计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(10):140−146 [doi: 10.15888/j.cnki.csa.009258]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于层次信息融合的声学场景分类①

江　港, 马忠臣

(江苏大学 计算机科学与通信工程学院, 镇江 212013)
通信作者: 马忠臣, E-mail: zhongchen_ma@ujs.edu.cn

摘　要: 声学场景分类技术可以通过在公共区域中录制的音频分析出它的录制环境, 在日常生活中发挥着重要的作用. 与传统分类问题类与类之间没有关系不同, 声学场景分类的类别间存在着层次结构关系 (父类与子类), 如机场和购物中心的父类为室内. 而现有的方法在设计时并未考虑声学场景分类任务的这一特性, 忽略了父类和子类间的依赖关系. 因此, 本文利用声学场景类别间的层次结构关系, 提出了一种基于层次信息融合的声学场景分类方法. 该方法为父类和子类分别设计了单独的分类器, 在子类分类的过程中融合了父类的信息, 并设计了层次依赖损失来对预测的父类和子类不匹配的情况进行惩罚. 在 TAU 城市声学场景 2020 移动开发数据集上的实验结果表明, 基于层次信息融合的方法有效地提升了声学场景分类模型的性能, 分类准确率提升了 1.1%.

关键词: 声学场景分类; 层次结构; 层次信息融合; 层次依赖损失

引用格式: 江港,马忠臣.基于层次信息融合的声学场景分类.计算机系统应用,2023,32(10):140–146. http://www.c-s-a.org.cn/1003-3254/9258.html

Acoustic Scene Classification Based on Hierarchical Information Fusion

JIANG Gang, MA Zhong-Chen

(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: Acoustical scene classification technology plays an important role in daily life by analyzing its recording environment through the audio recorded in public areas. Different from the traditional classification problem in which there is no relationship between classes, there is an implicit hierarchical structure relationship between the classes of acoustic scene classification (parent class and subclass). For example, the parent class of the airport and shopping mall is indoor. However, the existing methods do not consider this characteristic of acoustic scene classification task and ignore the dependency relationship between the parent class and the subclass. Therefore, an acoustic scene classification method is proposed, which is based on hierarchical information fusion by using the hierarchical structure relationship between acoustic scene classes. In this method, two separate classifiers are designed to classify the parent class and the subclass respectively. The information of the parent class is fused in the process of the subclass classification, and the hierarchical dependency loss is designed to punish the predicted mismatch between the parent class and the subclass. The experimental results on TAU urban acoustic scenes 2020 mobile development dataset show that the method based on hierarchical information fusion effectively improves the performance of the acoustic scene classification model with an increase of 1.1% in classification accuracy.

Key words: acoustic scene classification; hierarchical structure; hierarchical information fusion; hierarchical dependency loss

① 基金项目: 国家自然科学基金 (62006098); 中国博士后科学基金 (2020M681515)
收稿时间: 2023-03-22; 修改时间: 2023-04-28; 采用时间: 2023-05-06; csa 在线出版时间: 2023-07-21
CNKI 网络首发时间: 2023-07-21

1 引言

随着计算机听觉的发展, 声学场景分类逐渐成为一个重要的研究领域. 声学场景分类是指根据生活中的音频记录识别出其录制环境的任务[1−3]. 目前, 声学场景分类技术在智能穿戴设备, 残障辅助设备, 音频监控系统, 音频文件管理等领域中都具有重要的实际应用意义[4,5]. 而一段音频中的信息量是十分巨大的, 其中也包含着许多与场景无关的内容, 这就使得声学场景的预测变得困难. 因此, 设计一个性能出色的声学场景分类模型十分具有挑战性.

