基于多特征融合的 SVM 声学场景分类算法研究

赵薇[1]， 靳聪[1]， 涂中文[2]， SRIDHAR Krishnan[3]， 刘杉[1]
(1. 中国传媒大学 信息与通信工程学院，北京 100024；2. 中国传媒大学 播音主持艺术学院，北京 100024；
3. 加拿大怀雅逊大学 电气和计算机工程系，多伦多 M5B 2K3，加拿大)

摘 要：针对 DCASE2017 挑战赛的声场环境数据集，提取梅尔频率倒谱系数(MFCC)、短时能量(SE)、声学事件似然特征(AELF)、静音时间(MT)特征，组成多特征融合矩阵，通过对比多种核函数和寻优算法，最终选取高斯径向基核函数(RK)建立支持向量机(SVM)模型，采用交叉验证(CV)方法进行 SVM 参数寻优，对 15 种声学场景进行分类. 实验结果表明，杂货店、办公室的分类准确性达到了 90%以上，平均分类准确性达到 71.11%，远高于挑战赛的基线系统 61%的平均分类准确性.

关键词：声学场景分类；支持向量机；参数寻优；特征融合
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-0645(2020)01-0069-07
DOI：10.15918/j.tbit1001-0645.2018.171

Support Vector Machine for Acoustic Scene Classification Algorithm Research Based on Multi-Features Fusion

ZHAO Wei[1], JIN Cong[1], TU Zhong-wen[2], SRIDHAR Krishnan[3], LIU Shan[1]
(1. School of Information and Communication Engineering, Communication University of China, Beijing 100024, China; 2. School of Broadcasting and Hosting Art, Communication University of China, Beijing 100024, China; 3. Department of Electrical and Computer Engineering, Ryerson University, Toronto M5B 2K3, Canada)

Abstract: For the sound environment dataset of the DCASE 2017 Challenge, Mel frequency cepstral coefficients (MFCC), short-time energy (SE), acoustic event likelihood features (AELF), and mute time (MT) features were extracted to form a multi-features fusion matrix. Comparing various kernel functions and optimization algorithms, radial basis function kernel (RK) was finally selected to establish the support vector machine (SVM) model, and cross validation (CV) method was utilized to optimize SVM parameters and to classify 15 acoustic scenes. The experimental results show that the classification accuracy of grocery store and office can reach more than 90%, and the average classification accuracy reaches 71.11%, which is much higher than the average classification accuracy of 61% of the baseline system given in the challenge.

Key words: acoustic scene classification; support vector machine; parameter optimization; feature fusion

声音中携带了日常生活环境和物理事件中的大量信息. 人类所固有的能力使得我们可以通过声音感知所处环境，比如嘈杂的街区、安静的图书馆. 人类还可以通过识别独立的音源来推断出所处的环境和即将发生的事情，比如汽车的刹车声、柔和的催眠曲. 随着手机等声音拾取设备的广泛使用，越来越多的声音被记录下来. 信号处理和人工智能技术的发展，让机器可以自动感知并提取声场环境有用信息，经过多维分析进行智能分类.

声学场景分类(acoustic scene classification,

收稿日期：2018-04-17
基金项目：国家自然科学基金资助项目(61631016，61901421)；中央高校基本科研业务费专项基金(CUC19ZD003)
作者简介：赵薇(1981—)，女，博士，高级工程师，E-mail：zhao_wei@cuc.edu.cn.

ASC)是指利用信号处理和机器学习算法,通过对输入音频信号的感知,对音频流识别出其产生环境,并标注语义标签[1]. 声学场景信号处理涉及到数字信号处理、声学、听觉心理学、人工智能等多学科领域,是一门综合性很强的交叉学科.

1997年由麻省理工学院(MIT)Sawhney和Maes发表的技术报告[2]最早记录了ASC的分类. 他们录制的音频来自于语音、地铁和交通,利用卷积神经网络(recurrent neural network,RNN)和最邻近规则(K-nearest neighbor,KNN)算法对特征和类别进行建模,实现总体分类正确率68%. Ballas[3]利用实验心理学研究发现,声学事件的识别速度和准确率与激励的声学属性有关系,例如激励出现的频率、是否有某种物理诱因或者声音先验知识的影响. Peltonen等[4]通过在25个音频场景中的实验统计,得出人类的分辨能力超过70%. 他们认为,人类通过典型声学事件的识别来认知声学场景. Eronen等[5]在MIT早期工作成果基础上,更关注于局部和全局特征. MFCC描述音频信号的局部频谱包络,GMMs来对MFCC的分布进行统计,继而用HMMs引导GMMs时间上的进化,这种算法在18个场景分类中取得了58%的正确率.

