arXiv:2303.03666v1 [cs.SD] 7 Mar 2023

Face：快速、准确且具上下文感知的音频标注与分类

预印本

M. Mehrdad Morsali
电气工程系
沙里夫理工大学
morsali@sharif.edu

Hoda Mohammadzade
电气工程系
沙里夫理工大学
hoda@sharif.edu

Saeed Bagheri Shouraki
电气工程系
沙里夫理工大学
bagheri-s@sharif.edu

2023年3月8日

摘要

本文提出了一个用于特征选择和分类流程的上下文感知框架，以实现快速、准确的音频事件标注与分类。上下文感知设计首先探索特征提取技术，以找出合适的组合，从而选出一组能够以最少计算开销实现显著分类准确率的特征。特征选择的探索还包括对音频节奏表示的研究；这种优势显著的特征提取方法此前在环境音频分类领域的研究中鲜少涉及。所提出的标注方法考虑离群点、内点和难以预测的数据样本，以实现上下文感知主动学习；在仅有15%的数据具有初始标注时，平均准确率可达90%。我们提出的声音分类算法在UrbanSound8K 数据集上获得了98.05%的平均预测准确率。包含源代码和实现结果的笔记本可通过 https://github.com/gitmehrdad/FACE 获取。

关键词 上下文感知机器学习 · 主动学习 · 声音分类 · 音频标注 · 音频节奏表示（节奏图） · 半监督音频分类 · 特征选择

1 引言

近年来，声音事件分类因其在音频分析、音乐信息检索、噪声监测、动物叫声分类和语音增强中的应用而受到关注 [1, 2, 3]。上述领域的发展需要已标注录音 [3] 和更大的数据集，而非较小的数据集 [4]。由于音频标注耗时且成本高昂，建立大规模已标注音频数据集并非易事 [5]。因此，开发可靠的数据集标注方法似乎十分必要。目前有多种方法可用于确定未标注数据的标签。一些论文，包括 [1, 4, 5, 6, 7, 8] ，采用半监督学习；另一些研究，包括 [2, 3, 9, 10, 11, 12] ，则尝试利用主动学习。鉴于基于主动学习的方法具有较高的标注准确率，本文也聚焦于主动学习，以确定音频样本的标签。在基于确定性的主动学习中，首先使用一小部分带有确定标签的数据训练分类器。随后，分类器为未标注数据赋予标签，并将分类确定性最高的音频样本加入训练集。分类器训练、分类和更新训练集这一循环会持续进行，直至所有音频样本均完成标注。

针对音频事件分类问题提出的解决方案探讨了多种方法。在 [1, 4, 5, 7, 8], 中，他们建议利用深度多层神经网络对数据进行分类。一些深度学习解决方案面临的主要挑战是：由于训练数据匮乏，模型发生过拟合时预测准确率较低。在音频分类任务中，通常有充足的训练样本，因此不会面临模型过拟合；然而，在仅有一小部分数据带有标签的音频标注任务中，过拟合问题会降低深度多层神经网络的吸引力。在 [2, 3, 6], 中，他们采用经典机器学习分类器对数据进行分类。这些方法可能较不易发生过拟合，因为训练过程中需要确定的未知参数较少。尽管如此，经典机器学习方法仍受固定输入大小问题的困扰。

鉴于上下文感知机器学习取得的显著成果 [13],，本文针对所讨论的问题提出了一些上下文感知解决方案。上下文感知设计会考虑当前情境来进行决策。举例来说，基于系统状态，上下文感知机器学习系统可能会动态更新模型微调和超参数调整。本文的第一项研究贡献是提出一种用于上下文感知特征选择的启发式方法，以提供经过汇总的固定大小输入向量，从而获得快速、准确且抗过拟合的分类模型。在音频信号分类任务中，对音频信号的循环节拍图特征进行研究，是本研究的另一项贡献。本文的第三项研究贡献是提出一种基于上下文感知主动学习的音频信号标注方法。第四项研究贡献是设计上下文感知分类器，以获得最佳分类结果。第2节介绍上述在数据准备与特征提取、数据标注和音频分类方面的贡献。第3节讨论所提出方法的实现结果。第4节总结全文。