近年来, 声学场景事件的检测与分类 (detection and classification of acoustic scenes and events, DCASE) 挑战赛的举办极大地推动了声学场景分类任务的发展. 因为在此之前, 声学场景分类领域并没有一个权威的数据集和统一的评判标准. 而 DCASE 挑战赛的组织者解决了这一问题, 为广大的研究者们提供了一个开放性的研究平台, 并逐年对数据集进行扩充. 而从近几年 DCASE 的挑战赛结果来看, 表现较好的声学场景分类模型大多从音频中提取对数梅尔声谱图作为网络模型的输入, 并采用卷积神经网络及其变体的结构[6]. 再结合一些先进的深度学习技术, 如注意力机制[7,8], 数据增强[9,10] 等, 声学场景分类模型的分类准确率逐步提升.

但是, 声学场景分类问题与传统的分类问题存在着一些不同. 在传统的分类问题中, 类别与类别之间没有关系, 各自独立. 但在声学场景分类中, 类别之间存在着一个层次结构关系, 即父类与子类的关系. 在日常生活中有各种各样具体的声学场景, 而这些具体的声学场景可概括为 3 个更具广泛代表性的类别: 室内、室外和交通. 本文将具体的声学场景称为声学场景子类, 而将 3 个更加具有广泛代表性的类别称为声学场景父类.

例如, 在 DCASE 官方提供的 TAU 城市声学场景 2020 移动开发数据集中, 包含了机场, 购物中心, 地铁站, 人行街道, 公共广场, 交通街道, 公园, 有轨电车, 公交车和地铁 10 个具体的声学场景类别. 它们之间的层次结构关系如图 1 所示. 机场, 购物中心, 地铁站 3 个子类的父类为室内. 人行街道, 公共广场, 交通街道, 公园 4 个子类的父类为室外, 而地铁, 有轨电车, 公交车 3 个子类的父类为交通.

然而, 现有的声学场景分类方法在设计模型的时候往往忽略了声学场景类别间的这种层次结构关系, 将它视为一个传统的分类问题进行处理. 为了充分利用声学场景类别间的层次结构关系信息, 提升声学场景分类模型的性能, 本文提出一种基于层次信息融合的声学场景分类方法. 在这项工作中, 本文的贡献主要有以下几个方面.

(1) 本文设计了一个声学场景层次分类框架, 使用两个分类器分别进行父类与子类分类, 并且在子类分类的过程中融合父类的信息.

(2) 本文引入了一个层次依赖损失来对模型预测出的父类类别与子类类别不匹配的情况进行惩罚, 训练模型预测出的子类能够隶属于预测出的父类, 符合层次结构关系.

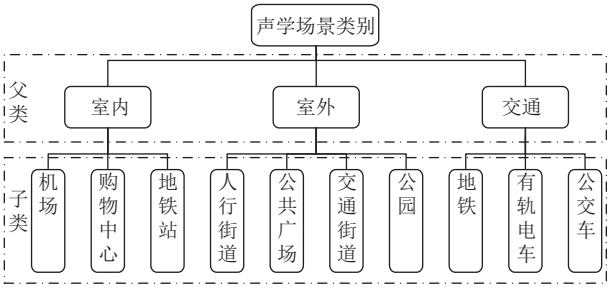

图 1 声学场景类别间的层次结构关系

本文其余部分结构如下. 第 2 节介绍了相关工作. 在第 3 节中提出了基于层次信息融合的声学场景分类方法, 介绍了层次信息融合模型的框架结构和层次依赖损失. 在第 4 节中描述相关了实验设置, 数据集的情况等, 展示了实验结果并对结果进行了分析. 在第 5 节中对本文进行了总结.

2 相关工作

声学场景分类是一个典型的机器学习分类问题, 在日常生活中的重要作用使其成为一个日渐活跃的研究领域. 随着计算机技术的发展, 基于深度学习的声学场景分类模型展示出了比基于传统机器学习声学场景分类模型更好的性能. 目前, 主流的声学场景分类模型大都从音频中提取贴近人耳的响应特性的对数梅尔声谱图作为输入, 采用基于卷积神经网络及其变体的结构, 再结合注意力机制, 数据增强等先进的深度学习技术来搭建[11−13].

