基于联合学习框架的音频场景聚类

张聿晗,李艳雄,江钟杰,陈 昊
(华南理工大学电子与信息学院,广东广州 510640)

摘 要: 音频场景聚类的任务是将属于相同音频场景的音频样本合并到同一个类中.本文提出一种基于联合学习框架的音频场景聚类方法.该框架由一个卷积自编码网络(Convolution Autoencoder Network, CAN)与一个判别性聚类网络(Discriminative Clustering Network, DCN)组成.CAN包括编码器和译码器,用于提取深度变换特征,DCN用于对输入的深度变换特征进行类别估计从而实现音频场景聚类.采用DCASE-2017和LITIS-Rouen数据集作为实验数据,比较不同特征与聚类方法的性能.实验结果表明:采用归一化互信息和聚类精度作为评价指标时,基于联合学习框架提取的深度变换特征优于其他特征,本文方法优于其他方法.本文方法所需要付出的代价是需要较大的计算复杂度.

关键词: 音频场景聚类; 联合学习框架; 卷积自编码网络; 判别性聚类网络
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112(2021)10-2041-07
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200573

Audio Scene Clustering Based on Joint Learning Framework

ZHANG Yu-han, LI Yan-xiong, JIANG Zhong-jie, CHEN Hao
(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong* 510640, *China*)

Abstract: Audio scene clustering (ASC) is a task to merge audio samples belonging to the same type of acoustic scene into a single cluster. This paper proposes a method of ASC based on joint learning framework. The proposed framework consists of a convolution autoencoder network (CAN) and a discriminative clustering network (DCN). The CAN is used to extract deep transformed feature (DTF), while the DCN is used to do cluster estimation on the input DTF for realizing ASC. Two datasets, DCASE-2017 and LITIS-Rouen, are used as experimental data, and the performance of different features and clustering methods are compared. Experimental results show that the DTF extracted by the joint learning framework outperforms other features, and our method is superior to other methods, in terms of the metrics of both normalized mutual information and clustering accuracy. The cost of the proposed method is the higher computational complexity.

Key words: audio scene clustering; joint learning framework; convolutional autoencoder network; discriminative clustering network

1 引言

随着智能手机等便携式记录设备的普及和移动互联网的广泛应用,各种音频场景数据爆炸式增长[1~3].面对海量音频样本,人工标注成本昂贵且具有很大的主观性,在实际应用中存在大量的无标签、不可靠标签的音频样本.如何将无标签、不可靠标签的音频样本聚集成不同类别的音频场景,即如何解决音频场景无监督聚类问题,是当前音频处理领域的一个研究热点.

从2013年开始,由美国电子电气工程师协会主办,英国玛丽女王大学、芬兰坦佩雷理工大学、美国卡耐基梅隆大学等科研单位承办的音频场景分类与事件检测(Detection and Classification of Acoustic Scenes and Events, DCASE)比赛吸引了全球众多知名单位的研究者,使得音频场景分类获得广泛关注[4].然而,由于强背景噪声、各种音频场景类内差异变化大等因素,音频场景分类目前并没有被有效解决[5].音频场景分类主要包括两个模块:特征提取和分类器构建.研究人员主要围绕这两个模块开展相关工作.常用的声学特征包括梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)、对数梅尔谱(Log-Mel Spectrum, LMS)、Gabor滤波器组、梯度特征直方图及局部二值模式[6~9].此外,基于矩阵分解和神经网络的变换特征也被用来

收稿日期:2020-06-15;修回日期:2020-10-04;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61771200)