2 所提出的方法

2.1 数据准备和特征提取

假设有一个音频信号数据集，其中音频信号的时长为四秒，采样率为 $22.05KHz$。由于每个音频信号由 4×22050 个采样点组成，在该数据集上运行的分类器的输入向量应包含 88200 个变量。这些变量或许能够区分不同的音频信号；然而，处理如此长的输入向量需要大量内存和昂贵的计算资源。此外，随着分类器输入向量长度的增加，训练得到的模型往往更容易出现过拟合 [14]。因此，与其将音频信号的时间内容视为分类器的输入向量，不如从信号的时间内容中提取一些区别性特征，并以此构成输入向量。上下文感知特征提取能够提供较高的分类准确率，并降低模型过拟合的可能性。

一些先前的研究，包括 [15, 16, 17, 18], ，讨论了适用于声音事件分类任务的各种特征提取方法。遵循上下文感知的方法，我们评估了不同的特征提取方法，以从这些方法中找出最有效的选择。由于本文追求快速分类，我们没有评估一些能够显著提高分类准确率的方法，例如散射变换，这是因为它们的特征提取速度较低。针对特征的消融研究设置包含一个数据集、一种特征提取方法和一个分类器。

本文采用的数据集是 [19], 中介绍的 UrbanSound8K（US8K），其中包含 10 类城市环境音频事件，包括空调、汽车喇叭、儿童玩耍、狗吠、钻孔、发动机怠速、枪声、风镐、警笛和街头音乐。每个类别包含数量不等的音频样本，且至少有一半类别的数据难以区分。该数据集包含 8732 个音频样本，采样率各不相同，最长时长为 4 秒。本文进行数据准备的第一步是将音频信号重采样至 $22.05KHz$ 采样率，以获得统一的数据集。由于音频信号的采样率通常高于 $22.05KHz$，重采样后往往需要进行数据压缩。由于 US8K 数据集未指定验证集，我们将 10% 的训练样本作为验证样本。本文中所有上下文感知决策均基于上述验证集。研究的特征包括起始强度包络的局部自相关（节奏图）、色度图、梅尔尺度谱图、梅尔频率倒谱系数（MFCCs）、频谱对比度、谱平坦度、谱带宽、谱质心、每帧音频样本的滚降频率和均方根值、音调质心特征（音调网络），以及音频时间序列的过零率。鉴于环境声音分类论文中缺乏相关研究历史，本文首次研究了 [20], 中提出的节奏图。

大多数上述特征提取方法都基于对音频信号进行加窗，因此提取出的特征被称为静态特征，因为它们表示信号特定窗口的信息。音频信号特征的差分系数和双重差分系数分别表示特征的一阶导数和二阶导数，被称为动态特征。差分特征体现语速，而双重差分特征

