音乐信息检索效率具有重要的意义。目前,音乐分类主要包括文本分类和基于音乐内容^[2]的分类。文本分类主要是根据音乐的元数据信息,如音乐名称,歌手,歌词,词曲作者,年代等标注的文本信息进行分类。这种分类方式的优点是易于实现,操作简单,检索速度快,但缺陷也很明显,首先,这种方式依赖于人工标注^[3]的音乐数据,需要耗费大量的人力,并

收稿日期:2018-12-10

修回日期:2019-04-11

网络出版时间:2019-06-26

基金项目:国家自然科学基金(61371143);北京市教委科研计划面上项目(KM201510009008)

作者简介:何 丽(1977-),女,硕士,研究方向为数据库;袁 斌(1991-),男,研究生,研究方向为大数据分析。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190626.0833.058.html>

且人工标注很难避免音乐信息标注错误的问题。其次,这种文本方式并没有涉及到音乐本身的音频数据,音频数据包括音乐的很多关键特性,比如音高、音色、旋律等,这些特性用文本是很难标注的;而基于内容的分类正是对音乐的原始数据进行特征抽取,用抽取的特征数据训练一个分类器,从而达到音乐分类的目的。因此,基于内容的音乐分类也成为近年来研究的热点。

音乐流派识别^[4]作为音乐信息检索领域中非常重要的组成部分,已经受到越来越多的关注。快速准确的分类将有助于音乐推荐系统进行更精准的推荐。例如,国外的音乐巨头公司 Spotify,有一个全职的“数据炼金”团队,他们致力于将 6 000 万首歌曲分为大约 1 000 个子类型。歌曲流派分类传统上依赖于特征工程,这些特征工程可以分为三大类,即音色纹理特征、节奏特征和音高内容特征。近年来,在基于音乐内容的研究中,使用频率最高的特征是梅尔频率倒谱系数^[5](Mel frequency cepstral coefficient, MFCC),研究者们通常使用的是传统的机器学习方法,通过特征工程将提取的特征数据输入到分类器中(如支持向量机^[6](support vector machine)、高斯混合模型^[7](Gaussian mixture model)、决策树^[8](decision tree))进行训练,输出分类结果。而关于深度学习在音乐分类中的进展,Feng Tao^[9]将玻尔兹曼机运用到音乐流派分类中,但是只能对四种流派进行分类。Dong M^[10]

将卷积神经网络(convolutional neural network, CNN)作为分类器,将原始音乐数据降维转换为声谱图,利用 CNN 在图片处理上的优势,让 CNN 自动学习声谱图上的特征。这种方法虽然可以进行 10 种流派的音乐分类,但分类的准确度不高,只有 61% 左右。鉴于此,文中在前人研究的基础上,提出将深度学习中的 LSTM 网络应用到音乐流派的识别上。这种方法不仅能进行 10 种音乐流派的分类,而且在准确率上也有所提高。

1 研究方法

由于该数据集是原始的音频数据,需要手动提取特征,所以在整个音乐流派识别过程中,采用的方法可以分为两部分。第一部分是特征工程,即特征抽取,从音频的原始数据中抽取出三个最能代表音频内容的特征,它们分别是梅尔倒谱系数、频谱质心^[11]、频谱对比度^[12],然后将抽取的特征数据输入到设计的 LSTM 网络模型中进行训练学习,最后输出分类结果。在整个过程中,首先会将梅尔倒谱系数、频谱对比度分别与频谱质心进行融合一起输入到 LSTM 中,观测实验结果,然后再将三个特征融合到一起,作为输入数据输入到 LSTM 网络中,将得到的结果与只有两个特征作为输入数据的分类结果进行比较,从而得出最佳特征的组合。整个过程如图 1 所示。

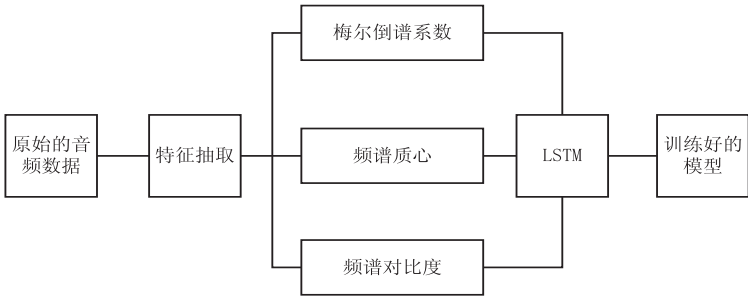


图 1 音乐流派识别过程

2 相关理论

2.1 特征工程

(1) 梅尔倒谱系数。

在声音处理中,梅尔频率倒谱是基于声音频率的非线性梅尔刻度(Mel scale)的对数能量频谱的线性变换,而梅尔倒谱系数是组成梅尔频率倒谱的系数。梅尔频率倒谱系数是语音识别领域中最重要特征,所以在用音频数据对音乐进行分类中也是非常重要的一个特征,不容忽视。关于梅尔频率倒谱系数,很多文献中都有提及,比如文献[5, 13],这里不在赘述。重点是从音频数据中抽取 MFCC,采用 Python 中的 Librosa 这个专门用于音频数据处理的库从音频数据中抽取