弥补常用声学特征的不足[10~12]. 分类器主要包括卷积神经网络[13](Convolution Neural Network, CNN)、门控循环神经网络[14](Gated Recurrent Neural Network, GRNN)、双向长短时记忆网络[5](Bidirectional Long Short Term Memory Network, BLSTMN)、高斯混合模型(Gaussian Mixture Model, GMM)、支持向量机(Support Vector Machine, SVM)和隐马尔科夫模型(Hidden Markov Model, HMM)[6,7]. 例如,Wu等人采用高斯差分、Sobel算子及背景漂移处理对数梅尔谱特征,增强时频谱图的边缘信息,然后将处理完的频谱特征输入卷积神经网络进行分类[15]. Singh等人采用深度卷积神经网络进行音频场景分类,通过组合神经网络每层的得分进行判决[16]. McDonnell等人采用深度残差网络、对数梅尔谱及其一阶和二阶差分进行音频场景分类,并在残差网络中使用两个通路分别处理低频信息和高频信息[17]. Chen等人提出一个混合框架,联合训练前端滤波器和后端深度卷积神经网络,同时,在网络的高层特征谱图之后插入一个基于离散余弦变换的时域模块[18]. Zhang等人提出一种可以使神经网络快速实现傅里叶变换的模块,在多个实验数据集的测试结果表明,所提出的变换模块可以同时提高音频场景分类速度和精度[19]. Bai等人提出一种声学段建模的高分辨率注意力网络用于音频场景分类,首先采用卷积神经网络获得高层语义信息,然后采用两步注意力策略选择相关音频场景段[20]. Zhang等人为了获取整个音频样本的时间信息,提出一种基于分层金字塔状池化层的音频表征学习方法,在全局时间池化层,他们联合优化一个可学习的区分性映射和Softmax分类器[21]. Phaye等人提出一种基于子谱图的卷积神经网络,通过将频带层差异并入模型子空间的方式获取区分性特征[22].

从上述介绍可知,目前的研究工作主要研究音频场景分类问题,将某个测试样本判别成已知音频场景类别中的某一种. 然而,在实际应用中,由于标签丢失及人工标注成本昂贵等原因,绝大部分待处理音频样本的音频场景类别及种类并不可知. 当处理海量音频样本时,初始任务可能是确定哪些样本属于同一种音频场景类别,而不是辨识它们的具体类别标签,也就是需要解决音频场景聚类问题.

目前音频场景聚类的研究报道非常少. 例如,Li等人提出一种基于稀疏子空间的音频场景聚类方法[23]和基于增强流的子空间聚类方法[24]. 他们所采用的输入特征都是梅尔频率倒谱系数. 上述音频场景聚类方法存在以下不足:首先,所采用的常规声学特征并不能有效刻画各种音频场景之间的特性差异;其次,特征提取与类别估计过程是独立顺序进行的,没有联合迭代优化. 因此,前端提取的特征对后端的类别估计可能并不友好.

为了克服目前音频场景聚类方法存在的上述不足,本文提出一种基于联合学习框架的音频场景聚类方法. 该框架由一个卷积自编码网络与一个判别性聚类网络组成,前者用于提取深度变换特征,后者用于对输入的深度变换特征进行类别估计从而实现音频场景聚类.

2 聚类方法

假设N_s个待聚类样本的集合$\boldsymbol{S}=\{\boldsymbol{s}_i\}$,$1\leqslant i\leqslant N_s$;从各音频样本提取对数梅尔谱特征的集合$\boldsymbol{X}=\{\boldsymbol{x}_i\}$,$1\leqslant i\leqslant N_s$,并作为卷积自编码网络的输入;从卷积自编码网络提取的深度变换特征集合$\boldsymbol{Z}=\{\boldsymbol{z}_i\}$,$1\leqslant i\leqslant N_s$. 音频场景聚类的目标是要将$\boldsymbol{Z}$合并为$N_c$个不同的类$\boldsymbol{C}=\{\boldsymbol{c}_k\}$,$1\leqslant k\leqslant N_c$,其中,$\boldsymbol{c}_k=\{\boldsymbol{z}_i|y_i=k;\ 1\leqslant i\leqslant N_s,\ 1\leqslant k\leqslant N_c\}$,$y_i$表示样本$\boldsymbol{s}_i$的音频场景类别预测值.

2.1 联合学习框架

本文提出的联合学习框架如图1所示,由卷积自编码网络与判别性聚类网络构成. 前者包括编码器和译码器;后者包括全连接层和Softmax层.