不同于传统的分类问题, 一些特殊的分类任务的

类别间存在着层次关系, 这些类别间的层次关系往往以结构树的形式出现[14,15]. 类别间的层次结构有显示和隐式之分. 显式的层次关系要求处于不同层次的父类和子类的每一个类别都对应着原始数据中的一个类别. 而隐式的层次关系则表示原始数据中只有子类类别, 而父类类别由这些子类类别集合的子集构成. 声学场景分类的类别间的层次结构关系即为隐式的结构关系. 如图 1 所示, 声学场景分类的原始数据中只有人行街道, 有轨电车等子类的样本, 而声学场景父类则为隶属于对应父类下的几个子类的集合.

而对于声学场景类别间隐式的层次结构关系, 目前的方法往往选择忽略类之间的关系信息, 将声学场景分类任务当做传统的分类问题进行处理. 如 McDonnell 等人考虑到声谱图不同于普通图像, 时间轴和频率轴与图像中两个空间轴具有和不同的性质, 同样的特征出现在声谱图的低频区域或者高频区域可能代表着完全不同的物理意义, 设计了一种高低频路径分离与后期融合的残差网络模型[16]. Suh 等人在此基础上设计了一种 3 个残差网络分支并行的分类模型, 来分别针对声谱图的不同频率范围进行学习[17], 获得了较好的结果. Liu 尝试了不同的注意力模块, 如压缩和激励结构 (squeeze-and-excitation, SE), 卷积块的注意力机制模块 (convolutional block attention module, CBAM) 等[18]. Gao 等人用焦点损失替代交叉熵损失, 重点关注分类差的样本, 同时减少高概率分类好的样本的损失, 另外添加一个辅助的二进制分类器来服务于域适应的目的[19]. 这些方法简单且易于实现, 但由于忽略了类别之间的层次结构信息, 在准确性方面仍有所欠缺.

本文为声学场景类别间的层次结构关系, 设计了基于层次信息融合的声学场景分类模型. 一方面, 分别使用两个分类器对父类和子类声学场景进行分类. 另一方面, 由于同一父类下的兄弟子类之间存在着一些共同的特性, 在子类分类的时候将父类信息融合进子类信息中, 让父类的信息在兄弟子类之间共享, 迫使模型学习声学场景类别间的层次结构关系信息. 然而, 如果两个分类器预测的父类与子类不一致, 子类分类时将获得错误的父类信息, 对层次信息融合模型造成误导. 如父类分类器预测的类别是室外, 而子类分类器预测的类别是属于室内这个父类的购物中心. 因此, 本文还使用了一个层次依赖损失对这种预测的父类与子类不匹配的情况进行惩罚, 迫使模型的预测类别符合声学场景类别间的层次结构关系.

3 基于层次信息融合的声学场景分类方法

3.1 基于层次信息融合的声学场景分类模型框架

由于声学场景类别间存在着一种隐式的层次结构关系, 本文提出了层次信息融合的声学场景分类方法学习声学场景类别间的层次结构关系信息. 该方法主要包含两个模块: 声学场景特征学习模块, 层次信息融合模块. 图 2 给出了本文提出的层次信息融合的声学场景分类模型框架图. 在声学场景特征学习模块中, 首先从音频文件中提取 FBank 声学特征作为输入, 然后送入主干网络中学习高级特征. 主干网络部分采用 Suh 等人提出的 3 个残差网络并行 (Trident ResNet) 的结构[17]. 首先按频率轴将声谱图分成低频, 中频, 高频 3 个部分, 分别送入 3 个相同结构的残差网络进行学习. 残差网络中不会对频率轴进行下采样, 保证了频率轴的维度不被改变, 同时采用了空洞卷积来扩大感受野, 这在 Suh 等人的工作中已经被证明是有意义的. 最后将 3 个路径的输出按频率轴分裂的方式进行拼接, 获得网络学习到的高级特征. 接下来将学习到的高级特征送入层次信息融合模块. 首先将高级特征分别进行 1D 卷积, 得到父类表示和子类表示. 将父类表示送入父类分类器, 得到预测的父类类别. 同样的子类表示也将被送入子类分类器进行分类, 不同的是父类表示将会与子类表示拼接融合, 然后一同送入子类分类器. 通过这样的方式, 让父类信息在属于这一父类的兄弟子类之间共享, 迫使模型学习声学场景类别间的层次结构关系信息.