声学场景分类的研究具有重要的应用价值. 通过麦克风实时录入音频,识别声场环境,自动对音频文件分配对应元数据,进而对声学场景进行渲染,作为视频的必要补充,可以提高VR/AR的沉浸感[6]. 数字音频档案迅速增长,这些海量的数据中包含了各种各样的语音、音乐、动物声、城市环境声等等,目前这些音频档案的利用率远远低于文本和图像档案,利用机器学习方法对其中的声学场景进行分类,可以挖掘出大量的有用信息[7]. 智能手机[8]、导航机器人[9]可以持续感知周围声场变化,从而自动切换场景模式或提供定制信息. 智能助听器[10]、智能轮椅[11]可以基于室内室外的环境来调节相应的功能,帮助残障人士更方便的生活.

声学场景相对于语音和音乐,没有持续和清晰的谐波成分,包含了更加广泛的声学事件以及相当丰富多样的信号特征,因此对于机器监听系统来说,目前还无法精确实现对环境的分类. 声学场景分类最常见的方法是提取声学场景本身的时域、频域特征,例如基于人耳听觉特性梅尔频谱系数(Mel-frequency cepstral coefficients,MFCC)特征[12]、能量特征、频谱特征[13]等,以及基于这些特征的融合与改进[14],但是寻找新的特征非常困难,需要不断的进行尝试. 此外还有利用典型声学事件及其特征[15]的识别作为分类依据的方法,但是该方法需要对大量的声学环境中包含的事件进行统计分析[16],从而得到典型声学事件及其占比关系[17]. 当多种声音同时出现或者声音被环境扭曲时,声学事件的识别本身就是个难以解决的问题[18]. 深度神经网络也是近年来较为流行的方法,构建更大、更复杂的神经网络,对大数据集进行训练和测试. 但是这种方法需要很高的硬件配置,而且模型处于"黑箱状态",难以理解内部机制,提升系统性能存在困难. 针对普遍使用的声学场景特征,采用隐马尔科夫模型(hidden Markov model HMMs,HMMs)[13]、高斯混合模型(Gaussian mixture models,GMMs)[19]、支持向量机(support vector machine,SVM)[20]等机器学习方法建立声学模型作为分类器,也是场景分类的主流方法.

本文提出了一种基于多特征融合的SVM算法模型,该算法针对声场环境数据集,提取MFCC、SE、AELF、MT特征,组成多特征融合矩阵,通过对比多种核函数和寻优算法,最终选取RK核函数建立SVM模型,采用CV方法进行SVM参数寻优,对15种声学场景进行分类.

1 多特征提取及融合

1.1 梅尔频率倒谱系数(Mel frequency cepstral coefficients,MFCC)

MFCC是一种广泛应用于音频场景和语音识别的特征参数[21]. Mel标度描述了人耳频率的非线性特性,它与频率f的关系可用式(1)近似表示为

$$f_{\mathrm{mel}}(f)=2\,595\lg\left(1+\frac{f}{700}\right). \tag{1}$$

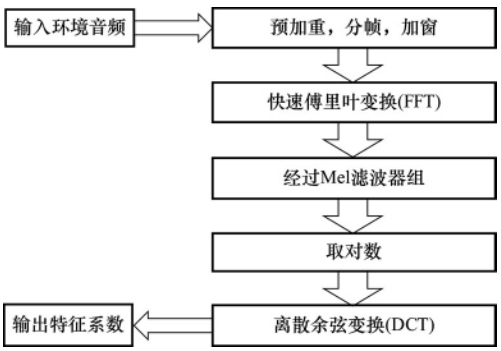

图1 MFCC提取流程图

Fig. 1 Extraction flowchart of MFCC

输入的声学场景音频信号进行预加重、分帧、加窗等处理后,对各帧信号进行快速傅里叶变换,并对

得到的频谱取模平方,即可得到信号的功率谱. 然后将其通过梅尔滤波器组. 将频谱按人耳敏感程度分为多个 Mel 滤波器组,采用的滤波器为三角滤波器,中心频率 $f(m)$,$m=1,2,\cdots,M$,M 通常取 22～26. 如式(2)所示.

$$H_m(k)=\begin{cases}0 & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1)\leqslant k\leqslant f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & f(m)\leqslant k\leqslant f(m+1)\\ 0 & k>f(m+1)\end{cases}. \tag{2}$$

式中:m 为滤波器数目;$f(m)$为 $m+2$ 个 Mel 间隔频率[22].

