基于新型损失函数DV-Softmax的声纹识别方法

曹 毅，李 平，吴伟官，夏 宇，高清源

（江南大学 机械工程学院，江苏 无锡 214122）

摘 要：针对声纹识别领域中现有模型分类损失函数无法有效区分类别之间的可分性与缺乏对声纹数据质量关注的问题，本文提出一种新的分类损失函数DV-Softmax。首先，介绍了声纹领域现有边界损失函数工作原理；其次，介绍目标检测领域的挖掘损失函数，并在其基础上提出模糊样本的概念；再次，引入人脸识别领域的MV-Softmax损失函数，并加入模糊样本，使其能自适应强调不同样本之间的区别并指导特征学习；最后，分别在Voxceleb1和SITW数据集进行声纹识别的研究。实验结果表明，DV-Softmax损失函数相较于现有边界损失函数，等错误率分别下降8%和5.4%，其验证了该损失函数有效解决类别之间的可分性及对样本声纹数据质量的关注，并在声纹识别领域具有良好的性能。

关键词：深度学习；声纹识别；损失函数；信息挖掘

中图分类号：TN912.34 **文献标志码**：A **文章编号**：1671-5497(2024)11-3318-09

DOI：10.13229/j.cnki.jdxbgxb.20221635

Voiceprint recognition method based on novel loss function DV-Softmax

CAO Yi，LI Ping，WU Wei-guan，XIA Yu，GAO Qing-yuan

（*School of Mechanical Engineering*，*Jiangnan University*，*Wuxi* 214122，*China*）

Abstract：In view of the problems that the classification loss function of existing models in the field of voiceprint recognition cannot effectively distinguish the separability between categories and lack of attention to the quality of voiceprint data，a new classification loss function DV-Softmax is proposed in this paper. Firstly，the working principle of the existing boundary loss function in voiceprint field is introduced. Secondly，the mining loss function in the field of object detection is introduced，and the concept of fuzzy sample is proposed based on it. Then，the MV-Softmax loss function is introduced in the field of face recognition，and fuzzy samples are added to make it adaptive to emphasize the difference between different samples and guide the feature learning. Finally，the voicing recognition was studied on Voxceleb1 and SITW data respectively. The experimental results show that compared with the existing boundary loss function，the equal error rate of DV-Softmax is reduced by 8% and 5.4%，respectively，which verifies

收稿日期：2022-12-28.

基金项目：国家自然科学基金项目(51375209)；江苏省“六大人才高峰”计划项目(ZBZZ-012)；江苏省研究生创新计划项目(KYCX18_0630，KYCX18_1846)；高等学校学科创新引智计划项目(B18027).

作者简介：曹毅(1974-)，男，教授，博士. 研究方向：机器人，深度学习. E-mail：caoyi@jiangnan.edu.cn

that the DV-Softmax loss function effectively solves the separability between categories and concerns the quality of sample voice print data, and has a good performance in the field of voice print recognition。
Key words: deep learning; voiceprint recognition; loss function; information mining

0　引　言

声纹是指原始音频数据中包含能表征该说话人的音频特征，进而实现对不同说话人进行辨别的方法。声纹识别是语音处理领域的热点研究方向之一，与人脸识别或指纹识别等生物特征类似，每个人的声纹都是独一无二的，无法通过模仿达到相同的发音特性和声道特征。因此，声纹识别技术被广泛应用于银行交易和远程支付的信息安全、调查嫌疑人是否有罪、自动身份标记等领域。