3.2 层次依赖损失

基于层次信息融合的声学场景分类模型中包含两个分类器: 父类分类器和子类分类器. 这样就会得到两个分类结果. 显然, 父类与子类的隶属关系是固定的, 即对于某一个子类, 当知道了它的类别之后, 它的父类也就确定了. 但是, 模型的预测并不能保证这样的隶属关系, 因此就会出现模型预测的父类类别与子类类别不匹配的情况出现, 即预测的子类不属于预测的父类. 为了解决这一问题, 引入了一个层次依赖损失来对这一情况进行惩罚, 迫使模型预测出的父类类别与子类类别符合它们的隶属关系[20]. 层次依赖损失的形式如式 (1).

$$\mathcal{L}_d = (ploss)^{\mathbb{D}\mathbb{I}_p}(ploss)^{\mathbb{D}\mathbb{I}_c} - 1 \tag{1}$$

其中, $\mathbb{D}$表示父类与子类的隶属关系是否正确. $\mathbb{I}$表示模型的预测结果是否正确, $\mathbb{I}_p$和$\mathbb{I}_c$分别表示父类与子类的预测结果是否正确. 而$ploss$是一个依赖惩罚, 设置为常数.

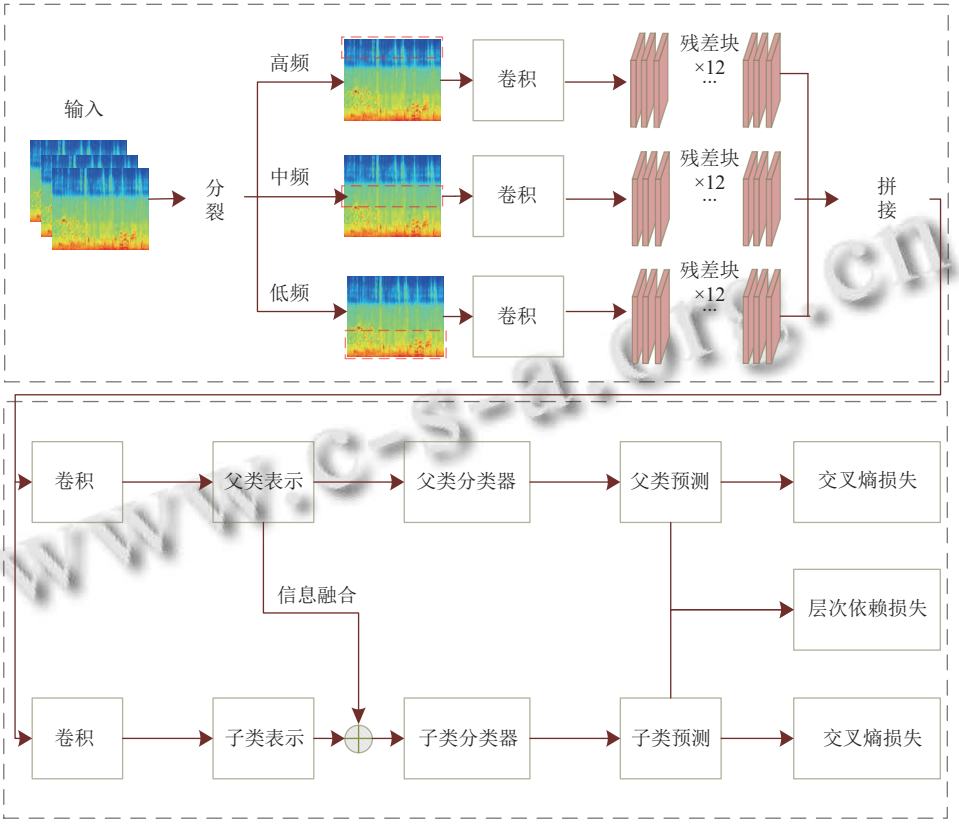

图 2 基于层次信息融合的声学场景分类模型框架图

$\mathbb{D}$的形式如式 (2). 若模型预测的子类确实属于预测的父类, 则$\mathbb{D}$的值为 0, 否则$\mathbb{D}$的值为 1. $\hat{y}_p$表示预测的父类类别, $\hat{y}_c$表示预测的子类类别.

