基于机器学习的音频分类

熊华煜，余 勤+，任 品，雒瑞森

（四川大学 电气工程学院，四川 成都 610065）

摘 要：为施行有效的音频分类以高效率处理日渐复杂的音频信息，研究采用包含多种神经网络在内的 5 种机器学习模型，实现多种决策下的音频分类以寻找最优模型，基于分类准确度对各模型分类效果进行评估，在使用正则化方法保证模型泛化能力的条件下，通过比较和实验，挖掘并验证出了相对最优的模型——卷积神经网络音频分类模型及对应参数，为现有音频分类模型的进一步优化提供了参考方向。

关键词：多媒体技术；机器学习；音频分类；神经网络；正则化

中图法分类号：TP183 **文献标识号**：A **文章编号**：1000-7024（2021）01-0156-05

doi：10.16208/j.issn1000-7024.2021.01.023

Audio classification based on machine learning

XIONG Hua-yu, YU Qin+, REN Pin, LUO Rui-sen

(College of Electrical Engineering, Sichuan University, Chengdu 610065, China)

Abstract: To implement effective audio classification for efficient audio information process which is becoming increasingly complex, five kinds of machine learning models including neural network were built respectively to find the best classification model, and the accuracy of classification of each model was evaluated. The relatively better model, convolutional neural network for audio classification, and its corresponding best parameters were recommended and verified based on experiment and comparison under dropout regulation to guarantee the ability of generalization. The new reference orientation was provided to optimize existing classification model.

Key words: multi-media technology; machine learning; audio classification; neural network; dropout regularization

0 引 言

音频信息的种类在多媒体技术发展下变得越发多样化，如不同风格、乐器演奏的音乐，各地方言人声等。因而语料库质量的好坏对系统的影响巨大。标准语料库训练出的语音模型在这些干扰下识别率会大大降低，需适当的预处理技术进行改善。

因此，音频分类技术具有非常巨大的理论价值和实际意义[1]。预分类为对应的音频处理算法模型的使用提供指导，能提高处理的效率和准确度；音频检索方面，已分类音频可作为带标签音频数据库，方便用作音频处理模型训练素材；日常生活中，音频分类可以为实时语音进行识别、目标场景分析等应用提供精确化预处理支持，可驱动更多音频信息处理任务的优化发展。

1 相关方法简述

本文所涉及分类模型使用音频的 MFCC（melscale frequency cepstral coefficients）作为特征向量，分别输入到支持向量机、贝叶斯分类器、全连接神经网络、卷积神经网络、循环神经网络中以探究各个模型的二分类和多分类准确度。

1.1 MFCC 特征参数

梅尔倒谱系数（MFCC）是用数字形式表征音频信号的常用特征，其详细定义请参见文献[2]，运用在语音相关系统中时可赋予系统良好的鲁棒性与识别率，与基于线性预测的倒谱相比较可以更好模拟人的听觉感知效果。

1.2 支持向量机分类器

支持向量机（support vector machine，SVM）是一种

收稿日期：2019-07-01；修订日期：2019-12-05

基金项目：校企合作基金项目（17H1199、19H0355）；中国博士后科学基金面上基金项目（2017M612958）

作者简介：熊华煜（1995-），男，湖北武汉人，硕士研究生，研究方向为语音信号处理、模式识别；+通讯作者：余勤（1968-），女，四川成都人，副教授，研究方向为智能系统与模式识别；任品（1997-），男，重庆人，本科，研究方向为信号处理；雒瑞森（1989-），男，甘肃酒泉人，博士，讲师，研究方向为信号处理与机器学习。E-mail：563783915@qq.com

经典的监督式学习模型及相关的分类算法。以二分类为例，其原理是在得到一些待分类数据点的情况下，找到一个满足分类要求的最优超平面，使所有正分类点到该平面的距离与所有负分类点到该平面的距离的总和达到最大，这个平面就是最优分类超平面。

1.3　朴素贝叶斯分类器

朴素贝叶斯分类器（简称 BYS）[3]具有将先验知识综合的特性，它提供了推理的一种概率手段，即基于待考察的量遵循某概率分布且根据这些概率以及已观察数据进行推理这一假定，以求作出最优的决策。

1.4　人工神经网络

人工神经网络（artificial neural network，ANN）是一种类人脑神经连接结构的计算模型，其详细定义参见文献［4］。结构中的每个神经元都是对信息进行处理的最小结构，其经过组合形成的系统具有非线性、自适应[5]信息处理特征。神经元布局如图 1 所示。