经过梅尔滤波器后得到了平滑化的音频信号,对其进行梅尔倒谱分析. 对每个 Mel 滤波器的输出取对数,得到对应的对数功率谱,然后对其作离散余弦变换,如式(3)所示.

$$C(n)=\sum_{m=0}^{N-1}s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right),\quad n=1,2,\cdots,L. \tag{3}$$

最后与归一化倒谱提升窗口相乘,求取一阶与二阶差分参数,合并后去除首尾一阶差分系数为 0 的两帧,即得到声学场景音频的 MFCC 特征系数.

1.2　短时能量(short-time energy,SE)

音频信号的能量随着时间的变化而变化,环境中主要包含的声音事件和噪声的区别可以体现在他们的能量上. 由于不同的声音类型具有不同的能量,且音频信号是非平稳的,可以引入短时能量作为音频信号幅度及能量上的一个特征,其表达式如式(4)所示.

$$E_n=\sum_{m=0}^{N-1}x_n^2(m). \tag{4}$$

式中:$x_n(m)$为输入的第 n 帧信号;N 为示帧长;E_n 为第 n 帧信号的能量,即短时能量.

1.3　声学事件似然特征(acoustic event likelihood features,AELF)

声学事件是指能够引起人们感知注意的一段单一完整的短时连续声音信号. 在特定的场景中会有特定的声学事件出现,并且每种场景会有出现可能性较高或次数较多的某些声学事件,因此可以提取声学事件的特征,并将其与全部场景的音频进行比对,得到一个似然矩阵作为新的音频特征.

对 15 种场景中的主要声学事件进行统计,共有 24 种,分别为:海浪声、风声、鸟叫声、说话声、音乐声、汽车引擎声、提示音声、汽车转向灯声、汽车经过声、溪水声、脚步声、物品碰撞声、自来水声、餐具碰撞声、洗衣机声、洗餐具声、树叶声、翻书声、地铁运行声、鼠标点击声、鸡鸣声、雨声、汽车经过声、车厢内平稳行驶声,其余声音基本可视为噪声.

由于声学事件音频信号样点之间存在相关性,因此一个采样值可利用若干个过去的声学事件音频采样值的线性组合来逼近,得到一组唯一的预测系数,即线性预测系数(LPC). 为了提高特征参数的稳定性,本文对声学事件音频信号求倒谱,用线性预测倒谱系数(LPCC)提取声学事件特征.

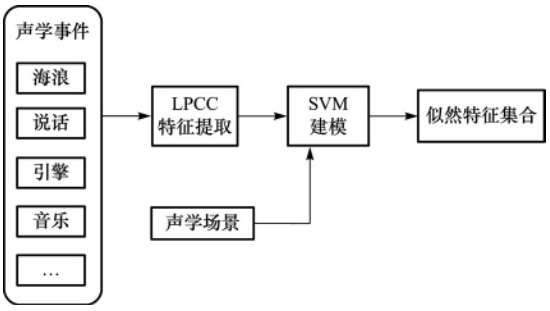

图 2　声学事件似然特征提取流程

Fig. 2　Extraction process of acoustic event likelihood features

1.4　静音时间(mute time,MT)

实验过程中发现,在一些室内场景中,整体较安静,音频背景声音的静音时间较长,可能造成某些音频特征的不准确. 统计 15 种场景的静音情况,如表 1 所示.