$$\mathbb{D} = \begin{cases} 1, & \text{if } \hat{y}_c \nRightarrow \hat{y}_p \\ 0, & \text{else} \end{cases} \tag{2}$$

$\mathbb{I}$的形式如式 (3). 如果模型的预测结果正确, 则$\mathbb{I}$的值为 0, 若不正确则为 1. $\hat{y}$表示模型预测的类别, y 表示数据真实的类别.

$$\mathbb{I} = \begin{cases} 1, & \text{if } \hat{y} \neq y \\ 0, & \text{else} \end{cases} \tag{3}$$

因此, 根据模型预测结果的不同, 层次依赖损失会出现以下几种情况.

(1) 父类与子类都预测正确. 这种情况下它们的隶属关系也一定是正确的, 此时$\mathbb{D}$, $\mathbb{I}_c$, $\mathbb{I}_p$都为 0, 所以不会被惩罚.

(2) 父类预测正确, 子类预测错误. 此时若隶属关系正确, 则$\mathbb{D}$为 0, $\mathbb{I}_c$为 1, $\mathbb{I}_p$为 0, 所以不会被惩罚. 若隶属关系错误, $\mathbb{D}$为 1, $\mathbb{I}_c$为 1, $\mathbb{I}_p$为 0, 这种情况则会被惩罚.

(3) 父类预测错误, 子类预测正确. 这种情况下它们的隶属关系必然是错误的, 因为一个子类只隶属于一个父类, 此时$\mathbb{D}$为 1, $\mathbb{I}_c$为 0, $\mathbb{I}_p$为 1, 所以会被惩罚.

(4) 父类与子类都预测错误. 此时若隶属关系正确$\mathbb{D}$为 0, $\mathbb{I}_c$为 1, $\mathbb{I}_p$为 1, 这种情况下会不被惩罚. 若隶属关系错误, 此时$\mathbb{D}$为 1, $\mathbb{I}_c$为 1, $\mathbb{I}_p$为 1, 这种情况下则会被惩罚.

3.3 基于层次信息融合的声学场景分类的联合损失函数

最终, 如式 (4), 基于层次信息融合的声学场景分类模型的联合损失函数为父类分类的交叉熵损失、子类分类的交叉熵损失与层次依赖损失的加权求和.

$$\mathcal{L} = \mathcal{L}_{CE_P} + \mathcal{L}_{CE_C} + \alpha\mathcal{L}_D \tag{4}$$

4 实验结果与分析

4.1 数据集

本文的实验在 DCASE 2020 提供的 TAU 城市声学场景移动开发数据集 2020[21] 上进行. 该数据集中包含 23 040 条来自在 10 个欧洲城市录制的 10 个不同声学场景的录音数据. 10 个声学场景为: 机场、购物中心、地铁站、人行街道、公共广场、交通街道、公园、有轨电车、公交车和地铁. 所有的录音数据由 3 台不同的设备录制, 包含设备 A, B, C. 设备 A 为主录音设备, 包括一个 Soundman OKM II Klassik/studio A3、驻极体双耳麦克风和一个采用 48 kHz 采样率和 24 位分辨率的 Zoom F8 录音机. 设备 B 为三星 Galaxy S7, 设备 C 为是 iPhone SE. 此外, 基于以上 3 种设备录制的数据, 创建了 6 个移动设备的合成数据, 这 6 个模拟的设备分别成为 S1, S2, S3, S4, S5, S6. 其中设备 S4, S5, S6 的数据不会用于训练. 23 040 条录音中, 来自设备 A 的数据最多, 有 14 400 条. 其余 8 个设备各有 1 080 条录音. 每条音频数数据的长度均为 10 s, 且所有的数据都重采样为 44.1 kHz.