近年来，高级声纹识别模型通常建立在深度卷积神经网络上，其学习到的辨别特征起重要作用。为训练声纹模型，神经网络通常采用分类损失函数以使模型收敛。因此，越来越多的研究人员将重心转移到重新设置经典的分类损失函数构建深度声纹识别模型上。从根本上讲，如果声纹特征的类内紧凑性和类间可分性都能得到最大化，则声纹特征是可区别的。然而，主流的Softmax损失函数[1]缺乏用于深度声纹识别的特征辨别能力。为解决这个问题，2017年Liu等[2]提出在真实类别与其他类别之间引入一个角裕度的A-Softmax损失函数，以此鼓励更优的类间方差；2018年Wang等[3]提出附加裕度的AM-Softmax损失函数，进而提高角裕度损失的稳定性；2019年Deng等[4]提出加性角边缘的AAM-Softmax损失函数，其具有更加清晰的几何解释；2020年Thienpondt等[5]提出在AAM-Softmax上进行补偿偏移的改进；2020年Li等[6]提出广义焦点损失GFL，将离散形式推广到连续形式；2021年Ma等[7]提出通过特征范数逼近图像质量的NMM-Softmax损失函数；2022年Lee等[8]提出通过引入线性角度裕度的AMM-Softmax损失函数；2022年Boutros等[9]提出弹性边缘损失，最大化类间差异。

综上所述，针对声纹识别而言，尽管国内外诸多学者开展了大量实验研究并取得一定的研究成果，但不难发现以下问题对声纹识别研究的影响：(1) 目前的损失函数研究中，通常是基于音频处理良好的训练集的假设，这是不切实际的；(2) 忽视了不同样本的信息特征挖掘对辨别学习的重要性；(3) 对所有类别使用相同的固定值扩大特征裕度，忽视了不同类别间的可分性。

针对上述问题，首先，引入了挖掘损失函数，在其将样本分为简单样本和硬样本的基础上，提出模糊样本的概念，并对三类样本给予不同的权重以进行信息挖掘；其次，引入MV-Softmax损失函数[10]，指导区分性特征学习，从而达到对不同类别之间的可分性学习；再次，结合上述内容提出新的损失函数；最后，基于Voxceleb1[11]和SITW[12]声纹数据集进行声纹识别的研究。研究结果表明，该损失函数能有效提取特征信息并提升模型的性能。

1　损失函数

近年来，在声纹识别领域，利用分类损失函数对声纹模型进行优化已成为重要组成部分之一，损失函数主要用来衡量模型训练值与真实值之间的差距，为模型的优化提供方向。该领域现有损失函数大都基于Softmax损失函数及其变体的边界损失函数，必须指出的是，现有损失函数在一定程度上解决了样本难分类问题，然而未能对特征信息充分挖掘，缺乏对样本自身的关注。近些年在目标检测和人脸识别等领域，通过对不同样本之间的关系提出挖掘损失函数，并取得较理想的结果。因此，为实现特征信息的充分挖掘，在声纹领域中开展挖掘损失函数的研究是非常有必要的。

1.1　边界损失函数

边界损失函数是以Softmax损失函数为原型，采取不同形式增强特征识别的一类损失函数。其中，Softmax损失函数是由最后一个全连接层、Softmax激活函数和交叉熵损失函数的组合。交叉熵损失函数表达式为：

$$L_1=-\sum_{j=1}^{J}t_j\log p_j \tag{1}$$

式中：$[t_1,t_2,\cdots,t_J]$为样本x的标签l对应的one-hot编码，当x属于第j类别时，t_j为1，其余为0；p_j为样本x通过Softmax激活函数获取属于第j类的后验概率。Softmax函数表达式为：

$$p_j=\frac{\exp(w_j^T x+b_j)}{\sum_{j=1}^{J}\exp(w_j^T x+b_j)} \tag{2}$$

由式(1)、式(2)可得Softmax损失函数为:

$$L_2=-\log\frac{\exp(w_l^T x+b_l)}{\exp(w_l^T x+b_l)+\sum_{j\neq l}^{J}\exp(w_j^T x+b_j)} \tag{3}$$

式中:W_l为样本x属于第l个类的权重;b_j为第j个类别的常数项;x为样本。

为进一步最大化类间距离、最小化类内距离,式中:$sf(m,\theta_{l,l})$为边界损失函数的裕度函数。文献[2]提出以角度距离进行优化的A-Softmax,其$sf(m_1,\theta_{l,l})=\cos(m_1\theta_{l,l})$,裕度$m_1$为大于等于1的整数。文献[3]提出以余弦距离进行优化的AM-Softmax,其$f(m_2,\theta_{l,l})=\cos(\theta_{l,l})-m_2$,裕度$m_2>0$。文献[4]提出加性角边缘损失AAM-Softmax,其$f(m_3,\theta_{l,l})=\cos(\theta_{l,l}+m_3)$,裕度$m_3>0$。