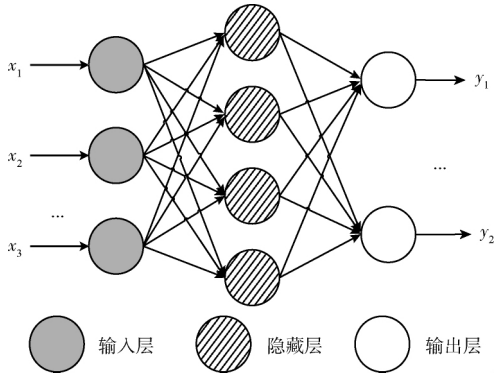

图 1　神经网络基本结构[6]

在训练多层神经网络时常采用逆误差传播算法（error backpropagation，BP），其执行一般包含以下 4 步：

（1）按输入层到输出层方向，计算输出值；

（2）按输出层到输入层方向，计算误差值；

（3）计算每个权重的梯度；

（4）使用梯度下降算法更新权重。

全连接神经网络（fully connected neural network，FC）即神经网络中除输出层外每层每个神经元都与上一层所有神经元相连，每个连接都有一个权值，第 N－1 层神经元输出是第 N 层神经元的输入。

卷积神经网络（convolutional neural network，CNN）是一种利用卷积突出数据特征的神经网络，详细定义参见文献［7］。它使用 3 种方法强化模型训练效果：一是局部连接，神经网络的结构不再是每一层的所有神经元都与上一层所有神经元相连；二是权值共享，在神经网络中一组连接可以用同样的权重，不再是每一个连接都有一个不同的权重；三是通过池化层的下采样，这种方式不仅可以减少该层的样本数量，而且还有助于该模型鲁棒性的提高。

循环神经网络（recurrent neural network，RNN）的详细定义参见文献［8］，与卷积神经网络最大的不同在于，前者输入数据都是独立的，在训练过程中彼此之间不存在联系，而循环神经网络先后输入数据在训练过程中存在联系，因此可更好地处理包含顺序信息的数据。

1.5　实验总体设计

本文设计以各种分类算法模型的基本框架为基础，不断调整参数以观察音频的分类准确度。在经典的机器学习算法中，调整输入特征尺寸，不断改变训练与测试样本的比例，并反复进行测试观察模型的准确度波动，在先进的神经网络算法中，不断调整网络层数以及迭代次数以观察准确度波动，之后在同样的样本训练及测试条件下横向对比各个模型的准确度以寻求出最适合音频分类的模型，最后研究了正则化参数对以上得出的最优模型准确度的影响，得到了最合适的参数。

2　模型具体实现

2.1　所用语言及开发环境

本文的模型构建使用 Python 语言完成，采用 PyCharm 作开发环境，音频库采用 GTZAN 的 genres 数据集，包括 5 种不同风格（布鲁斯、古典、乡村、迪斯科、嘻哈）音乐，每种 100 条，每条 30 s。

由于样本数量较少，考虑通过对音频进行剪切实现数据扩充，因此将音频每 9 s 剪成一段，每两段间有 4 s 的重合，从而将每段 30 s 的音频剪切成为 5 段 9 s 的音频，成功将样本数量扩大了 5 倍。最终扩展成每类 500 条数据。

2.2　特征提取

使用 Python 库 librosa 中 feature.mfcc 函数提取音频的前 20 维 MFCC 特征参数，通过 numpy 中 ndarry.flatten 将 20＊388 的特征参数平铺成一维长度为 7760 的向量。当使用支持向量机和贝叶斯分类器进行分类时，可仅使用一位数作为标签代表音频的种类，当使用神经网络进行分类时，使用 One-shot 编码作为特征向量的标签，以匹配神经网络输出层神经元的个数。最终得到带标签的特征向量矩阵。

2.3　支持向量机分类器的实施

2.3.1　模型实现

支持向量机模型主要使用了 sklearn 库，在得到带标签的特征向量矩阵之后，先将各特征向量打乱顺序，再使用 sklearn 中的 train_test_split 函数选择一部分数据作训练集，剩下部分作测试集。使用 svm.SVC 构建支持向量机模型，使用 fit 函数进行训练，predict 函数进行测试，最后通过 classification_report 函数和 accuracy_score 函数得到支持向量机的分类效果。