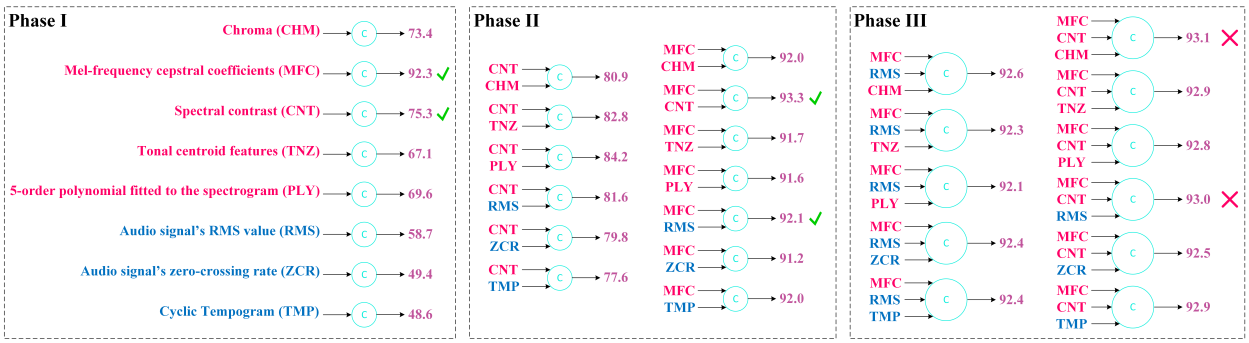

图1：基于所提出的启发式方法的特征选择结果

体现语音的加速度 [21]。一些特征提取方法会生成频谱图，即音频信号频谱的可视化表示。将频谱图视为图像后，我们可以运用计算机视觉领域中的一些特征提取属性，因此在求导之前，我们会对特征向量应用高斯滤波器。在 [14], 中，特征选择通过减少输入向量的维度，帮助克服模型的过拟合问题，并解决上一节提到的固定输入大小问题。输入向量的长度取决于特征提取过程中使用的窗口数量。窗口数量过多时，相邻窗口所表示的信息不会有显著差异；窗口数量过少时，提取出的特征量又不足以轻易区分不同的音频样本。为降低数据维度，可以用每个窗口提取特征的均值和方差替代原始数据。差分向量和双重差分向量也采用类似的替代方式。因此，最终的特征向量由六个浮点数向量组成，包含原始特征及其一阶和二阶导数的均值与方差值。本文使用了 Python 软件包 Librosa 提供的特征提取工具，其中大多数特征提取函数的参数均保留默认值。不过，我们根据验证数据调整了部分函数参数，包括窗口长度和每个窗口的样本数。实验开始前，我们利用分位数信息对特征进行变换，将其转换为标准向量，使所有元素的值都落在 (-1,1) 范围内。

具备上下文感知能力的分类器，应当是最善于区分数据的分类器。特征选择与分类器选择之间存在两难：分类器可能会掩盖某些特征所提供的区分能力，而特征集也可能会限制某些分类器所能达到的分辨能力。因此，我们首先根据对数据分布的推测，选用特定分类器来探索各种提取出的特征。完成特征选择后，我们再尝试找出最适合区分数据的分类器。

如 [22] 中 UrbanSound8K 数据集数据样本的 t-SNE图所示，特征空间中的数据分布具有异质性且稀疏。在异质性数据分布中，每个数据样本的邻域都包含一些类别标签不同的数据样本。稀疏数据的分布特点是，同一类别的数据样本不会形成密集区域。假设我们所做的特征选择也会产生异质且稀疏的数据分布，我们便需要寻找能够分离各类别数据的分类器。绘制线性决策边界的分类器无法分离稀疏且异质的数据，而采用非线性核的分类器则需要付出高昂的计算代价。因此，使用适当深度的决策树分类器似乎是不错的选择。使用决策树分类器容易导致模型过拟合，因此有必要采用树分类器集成。尽管有所改进，模型过拟合仍然会影响上述分类器的准确率，因此，XGBoost分类器通过正则化移动决策边界，从而提供了一种解决方案 [14]。受一对多方法启发，我们提出采用一对多XGBoost分类器；该方法可将二元分类器用于多类分类。实际上，我们不使用单个多类XGBoost分类器，而是由多个二元XGBoost分类器绘制决策边界，每个分类器都尝试将特定类别的成员与所有其他数据区分开来。鉴于二元XGBoost的计算负担比多类XGBoost轻，所提出的方法不会面临计算开销问题。

图1展示了特征选择结果。为避免检查所有特征排列组合，我们采用了一种启发式方法：每个阶段都向每个特征集添加一个新特征，报告验证样本的分类准确率，将排名前两位的特征集推进到下一研究阶段，并在分类准确率不再提高时停止进一步研究。图1中，第三阶段的结果相比第二阶段没有改善；因此，我们选择第二阶段的 MFC+CNT 组合为最终方案。图1展示了实验第二阶段和第三阶段中各方法的缩写，并以蓝色表示时域特征、红色表示频域特征。图1未展示梅尔尺度频谱图特征

