人声音频信号处理与多算法情绪识别分类

目录

摘要

抑郁症作为世界第四大疾病，近年来患病人数增长将近 20 倍，但现用的治疗方案还相对落后，在长期治疗过程中复诊和日常情况检测仍是靠医生问询和填写问卷来显示患者的情绪状态，这样的方案常有患者说谎或无法正确表达情绪的误差情况，不具有日常实时性和准确性。为了即时检测患者情绪，我们提出实时录音检测情绪的方案，不需要识别语言文字而是提取音频信号的 MFCC 系数特征进行基本情绪分类，实现跨语言跨人种跨地区的音频情绪检测。我们对比了 MLP、KNN、随机森林、SVM、2D CNN 等八种算法的准确率与训练用时，使用时间缩放和音高变换两种数据增强方式进一步提高算法的准确率和泛化能力。准确率与训练用时综合指标最好的算法为随机森林，准确率最高的算法为 CNN，达到 72%。模型精度仍需后续提高，且由于训练集均为英语采集，在实际部署方面仍然需要克服源数据的偏向性。

引言

根据全球顶尖综合大学之一的伦敦大学学院（UCL）脑科学院部官网介绍，全球患有抑郁症或焦虑症的人口总数超过了 6.15 亿人[1]。UCL 作为英国疾病研究中心和心理学研究领军学院，正在结合人工智能技术用于临床诊断与治疗。抑郁症作为世界第四大疾病，在近十几年间报告的抑郁症患病人数暴增将近 20 倍，在 2020 年被预计成为世界第二大疾病。心理疾病的治疗过程包括药物治疗、物理治疗等，患者须定期复诊。目前我国对心理疾病的治疗方案较为落后，复诊和日常情况检测靠医生面谈和填写在线问卷得到，故患者可能会有所隐瞒或无法正确表达自己的情绪情况。

本项目通过各种算法检测声音的情绪，希望能对患者的治疗过程有所启发，例如通过手机实时录音代替填问卷采集患者情绪情况。

现有检测声音情绪的方式主要有两种，最常用的一种是识别语音中的文本内容，整段或提取关键词进行语义分析，对整句话进行自然语言处理（NLP）得出语音的情感倾向。NLP 模型较复杂但整体识别效果较好，准确率一般在 70%-80%，在实际运用中，众多开发者仍在寻找提高文本情感分析准确度的方法。Kaushik[2]提出了一种基于一个关键字代表整体的语音情感分析系统，降低了模型复杂度且提高了实际运用中的表现。另一种是通过提取音频信号特征来获得情感倾向（SER），此方法研究热度比 NLP 小很多，模型简单但目前普遍准确率在 50%-60%左右。Daneshfar[3]使用 MFCC，PLPC，PMVDR 和音高 pitch 作为特征向量，采用降维（PCA,FA,PPCA）和比较各种分类器算法，得到较好的几种算法结果准确率在 65%-75%。

设计方案

数据集选取

本项目采用四个开源数据集 CREMA-D[4]，TESS[5]，SAVEE[6]，RAVDESS[7]相结合作为整体音频样本。

CREMA-D 数据集是由不同种族的 20 至 74 岁的 48 名男士和 43 名女士的原始声音片段组成，总共有 7442 条数据。演员们用六种不同的情绪说出给定的 12 句话，情绪包括愤怒、厌恶、恐惧、开心、中性和悲伤。

TESS 数据集是由一名 26 岁的女演员和一名 64 岁的女演员用愤怒、厌恶、恐惧、开心、惊喜、悲伤和中性七种情绪读出包含指定单词的句子的原始声音片段组成，总共有 2800 条数据。

SAVEE 数据集是由四名 27 至 31 岁的男性读出文本材料中的句子，并将情绪分为愤怒、厌恶、恐惧、开心、中性、悲伤和惊讶七种的声音片段组成，总共有 480 条数据。

RAVDESS 数据库由 12 名女性和 12 名男性用平静、快乐、悲伤、愤怒、恐惧、厌恶和惊讶七种情绪读出句子的声音片段组成，总共有 1440 条音频数据。

已有情绪分类算法多数是集中在一个数据集上进行训练，虽然效果尚可但泛化能力较差。为了消除数据集集中偏向的人为误差，提高模型的泛化能力，我们合并四个情绪识别领域常用的开源数据集并将 RAVDESS 数据集中的平静情绪改为中性情绪分类，TESS 数据集中的惊喜情绪改为惊讶情绪分类，得到按七种情绪分成七个文件夹的合并数据集。

特征提取

在语音识别领域，梅尔倒频谱系数 (Mel-Frequency Cepstral Coefficients, MFCC)是最为广泛运用的人工特征之一[8]。将输入的音频信号进行预加重、分帧和加窗，再进行 FFT 快速傅里叶变换，取绝对值或平方值后使用 Mel 滤波器组过滤，再对得到的能量值取对数 log 并进行离散余弦变换 DCT，即可输出每帧的特征向量 MFCC[9]。