4.2 数据处理

基于层次信息融合的声学场景分类模型从音频文件中提取对数梅尔声谱图作为输入. 首先对每条音频数据进行预加重, 分帧, 加窗. 然后采用 2 048 个 FFT 点的短时傅里叶变换. 将每个频谱压缩到 256 个 Mel 频率标度, 然后取对数得到对数梅尔声谱图. 此外, 从对数梅尔声谱图中计算 deltas 和 delta-deltas, 并沿通道维度进行叠加获得最终模型的输入.

4.3 主干网络结构

主干网络中包含 3 条残差网络路径, 每条路径的网络结构具体配置如表 1 所示. 每条残差路径会先进行一次卷积, 频率维度的步长为 1, 时间维度的步长为 2. 然后会连接 12 个残差块. 每个残差块包含两次卷积, 在每次卷积之前都会进行批归一化处理, 并用线性整流函数激活. 卷积核的尺寸为 3×3.

对于时间维度, 在第 4, 第 7, 第 10 个残差块的第 1 次卷积中采用了步长 2 来对时间维度进行下采样. 而在频率维度则不会进行下采样, 所有的卷积步长均为 1. 同时对于频率维度, 在第 2, 第 3, 第 5, 第 6, 第 8, 第 9, 第 11, 第 12 个残差块的第 1 次卷积中采用了扩张率为 2 的扩张卷积. 扩张卷积可以在不改变特征图尺寸的基础上扩大感受野, 从而使得频率维度的大小在整个嵌入网络中的大小不会被改变, 确保了 3 条残差路径的输出可以按原来的分裂方式重新进行拼接.

表 1 残差网络结构配置

结构名	配置
输入	—
批归一化	学习γ和β
卷积	(1×2) 步长
残差块1	—
残差块2	第1次卷积使用 (2×1) 扩张率
残差块3	第1次卷积使用 (2×1) 扩张率
残差块4	第1次卷积使用 (1×2) 步长
残差块5	第1次卷积使用 (2×1) 扩张率
残差块6	第1次卷积使用 (2×1) 扩张率
残差块7	第1次卷积使用 (1×2) 步长
残差块8	第1次卷积使用 (2×1) 扩张率
残差块9	第1次卷积使用 (2×1) 扩张率
残差块10	第1次卷积使用 (1×2) 步长
残差块11	第1次卷积使用 (2×1) 扩张率
残差块12	第1次卷积使用 (2×1) 扩张率
输出	—

4.4 其他实验设置

实验中, 使用 TensorFlow 和 Keras 深度学习框架来搭建模型, 使用反向传播和动量为 0.9 的随机梯度下降算法, 批大小设置为 16. 超参数 α 设置为 0.8. 每次训练持续 510 个周期, 并在 2、6、14、30、62、126 和 254 个周期后将学习率重置为最大值 0.1, 然后根据余弦模式衰减到 1×10^{-5}.

4.5 超参数 α 的取值

图 3 展示了不同 α 的值, 模型的分类准确率情况. 从图 3 可以观察到, 当 α 取 0.8 时, 模型的分类准确率最高.

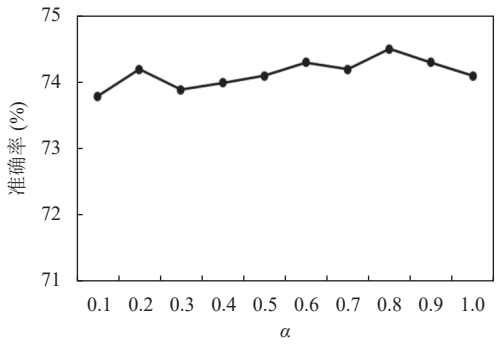

图 3 不同 α 值的模型分类准确率

4.6 消融实验

本节在 TAU 城市声学场景移动开发数据集 2020 上进行了消融实验.