表 1　所有场景静音情况

Tab. 1　Mute condition of all scenes

有静音时间的场景	无静音时间的场景
咖啡厅/餐厅、汽车、森林小路、家、图书馆、办公室、公园	海滩、公交车、市中心、杂货店、地铁站、住宅区、火车、有轨电车

本文引入静音时间的检测,将其作为一个描述音频的新特征. 在训练集每一类室内场景音频中,分别截取 50 段音频没有任何明显声学事件的静音片段,将其幅值取平均值,记作 U_{level}. 检测全部声音片段中小于或等于此 U_{level} 值的时刻,作为对音频的静音时刻的描述,然后将所有音频中低于 U_{level} 的值置为 0,得到新的音频描述矩阵. 其公式表达如下.

$$\boldsymbol{X}(t)=\begin{bmatrix}a_{11} & \cdots & a_{1j}\\ \vdots & \ddots & \vdots\\ a_{i1} & \cdots & a_{ij}\end{bmatrix}=\begin{cases}a_{ij}, a_{ij}>U_{\text{level}}\\ 0, a_{ij}\leqslant U_{\text{level}}\end{cases}. \tag{5}$$

式中:$\boldsymbol{X}(t)$为信号幅值矩阵;a_{ij} 为每一帧的幅值,检

测其中0的个数,即可得到MT参数.

1.5 特征融合

以上4个特征中,MFCC代表频域特征,SE是时域的典型特征,两者互相补充. AELF反映了声学场景中所包含事件,MT则描述了音频中安静程度,这两个特征各自独立,分别反映了声学场景特征不同的方面. 将各个特征提取结果分别归一化,并以列向量形式进行组合,从而实现对声学场景音频特征更加全面的描述.

$$\boldsymbol{F}=[\mathrm{MFCC}\quad \mathrm{SE}\quad \mathrm{AELF}\quad \mathrm{MT}]. \tag{6}$$

式中:MFCC取52维,SE取1维,AELF取16维,MT取1维,共同构成70维的特征融合矩阵$\boldsymbol{F}$.

2 支持向量机分类器及其参数选择

2.1 支持向量机分类器原理

支持向量机(support vector machine,SVM)作为机器学习方法的主流技术之一,具有较好的分类性能.

声学场景音频数据属于复杂的非线性分布,无法在低维度找出一个线性决策边界,需要通过向量积的方法将数据从低维度映射到高维度,进而在高维度中寻找一个最优分类函数. 计算两向量内积的方法称为核函数[23],本文采用训练集和测试集的内积$\langle\Phi(x)\times\Phi(Z)\rangle$来作为决策规则,核函数表达式如式(7).

$$K(x,z)=\langle\Phi(x)\times\Phi(z)\rangle. \tag{7}$$

非线性分类下的优化问题如式(8)所示,

$$\min_{\alpha}\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,z_i)-\sum_{i=1}^{n}\alpha_i. \tag{8}$$

最优分类决策函数用符号函数判定为

$$f(x)=\mathrm{sgn}\left\{\sum_{i=1}^{n}\alpha_i y_i K(x_i,y_i)+b\right\}. \tag{9}$$

常用的核函数$K(x,z)$有以下几种.

① 线性核函数(linear Kernel,LK):

$$K(x,z)=\langle x,z\rangle. \tag{10}$$

② 多项式核函数(polynomial Kernel,PK):

$$K(x,z)=(\langle x,z\rangle+R)^d. \tag{11}$$

③ 高斯径向基核函数(RBF Kernel,RK):

$$K(x,z)=\exp(-g\parallel x-z\parallel^2). \tag{12}$$

式中d,g为支持向量机核函数的参数.

2.2 SVM三种寻优算法

合理的设置支持向量机的参数,能有效地提高SVM分类器的分类精度. 在核函数中主要有2个参数,即惩罚参数c和核函数中的gamma参数g. 进行参数寻优的目的是找到能使分类效果较好的参数对(c,g),使得分类器能精确预测未知的数据.

① 交叉验证寻优算法(cross validation,CV).

本文将声学场景的训练素材按3∶1分成两部分,其中3份是已知声学场景类型,进行SVM模型的训练,而一份被认为是未知的,利用这个未知部分的音频数据预测分类器的性能[24]. 本文采用网格搜索,通过对各种可能的(c,g)组合来寻找最优的声学场景分类性能. 通过这种交叉验证方法,得到预测精确度最高的参数对(c,g).

② 遗传学寻优算法(genetic algorithm,GA).