分类器算法对比

简单算法

对提取的音频特征进行降维，得到轻量级特征数据。将数据集分割为训练集与测试集后选取 SVM、MLP、决策树、KNN、朴素贝叶斯、集成算法中的 Bagging&KNN 和随机森林等常见算法进行训练与检验，对比其准确率与时间成本。

复杂算法

保留提取的音频特征的全部维度，分割数据集后进行 2D CNN 算法训练与测试，记录其准确率与时间成本。

音频信号数据增强

对音频信号进行数据增强可以成倍扩充数据集并进一步提高算法性能与泛化能力。音频信号的数据增强一般是对音频的时频谱进行增强，常见的有噪声增强、时移增强和音高变换增强。噪声增强是在样本中加入白噪声，白噪声是服从给定的概率分布的随机变量，常见的有标准正态分布与均匀分布。本项目采用的是时间缩放和音高变换增强。时间缩放是在不改变音高的条件下改变声音的持续时间，音高变换是在不改变音速的条件下改变音高。用数据增强后的数据集重新训练算法，对比其准确率与时间成本。

实验建立

数据集可靠性验证

对于四个数据集中的数据，我们希望不同情绪之间的音频信号有显著差别，且不同数据集间的音频也有一定差别，从而有助于算法在不同情形下的泛化。我们从四个数据集愤怒情绪音频中各选一例，读取其音频文件得到音频的时间序列，再绘制波形的幅度包络线，如图 1 所示：