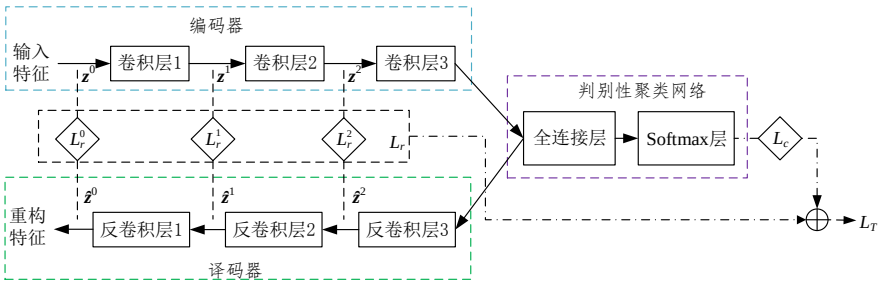

图1 联合学习框架

编码器由多个卷积层组成,每个卷积层对输入的特征进行变换. 译码器由多个反卷积层组成,每个反卷积层对输入矢量进行特征重构. 编码器每一层最后进行Dropout操作[10],其输出为

$$\boldsymbol{z}_i^m=\mathrm{Dropout}\left(g\left(\boldsymbol{W}_e^m\boldsymbol{z}_i^{m-1}\right)\right) \tag{1}$$

其中,$\boldsymbol{z}_i^m$表示第i个音频样本在编码器第m层的特征矢量;$\boldsymbol{W}_e^m$表示编码器第m层的参数;$g(\cdot)$是激活函数,本

文采用整流线性单元(Rectified Linear Unit, ReLU);Dropout(·)是一个随机掩码函数,随机将输入的部分元素置0.

译码器各层的输出表示为

$$\hat{\boldsymbol{z}}_i^{m-1}=g\left(\boldsymbol{W}_d^m\hat{\boldsymbol{z}}_i^m\right) \tag{2}$$

其中,$\hat{\boldsymbol{z}}_i^{m-1}$为第$i$个音频样本在译码器第$m$个反卷积层的输出;$\boldsymbol{W}_d^m$为译码器第$m$层的参数.

2.2 损失函数

用于指导联合学习框架训练的损失函数L_T包含重构损失函数L_r和聚类损失函数L_c. 前者用于指导卷积自编码网络的训练,后者用于指导判别性聚类网络的训练. 当损失函数L_T值到达设定值时停止迭代优化过程,并从Softmax层输出类别估计结果.

损失函数L_T定义为

$$L_T=L_r+L_c \tag{3}$$

卷积自编码网络的重构损失函数L_r定义为

$$L_r=\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{m=0}^{M-1}\frac{1}{\left|\boldsymbol{z}_i^m\right|}\left\|\boldsymbol{z}_i^m-\hat{\boldsymbol{z}}_i^m\right\|_2^2 \tag{4}$$

其中,N_s为音频样本总数;M为编(译)码器的层数;$\boldsymbol{z}_i^m$和$\hat{\boldsymbol{z}}_i^m$分别表示第i个音频样本在编码器第m层和译码器第m+1层的特征输出;$\left|\boldsymbol{z}_i^m\right|$表示第$m$层的输出尺寸.

判别性聚类网络对输入的深度变换特征$\boldsymbol{Z}$进行类别估计. 它的Softmax层的输出概率表示为$\boldsymbol{P}=\{p_{ik}\}$, $1\leqslant i\leqslant N_s$, $1\leqslant k\leqslant N_c$. p_{ik}表示第i个音频样本的深度变换特征$\boldsymbol{z}_i$属于第k个类的概率,定义为

$$p_{ik}=P(y_i=k|\boldsymbol{z}_i,\boldsymbol{\Lambda})=\frac{\exp\left(\boldsymbol{\lambda}_k^{\mathrm{T}}\boldsymbol{z}_i\right)}{\sum_{k'=1}^{N_c}\exp\left(\boldsymbol{\lambda}_{k'}^{\mathrm{T}}\boldsymbol{z}_i\right)} \tag{5}$$

其中,$\boldsymbol{\Lambda}=[\boldsymbol{\lambda}_1,\cdots,\boldsymbol{\lambda}_{N_c}]$表示判别性聚类网络的Softmax层的参数;T表示矩阵(矢量)转置.