1.2 挖掘损失函数

挖掘损失函数目前主要应用于人脸识别和目标检测等领域,其思想是将样本分为简单样本和硬样本,通过强调富含特征信息的硬样本权重,同时降低对简单样本的权重,因此,也会产生更多具有区别性的特征。挖掘损失函数的一般形式为:

$$L_4=-g(p_l)\log\cdot\frac{\exp(s\cos(\theta_{l,l}))}{\exp(s\cos(\theta_{l,l}))+\sum_{j\neq l}^{J}\exp(s\cos(\theta_{j,l}))} \tag{6}$$

式中:$p_l=\exp(s\cos(\theta_{l,l}))/(\exp(s\cos(\theta_{l,l}))+\sum_{j\neq l}^{J}\exp(s\cos(\theta_{j,l}))$为预测真值对应的概率值;$g(p_l)$为权重指示函数。2016年Shrivastava等[13]提出HM-Softmax损失函数,当其样本为简单样本时,$g(p_l)=0$;当样本为硬样本时,$g(p_l)=1$。HM-Softmax采用固定的网络对数据训练,将表现好的样本归为简单样本,其余归为硬样本。2017年Lin等[14]提出通过增加调制因子的F-Softmax区分简单样本和硬样本,其$g(p_l)=(1-p_l)^{\gamma}$,$0\leqslant\gamma\leqslant5$,通常取$\gamma$值为2。

1.3 MV-Softmax损失函数

MV-Softmax损失函数在挖掘损失函数基础上提出一种以语义指导对简单样本和硬样本进行区分并结合边界的损失函数,将训练集中于硬样本上。其中,基于边界损失函数,定义一个二进制近些年提出各类边界损失函数,采用优化角度距离和余弦距离替代优化内积,w_j和x的内积可重写为:

$$w_j^T x=\|w_j\|\,\|x\|\cos(\theta_{j,l}) \tag{4}$$

式中:$\theta_{j,l}$表示w_j和x的夹角。基于式(4),取$s=\|w_j\|\,\|x\|$,可得各类边界损失函数的一般形式为:

$$L_3=-\log\frac{\exp(sf(m,\theta_{l,l}))}{\exp(sf(m,\theta_{l,l}))+\sum_{j\neq l}^{J}\exp(s\cos(\theta_{j,l}))} \tag{5}$$

指示器I_j,以自适应地指示样本是否属于硬样本,具体形式为:

$$I_j=\begin{cases}0, & f(m,\theta_{l,l})-\cos(\theta_{j,l})\geqslant 0\\ 1, & f(m,\theta_{l,l})-\cos(\theta_{j,l})<0\end{cases} \tag{7}$$

从式(7)的定义中可以看出,当样本被误分为别的类别时,即$f(m,\theta_{l,l})-\cos(\theta_{j,l})<0$,该损失函数将其判定为硬样本,通过加强对硬样本的训练提升模型精度。基于该指示器I_j,MV-Softmax损失函数定义如下:

$$L_5=-\log\cdot\frac{\exp(sf(m,\theta_{l,l}))}{\exp(sf(m,\theta_{l,l}))+\sum_{j\neq l}^{J}h(t,\theta_{j,l},I_j)\exp(s\cos(\theta_{j,l}))} \tag{8}$$