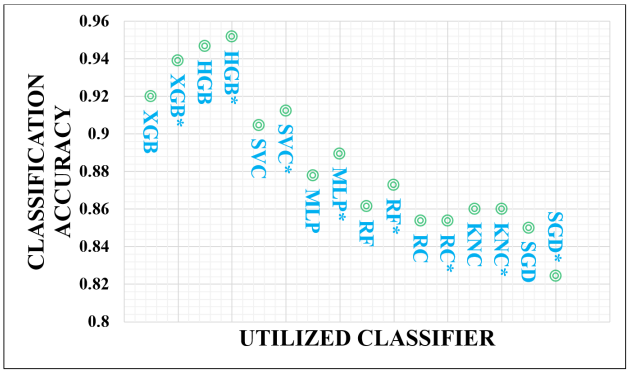

图 2：分类准确率与所用分类器的关系

提取结果，因为梅尔频率倒谱系数特征不仅可以转换为梅尔尺度频谱图特征，还能提供更高的分类准确率。由于分类准确率结果不佳，谱平坦度、谱带宽、谱质心和滚降频率等方法未纳入图1。请注意，所报告的准确率来自简单的分类流程，其中未应用主动学习或其他准确率优化技术。如图1所示，梅尔频率倒谱系数特征单独使用时表现出的出色准确率，说明了该特征在以往大多数研究中被广泛采用的原因。所选组合的特征向量由810个浮点数组成，大幅减少了标注和分类任务所需的计算资源，因为只需检查810个特征，而不是88200个特征。

图 2 展示了采用不同分类器得到的分类结果，过程中未使用主动学习或其他准确率优化技术。本实验旨在为标注任务找到最佳分类器，因此仅考察经典机器学习方法。图 2 中，XGB、HGB、SVC、MLP、RF、RC、KNC 和 SGD 分别表示 XGBoost、基于直方图的梯度提升分类器、支持向量机、多层感知机、随机森林、岭分类器、K近邻分类器，以及采用随机梯度下降进行学习的正则化线性支持向量机。图 2 中，分类器名称后的星号表示该分类器采用一对多方案。每个分类器均使用我们定义的验证集作为测试集，并将 US8K 数据集中其余样本作为训练集。图 2 展示了分类准确率至少达到 80% 的分类器结果，因此未包含决策树、高斯朴素贝叶斯、AdaBoost 和二次判别分析等未能达到 80% 分类准确率的分类器。图 2 的结果印证了我们对数据分布的推测，也支持通过具上下文感知能力的分类器选择来区分类别。如图 2 所示，HGB 分类器的分类准确率优于 XGB 分类器，但 XGB 分类器运行速度更快。因此，对于音频标注任务，我们将使用一对多 XGBoost 分类器；这是一个具上下文感知能力的选择，能够在验证集上实现令人满意的分类准确率。

2.2 数据标注

对于标注任务，通常假设人工标注员为少量数据提供标签。这样的假设使我们避免使用深度神经网络分类器来标注新样本，因为训练样本数量少而模型参数数量多，会使模型容易过拟合，进而导致预测准确率下降。因此，基于第2.1节中观察到的结果，我们将使用一对多XGBoost分类器对数据进行标注；这是一种上下文感知的选择，能够提供良好的分类准确率。上下文感知框架在我们所提出的标注方法中的另一种体现，是指定人工标注员应标注的数据。算法1展示了我们提出的标注方法。 [23], 中提出的局部离群因子（LOF）方法会找出与其最近邻之间欧氏距离最小的内点样本。主动学习器是modAL软件包中的一个Python工具，用于检测难预测数据，即预测确定性最低的样本。根据算法1第1行至第8行所示内容，人工标注员为内点样本、难预测样本以及从每个不同类别中均匀选取的一些随机样本确定标签。选取的内点、难预测样本和随机样本的最大数量，分别限制为数据集总样本数的2.5%、5%和7.5%。选择内点样本进行初始标注的理念是，内点样本的邻域很可能包含同一类别的样本。将一些随机数据样本加入内点样本的原因在于，不同类别的内点样本数量可能存在不均衡。添加随机样本的另一个原因，是寻找与内点样本相邻的不同类别数据样本。