美国Michigan大学的Holland教授提出了基于自然进化理论的遗传学算法[24]. 本文将声学场景训练音频按3∶1分成两部分,其中未知的一部分设为种群,设定种群的进化代数为100,种群数量为20,参数c和g是算法中的交叉概率和变异概率,取值范围为[0,100]. 按照适应度计算法则,计算初始参数c和g下种群中每个个体的初始适应度[20]. 根据进化的代数,执行选择算子,交叉算子和变异算子的操作后产生最优的个体,将个体重新插入到种群中产生新的后代,若当代种群个体的适应度高于上一代个体,则更新个体的适应度,最后得到最优适应度下的(bestc,bestg).

③ 粒子群寻优算法(particle swarm optimization,PSO).

粒子群寻优算法由Kennedy和Eberhart于1995年提出,它的基本概念源于对人工生命和鸟群捕食行为的研究[25]. 首先设定种群的数量和进化代数初始值,参数c和g最大值为100,最小值为0.1,使用参数c和g初始化种群的位置和粒子的速度及位置,计算此时种群和个体的初始适应度. 根据设定的进化代数和种群数量进行迭代更新,判断粒子的速度和种群当前的位置是否高于上次迭代得到的值,若高于则更新该值,即更新参数c和g,计算此时的适应度. 遍历上述结果中得到的适应度,确定最高适应度下的最优解(bestc,bestg).

3 实验数据集及基线系统

3.1 实验数据集

本文所用训练和测试集来自于IEEE音频与声学协会举办的DCASE2017(声学场景和事件检测分类挑战赛)公开数据集. 该数据集由芬兰坦佩雷理工大学(Tampere University of Technology,2015.06-2016.01)收录,包括15类日常生活的环境和测试集两部分. 相对于2016年挑战赛的30 s

素材,2017 年每段的素材长度缩减到 10 s,这导致其中所包含的音频信息大大减少,无论对人类还是机器都是更难的挑战. 录音中包含了大量的声学事件,其中包括鸟叫、风声等自然音源,人的谈笑声、脚步声等人类活动音源,清洗杯盘、推拉抽屉、汽车马达声等物品发出的音源[19]. 声学场景类型如图 3 所示.

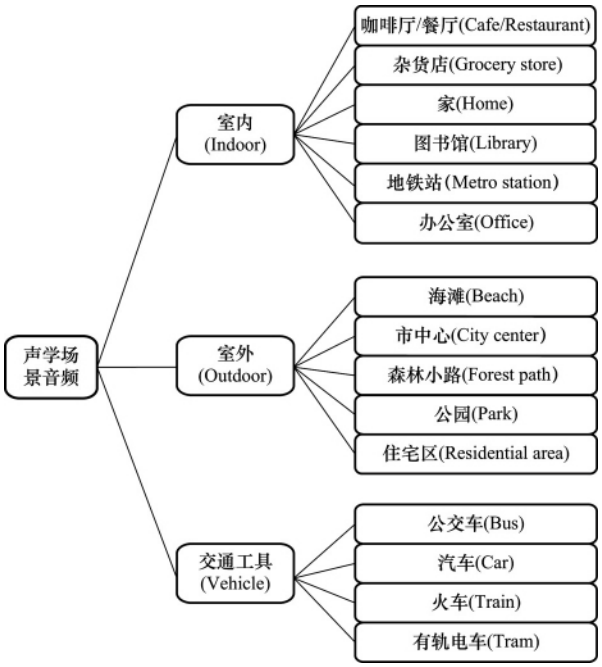

图 3　音频数据集结构图

Fig. 3　Structure diagram of audio dataset

3.2　基线系统(baseline system)

在 DCASE2017 官方网站中给出了一个采用深度学习的基线系统(baseline system),这也是目前较为常用的方法之一. 该基线系统采用多层感知器进行架构,将对数梅尔能量作为特征,矢量长度为 200,使用包含两层密集层(每层 50 个隐藏单元和 20% dropout)的神经网络训练 200 个 epoch,分类决策为基于 softmax 类型的网络输出层. 最终得到测试集的平均准确率为 61%.

4　实验结果及分析

4.1　核函数选取实验结果及分析

利用 3 种寻优算法对比 SVM 的常用核函数,得到的平均预测准确率如图 4 所示. 在每一种寻优算法下,RK 核函数得到的平均分类预测准确度都是最高,达到 75%以上,说明对于声学场景,RK 核函数的分类性能最好,因此本文优先选择 RK 作为 SVM 模型的核函数.