式中:$h(t,\theta_{j,l},I_j)\geqslant1$为样本的权重函数,针对指示器$I_j$区分后的不同样本给予不同权重,以此强调硬样本。以下两种形式分别为固定权重函数和自适应权重函数:

$$h(t,\theta_{j,l},I_j)=\exp(stI_j) \tag{9}$$

$$h(t,\theta_{j,l},I_j)=\exp(st(\cos(\theta_{j,l})+1)I_j) \tag{10}$$

其中,超参数$t\geqslant0$。当$t=0$时,MV-Softmax损失式(8)等同于边界损失式(5)。

综上所述,边界损失函数通过引入裕度扩大类间距离、缩小类内距离,但未能重视不同样本的差别,挖掘损失函数通过将样本划分为简单样本和硬样本,并强调硬样本的权重以提升训练效果,但通常是通过经验或复杂的预训练来完成。MV-Softmax损失函数采用语义指导自适应地区分样本,从而实现更理想化的训练效果。然而,上述损失函数通常以样本都是纯净样本为前提,即不包含噪声。而在实际应用中,样本采集通常难以去除外部噪声,这部分样本对训练效果产生负面影响,应降低对该部分样本的权重。

2　DV-Softmax损失函数

针对上述样本问题，在简单样本和硬样本的基础上，提出模糊样本的概念。对声纹识别系统中的样本，将纯净且容易判别类别的样本归为简单样本；纯净且包含重要声纹信息但难以判别类别的样本归为硬样本；包含噪声的样本归为模糊样本。从定义可知，简单样本对模型训练仅起微弱的正作用，硬样本训练困难但对训练起较大的正作用，模糊样本所包含的噪声对训练起反作用。因此，通过加强硬样本的权重，降低简单样本和模糊样本的权重可有效提升训练效果。

2.1　DV-Softmax损失函数定义

针对某一类别样本，由式(2)可知$0\leqslant p_j\leqslant 1$，当$p_j=1$时，模型可直接判别该样本类别，对后续训练效果基本不起作用。当$p_j=0$时，该样本完全不包含该类别的特征信息，对训练效果起反作用。当介于两者之间时，则其样本包含更多该类别样本信息，应着重训练。因此，提出样本权重指示函数$d(p_l)$，其表达式为：

$$d(p_l)=\frac{6}{\sqrt{2\pi}}\exp(-18(p_l-0.5)^2)+1 \tag{11}$$

式中：p_l为式(2)中样本类别l的概率。如图1所示为权重指示函数示意图。

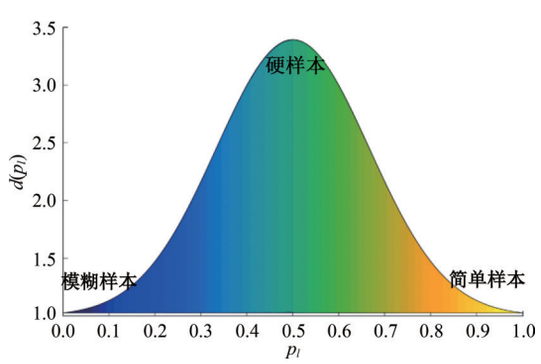

图 1　权重指示函数示意图

Fig. 1　Schematic diagram of weight indicator function

显然，当靠近两端端点时，$d(p_l)$的值小，当远离两端时，$d(p_l)$的值增大，且呈现先快速增大、后缓慢增大，从而强调不同样本的重要性。根据式(8)和式(11)可知，提出一个新的损失函数DV-Softmax，其表达式为：

$$L_5=-d(p_l)\log\cdot\frac{\exp(sf(m,\theta_{l,l}))}{\exp(sf(m,\theta_{l,l}))+\sum_{j\neq l}^{J}h(t,\theta_{j,l},L_j)\exp(s\cos(\theta_{j,l}))} \tag{12}$$

式中：$L_j=d(p_j)-1$；$h(t,\theta_{j,l},L_j)\geqslant 1$为重加权函数，用于强调不同的样本的权重，分别有以下两种形式：

$$h(t,\theta_{j,l},L_j)=\exp(stL_j) \tag{13}$$

$$h(t,\theta_{j,l},L_j)=\exp(st(\cos(\theta_{j,l})+1)L_j) \tag{14}$$

式中：超参数$t\geqslant 0$；$\cos(\theta_{j,l})$为类别j和类别l之间的余弦相似度。

本文设计新的损失函数，其框架如图2所示，主要包含两个关键部分：(1)利用权重指示函数分布作为线索估计三类样本标签，强调硬样本的信息量；(2)通过MV-Softmax损失函数指导的可分性学习，从而实现对不同类别之间的区分。综上所述，本文提出的损失函数可以动态区分信息量不同的样本，并明确强调硬样本中的信息向量，同时吸收不同类别间的可辨别性，以指导区分性特征学习。