2.3.2　多分类效果

当使用 20％的数据为测试集，剩下的数据为训练集时，

该模型的分类准确度为 0.837。

重复执行程序，得到分类结果见表 1。

表 1 重复运行支持向量机模型所得数据

序数	0	1	2	3	4	平均值
准确度/%	0.837	0.707	0.888	0.595	0.96	0.797

观察到分类准确度在 0.8 左右，但波动较大，推测原因是测试样本量过大，同支持向量机小样本训练产生过拟合的特性相矛盾。

2.4 贝叶斯分类器的实施

2.4.1 模型实现

贝叶斯分类器的实现通过 sklearn 库实现，使用 sklearn.naive_bayes 中的 GaussianNB 搭建一朴素贝叶斯分类器，其它与支持向量机相同。

2.4.2 多分类效果

使用 20%数据为测试集，剩下的为训练集时，模型分类准确度为 0.84。重复执行结果见表 2。

表 2 重复运行贝叶斯分类器模型所得数据

序数	0	1	2	3	4	平均值
准确度/%	0.84	0.81	0.798	0.774	0.772	0.799

观察到分类准确度在 0.8 左右，且波动较小。

2.5 全连接神经网络的实施

2.5.1 模型实现

全连接神经网络模型的实现基于 Keras 框架，使用 keras.models 中 Sequencial 模块建立序列模型，通过 keras.layers 中 Dense 和 Activation 函数添加全连接层和激活函数，通过 model.compile 函数添加 adam 优化器，使用交叉熵损失函数，以分类准确度作衡量指标，使用 model.fit 函数进行训练，最后通过 model.evaluate 输入测试集数据检验分类效果。

2.5.2 多分类效果

使用含 5 层隐藏层的神经网络，输入 genre 数据集，选择 64%的数据作为训练集，16%的数据作验证集，20%的数据作为测试集，迭代 35 次，得到分类准确度为 0.709。

更改迭代次数，得到程序分类准确度随迭代次数的变化如图 2 所示。

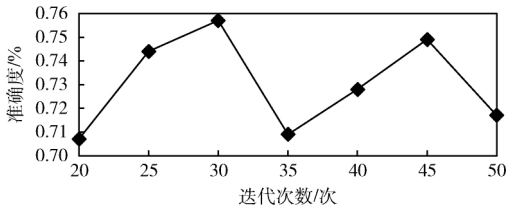

图 2 不同迭代次数对应的分类准确度

观察分类准确度的变化，可以发现迭代 30 次时分类准确度最高，为 0.757，控制迭代次数为 30 次，修改隐藏层的层数，对每种情况重复 3 次实验取平均值，观察分类准确度的变化如图 3 所示。

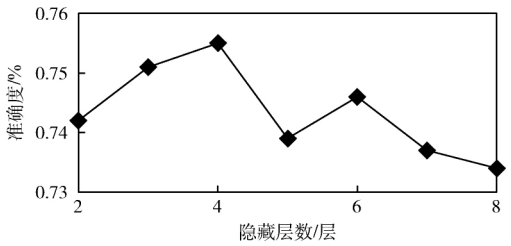

图 3 不同隐藏层层数对应的平均分类准确度

观察到神经网络层数的变化对分类准确度的影响并不大，其中当隐藏层层数为 4 时，分类准确度相对较高，为 0.755。

2.6 卷积神经网络的实施

2.6.1 模型实现

卷积神经网路模型构建亦基于 Keras 框架，使用 keras.models 中的 Sequencial 模块建立序列模型，不同之处在于使用卷积神经网络时，先将输入数据转化为三维形式，其深度设为 1，然后通过 Conv2D 函数实现卷积操作，MaxPooling2D 函数实现池化层下采样，最后通过全连接层得到分类结果。

2.6.2 多分类效果

输入 genres 数据集，让数据先通过两个卷积层加池化层的组合，再通过一个全连接层，迭代 20 次，得到分类准确度为 0.758。改变迭代次数，得到分类效果及相应花费时间见表 3。

表 3 不同迭代次数对应的分类准确度及时间

迭代次数	15	20	25	30	35	40	45	50	55
准确度/%	0.734	0.758	0.784	0.823	0.826	0.828	0.82	0.83	0.834
耗时/s	585.5	769.7	998.9	1245.4	1396.6	1643.7	1718.5	1823.9	2094