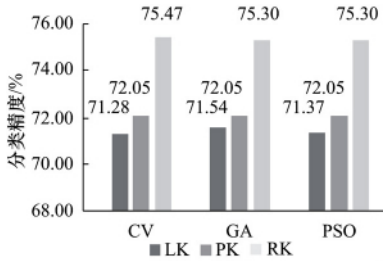

图 4　3 种寻优算法下不同核函数 SVM 分类预测准确率

Fig. 4　Prediction accuracy of SVM classification with different Kernel functions under three optimization algorithms

4.2　SVM 模型参数寻优实验结果及分析

以 RK 为核函数构造 SVM 模型,利用训练集比较 3 种寻优方法,得到最优(c,g)如表 2 所示,训练数据的预测最高准确率都可以达到 90%以上.

表 2　3 种寻优算法下最优参数对(c,g)

Tab. 2　Optimal parameter pair(c,g) under three optimization algorithms

寻优算法	c	g	训练集准确率/%
CV	5.278 0	0.329 9	93.82
GA	9.861 9	1.773 7	93.70
PSO	100	0.850 0	91.30

3 种寻优算法的预测平均准确率相差较小,尤其是 CV 算法和 GA 算法,但是 CA 算法用时较短,PSO 算法其次,GA 算法耗时最长. CA 的网格寻优过程中,随机数据集分成两部分作为训练和预评估,最后记录每种最高准确率下的参数对(c,g),处理和寻优的思路和过程较为简单;GA 算法寻优时需要进行多次迭代,迭代过程中还会有选择、交叉、变异等操作,参数选择需要大量经验,从而使寻优的过程变得复杂,因此 GA 算法编程复杂,耗时长;而 PSO 算法在处理本文所用声学场景音频时,分类准确性比前两者略低. 因此在综合考虑预测准确性、算法实现复杂程度、程序运行时间等情况下,本文优先采用 CV 算法进行参数寻优.

4.3　声学场景测试集分类结果及分析

统计每种类别的最高的识别准确率如图 5 所示. 本文针对声场环境提取 MFCC、SE、AELF、MT 特征,组成多特征融合矩阵,选取 RK 核函数建立 SVM 模型,采用 CV 方法进行 SVM 参数寻优. 由图 5 可知,通过本文的方法进行声学场景分类,杂货店、办公室的分类正确性达到了 90%以上,平均分类准确性达到 71.11%,远高于挑战赛给出的基线系统(baseline system)61%的平均分类准确性.

为了能够深入分析分类结果,本文将每一类训练集的分类结果进行逐一统计,表格颜色越深,代表

判断为此类的个数越多，表格中的数字即该类声学场景音频的具体判断个数，如图6所示.

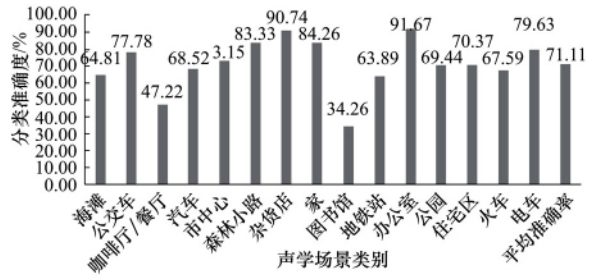

图5 声学场景测试集分类结果

Fig. 5 Acoustic scene classification results of the test set

判断结果 / 原场景	海滩	公交车	汽车	市中心	森林	杂货店	家	图书馆	地铁站	办公室	公园	住宅区	餐厅	火车	电车
海滩	70				3	4	2	1		1	2	22		3	
公交车	1	84	5			3			1				6	2	6
汽车		1	74						4					12	17
市中心				79		5	1		9		4	5	5		
森林	1	1	1	1	90			6	3		1	2		1	1
杂货店					1	98			6				3		
家	2				1	2	91	8	2	2					
图书馆		1	2		6	12	23	37	10	10	2	5			
地铁站				5	1	11		3	69				17	1	1
办公室					1		7		1	99					
公园	3				1	9					75	17		3	
住宅区	2			10	7		1	6		1	4	76	1		
餐厅				1		15	4	11	22		1		51	3	
火车		3	9				1	2	8		1		1	73	10
电车		2				1			8					11	86

图6 声学场景测试集分类结果混淆矩阵

Fig. 6 A confusion matrix of acoustic scene classification test set result

图6可以明显观察到，每一类声学场景的大部分音频都得到了正确的分类，整体效果较好，说明选用的针对声学场景的特征融合矩阵、建立的SVM模型及其参数是非常有效的. 分析具体原因如下.