算法1： 所提出的音频事件标注方法

数据： $dataset$：包含未标注数据的数据集
结果： $Label$：该数据集的预测标签

1 $Label \leftarrow null$;
2 $L \leftarrow dataset.SamplesCount$;
3 $LocalOutlierFactor(neighbors = 1).Fit(dataset)$;
4 $I \leftarrow LocalOutlierFactor.Inlier(size = L/40)$;
5 $R \leftarrow RandomSelection(data = dataset,\ \ size = 3L/40)$;
6 $Q \leftarrow ActiveLearner.query(data = dataset,\ \ size = L/20)$;
7 $T \leftarrow Concatenate(I, R, Q)$;
8 $Label[T] \leftarrow AskLabel(T)$;
9 **从** $i \leftarrow 1$ **到** 4 **执行**
10 $Labeller.fit(dataset[T], P[T])$;
11 $Prediction \leftarrow Labeller.Predict(dataset)$;
12 $Score \leftarrow Labeller.PredictionScore(dataset)$;
13 **从** $j \leftarrow 1$ **到** L **执行**
14 $Outlier \leftarrow LocalOutlierFactor.IsOutlier(dataset[j])$;
15 **如果** $Label[j] = null$ **且** $Outlier = False$ **且** $Score[j] > 0.7$ **则**
16 $Label[j] \leftarrow Prediction[j]$;
17 $T.Append(j)$;
18 **end**
19 **end**
20 **结束**
21 $Labeller.fit(dataset[T],\ \ P[T])$;
22 $Prediction \leftarrow Labeller.Predict(dataset)$;
23 **对于** $j \leftarrow 1$ **到** L **执行**
24 **如果** $Label[j] = null$ **那么**
25 $Label[j] \leftarrow Prediction[j]$;
26 **end**
27 **结束**

算法1第9行至第27行说明了所提出的上下文感知主动学习方法。该方法包含五个数据标注阶段，其中具有可靠预测准确率的预测样本也有助于确定标签。在本研究中，可靠预测至少具有70%的预测准确率分数。主动学习阶段的数量以及接受某个预测为可靠预测的阈值，是该问题的超参数，取决于多种因素，包括音频样本数量、数据可分性、分类器的准确率以及分类的确定性。我们使用验证集对该问题的超参数进行了调整，验证集占未标注样本的10%。我们方法的主要创新在于，在最后一个标注阶段之前不对离群样本进行标注，因为我们认为，训练样本中存在离群数据样本会降低标注准确率。在基于主动学习的标注第一阶段，我们使用人工标注数据训练分类器。随后，训练好的分类器预测数据集样本的标签。接下来，将可靠预测加入最终预测列表，并将其加入后续标注阶段分类器的训练样本集合。在标注流程的最后阶段，为离群样本和其余未标注数据预测的标签将无条件地加入最终预测列表。

2.3 数据分类

图3展示了所提出的音频分类神经网络。图3中，蓝色部分表示卷积块，绿色部分表示局部连接卷积块。局部连接卷积块的工作方式与卷积块类似；然而，与卷积块不同的是，卷积核窗口会在每个选定的输入块上应用不同的滤波器组。每个块上的数字表示该块的卷积核尺寸，每列块上方的数字表示滤波器数量。所有块均采用批归一化，之后接 LeakyReLU 激活函数。我们通过加法运算合并图3所示的五个并行流的输出。神经网络的最后几层由两个全连接层组成，两者通过一个 Dropout层连接。我们使用一个包含 10% 训练样本的验证集调整神经网络的超参数，包括滤波器数量、卷积核尺寸、Dropout比例和训练轮数。在神经网络设计中，我们倾向于宽度扩展而非深度扩展，这使神经网络能够通过为五个并行流提供并行计算的机会而更快地运行。