考察分类准确度随迭代次数的变化，可得如图 4 所示。

一开始分类准确度随迭代次数增加而增加，30 次之后渐趋于平稳，最终分类准确度在 0.83 左右。

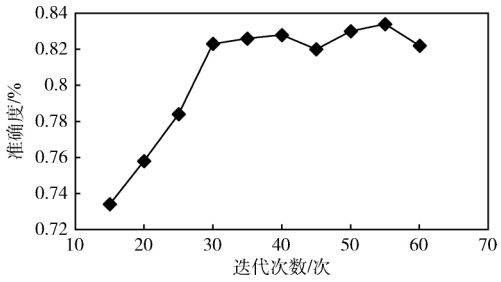

图 4　卷积神经网络分类准确度随迭代次数的变化

2.7　循环神经网络的实施

2.7.1　模型实现

循环神经网络模型的构建同样使用 Keras 框架实现，在建立模型时使用了 SimpleRNN 函数建立循环层，然后再接入全连接层，输出分类结果。

2.7.2　多分类效果

输入 genres 数据集，调整迭代次数，每种次数运行 3 次，得到平均分类效果如图 5 所示。

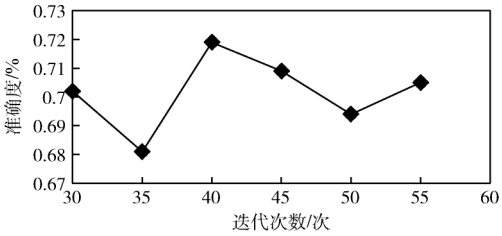

图 5　多分类准确度随迭代次数的变化

观察到随着迭代次数的变化，分类准确度在 0.68 到 0.72 之间上下波动，总体变化不大。

2.8　分类效果对比

综合结果，得各模型分类准确度如图 6 所示。

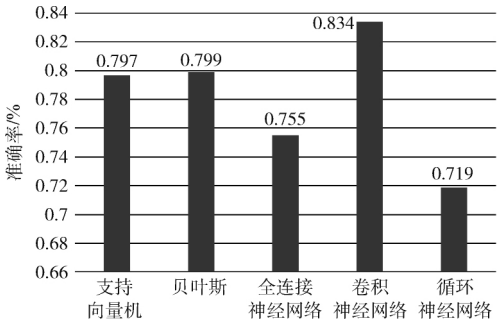

图 6　5 种模型多分类准确度对比

从图 6 可观察到，使用 genres 数据集考察各模型多分类效果时，准确度最高的是卷积神经网络，其次是支持向量机和贝叶斯分类器，其中两者的平均分类准确度相仿，但多次执行程序发现，支持向量机分类准确度方差高于贝叶斯分类器，因此综合来看，贝叶斯分类器分类优于支持向量机，之后是全连接神经网络，最后是循环神经网络。可发现卷积神经网络分类效果最好，因此接下来将基于卷积神经网络处理 genres 数据集使用 Dropout 正则化对实验过程进行优化，以达到更好的分类效果。

3　Dropout 正则化

3.1　Dropout 正则化简介

观察模型训练过程可发现，训练集与验证集的分类效果良好，但测试集效果较差，这是泛化能力不足的体现，如何减少泛化误差则成为一大关键问题。正则化可通过使模型复杂度降低从而具有更好泛化能力，Dropout 正则化便是其中一种方法，可通过每个训练案例中随机省略一半特征检测器来减少神经网络过度拟合[9]。即将一半隐藏层节点设置为 0，降低隐藏层节点间影响，让模型具有更强泛化性。使用 Dropout 正则化时，并不局限于省略一半隐藏层节点，而可根据实际需求人为确定省略掉的隐藏层节点数量的比例，以提高模型泛化性。

3.2　Dropout 正则化实现及效果

在卷积神经网络中仍通过 Keras 实现 Dropout，在模型训练时屏蔽掉一些神经元，让神经元输出乘以 $1/(1-p)$ 进行放大，p 为神经元被屏蔽概率（Dropout 率）。该处放大操作是防止训练时结果不稳定增加的补偿，使输入同样数据情况下输出期望也相同。

分别在卷积神经网络的隐藏层中最后一层池化层及全连接层激活函数后面加上 Dropout 语句，设置迭代次数为 40，更改 p 值，得到效果如图 7 所示。

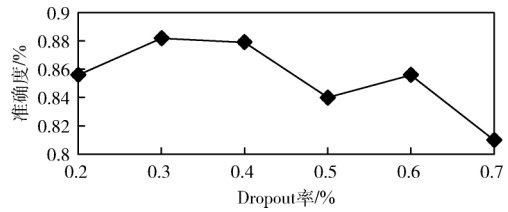

图 7　分类准确度随 Dropout 率的变化

可观察到，当 Dropout 率设置为 0.3 时 5 类音频的分类准确度最高，达到 0.882。设定 Dropout 率为 0.3，增加迭代次数，得到效果如图 8 所示。