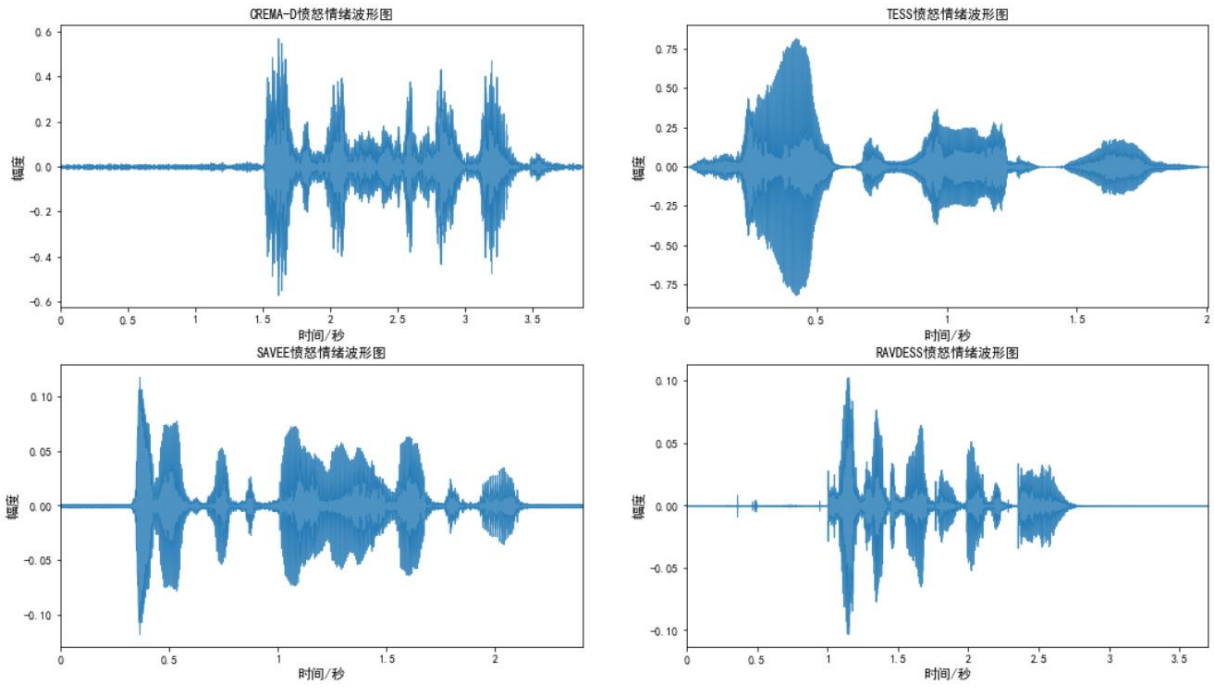

图 1

可见来自不同数据集同一情绪的波形图也有一定区别，开始时间、持续时间等都和其自身的采集数据方式有关。

我们从 RAVDESS 数据集的七种情绪中各取一例绘制波形图，如图 2 所示：

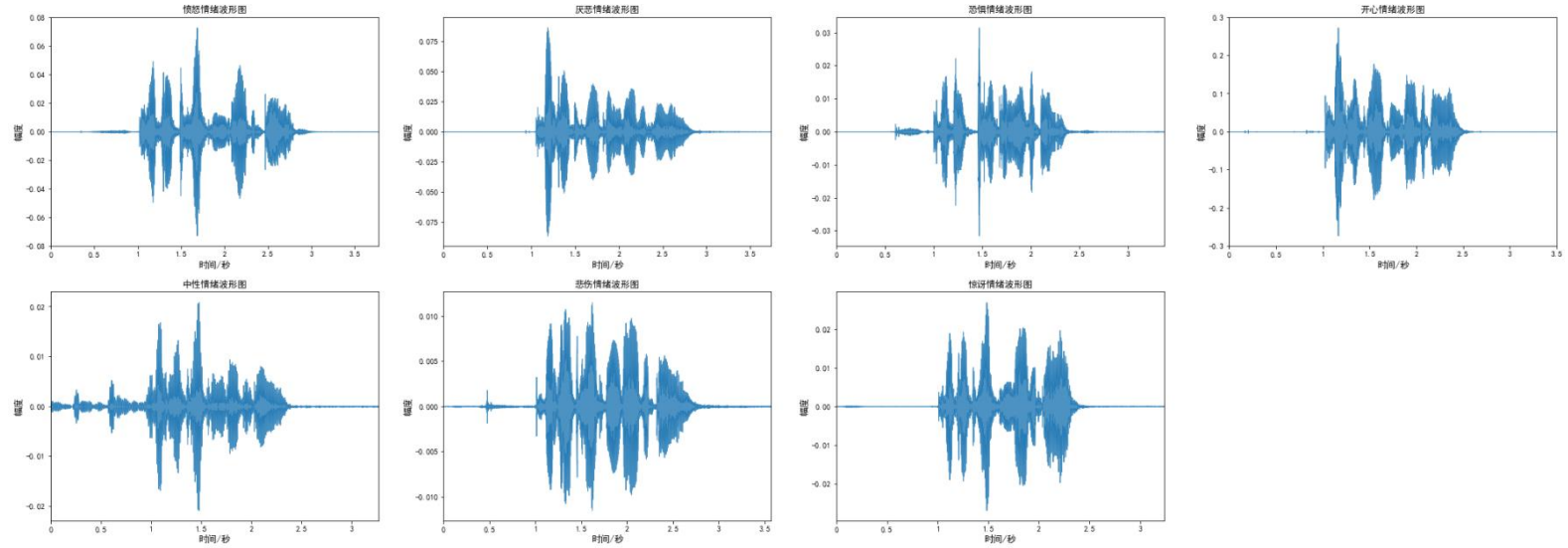

图 2

从图 2 可知音频数据在前 0.5 秒内几乎是环境噪声，从 0.5 秒后大概持续 2 至 2.5 秒的有效人声，故特征提取适合选择提取中间 2.5 秒包含信息最多的有效音频特征。