引入辅助目标变量$\boldsymbol{O}$表示类别的真实概率值,用于迭代更新类别预测值$\boldsymbol{P}$. 采用KL散度度量判别性聚类网络输出的类别预测值$\boldsymbol{P}$与类别的真实概率值$\boldsymbol{O}$之间的距离,定义为

$$\mathrm{KL}(\boldsymbol{O}\|\boldsymbol{P})=\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{k=1}^{N_c}o_{ik}\log\frac{o_{ik}}{p_{ik}} \tag{6}$$

且满足约束条件$\sum_k o_{ik}=1$,即音频样本$\boldsymbol{s}_i$属于所有音频场景类别的概率之和等于1. 进一步引入一个经验分布$\boldsymbol{F}=\{f_k\}$, $1\leqslant k\leqslant N_c$,定义为

$$f_k=\frac{1}{N_s}\sum_{i=1}^{N_s}o_{ik} \tag{7}$$

定义一个均匀分布$\boldsymbol{U}=\{u_k\}$, $1\leqslant k\leqslant N_c$. $\boldsymbol{F}$与$\boldsymbol{U}$之间的距离定义为

$$KL(\boldsymbol{F}\|\boldsymbol{U})=\sum_{k=1}^{N_c}f_k\log\frac{f_k}{u_k} \tag{8}$$

为了防止大多数音频样本在类别估计时被分配给某几个类别,在式(6)的基础上加入式(8)的正则项,从而定义一个距离函数,即

$$\begin{aligned}\Psi&=KL(\boldsymbol{O}\|\boldsymbol{P})+KL(\boldsymbol{F}\|\boldsymbol{U})\\&=\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{k=1}^{N_c}o_{ik}\log\frac{o_{ik}}{p_{ik}}+\sum_{k=1}^{N_c}f_k\log\frac{f_k}{u_k}\\&=\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{k=1}^{N_c}\left(o_{ik}\log\frac{o_{ik}}{p_{ik}}+o_{ik}\log\frac{f_k}{u_k}\right)\end{aligned} \tag{9}$$

式(9)第一项最小化真实概率值与预测概率值之间的距离,第二项平衡了各类所分配的音频样本个数.

求Ψ关于o_{ik}的偏导数,即

$$\frac{\partial\Psi}{\partial o_{ik}}\propto\log\left(\frac{o_{ik}f_k}{p_{ik}}\right)+\frac{o_{ik}}{\sum_{i'=1}^{N_s}o_{i'k}}+1 \tag{10}$$

在实际聚类时音频样本数N_s一般非常大,式(10)右边第二项非常小,可以忽略. 再令式(10)偏导数等于0,化简整理后可得

$$\frac{e}{N_s}o_{ik}\sum_{i'=1}^{N_s}o_{i'k}=p_{ik} \tag{11}$$

两边同时求和Σ_i,化简整理后可得

$$\sum_{i=1}^{N_s}o_{ik}=\left(\frac{N_s}{e}\right)^{\frac{1}{2}}\left(\sum_{i=1}^{N_s}p_{ik}\right)^{\frac{1}{2}} \tag{12}$$

将式(12)代入式(11),且有约束条件$\sum_k o_{ik}=1$,变换整理后可得o_{ik}的表达式,即

$$o_{ik}=\frac{p_{ik}\Big/\left(\sum_{i'=1}^{N_s}p_{i'k}\right)^{\frac{1}{2}}}{\sum_{k'=1}^{N_c}\left(p_{ik'}\Big/\left(\sum_{i'=1}^{N_s}p_{i'k'}\right)^{\frac{1}{2}}\right)} \tag{13}$$

聚类损失函数L_c定义为

$$L_c=-\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{k=1}^{N_c}o_{ik}\log p_{ik} \tag{14}$$

2.3 方法小结

综上所述,本文方法的流程如下.

(1)从音频样本提取对数梅尔谱特征作为卷积自编码网络的输入. 通过最小化重构损失函数L_r,构建卷积自编码网络,并提取深度变换特征.

(2)采用初始聚类算法(例如K均值算法、谱聚类算法)对上述深度变换特征$\boldsymbol{z}_i$进行聚类,得到初始类.

(3)分别从编码器和译码器的各层提取深度变换特征$\boldsymbol{z}_i^m$和$\hat{\boldsymbol{z}}_i^m$,从判别性聚类网络的Softmax层输出预测

概率值p_{ik}.

(4)基于p_{ik}、o_{ik}、$\boldsymbol{z}_i^m$和$\hat{\boldsymbol{z}}_i^m$,计算重构损失函数L_r、聚类损失函数L_c和总的损失函数L_T.

(5)通过最小化损失函数L_T,更新联合学习框架参数. 聚类类别估计与联合学习框架参数更新迭代进行,直到联合学习框架收敛.

3 实验及结果分析

本节首先介绍实验数据及设置;其次,讨论不同初始聚类算法对聚类结果的影响;最后,比较分析不同特征与不同聚类方法的性能.