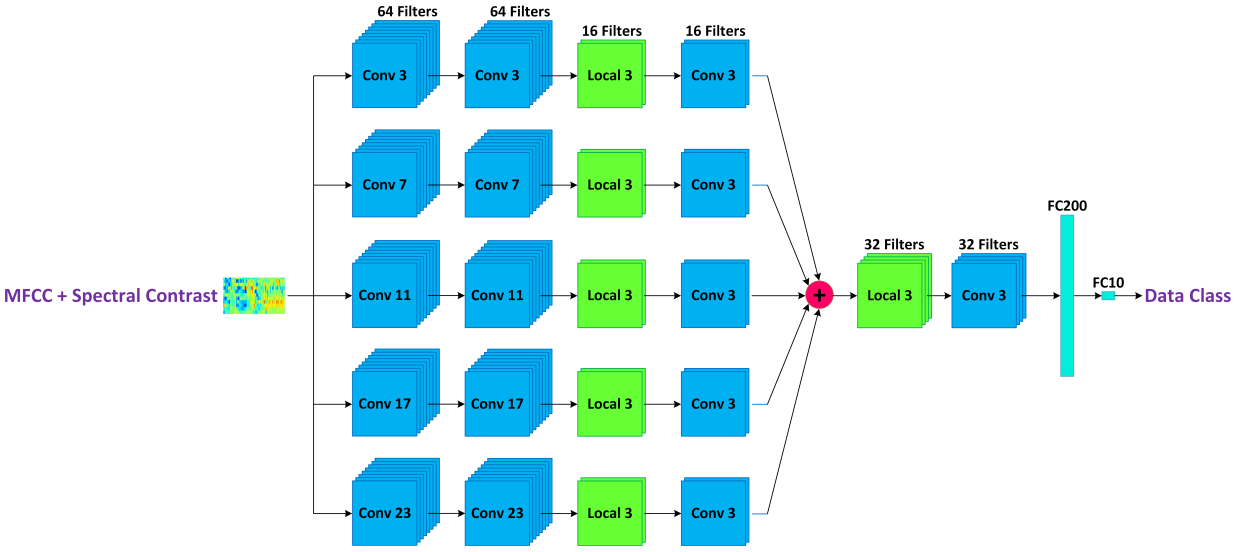

图3：所提出的音频分类任务神经网络

表1：标注预算为15%时的标注准确率结果

方法	方法	准确率
SSCDE[6]	半监督学习	35
CRTAL+SSL[9, 10]	主动学习和半监督学习	50
LDAL[11]	基于字典的主动学习	55
DBAL[12]	基于字典的主动学习	57
MAL[3]	主动学习	61
本研究（FACE）	**上下文感知主动学习**	**90**

3 实施结果

按照本文参考文献的惯例， [3, 6, 9, 10, 11, 12],我们在 UrbanSound8K 数据集上测试了所提出的上下文感知标注方法。如第2节所述，我们假设标注预算为 15%，这意味着在标注过程开始前，只有 15% 的数据具有确定标签。表1展示了标注准确率结果，准确率达到 90%，较此前报道的最佳结果提高了 29%。

对于声音分类任务，我们使用 SGD 优化器训练神经网络 25 轮，学习率为 0.01。我们确定了训练轮数的最大值和优化器设置，使验证数据（训练集的 10%）上的分类准确率达到最高。鉴于 UrbanSound8K 数据集在音频分类任务中的广泛应用，表2展示了 UrabanSound8K 数据集上的分类准确率结果。如表2所示，我们提出的方法实现了 98.05% 的分类准确率，高于迄今为止报告的任何其他研究结果。作为对我们分类方法进行质量评估的另一项指标，图4展示了 UrbanSound8K 数据集的分类混淆矩阵。为研究主动学习对提升分类准确率的作用，我们让主动学习器（modAL 软件包中的 Python 工具）查询部分测试集样本的标签。我们发现，只需提供 0.9% 测试数据的标签，就能达到 100% 的分类准确率。

包含我们源代码和实现结果的笔记本可在 https://github.com/gitmehrdad/FACE 获取。

4 结论

本文提出了一种快速、准确且具备上下文感知能力的方法，可用于标注和分类音频信号。在特征选择、分类器设计和标注算法设计中采用上下文感知方法

表2：分类准确率结果

方法	准确率
MAL[3]	65
CRTAL+SSL[9, 10]	65
DBAL[12]	67
LDAL[11]	68
CDSS[5]	78
AemNet-DW[24]	83.6
不含 FF本体的图[22]	89.0
TFCNN[25]	93.1
SacNet-8[18]	95.3
TSCNN[26]	95.7
TSCNN-DS[17]	97.2
MHAT[15]	97.5
本工作（FACE）	**98.05**
本研究（FACE + 主动学习）	**100**