消融实验的结果如表 2. 第 1 行为本文的方法所使用的基线模型 Trident ResNet, 它的分类准确率达到了 73.4%. 第 2 行为本文提出的层次信息融合声学场景分类方法, 但没有使用依赖损失, 它的分类准确率达到了 74.2%, 相较于基线模型 Trident ResNet 提升了 0.8%. 第 3 行是使用了层次依赖损失后的层次信息融合方法, 准确率达到了 74.5%. 相较于不使用层次依赖损失, 准确率继续提升了 0.3%. 由此可见, 两个分类器分类结果不匹配的情况确实会影响模型的性能, 而层次依赖损失的引入改善了这一现象. 上述实验结果充分证明了本文提出的基于层次信息融合的声学场景分类方法的有效性.

表 2 消融实验结果 (%)

模型	准确率
Trident ResNet (基线模型)	73.4
本文方法 (不使用层次依赖损失)	74.2
本文方法 (使用层次依赖损失)	74.5

4.7 基于层次信息融合的声学场景分类方法性能评估

为了对基于层次信息融合的声学场景分类方法的性能进行评估, 将该方法与其他先进的声学场景分类方法进行了比较.

表 3 展示了在 TAU 城市声学场景移动开发数据集 2020 上的实验结果. 将所有的设备分为 4 组. 第 1 组为主录音设备 A, 第 2 组为真实录音设备 B 和 C, 第 3 组为模拟的录音设备 S1, S2, S3, 第 4 组为模拟的录音设备 S4, S5, S6.

表 3 本文方法与其他方法在 TAU 城市声学场景移动开发数据集 2020 上的结果 (%)

模型	A	B&C	S1&S2 &S3	S4&S5 &S6	准确率
官方基线系统[22]	70.6	61.6	53.3	44.3	54.1
Suh等人[17]	82.5	75.6	72.3	70.3	73.4
Liu[18]	—	—	—	—	72.1
Gao等人[19]	79.7	76.4	70.8	69.2	72.5
本文方法	**83.4**	**76.5**	**73.3**	**71.1**	**74.5**

DCASE 官方提供的基线系统使用两个完全连接的层神经分类器, 并使用 OpenL3 提取输入音频嵌入. 它的平均分类准确率为 54.1%. 其中设备 A 上的效果最好, 这是因为设备 A 为主设备, 拥有最多的音频数据, 约 75%, 且设备 A 为专业的录音设备, 录制的音频数据质量最好. 但在其他设备上均观察到了严重的性能下降, 尤其是在不可见设备 S4, S5, S6 上. Suh 等人[17] 的方法平均分类准确率达到了 73.4%. Gao 等人[19] 的方法平均分类准确率达到了 72.5%. Liu[18] 的方法平均分类准确率达到了 72.1%. 本文提出的基于层次信息融合的声学场景分类方法获得了 74.5% 的平均分类准确率, 具体在各个设备上都获得了很有竞争力的结果.

5 总结

针对当前声学场景分类模型往往忽略声学场景类别间的层次结构关系信息的问题, 本文提出了基于层次信息融合的声学场景分类方法. 该方法设计了一个声学场景层次分类框架, 其中包含一个父类分类器, 一个子类分类器. 在子类分类的过程中, 将父类信息融合进子类信息中, 让父类信息在兄弟子类之间共享, 迫使模型学习类别间的层次结构关系. 我们还使用了一个层次依赖损失来对两个分类器预测的类别不匹配的情况进行惩罚. 在 TAU 城市声学场景移动开发数据集 2020 上的实验结果充分证明了基于层次信息融合的声学场景分类方法的有效性.

参考文献

1 Hou YB, Kang B, Van Hauwermeiren W, *et al*. Relation-guided acoustic scene classification aided with event embeddings. Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN). Padua: IEEE, 2022. 1–8. [doi: 10.1109/IJCNN55064.2022.9892893]

2 Kacprzak S, Kowalczyk K. Adversarial domain adaptation with paired examples for acoustic scene classification on different recording devices. Proceedings of the 29th European Signal Processing Conference (EUSIPCO). Dublin: IEEE, 2021. 1030–1034. [doi: 10.23919/EUSIPCO54536.2021.9616321]