3.1 实验数据

实验数据集包括DCASE-2017[25]和LITIS- Rouen[11]数据集. 前者是2017年DCASE比赛提供的公开数据集,后者是法国鲁昂大学ITIS实验室录制的公开数据集. 它们的相关信息如表1所示. 两个数据集中的调参样本量和测试样本量各占50%.

表1 实验数据集

样本参数	DCASE-2017	LITIS-Rouen
音频场景类别数	15	19
样本总个数	4680	3026
样本平均时长	10s	30s

DCASE-2017数据集包含15个音频场景类别,分别为公交车、小汽车、火车、有轨电车、咖啡厅、杂货店、图书馆、地铁站、办公室、家庭室内、住宅区、公园、湖边沙滩、林间小径、市中心. 采用的录音设备是Soundman OKM II Klassik/studio A3,驻极体双耳麦克风和Roland Edirol R-09波录仪,采样频率为44.1kHz,量化位数为24位,采用双声道数据. 每个场景的样本数相等,均为312个.

LITIS-Rouen数据集包含19个音频场景类别,分别为飞机、公交车、小汽车、普通火车、高速火车、巴黎地铁、鲁昂地铁、咖啡厅、火车站大厅、儿童游戏厅、市场、台球厅、学生礼堂、餐厅、商店、繁华的街道、安静的街道、步行街、地铁站. 采用的录音设备是装有Android系统Hi-Q MP3录音机应用程序的Galaxy S3智能手机,采样频率为44.1kHz,量化位数为64位,采用双声道数据.

3.2 实验设置

采用归一化互信息(Normalized Mutual Information, NMI)和聚类精度(Clustering Accuracy, CA)作为性能评价指标. 它们被广泛用于聚类算法的性能评价[26].

假设n_{ij}表示第i类中属于第j个音频场景的音频样本数,$n_{\cdot j}$表示第j个音频场景的样本总数,$n_{i\cdot}$为第i类中的音频样本数,N_g表示音频场景类别数(真实值),N_c为聚类类别数(预测值),N_s为音频样本总数. 上述变量之间的关系为

$$n_{i\cdot}=\sum_{j=1}^{N_g}n_{ij} \tag{15}$$

$$n_{\cdot j}=\sum_{i=1}^{N_c}n_{ij} \tag{16}$$

$$N_s=\sum_{i=1}^{N_c}\sum_{j=1}^{N_g}n_{ij} \tag{17}$$

归一化互信息NMI定义为

$$\mathrm{NMI}=\frac{\sum_{i=1}^{N_c}\sum_{j=1}^{N_g}n_{ij}\log\left(\frac{N_s\times n_{ij}}{n_{i\cdot}\times n_{\cdot j}}\right)}{\sqrt{\left(\sum_i n_{i\cdot}\log\frac{n_{i\cdot}}{N_s}\right)\left(\sum_j n_{\cdot j}\log\frac{n_{\cdot j}}{N_s}\right)}} \tag{18}$$

当NMI的值为1时,真实标签与聚类标签全部匹配正确. 反之,当NMI的值为0时,聚类标签与真实标签没有正确匹配.

假设y_j和c_j分别为第j个音频样本的真实类标签和预测类标签,定义函数$\delta(\cdot)$为

$$\delta\left(y_j,\mathrm{map}(c_j)\right)=\begin{cases}1, & \text{if } y_j=\mathrm{map}(c_j)\\ 0, & \text{if } y_j\neq\mathrm{map}(c_j)\end{cases} \tag{19}$$

其中,map(·)是将预测类标签映射到真实类标签的映射函数. 聚类精度CA定义为

$$\mathrm{CA}=\frac{\sum_{j=1}^{N_s}\delta\left(y_j,\mathrm{map}(c_j)\right)}{N_s} \tag{20}$$

其中,NMI和CA的取值范围都是[0,1]. 它们的值越大,则说明聚类算法性能越好.

实验硬件平台如下:Intel(R) Core(TM) i7-6700@3.10GHz的CPU,48GB RAM,4块NVIDIA GTX 1080Ti显卡的服务器. 实验工具箱包括Theano、Python和Librosa等. 采用汉明窗进行加窗分帧,帧长为40ms,帧移为20ms. 卷积自编码网络参数如下:批大小(Batch Size)为64,输入层的通道数为4,学习率设置为0.003,编码器和译码器都包含两个隐含层(隐含层1的节点数为64,卷积核大小为5*5;隐含层2的节点数为128,卷积核大小为3*3). 判别性聚类网络参数如下:全连接层节点数为32;Softmax层的节点数等于音频场景类别数. 深度变换特征由全连接层输出,维数为32.