① 本文选用了基于频谱能量的MFCC参数和基于时域的SE参数，两者之间相关性不大，体现了声场环境的不同特征.

② 通过提取24个声学事件的特征，并将其在15个声学场景求取似然估计距离，得到AELF特征. 分类准确度高的类别，如办公室、杂货店，其AELF较独特. 办公室整体较安静，杂音较少；而杂货店则相反，整体环境较为嘈杂，声学事件丰富，远近不同的说话声、音乐声，以及物体碰撞声、提示音等遍布整个音频. 该特征很好的利用了声学事件对分类特性的影响.

③ 由于测试集音频时长只有10 s，因此音频中的大量的静音片段不利于声学场景特征的提取. 本文通过统计500个片段的静音部分平均幅值，将原始音频信号重新清洗，MT特征有效弥补了静音的影响.

④ 充分利用4 680段训练集来对比CV、GA、PSO三种寻优方法，进行以LK、PK、RK为核函数的SVM模型比较，最终选择CV寻优方法和RK核函数，测试数据规模大，选用的方法更有说服力.

但其中仍有误判的情况存在，分析具体原因如下.

① 当不同的场景中含有相似的声学事件时，例如火车车厢与有轨电车内有着相似的平稳行车声、说话声以及一些物品碰撞的声音，造成这两者相互误判的个数比其它类别略多.

② 分类准确度较低的类别，如餐厅/咖啡厅、图书馆，声学特征不明显. 餐厅/咖啡厅中整体环境嘈杂，含有大量人们的交谈说话声、背景音乐声和杂音，而这与地铁站在没有地铁通过时的声学环境较为类似. 而图书馆这一类别的音频嘈杂程度不同，安静的部分与家、办公室的整体环境较相似，但嘈杂的图书馆环境会有说话声，可能与地铁站相似度较高，因此易被误判.

③ 当测试集音频与训练集内容差别较大，也会造成误判. 例如海滩类别，在训练集中含有的声学事件大多为海浪声，但是测试集的某些音频中儿童嬉闹的声音更为明显，海浪声音被弱化，这就可能使其与住宅区的声学元素更相似，造成误判.

5 结 论

本文将DCASE2017挑战赛的声学场景音频作为数据集，提出了一种基于多特征融合的SVM算法模型. 对全部15种场景音频提取MFCC、SE、AELF、MT 4种具有独立性和互补性的特征，将其组合为多特征融合矩阵. 并且讨论了多种核函数和寻优算法，进行分类效果对比后，选取RK核函数建立SVM模型，CV方法进行SVM参数寻优，最终使得平均分类准确性达到了71.11%，远高于挑战赛给出的基线系统61%的平均分类准确性，其中，杂货店、办公室的分类准确性达到了90%以上. 针对几种准确性较差的声场，未来的研究要对其声学特征进一步深入挖掘，寻找新的方法提高算法分类的正确率.

参考文献：

[1] Barchiesi D, Giannoulis D, Stowell D, et al. Acoustic scene classification: classifying environments from the sounds they produce[J]. IEEE Signal Process. Mag., 2015, 32(3): 16-34.

[2] Sawhney N, Maes P. Situational awareness from environmental sounds[R]. Boston, USA: Massachusetts Institute of Technology, 1997.

[3] Ballas J. Common factors in the identification of an assortment of brief everyday sounds[J]. J Exp Psychol,

1993,19(2):250 - 267.

[4] Peltonen V T, Eronen A J, Parviainen M P. Recognition of everyday auditory scenes: Potentials, latencies and cues[C]// Proc. 110th Audio Engineering Society Convention. [S. l.]: AES, 2001: 1 - 5.

[5] Eronen A, Tuomi J, Klapuri A, et al. Audio-based context awareness-acoustic modeling and perceptual evaluation[C]// JCASSP. [S. l.]: IEEE, 2003: 529 - 532.