2.2　与现有损失函数对比

2.2.1　对比边界损失函数

为说明DV-Softmax损失函数相对于传统的边界损失函数具有优势，如图3所示中左图的示例。假设有5个样本x_1、x_2、x_3、x_4、x_5，它们都来自类别1，其中只有x_1被很好地分类，其余没有。由式(5)可知，边界损失函数对不同的类别有固定的裕度，其理想状态为样本x_5与类别1的距离相较

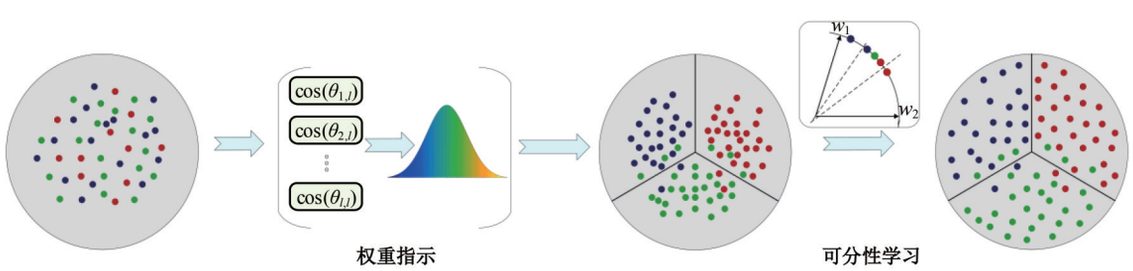

图 2　DV-Softmax框架图

Fig. 2　The framework of the proposed DV-Softmax

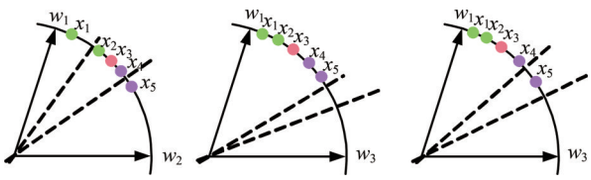

图3 特征角度下的几何解释

Fig. 3 Geometric interpretation from feature perspective

于类别2和类别3的距离都要小。由于其固定的裕度,训练结果会选取类别1和其他两个类别的裕度中较大的一个,由图2可知,样本已经被很好地分类,不需要额外限制。该损失函数也未考虑到样本的质量,部分含有噪声的模糊样本(如x_5)将会对结果起反作用。DV-Softmax损失函数采用$h(t,\theta_{j,l},L_j)$重加权函数将会对不同的类别产生不同的裕度,权重指示函数$d(p_l)$使训练更集中于硬样本,很好地强调了不同样本之间和不同类别之间的关系。

2.2.2 对比挖掘损失函数

类似地,仍以图2中左图为例,假设5个样本都属于类别1。HM-Softmax根据经验辨别简单样本和硬样本,图中的样本x_1会被舍弃,采用其余样本进行训练。Focal-Softmax对所有样本进行加权,样本x_1有较小的损失值,相对地,样本x_5有较大的损失值。前者直接将部分样本舍弃,可能导致部分信息丢失,后者未考虑硬样本中是否包含噪声,过于强调样本5将导致模型更注重于噪声,不能很好地进行分类。DV-Softmax损失函数采用$h(t,\theta_{j,l},L_j)$重加权函数通过对样本与分类器之间的局部关系学习特征,并结合权重指示函数$d(p_l)$在全局关系上区分简单样本、硬样本和模糊样本。这种相较于挖掘损失函数单一的全局关系区分方式更具优越性。