3.3 不同特征的比较

本小节比较深度变换特征(Deep Transformed Feature, DTF)与主流声学特征的音频场景聚类性能. 所对比的声学特征包括梅尔频率倒谱系数MFCC[27]、对数梅尔谱LMS[28]和Gabor滤波器组[29]. 声学特征维数设置为32. 采用凝聚层次聚类(Agglomerative Hierarchical Clustering, AHC)算法对输入特征进行聚类. 表2给出了不同特征在音频场景聚类时的表现.

表2　不同特征的音频场景聚类性能

特征	DCASE-2017		LITIS-Rouen	
	NMI/%	CA/%	NMI/%	CA/%
DTF	**61.66**	**52.83**	**58.57**	**50.25**
MFCC	47.23	43.10	44.33	42.90
LMS	57.02	48.01	56.03	48.51
Gabor	46.86	45.65	45.20	44.58

注:加粗表示最高值

从表2可以看出:在DCASE-2017数据集上,深度变换特征DTF取得的NMI值为61.66%、CA值为52.83%,DTF取得的NMI值分别比MFCC特征、LMS特征和Gabor特征取得的NMI值高14.43%、4.64%和14.80%,DTF取得的CA值分别比MFCC特征、LMS特征和Gabor特征取得的CA值高9.73%、4.82%以及7.18%;相似地,在LITIS-Rouen数据集上,深度变换特征DTF也取得了最高的NMI值(58.57%)和CA值(50.25%). 在上述两个实验数据集上的测试结果表明:深度变换特征DTF在音频场景聚类时都优于主流声学特征.

3.4　不同初始聚类算法的影响

初始聚类算法用于产生待聚类样本的初始类,并对联合学习框架中的Softmax层的参数进行初始设置. 本小节讨论各种初始聚类算法对基于联合学习框架的音频场景聚类结果的影响. 采用对数梅尔谱作为卷积自编码网络的输入特征.

表3给出了采用不同初始聚类算法时,基于联合学习框架的音频场景聚类结果. 所采用的初始聚类算法包括谱聚类[30](Spectral Clustering, SC)、K-means[31]、利用层次方法的平衡迭代规约与聚类[32](Balanced Iterative Reducing and Clustering using Hierarchies, BIRCH)、AHC[30]和高斯混合模型[33](Gaussian Mixed Model, GMM). 表3中的Random表示采用随机初始化方法(不采用任何初始聚类算法)生成初始类.

表3　不同初始聚类算法的影响

初始聚类算法	DCASE-2017		LITIS-Rouen	
	NMI/%	CA/%	NMI/%	CA/%
SC	61.31	50.68	55.50	48.87
K-means	62.51	54.07	52.31	45.90
BIRCH	**67.12**	**56.54**	**60.30**	**55.68**
GMM	64.96	55.68	56.78	50.13
AHC	66.20	54.26	58.53	50.30
Random	60.11	49.52	50.46	42.87

注:加粗表示最高值

从表3可以看出:在两个数据集上,采用BIRCH算法生成初始类时,基于联合学习框架的音频场景聚类得到的NMI值与CA值最高(表中黑体所示);而采用Random(随机初始化)方法生成初始类时,基于联合学习框架的音频场景聚类得到的聚类效果最差. 因此,初始聚类算法可以提供合理的初始类分配,有助于提高后续基于联合学习框架的音频场景聚类性能.

3.5　不同聚类方法的比较

本小节比较不同音频场景聚类方法的性能. 表4给出了不同方法的音频场景聚类结果. 所采用的特征是对数梅尔谱LMS或深度变换特征DTF,所有聚类方法的参数都在数据集上设置为最优值.