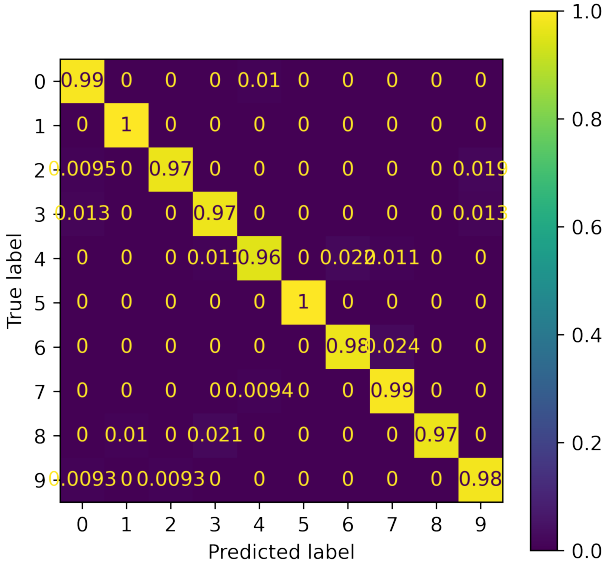

图 4：US8K 数据集上的分类混淆矩阵

显著提高了我们的方法在标注和分类任务中的准确率。紧凑的提取特征向量和计算效率高的分类方案使标注器能够快速完成标注：借助免费的 Google Colaboratory 计算平台，在不到一分钟内以 90% 的准确率标注近 8000 个音频样本。我们提出的方法在 US8K 数据集上取得了 98.05% 的平均分类准确率。与先前的研究相比，所取得的分类准确率为目前报道的最高水平。主动学习流程只需询问 0.9% 的测试数据的标签，即可将分类准确率提升至 100%。

参考文献

[1] Siddharth Gururani 和 Alexander Lerch。利用部分标注数据进行半监督音频分类。载于 *2021*年 *IEEE* 国际多媒体研讨会（*ISM*），第111–114页，2021年。doi:10.1109/ISM52913.2021.00027。[2] Yu Wang、Ana Elisa Mendez Mendez、Mark Cartwright 和 Juan Pablo Bello。利用大量未标注数据进行高效音频标注与分类的主动学习。载于 *ICASSP 2019 - 2019*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第880–884页，2019年。doi:10.1109/ICASSP.2019.8683063。