2.2.3 对比MV-Softmax损失函数

MV-Softmax损失函数继承了边界损失函数的固定裕度(例如,对于真值类别1,其与类别2和类别3的裕度相同),并在其基础上学习不同类别之间的潜在可分性(例如,类别2和类别3之间的可分性),使其具有自适应的裕度。其有效地解决了分类器与分类器之间的局部关系,但由式(8)可知,该损失函数仍采用边界损失函数的逻辑边界,对正确分类样本和错误分类样本采用相同的挖掘条件,这将出现难以适应数据样本不平衡的情况。DV-Softmax损失函数采用$h(t,\theta_{j,l},L_j)$重加权函数对某一样本属于不同类别进行判别,以此获取不同类别之间的自适应裕度。如图2中间图所示,MV-Softmax损失函数对图中5个样本的挖掘条件一致,致力于将5个样本全部归为类别1。然而,模糊样本包含一些噪声,强调该样本将降低训练效果。DV-Softmax损失函数采用权重指示函数$d(p_l)$对不同样本进行挖掘,如图2右图所示,其重点强调硬样本(x_3),并降低对简单样本(x_1、x_2)和模糊(x_4、x_5)样本的重视,以此适应不同样本之间的关系。

本节对新的损失方程DV-Softmax进行了可行性分析及与现有损失函数性能作对比,并且可以通过典型的Adam优化器进行训练,与传统声纹识别中的损失函数仅在最后一个全连接层的计算上有差异,算法的伪代码如表1所示。

表1 DV-Softmax伪码表

Table 1 Pseudocode table of DV-Softmax

Algorithm:DV-Softmax
Input: Training set x with its annotated label y. Epochs, Adam, The hyper-parameter t.
Initialization: Randomly initialize the parameter $\boldsymbol{\Theta}$ in convolution layers and $\boldsymbol{W}$ in the last fully connected layer.
function GO ( ):
for epoch in Epochs to **do**
Training model use Adam
Forward: According to the indication of different examples Eq. (11), we compute the DV-Softmax loss by Eq. (12);
Backward: Update the parameters $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ by Adam
end for
return W and Θ
end function
Output: Parameters $\boldsymbol{\Theta}$ and $\mathbf{W}$.

3　实验设计与结果分析

3.1　实验数据集及评价指标

为进一步验证MV-Softmax损失函数在声纹识别领域应用的有效性,利用Voxceleb1数据集、SITW数据集在声纹识别网络ECAPA-TDNN[15]上开展损失函数的实验研究。其中,以Voxceleb1训练集作为本实验的训练集,分别在Voxceleb1测试集、SITW的core-core测试集上进行测试,其中,Voxceleb1训练集包含1 211名说话人共计148 642条语音。Voxceleb1测试集、SITW的core-core测试集分别包含40名说话人共计4 874条语音和180名说话人共计2 883条语音。

为评价损失函数对声纹识别的效果,设置等错误率(Equal error rate, EER)、最小检测代价函数(Minimum normalized detection cost, minDCF)和检测错误权衡曲线(Detection error tradeoff, DET)作为评价指标[16]。定义错误接受率(false acceptance rate, FAR)和错误拒绝率(False rejection rate, FRR)两个参数如下:

$$\mathrm{FAR}=\frac{N_{fa}}{N_{\mathrm{impostor}}}\times 100\% \tag{15}$$

$$\mathrm{FRR}=\frac{N_{fr}}{N_{\mathrm{target}}}\times 100\% \tag{16}$$

式中:N_{fa}表示声纹系统将非目标类别的测试样本错判为目标类别的样本数,N_{impostor}表示数据集中非目标类别的样本总数,N_{fr}表示系统将目标类别的测试样本错判为非目标类别的样本数,N_{target}表示数据集中目标类别的样本总数。

等错误率是指当错误拒绝率FRR和错误接受率FAR相等时的值,由定义可知,等错误率的值越小,则系统的性能越好。最小检测代价函数定义为:

$$\min \mathrm{DCF}=C_{\mathrm{FR}}\times \mathrm{FRR}\times P_{\mathrm{target}}+C_{\mathrm{FA}}\times \mathrm{FAR}\times (1-P_{\mathrm{target}}) \tag{17}$$