表4　不同聚类方法的性能对比

聚类方法	DCASE-2017			LITIS-Rouen		
	NMI/%	CA/%	耗时/s	NMI/%	CA/%	耗时/s
本文方法	**67.12**	**56.54**	1474	**60.30**	**55.68**	983
DTF+AHC	61.66	52.83	1008	58.57	50.25	733
LMS+SC	57.44	50.68	153	52.86	48.88	62
LMS+K-means	53.44	48.33	52	40.64	35.83	49
LMS+BIRCH	60.77	51.42	78	55.83	42.63	40
LMS+GMM	59.19	50.30	188	53.45	42.10	96
LMS +AHC	55.98	41.23	132	49.16	41.90	54

注:加粗表示最高值

从表4可以看出:在DCASE-2017和LITIS-Rouen两个数据集上,本文方法得到的NMI值分别为67.12%和60.30%,CA值分别为56.54%和55.68%,获得了最好的聚类效果. 本文方法将深度变换特征提取(由卷积自编码网络实现)与类别估计过程(由判别性聚类网络实现)进行联合优化(基于损失函数L_T进行迭代优化),卷积自编码网络输出的深度变换特征对判别性聚类网络是友好的,因而本文方法聚类效果优于其他常规的聚类方法. 在运算速度方面,本文方法耗时最多,因为需要进行多次迭代更新神经网络参数和类别估计结果.

4　结论

本文提出一种基于联合学习框架的音频场景聚类方法,将深度变换特征提取与类别估计过程联合起来优化,取得了较好的效果. 采用归一化互信息和聚类精度作为评价指标时,本文提取的深度变换特征优于其他特征,本文提出的聚类方法优于其他方法. 本文方法所需要付出的代价是需要较大的计算复杂度.

参考文献

[1] 金海. 基于深度神经网络的音频事件检测[D]. 广州: 华南理工大学, 2016.
Jin H. Audio events detection based on deep neural network[D]. Guangzhou, China: South China University of Technology, 2016. (in Chinese)

[2] 李应, 印佳丽. 基于多随机森林的低信噪比声音事件检

测[J]. 电子学报, 2018, 46(11): 2705 – 2713.
Li Y, Yin J L. Sound event detection at low SNR based on multi-random forests[J]. Acta Electronica Sinica, 2018, 46 (11): 2705 – 2713. (in Chinese)

[3] 李艳雄, 王琴, 张雪, 等. 基于凝聚信息瓶颈的音频事件聚类方法[J]. 电子学报, 2017, 45(5): 1064 – 1071.
Li Y X, Wang Q, Zhang X, et al. Audio events clustering based on agglomerative information bottleneck[J]. Acta Electronica Sinica, 2017, 45(5): 1064 – 1071. (in Chinese)

[4] Mesaros A, Heittola T, Benetos E, et al. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(2): 379 – 393.

[5] Li Y X, Li X K, Zhang Y H, et al. Acoustic scene classification using deep audio feature and BLSTM network[A]. 2018 International Conference on Audio, Language and Image Processing (ICALIP)[C]. Shanghai, China: IEEE, 2018. 371 – 374.

[6] Schröder J, Moritz N, Anemüller J, et al. Classifier architectures for acoustic scenes and events: Implications for DNNs, TDNNs, and perceptual features from DCASE 2016 [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(6): 1304 – 1314.

[7] Rakotomamonjy A, Gasso G. Histogram of gradients of time-frequency representations for audio scene classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 142 – 153.

[8] Yang W J, Krishnan S. Combining temporal features by local binary pattern for acoustic scene classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(6): 1315 – 1321.

[9] Abidin S, Togneri R, Sohel F. Spectrotemporal analysis using local binary pattern variants for acoustic scene classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 2112 – 2121.

[10] Bisot V, Serizel R, Essid S, et al. Feature learning with matrix factorization applied to acoustic scene classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(6): 1216 – 1229.

[11] Rakotomamonjy A. Supervised representation learning for audio scene classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(6): 1253 – 1265.

[12] Li Y X, Zhang X, Jin H, et al. Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection[J]. Multimedia Tools and Applications, 2018, 77(1): 897 – 916.

[13] Singh A, Thakur A, Rajan P, et al. A layer-wise score level ensemble framework for acoustic scene classification [A]. 2018 26th European Signal Processing Conference (EUSIPCO)[C]. Rome, Italy: IEEE, 2018. 837 – 841.

[14] Ren Z, et al. Deep sequential image features for acoustic scene classification [A]. Detection and Classification of Acoustic Scenes and Events[C]. Munich, Germany: Workshop, 2017. 113 – 117.