MFCC 特征。从抽取结果来看, MFCC 是一个二维数组,一个维度代表时间,另外一个维度代表不同的频率。

(2) 频谱质心。

频谱质心在数字信号处理中是用来表征频谱的度量,表示频谱的“质心”位于何处,是声音信号的频率分布和能量分布的重要信息。在主观感知领域,频谱质心描述了声音的明亮度,具有阴暗,低沉品质的声音倾向有较多的低频内容,频谱质心相对较低,具有明亮,欢快品质的声音多数集中在高频,频谱质心相对较高。

(3) 频谱对比度。

频谱对比度是 Jiang Danning^[12]在 2002 年提出来

的用于音乐流派分类的特征。频谱对比度表示为频谱中频峰与频谷之间的分贝差异,能够代表音乐的相对光谱特性,对于大部分音乐来说,强大的频峰差值几乎与谐波部分一致,而非谐波部分或者噪音,一般在频谱的谷值出现。因此频谱对比度可以大致反映出谐波和非谐波在频谱中的分布。另外,从文献[13]中可以看出,频谱对比度对音乐的流派具有良好的辨别能力。

2.2 长短时记忆网络(LSTM)

在深度学习中,卷积神经网络(CNN)和循环神经网络(RNN)是使用频率最高的两种神经网络。CNN

一般用于处理图片类型的数据,RNN 一般用来处理有连续时间序列的数据。LSTM 是一种特殊类型的 RNN,用来解决 RNN 不能长期依赖的问题。LSTM 由 Hochreiter & Schmidhuber^[14] (1997) 提出,并在近期被 Alex Graves^[15] 进行了改良和推广。LSTM 的具体设计细节可以参考文献[14]。一个典型的 LSTM 网络模型的结构如图 2 所示。

根据原始训练数据的特点,文中设计的网络结构如图 3 所示。

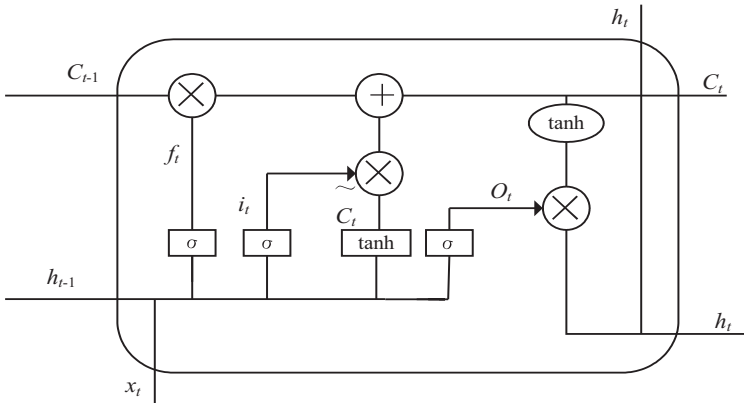


图 2 LSTM 网络模型

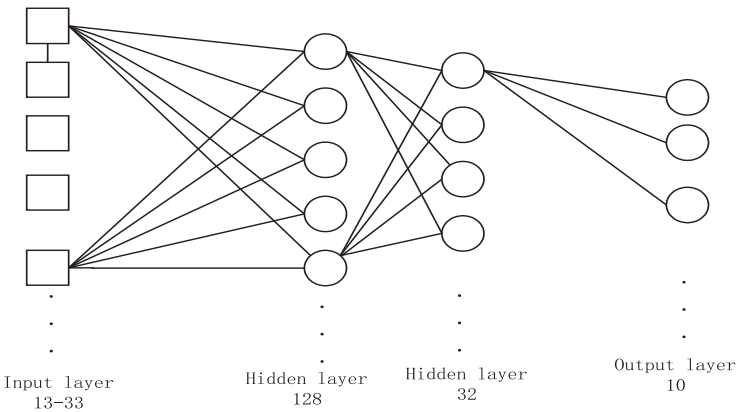


图 3 网络结构

模型的输入层是提取的特征数据 MFCC(0-13), 频谱质心 (14-26), 频谱对比度 (27-33), 第一个 LSTM 隐藏层的神经元个数为 128, 第二个 LSTM 隐藏层的神经元个数为 32, 最后输出层是 10 个神经元, 代表音乐 10 种流派的概率。