对音频信号进行 FFT 快速傅里叶变换，取绝对值后平方再除以样本长度即得到功率谱，如图 3 所示：

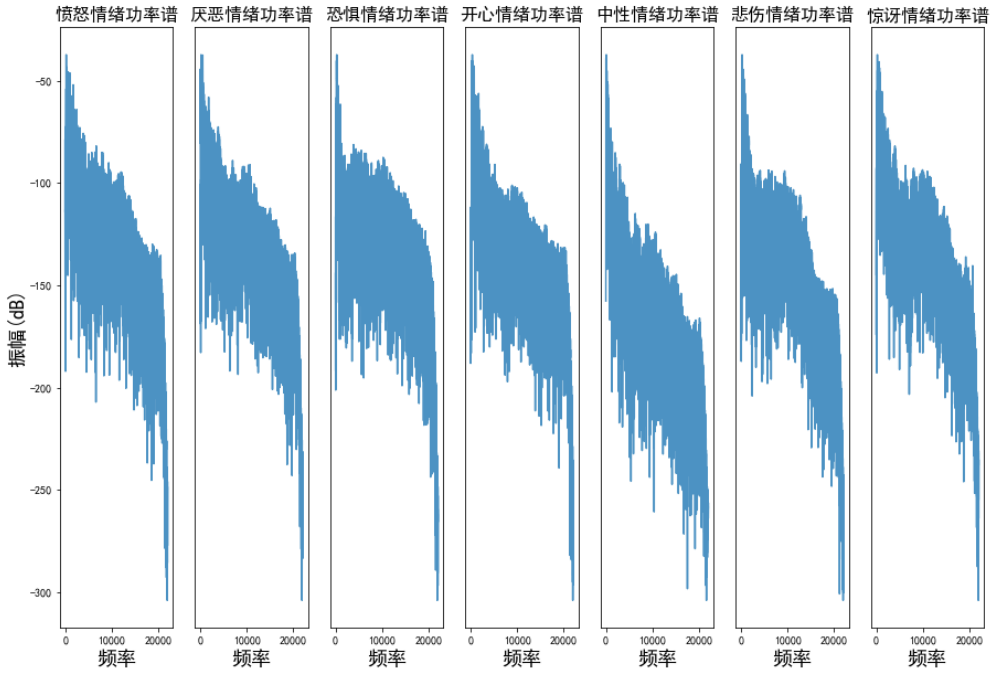

图 3

对音频信号进行 STFT 短时傅里叶变换，取绝对值后将振幅 S 转换为 $20*\log_{10}(S)$ dB 标度，得到七种情绪频谱，如图 4 所示：

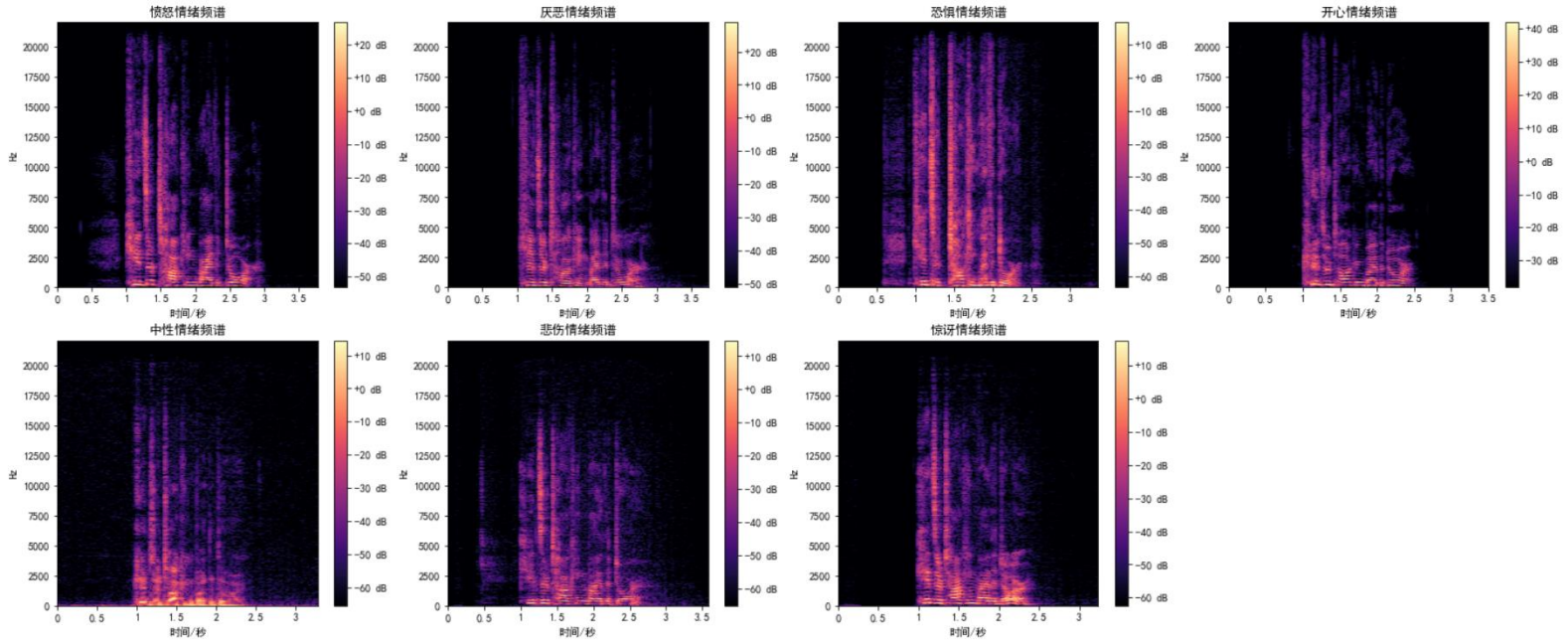

图 4

从波形图、功率谱和频谱中均能体现七种情绪之间的显著区别，这表明数据集的各类之间数据差别较大，利于分类器算法学习、分辨。

分类器算法对比

特征提取

MFCC 系数是语音信号处理领域常用的特征。将输入的音频信号进行预加重、分帧和加窗，再进行 FFT 快速傅里叶变换，取绝对值或平方值后使用 Mel 滤波器组过滤，再对得到的能量值取对数 log 并进行离散余弦变换 DCT，即可输出每帧的特征向量 MFCC。七种情绪样例以 4096 大小的窗口和 256 的相邻窗之间的距离