[6] Zeng Zhihong, Pantic M, Roisman G I, et al. A survey of affect recognition methods: audio, visual, and spontaneous expressions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(1): 39 - 58.

[7] Ranft R. Natural sound archives: past, present and future [J]. Anais daAcademia Brasileira de Ciências, 2004, 76(2): 456 - 460.

[8] Parviainen Jussi, Bojia Jayaprasad, Collins Jussi, et al. Adaptive activity and environment recognition for mobile phones[J]. Sensors, 2014, 14(11): 20753 - 20778.

[9] Hossain Md Arafat, Israt Ferdous. Autonomous robot path planning in dynamic environment using a new optimization technique inspired by bacterial foraging technique[J]. Robotics and Autonomous Systems, 2015, 64: 137 - 141.

[10] Sourabh Ravindran, Anderson David V. Audio classification and scene recognition and for hearing aids[C]// IEEE International Symposium Circuits and Systems, on. [S. l.]: IEEE, 2005.

[11] Hossain Md Arafat, Israt Ferdous. Autonomous robot path planning in dynamic environment using a new optimization technique inspired by bacterial foraging technique[J]. Robotics and Autonomous Systems, 2015, 64: 137 - 141.

[12] Aucouturier J J, Defreville B, Pachet F. The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music [J]. J. Acoust. Soc. Am., 2007, 112(2): 881 - 891.

[13] Eronen A J, Peltonen V T, Tuomi J T, et al. Audio-based context recognition [J]. IEEE Trans. Au-dio, Speech Lang. Processing, 2006, 14(1): 321 - 329.

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

[15] Heittola T, Mesaros A, Eronen A, et al. Audio context recognition using audio event histograms [C]// 18th European Signal Processing Conference. [S. l.]: IEEE, 2010: 1272 - 1276.

[16] **胡根生,张乐军,张艳. 结合张量特征和孪生支持向量机的群里行为识别[J]. 北京理工大学学报**, 2019, 39(10): 1063 - 1068.
Hu Gensheng, Zhang Lejun, Zhang Yan . Group activity rocognition based on tensor features and twin support vector machine[J]. Transactions of Beijing Institute of Technology, 2019, 39(10): 1063 - 1068. (in Chinese)

[17] Mesaros Annamaria, Toni Heittola, Tuomas Virtanen. TUT database for acoustic scene classification and sound event detection[C]// Signal Processing Conference (EUSIPCO), 2016 24th European. [S. l.]: IEEE, 2016.

[18] Mesaros Annamaria, Toni Heittola, Alek Sandr Dinent, et al. DCASE 2017 challenge setup: tasks, datasets and baseline system[C]// DCASE 2017 Workshop on Detection and Classification of Acoustic Scenes and Events. [S. l.]: DCASE, 2017: 1 - 8.

[19] Esmaili S, Krishnan S, Raahemifar K. Content based audio classification and retrieval using joint time-frequency analysis [C] // Proc of IEEE Int Conf Acoust, Speech, Signal Process. [S. l.]: IEEE, 2004: 665 - 668.

[20] **郑学恩,许承东,范国超,等. 应用遗传算法的多星座保护级别优化方法[J]. 北京理工大学学报**, 2018, 38(10): 60 - 64.
Zhang Xueen, Xu Chengdong, Fan Guochao, et al. Multi-constellation protection level optimization method using genetic algorithm[J]. Transactions of Beijing Institute of Technology, 2018, 38(10): 60 - 64. (in Chinese)

[21] Davis Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Trans. ASSP, 1980(28): 357 - 366.

[22] Sulistijono I A, Urrosyda R C, Darojah Z. Mel-frequency cepstral coefficient (MFCC) for music feature extraction for the dancing robot movement decision [M]. [S. l.]: Springer International Publishing, 2016.

[23] Hsu Chih-Wei, Chang Chih-Chung, Lin Chih-Jen. A practical guide to support vector classification[R]. Taipei: Department of Computer Science, National Taiwan University, 2003.

[24] Goldberg D E. Genetic algorithm in search, optimization, and machine learning [J]. Addison-Wesley Pub Co, 1989(7): 2014 - 2116.

[25] Kennedy J, Eberhart R. Particle swarm optimization [C]// Proceedings of the International Conference on Neural Networks. Australia: IEEE, 1995: 1942 - 1948.

(责任编辑:刘芳)