3 实验

3.1 数据集

使用的数据集是公开数据集 GTZAN, 该数据集包含 10 种音乐流派, 如表 1 所示。

每种流派的音乐包含 100 首, 时长为 30 s 的音乐文件, 文件格式以 .au 为后缀名。随机将数据集分成三份, 分别作为训练集、验证集和测试集, 所占比例分

别为 70%、20%、10%。

表 1 GTZAN 数据集流派中英文对照

英文	中文
bules	布鲁斯
classic	古典
country	乡村
disco	迪斯科
hip-hop	嘻哈
jazz	爵士
metal	金属
pop	流行
Reggae	雷鬼
rock	摇滚

3.2 数据预处理

由于 GTZAN 数据集中包含的数据都是音频原始数据,而音频所包含的数据信息太多,无法直接将原始数据作为训练数据使用。因此,从音频数据中提取出具有代表性的音乐特征,分别是 MFCC、频谱质心和频谱对比度。利用 Librosa 这个工具库,从每个音频文件中抽取出 13 个 MFCC,1 个频谱质心,7 个频谱对比度作为特征数据输入到设计的 LSTM 网络模型中进行训练。

3.3 实验结果分析

将整个实验分成三个实验进行。实验一是实验数据只选择 MFCC 和频谱质心,实验二是数据特征只选择频谱对比度和频谱质心,为了进行实验对比,实验三选择了 MFCC、频谱质心、频谱对比度三种特征,将三种特征融合在一起,作为输入数据输入到 LSTM 网络中。由于频谱质心中包含的数据信息太少,不适合作为单一的特征输入到 LSTM 中,所以选择将其作为 MFCC 和频谱对比度这两个特征的增强数据。三次实验中,LSTM 网络模型中的损失函数(loss function)采用的是 categorical_crossentropy,优化器(optimizer)采用的是 Adam,学习率(learning rate)为 0.001,批尺寸(batch_size)为 35,而 epoch 为 400,分类概率的输出采用的使 softmax 函数。三种实验分别进行了 5 次,每次实验测试集的误差和准确率如表 2~表 4 所示。

表 2 实验一中测试集的误差和准确率

实验次数	误差	准确率
1	1.896 3	0.566 6
2	2.627 4	0.466 6
3	2.374 2	0.533 3
4	2.822 8	0.416 6
5	2.274 3	0.550 0
平均	2.399 0	0.600 0

表 3 实验二中测试集的误差和准确率

实验次数	误差	准确率
1	0.956 7	0.583 3
2	0.892 0	0.633 3
3	0.902 6	0.600 0
4	0.844 0	0.600 0
5	1.018 4	0.583 3
平均	0.921 9	0.506 6

表 4 实验三中测试集的误差和准确率

实验次数	误差	准确率
1	0.968 0	0.650 0
2	0.911 2	0.683 3
3	0.960 4	0.700 0
4	0.953 3	0.716 6
5	0.978 0	0.666 6
平均	0.954 2	0.683 3

从实验结果中可以得出,实验一、实验二、实验三的平均正确率分别为 0.60,0.50,0.68。从实验一、实验二、实验三的 5 次实验中随机抽取 1 次,其在训练集的准确率和测试集上的准确率随 epoch 的变化如图 4~图 6 所示。

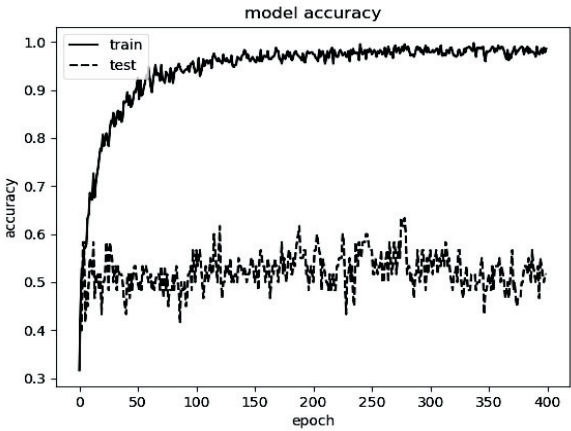


图 4 实验一训练集与测试集准确率随 epoch 的变化

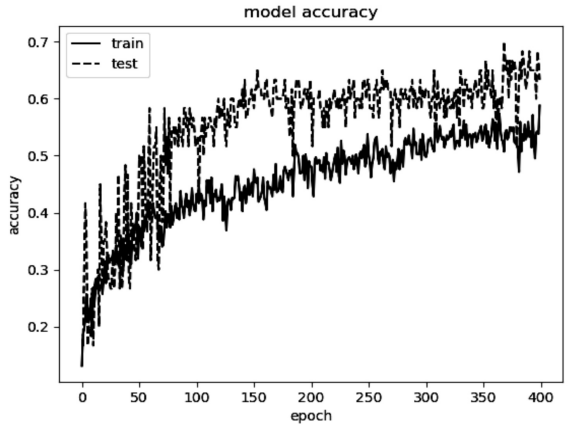


图 5 实验二训练集与测试集准确率随 epoch 的变化

可以看出,实验一在训练集上的准确率很高,但是在测试集上的准确率却很低,说明模型出现过拟合,泛化能力太差。实验二在训练集上与测试集上的准确率相差不大,说明模型的泛化能力很强。而对比实验一和实验三可以发现,在训练数据中有 MFCC 特征的时候,训练集上的准确率都很高,测试集上的准确率却不是很高,说明 MFCC 是三种特征中最能描述音频内容的特征,但由于训练的数据太少,导致模型过拟合。综