得到的 40 维特征向量 MFCC 如图 5 所示（Y 轴以对数刻度显示）：

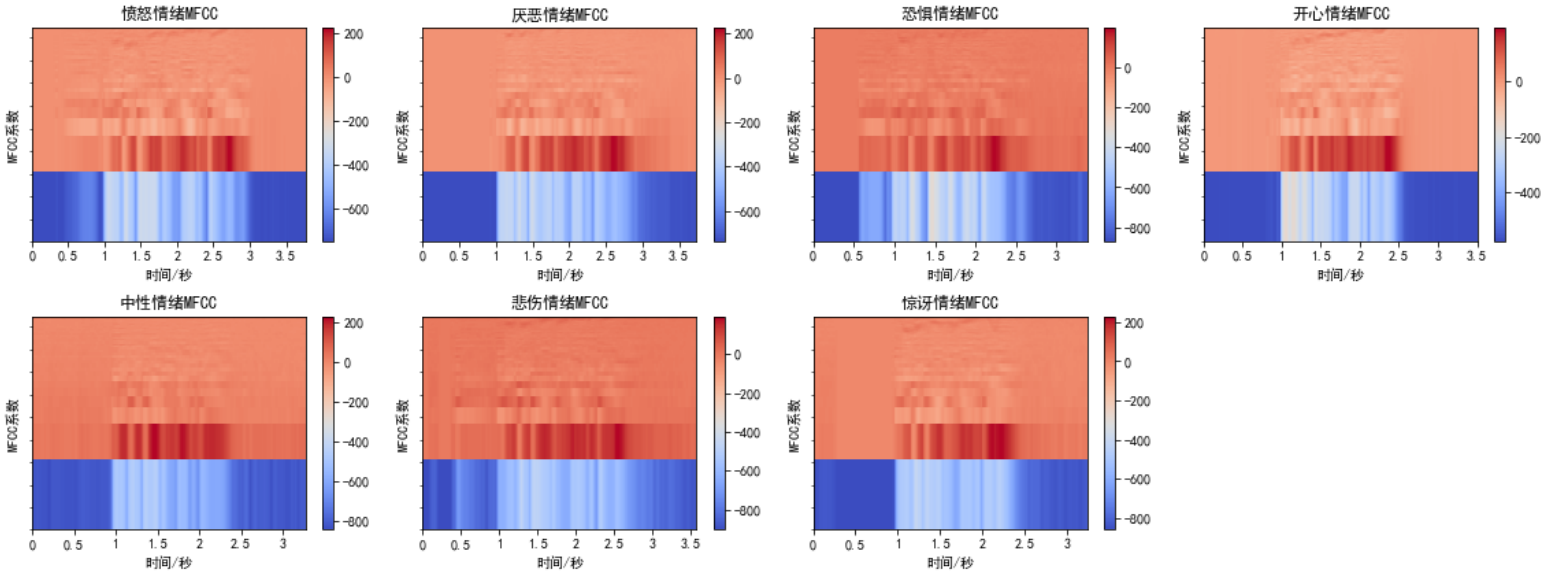

图 5

简单算法

提取每段音频每帧的 MFCC 系数特征向量，得到 431 个(40,1)的特征向量，对特征向量的各维度取平均值，得到一个(40,1)的最终 MFCC 特征，以其作为该音频对分类器算法的输入 X。打乱 X 的顺序后，将其按照 0.75 ： 0.25 的比例分为训练集与测试集。

MLP、SVM、朴素贝叶斯分类器、KNN、决策树、随机森林、Bagging&KNN 是常见的简单分类算法。七种情绪分类的准确率基线为 1/7=14.28%。我们选用两层隐藏层的 MLP 多层感知机，第一层隐藏层有 35 个神经单元，第二层隐藏层有 20 个神经单元，迭代训练 1000 次，选择激活函数为 ReLu，模型准确率为 60%。使用高斯概率密度公式进行分类训练的高斯朴素贝叶斯分类器准确率为 31%，SVM 支持向量机的模型准确率为 44%，决策树的准确率为 50%，设置 K=5 的 KNN 算法准确率为 55%。在集成算法方面，10 棵树的随机森林准确率为 58%，使用 Bagging 和 KNN 结合的模型准确率为 57%。