[3] Zhao Shuyang、Toni Heittola 和 Tuomas Virtanen。通过对未标注数据进行聚类的声音事件分类主动学习。载于 *2017*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第751–755页，2017年。doi:10.1109/ICASSP.2017.7952256。[4] Eduardo Fonseca、Shawn Hershey、Manoj Plakal、Daniel P. W. Ellis、Aren Jansen 和 R. Channing Moore。采用带有损失掩码的教师-学生框架解决大规模声音事件识别中的缺失标签问题。*IEEE* 信号处理快报，27:1235–1239，2020年。doi:10.1109/LSP.2020.3006378。[5] Donmoon Lee 和 Kyogu Lee。使用对比正则化的跨域半监督音频事件分类。载于*2021*年 *IEEE* 音频与声学信号处理应用研讨会（*WASPAA*），第196–200页，2021年。doi:10.1109/WASPAA52581.2021.9632721。[6] Jun Takagi、Yasunori Ohishi、Akisato Kimura、Masashi Sugiyama、Makoto Yamada 和 Hirokazu Kameoka。通过半监督典型密度估计进行自动音频标签分类。载于*2011*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第2232–2235页，2011年。doi:10.1109/ICASSP.2011.5946925。[7] Teck Kai Chan、Cheng Siong Chin 和 Ye Li。用于声音事件检测的半监督 NMF-CNN。*IEEE Access*，9:130529–130542，2021年。doi:10.1109/ACCESS.2021.3113903。[8] Diego De Benito-Gorrón、Daniel Ramos 和 Doroteo T. Toledano。一种基于多分辨率CRNN的 DCASE 2020挑战赛半监督声音事件检测方法。*IEEE Access*，9:89029–89042，2021年。doi:10.1109/ACCESS.2021.3088949。[9] Dilek Hakkani-Tür、Giuseppe Riccardi 和 Allen Gorin。用于自动语音识别的主动学习。载于 *2002*年 *IEEE* 国际声学、语音与信号处理会议，第4卷，第IV–3904–IV–3907页，2002年。doi:10.1109/ICASSP.2002.5745510。[10] Malte Darnstädt、Hendrik Meutzner 和 Dorothea Kolossa。通过主动学习和半监督学习降低破解音频验证码的成本。载于 *2014*年第*13*届机器学习与应用国际会议，第67–73页，2014年。doi:10.1109/ICMLA.2014.16。[11] Xiao Qin、Wanting Ji、Ruili Wang 和 ChangAn Yuan。用于环境声音事件标注的学习型基于字典的主动学习方法。多媒体工具与应用，78，2019年。doi:10.1007/s11042-018-7139- 2。[12] Wanting Ji、Ruili Wang 和 Junbo Ma。用于声音事件分类的基于字典的主动学习。多媒体工具与应用，78，2019年。doi:10.1007/s11042-018-6380-z。[13] Yun Zeng。上下文感知机器学习。2019年。doi:10.48550/ARXIV.1901.03415。[14] Trevor Hastie、Robert Tibshirani 和 Jerome Friedman。统计学习基础。统计学 Springer 丛书。Springer New York Inc.，2009年。[15] Shen Song、Cong Zhang 和 Zhihui Wei。基于自注意力机制的城市声音事件检测散射变换研究。*IEEE Access*，10:120804–120822，2022年。doi:10.1109/ACCESS.2022.3222495。[16] Theodoros Theodorou、Iosif Mporas 和 Nikos Fakotakis。城市环境事件的自动声音识别。载于 Andrey Ronzhin、Rodmonga Potapova 和 Nikos Fakotakis（编），语音与计算机，第129–136页。Springer International Publishing，2015年。doi:10.1007/978-3-319-23132-7_16。[17] Yu Su、Ke Zhang、Jingyu Wang 和 Kurosh Madani。采用基于决策级融合的双流CNN进行环境声音分类。传感器，19(7)，2019年。doi:10.3390/s19071733。[18] Minze Li、Wu Huang 和 Tao Zhang。采用多频率分辨率特征的基于注意力机制的卷积神经网络，用于环境声音分类。神经处理快报，2022年。doi:10.1007/s11063- 022-11041-y。[19] Justin Salamon、Christopher Jacoby 和 Juan Pablo Bello。用于城市声音研究的数据集和分类体系。载于 第*22*届 *ACM* 国际多媒体会议论文集。ACM，2014年。doi:10.1145/2647868.2655045。[20] Peter Grosche、Meinard Müller 和 Frank Kurth。循环节拍图——一种用于音乐信号的中层节奏表示。载于 *2010*年 *IEEE* 国际声学、语音与信号处理会议，第5522–5525页，2010年。doi:10.1109/ICASSP.2010.5495219。[21] K. Sreenivasa Rao 和 K E Manjunath。使用发音与激励源特征的语音识别。语音技术 SpringerBriefs 丛书。Springer Cham，2017年。doi:10.1109/ICASSP.2012.6288282。[22] Yiwei Sun 和 Shabnam Ghaffarzadegan。一种用于音频事件分类的本体感知框架。载于*ICASSP 2020 - 2020*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第321–325页，2020年。doi:10.1109/ICASSP40776.2020.9053389。

[23] Markus M. Breunig、Hans-Peter Kriegel、Raymond T. Ng 和 Jörg Sander。LOF：识别基于密度的局部离群点。*SIGMODRec.*，29(2):93–104，2000年5月。doi:10.1145/335191.335388.[24] Paulo Lopez-Meyer、Juan A. del Hoyo Ontiveros、Hong Lu 和 Georg Stemmer。面向目标应用的音频分类高效端到端音频嵌入生成。载于*ICASSP 2021 - 2021*年 *IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 601–605 页，2021年。doi:10.1109/ICASSP39728.2021.9414229.[25] Wenjie Mu、Bo Yin、Xianqing Huang、Jiali Xu 和 Zehua Du。使用基于时频注意力的卷积神经网络进行环境声音分类。*Scientific Reports*，11，2021年。doi:10.1038/s41598-021-01045-4.[26]Xifeng Dong、Bo Yin、Yanping Cong、Zehua Du 和 Xianqing Huang。采用双流卷积神经网络进行环境声音事件分类。*IEEEAccess*，8:125714–125721，2020年。doi:10.1109/ACCESS.2020.3007906。