合以上结果分析不难得出,当使用 MFCC、频谱质心、频谱对比度三种特征数据作为输入数据时,模型的泛化效果最好,分类的准确率最高。

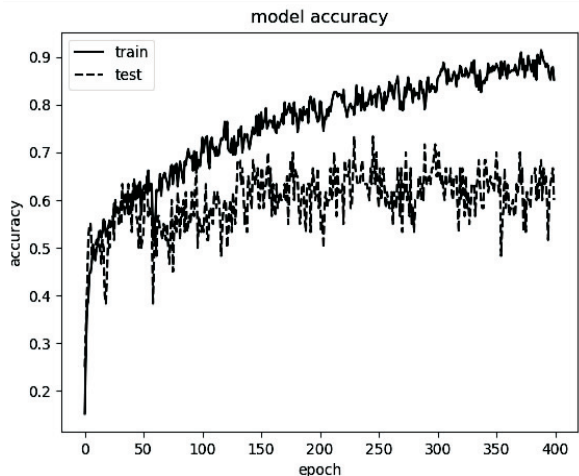


图 6 实验三训练集与测试集准确率随 epoch 的变化

4 结束语

针对传统的机器学习方法在音乐分类中随分类类别增多,分类准确率不断下降的问题,提出一种将深度学习中的 LSTM 网络应用到音乐流派分类中的方法。从 GTZAN 数据集的 1 000 首歌曲中,抽取出 MFCC、频谱质心、频谱对比度三种特征数据,将这些特征数据输入到 LSTM 中进行训练,从而得出音乐属于每种音乐流派的概率。从实验结果可以看出,对 10 种流派的音乐进行分类,在分类的准确率上比用卷积神经网络的方法要提高不少。

参考文献:

- [1] 邓璐华,董 讯. 音乐信息检索(高等学校现代文献信息检索系列教材)[M]. 北京:高等教育出版社,2006.
- [2] 王秀丽. 基于内容的音乐风格分类系统的研究与实现[D]. 南京:南京邮电大学,2014.
- [3] 陈荃有. 音乐文本的标注意义及方法探究[J]. 黄钟—中国·武汉音乐学院学报,2018(1):123-135.
- [4] 刘 丹,张乃尧,朱汉城. 音乐特征识别的研究综述[J]. 计算机工程与应用,2002,38(24):74-77.

- [5] LOUGHRAN R, WALKER J, O'NEILL M, et al. The use of mel-frequency cepstral coefficients in musical instrument identification[C]//International computer music conference. Belfast, N. Ireland: International Computer Music Association, 2008.
- [6] MANDEL M I, ELLIS D. Song-level features and support vector machines for music classification[C]//International conference on music information retrieval. [s. l.]: [s. n.], 2005.
- [7] CHANDANPREET K, KUMAR R. Study and analysis of feature based automatic music genre classification using Gaussian mixture model[C]//2017 international conference on inventive computing and informatics. [s. l.]: IEEE, 2017.
- [8] BERGSTRA J, CASAGRANDE N, ERHAN D, et al. Aggregate features and adaboost for music classification[J]. Machine Learning, 2006, 65(2-3):473-484.
- [9] FENG T. Deep learning for music genre classification[R]. Illinois: University of Illinois, 2014.
- [10] DONG Mingwen. Convolutional neural network achieves human-level accuracy in music genre classification[J]. IEEE Transactions on Multimedia, 2011, 13(2):303-319.
- [11] TZANETAKIS G, COOK P. Musical genre classification of audio signals[J]. IEEE Transactions on Speech and Audio Processing, 2002, 10(5):293-302.
- [12] JIANG Danning, LU Lie, ZHANG Hongjiang, et al. Music type classification by spectral contrast feature[C]//IEEE international conference on multimedia and expo. Lausanne, Switzerland: IEEE, 2002:113-116.
- [13] LEE C H, SHIH J L, YU K M, et al. Automatic music genre classification using - modulation spectral contrast feature [C]//IEEE international conference on multimedia and expo. Beijing, China: IEEE, 2007.
- [14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.
- [15] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6):602-610.