七种算法在测试集上的混淆矩阵热力图如图 6 所示：

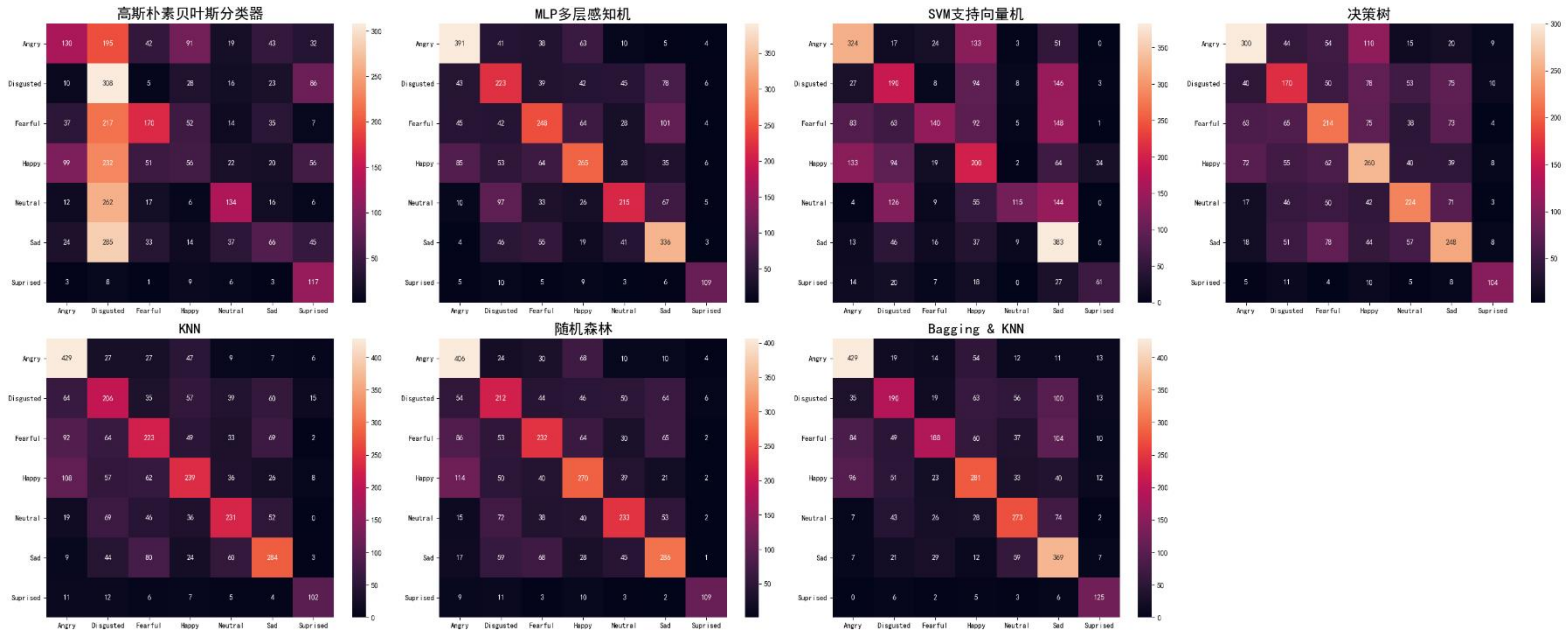

图 6

训练耗时方面，高斯朴素贝叶斯分类器用时 0.004534 秒，KNN 用时 0.08279 秒，Bagging&KNN 用时 0.178632 秒，随机森林用时 0.47729 秒，决策树用时 0.501679 秒，SVM 用时 6.384775 秒，MLP 用时 12.45273 秒。

复杂算法

简单算法中我们选择将每段音频的 MFCC 系数取平均值，压缩成一个特征向量作为输入，这样极大降低了模型的运算量，同时也保留了音频的基本特征，但在取平均值的过程中还是会有一定量的特征损失。

对每段音频提取每帧的 MFCC 系数特征向量，将得到的 431 个(40,1)的特征向量整体作为该音频对模型的输入 X，类似图片的 2D CNN 处理，将音频转化成(431, 40, 1)的 MFCC 系数矩阵。原数据集以 0.75 ： 0.25 的比例被分为训练集和测试集，模型采用 2D CNN，模型简介图如图 7 所示：

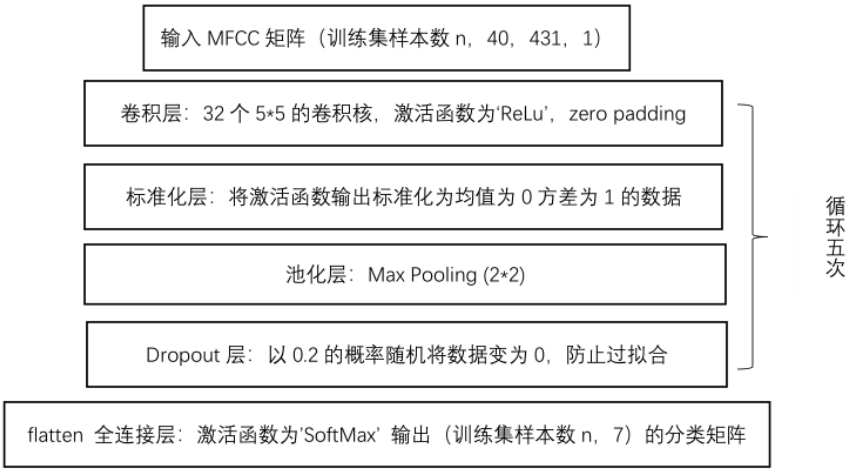

图 7

我们使用 5 对二维卷积层和最大值池化层，二维卷积层使用 32 个 5*5 的卷积核，激活函数为 ReLu，有补零 padding，并在每一层结束后添加标准化层，将数据转化成平均值为 0，方差为 1 的数据。为了防止过拟合，标准化层后再添加一个随机把数据设置为 0 的 Dropout 层，每个数据被设置为 0 的概率为 0.2。五层卷积池化后是 Flatten 和全连接层，激活函数为 SoftMax。训练 15 轮后准确率为 72%，训练 15 轮总时长 12964 秒，在测试集上的准确率热力图如图 8 所示：