3 Barchiesi D, Giannoulis D, Stowell D, *et al*. Acoustic scene classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine, 2015, 32(3): 16–34. [doi: 10.1109/MSP.2014.2326181]

4 Perera C, Zaslavsky A, Christen P, *et al*. Context aware computing for the internet of things: A survey. IEEE Communications Surveys & Tutorials, 2014, 16(1): 414–454. [doi: 10.1109/SURV.2013.042313.00197]

5 Martinson E, Schultz A. Robotic discovery of the auditory scene. Proceedings of the 2007 IEEE International Conference on Robotics and Automation. Rome: IEEE, 2007. 435–440. [doi: 10.1109/ROBOT.2007.363825]

6 Mesaros A, Heittola T, Benetos E, *et al*. Detection and

classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(2): 379–393. [doi: 10.1109/TASLP.2017.2778423]

7 Wang J, Li SC. Self-attention mechanism based system for DCASE2018 challenge task1 and task4. Technical Report, Beijing: Beijing University of Posts and Telecommunications. 2018. [doi: 10.13140/RG.2.2.28317.13281]

8 Guo JX, Xu N, Li LJ, *et al*. Attention based CLDNNs for short-duration acoustic scene classification. Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017. 469–473. [doi: 10.21437/Interspeech.2017-440]

9 Xu KL, Feng DW, Mi HB, *et al*. Mixup-based acoustic scene classification using multi-channel convolutional neural network. Proceedings of the 19th Pacific Rim Conference on Multimedia. Hefei: Springer, 2018. 14–23. [doi: 10.1007/978-3-030-00764-5_2]

10 Nguyen T, Pernkopf F. Acoustic scene classification with mismatched devices using cliqueNets and mixup data augmentation. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 2330–2334. [doi: 10.21437/Interspeech.2019-3002]

11 Dang A, Vu TH, Wang JC. Acoustic scene classification using convolutional neural networks and multi-scale multi-feature extraction. Proceedings of the 2018 IEEE International Conference on Consumer Electronics (ICCE). Las Vegas: IEEE, 2018. 1–4. [doi: 10.1109/ICCE.2018.8326315]

12 Hu H, Yang CHH, Xia JX, *et al*. A two-stage approach to device-robust acoustic scene classification. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 845–849. [doi: 10.1109/ICASSP39728.2021.9414835]

13 Pham LD, McLoughlin I, Phan H, *et al*. A robust framework for acoustic scene classification. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 3634–3638. [doi: 10.21437/Interspeech.2019-1841]

14 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

15 Larkey LS, Croft WB. Combining classifiers in text categorization. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich: ACM, 1996. 289–297. [doi: 10.1145/243199.243276]

16 McDonnell MD, Gao W. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 141–145. [doi: 10.1109/ICASSP40776.2020.9053274]

17 Suh S, Park S, Jeong Y, *et al*. Designing acoustic scene classification models with CNN variants. Technical Report, Daejeon: Electronics and Telecommunications Research Institute, 2020.

18 Liu J. Acoustic scene classification with residual networks and attention mechanism. Technical Report, Wuhan: Maxvision, 2020.

19 Gao W, McDonnell M, UniSA S. Acoustic scene classification using deep residual networks with focal loss and mild domain adaptation. Technical Report, Mawson: University of South Australia, 2020.

20 Gao DH, Yang W, Zhou H, *et al*. Deep hierarchical classification for category prediction in e-commerce system. Proceedings of the 3rd Workshop on E-commerce and NLP. Seattle: Association for Computational Linguistics, 2020. 64–68.

21 Heittola T, Mesaros A, Virtanen T. Acoustic scene classification in DCASE 2020 Challenge: Generalization across devices and low complexity solutions. Proceedings of the 5th the Workshop on Detection and Classification of Acoustic Scenes and Events. Tokyo: DCASE, 2020. 56–60.

22 Cramer AL, Wu HH, Salamon J, *et al*. Look, listen, and learn more: Design choices for deep audio embeddings. Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 3852–3856. [doi: 10.1109/ICASSP.2019.8682475]

(校对责编: 孙君艳)