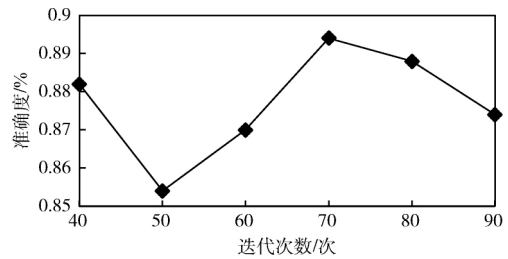

图 8　分类准确度随迭代次数的变化

可观察到，当迭代为 70 次时，分类准确度最佳，为 0.894，比使用 Dropout 之前提高了 6%。

4 结束语

音频分类，在生活中有巨大理论与实际意义，在音频检索或音频管控系统领域均有广泛的应用。

本文对常见分类模型的音频分类进行了探究，搭建了支持向量机、贝叶斯分类器、全连接网络、卷积网络、循环网络 5 种分类模型，输入数据集后得到每种模型分类效果，并对数据进行了对比与讨论，确定了卷积神经网络在用于音频多分类时是较为理想的机器学习模型，在对 5 类音频进行分类时，准确率接近 0.9，并通过实验得到了最佳的网络正则化优化参数 0.3 与迭代次数 70。

如何将最优模型的可分类数进一步扩展，探索平衡分类精度、分类速度的更加效率的模型结构，并将模型应用在更多更具体实用领域（例如音乐应用中的定向风格音乐推荐[10]，或是应用在非法广播监控中提前分离出音乐过多的条目以减少干扰等），将会是本文未来的工作目标。

参考文献：

[1] Feng Rong. Audio classification method based on machine learning [C] //IEEE International Conference on Intelligent Transportation, Big Data & Smart City. Changsha: IEEE Computer Society, 2016: 81-84.

[2] GAO Ming, SUN Rencheng. An efficient speaker feature parameter extraction method based on improved of MFCC [J]. Journal of Qingdao University (Natural Science Edition), 2019, 32 (1): 61-65 (in Chinese). [高铭，孙仁诚. 基于改进 MFCC 的说话人特征参数提取算法 [J]. 青岛大学学报：自然科学版，2019，32 (1)：61-65.]

[3] WANG Runfang, CHEN Zengqiang, LIU Zhongxin. Link prediction in complex networks with syncretic naive Bayes methods [J]. CAAL Transactions on Intelligent Systems, 2019, 14 (1): 99-107 (in Chinese). [王润芳，陈增强，刘忠信. 融合朴素贝叶斯方法的复杂网络链路预测 [J]. 智能系统学报，2019，14 (1)：99-107.]

[4] LI Hao, BAO Hong, ZHANG Jing. Research on speaker recognition model based on depth neural network [J]. Computer and Information Technology, 2018, 26 (5): 1-3 (in Chinese). [李浩，鲍鸿，张晶. 基于深度神经网络的说话人识别模型研究 [J]. 电脑与信息技术，2018，26 (5)：1-3.]

[5] LIU Fang, SUN Xiaoqi, WANG Linshan. Wellposedness and analysis of mean-square exponential attraction of stochastic neural networks with S-type distributed delays [J]. Journal of Binzhou University, 2014, 30 (6): 7-13 (in Chinese). [刘芳，孙小琪，王林山. S-分布时滞随机神经网络的适定性和均方指数吸引性 [J]. 滨州学院学报，2014，30 (6)：7-13.]

[6] LIU Songlin, QIN Xiaowei, DAI Xuchu. Downlink SINR prediction based on structural combination of neural networks [J]. Journal of Telemetry, Tracking and Command, 2018, 39 (4): 21-28 (in Chinese). [刘松林，秦晓卫，戴旭初. 基于神经网络结构化组合的下行链路 SINR 预测 [J]. 遥测遥控，2018，39 (4)：21-28.]

[7] FU Wei, YANG Yang. Audio classification method based on convolutional neural network and random forest [J]. Journal of Computer Applications, 2018, 38 (A02): 58-62 (in Chinese). [付炜，杨洋. 基于卷积神经网络和随机森林的音频分类方法 [J]. 计算机应用，2018，38 (A02)：58-62.]

[8] Feng W, Guan N, Li Y, et al. Audio visual speech recognition with multimodal recurrent neural networks [C] //IEEE International Joint Conference on Neural Networks. Alaska: IEEE Computer Society, 2017: 681-688.

[9] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15 (1): 1929-1958.

[10] GONG Zhi, SHAO Xi. A music recommendation system based on multi-modal fusion [J]. Journal of Nanjing University of Information Science & Technology, 2019, 11 (1): 68-76 (in Chinese). [龚志，邵曦. 基于多模态的音乐推荐系统 [J]. 南京信息工程大学学报：自然科学版，2019，11 (1)：68-76.]