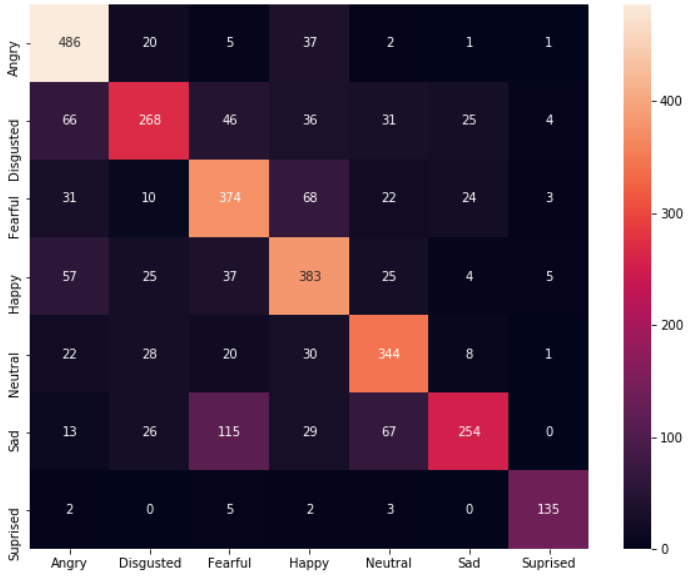

图 8

数据增强

数据增强是常用的成倍扩充数据集，提高模型准确率与泛化能力的方式。对于音频信号，我们采用时间缩放和音高变换两种数据增强方式，前者是在不改变音高的条件下改变声音的持续时间，后者是在不改变音速的条件下改变音高。对原数据集采用两种数据增强方式，使数据集扩充为原来的三倍。使用相同的特征提取方法和模型，高斯朴素贝叶斯分类器准确率从 31%上升为 32%，SVM 准确率从 44%上升为 46%，决策树准确率从 50%上升为 52%，KNN 准确率从 55%上升为 59%，随机森林准确率从 58%上升为 65%，Bagging&KNN 准确率从 57%上升为 63%，MLP 准确率从 60%下降为 59%。各模型在测试集上的准确率热力图如图 9 所示：