[15] Wu Y Z, Lee T. Enhancing sound texture in CNN-based acoustic scene classification[A]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)[C]. Brighton, UK: IEEE, 2019. 815 – 819.

[16] Singh A, Thakur A, Rajan P, et al. A layer-wise score level ensemble framework for acoustic scene classification [A]. 2018 26th European Signal Processing Conference (EUSIPCO)[C]. Rome, Italy: IEEE, 2018. 837 – 841.

[17] McDonnell M D, Gao W. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths[A]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)[C]. Barcelona, Spain: IEEE, 2020. 141 – 145.

[18] Chen H T, Zhang P Y, Yan Y H. An audio scene classification framework with embedded filters and a DCT-based temporal module[A]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)[C]. Brighton, UK: IEEE, 2019. 835 – 839.

[19] Zhang T, Wu J. Constrained learned feature extraction for acoustic scene classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1216 – 1228.

[20] Bai X, Du J, Pan J, et al. High-resolution attention network with acoustic segment model for acoustic scene classification[A]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [C]. Barcelona, Spain: IEEE, 2020. 656 – 660.

[21] Zhang L W, Shi Z Q, Han J Q. Pyramidal temporal pooling with discriminative mapping for audio classification [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 770 – 784.

[22] Phaye S S R, Benetos E, Wang Y. SubSpectralNet – using sub-spectrogram based convolutional neural networks for acoustic scene classification[A]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)[C]. Brighton, UK: IEEE, 2019. 825 – 829.

[23] Li S Y, Wang W W. Randomly sketched sparse subspace clustering for acoustic scene clustering[A]. The 26th European Signal Processing Conference (EUSIPCO) [C]. Rome, Italy: IEEE, 2018. 2489 – 2493.

[24] Li S Y, Gu Y T, Luo Y H, et al. Enhanced streaming based subspace clustering applied to acoustic scene data clustering[A]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [C]. Brighton, UK: IEEE, 2019. 11 – 15.

[25] Annamaria M, Toni H, Aleksandr D, et al. DCASE2017 challenge setup: Tasks, datasets and baseline system[A]. Detection and Classification of Acoustic Scenes and Events[C]. Munich, Germany: Workshop, 2017. 85 – 92.

[26] Li Y X, Zhang X, Li X K, et al. Mobile phone clustering from speech recordings using deep representation and spectral clustering[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(4): 965 – 977.

[27] Paseddula C, Gangashetty S V. DNN based acoustic scene classification using score fusion of MFCC and inverse MFCC[A]. 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) [C]. Rupnagar, India: IEEE, 2018. 18 – 21.

[28] Meng H, Yan T H, Yuan F, et al. Speech emotion recognition from 3D log-mel spectrograms with deep learning network[J]. IEEE Access, 2019, 7: 125868 – 125881.

[29] Schröder J, Goetze S, Anemüller J. Spectro-temporal Gabor filterbank features for acoustic event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(12): 2198 – 2208.

[30] Li Y X, Jin H, Li W, et al. Fast speaker clustering using distance of feature matrix mean and adaptive convergence threshold[J]. IET Signal Processing, 2014, 8(8): 844 – 851.

[31] MacQueen J. Some methods for classification and analysis on multivariate observations[A]. The Fifth Berkeley Symposium on Mathematical Statistics and Probability [C]. Durham, NC, USA: Project Euclid, 1967. 281 – 297.

[32] Nirmala G, Thyagharajan K K. A modern approach for image forgery detection using BRICH clustering based on normalised mean and standard deviation[A]. International Conference on Communication and Signal Processing (ICCSP)[C]. Chennai, India: IEEE, 2019. 441 – 444.

[33] Jing X X, Zhan L, Zhao H, et al. Speaker recognition system using the improved GMM-based clustering algorithm [A]. International Conference on Intelligent Computing and Integrated Systems[C]. Guilin, China: IEEE, 2010. 482 – 485.

作者简介

张聿晗　男,1995年6月出生,安徽黄山人.现为华南理工大学电子与信息学院硕士研究生,主要研究方向为语音及音频信号处理、机器学习.

李艳雄(通信作者)　男,1980年8月出生,湖南嘉禾人.现为华南理工大学电子与信息学院副教授、博士生导师,主要研究方向为语音及音频信号处理、机器学习、模式识别.

E-mail: eeyxli@scut.edu.cn