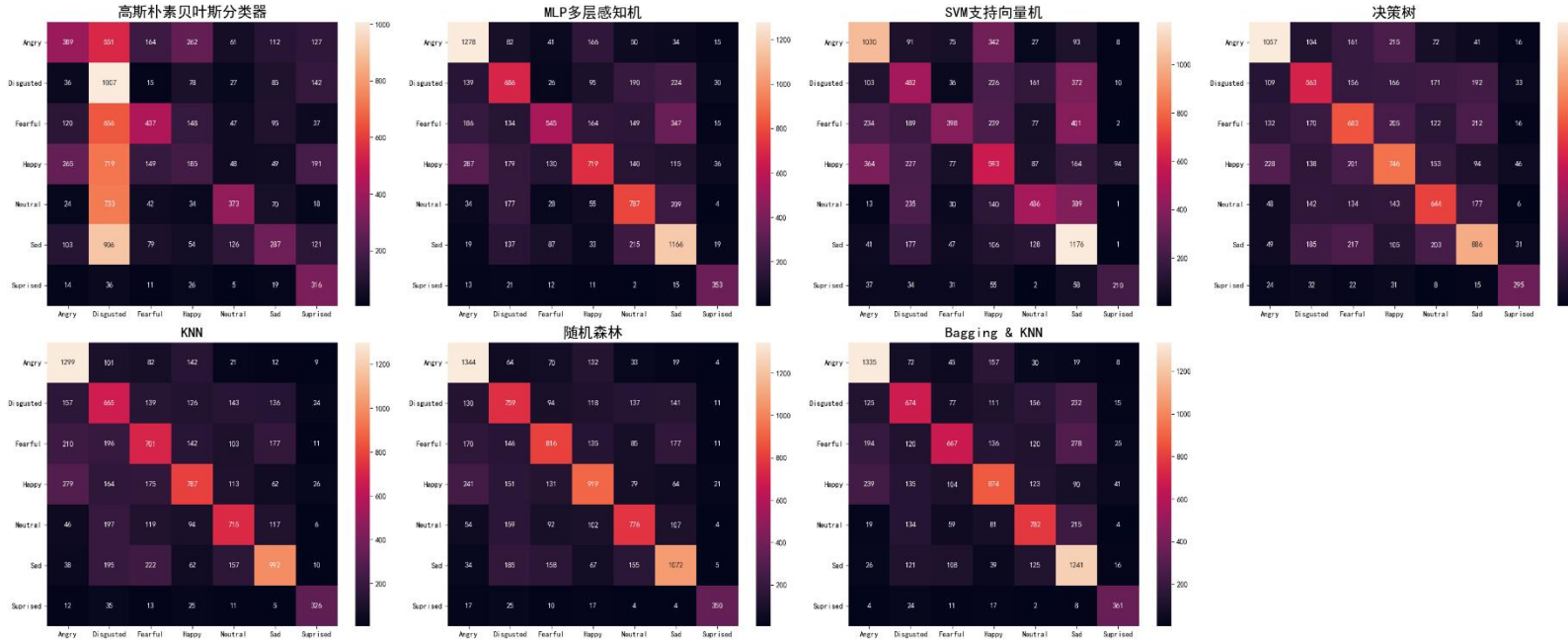

图 9

由于数据集变为原来的三倍，训练时长也成倍增加。高斯朴素贝叶斯分类器用时 0.031239 秒，KNN 用时 0.292218 秒，Bagging&KNN 用时 0.605522 秒，随机森林用时 1.560822 秒，决策树用时 1.998392 秒，MLP 用时 47.02697 秒，SVM 用时 62.56519 秒。

总结

模型准确率与模型训练耗时如下表所示，前两列为未进行数据增强时的模型数据，后两列为进行数据增强扩充数据集后的模型数据:

	模型准确率	训练用时/秒	模型准确率	训练用时/秒
高斯朴素贝叶斯分类器	31%	0.004534	32%	0.031239
MLP 多层感知机	60%	12.45273	59%	47.02697
SVM 支持向量机	44%	6.384775	46%	62.56519
决策树	50%	0.501679	52%	1.998392
KNN	55%	0.08279	59%	0.292218
随机森林	58%	0.47729	65%	1.560822
Bagging&KNN	57%	0.178632	63%	0.605522
CNN	72%	12964		

CNN 的模型准确率最高，达到 72%，超过大部分目前音频情绪分类算法的水准，但训练耗时将近 4 小时，时间成本极高，对算力要求较大。前七种轻量级分类算法中，朴素贝叶斯算法耗时最短但准确率最低，MLP 的准确率最高但耗时最长。随机森林和 Bagging&KNN 这两种集成算法的表现较好，耗时短且准确率较高，为 58%和 57%，仅次于 MLP，且在数据增强后准确率提高到 65%和 63%，为七种算法中的最高准确率。SVM 不适合做多维度大数据量的高负荷运算，故准确率较低且耗时偏长。

本项目聚焦信号处理中的音频信号处理领域，对 wav 音频文件进行分帧加窗读取，使用 FFT 和 STFT 绘制功率谱和频谱并提取 MFCC 系数作为音频的代表特征，输入多种模式识别分类器算法进行比较与总结，并采用时间缩放和音高变换两种音频信号数据增强方式扩充两倍数据集，进一步提高分类器算法的准确率和泛化能力。但该领域研究人员一直在探索克服音频情绪识别在实际应用中的多种阻碍，如数据集来源的欧美集中性等。本项目目前只利用了音频的 MFCC 系数，在未来算法的改进中，可以加入 Log Mel-spectrogram、HPSS、Chroma 等特征提高模型准确率。

附录

语音信号处理特征提取代码参考博客 https://www.cnblogs.com/LXP-Never/p/11561355.html 教程

软件环境版本

python 代码在实验过程中使用 pycharm 软件运行，其中 python 需要安装一些依赖包。特殊包的版本号如下：

librosa-0.7.2　　llvmlite-0.31.0　　numba-0.48.0
resampy-0.2.2　　scikit-learn-0.23.1　　soundfile-0.10.3.post1
python-3.6

引用

[1] UCL 脑科学学院官网介绍 https://www.ucl.ac.uk/brain-sciences/faculty-strengths/mental-health

[2] Kaushik, L., Sangwan, A., Hansen, J.H.L. Automatic audio sentiment extraction using keyword spotting(2015) Proceedings of the Annual Conference of the International Speech Communication

[3] Daneshfar, F., Kabudian, S.J., Neekabadi, A. Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier(2020) Applied Acoustics, 166

[4] CREMA-D 数据集 https://github.com/CheyneyComputerScience/CREMA-D

[5] TESS 数据集 https://tspace.library.utoronto.ca/handle/1807/24487

[6] SAVEE 数据集 http://kahlan.eps.surrey.ac.uk/savee/Database.html

[7] RAVDESS 数据集 https://zenodo.org/record/1188976#.XYP8CSgzaUk

[8] https://zh.wikipedia.org/wiki/%E6%A2%85%E7%88%BE%E5%80%92%E9%A0%BB%E8%AD%9C　　MFCC 系数维基百科

[9] MFCC 提取介绍 https://www.jianshu.com/p/24044f4c3531