式中:C_{FR}和C_{FA}分别表示错误拒绝和错误接受的惩罚代价;P_{target}表示目标类别的先验概率。检测错误权衡曲线是以错误接受率FAR为横坐标,错误拒绝率FRR为纵坐标的曲线。

3.2　平台及系统设置

本文声纹识别系统都是在Pytorch平台实现的,采用Adam优化器优化模型性能,batch size设为128,初始学习速率设为0.001,并采用余弦衰减的学习速率策略调整学习速率,训练轮次设为70,使用批量标准化和ReLU激活函数加速收敛。为验证损失函数之间的差别,特征采集及网络结构保持一致,其中语音特征采用梅尔频率倒谱系数,网络采用ECAPA-TDNN网络结构。

3.3　实验结果分析

为评估DV-Softmax损失函数在声纹识别领域的性能,基于Voceleb1和SITW数据集,开展声纹识别的等错误率、最小检测代价函数和检测错误权衡曲线对比实验。

3.3.1　Voxceleb1数据集下的实验对比

为验证DV-Softmax损失函数相较于传统的边界损失函数、挖掘损失函数以及MV-Softmax损失函数的优秀性能,在Voxceleb1数据集下,采用不同损失函数开展了声纹识别研究,其中最小检测代价函数指标中的P_{target}分别取0.1,0.01和0.001。将式(11)用于其他损失函数以D开头进行表示,f和a分别表示固定形式和自适应形式,实验结果如表2所示。

表 2　Voxceleb1测试集上不同损失函数的性能比较

Table 2　Performance comparison of different loss functions on Voxceleb1 test sets

损失函数	EER/%	minDCF(P=0.1)	minDCF(P=0.01)	minDCF(P=0.001)
Softmax	3.89	0.267	0.436	0.483
A-Softmax	3.02	0.193	0.352	0.405
AM-Softmax	2.72	0.150	0.308	0.350
AAM-Softmax	2.49	0.132	0.269	0.312
F-Softmax	3.93	0.243	0.418	0.463
MV-AAM-Softmax-f	2.38	0.126	0.258	0.298
MV-AAM-Softmax-a	2.34	0.121	0.254	0.288
D-AAM-Softmax	2.44	0.123	0.256	0.295
D-F-Softmax	3.71	0.218	0.410	0.443
DV-AAM-Softmax-f	2.32	0.118	0.242	0.278
DV-AAM-Softmax-a	**2.29**	**0.113**	**0.238**	**0.272**

由表2可知:

(1) D-F-Softmax损失函数相较于F-Softmax损失函数,EER下降5.6%,minDCF在P值为0.1、0.01、0.001分别降低10.3%、1.9%、4.3%,D-AAM-Softmax损失函数相较于AAM-Softmax,EER下降2%,minDCF在P值为0.1、0.01、0.001分别降低6.8%,4.8%,5.4%,进一步验证了权重指示函数$d(p_t)$的有效性。

(2) DV-AAM-Softmax损失函数相较于边界损失函数、挖掘损失函数和MV-Softmax损失函数,都取得了最佳性能。DV-AAM-Softmax损失函数相较于边界损失函数AAM-Softmax,EER下降8%,minDCF在P值为0.1、0.01、0.001分别降低14.4%、11.5%、12.8%,DV-AAM-Softmax损失函数相较于挖掘损失函数F-Softmax,EER下降41.7%,minDCF在P值为0.1、0.01、0.001分别降低53.5%、43.1%、41.3%。DV-AAM-Softmax损失函数相较于MV-AAM-Softmax损失函数,EER下降2.1%,minDCF在P值为0.1、0.01、0.001分别降低6.6%、6.3%、5.6%,进一步验证了DV-AAM-Softmax损失函数具有良好的分类性能。

(3) DV-AAM-Softmax损失函数自适应形式相较于固定形式,EER下降1.3%,minDCF在P值为0.1、0.01、0.001分别降低4.2%、1.7%、2.2%,以此表明自适应式(14)优于固定式(13)。这是因为在对更困难的样本分类时,区分特征学习更加重要。

由图4可知,DV-AAM-Softmax损失函数相较于现有边界损失函数、挖掘损失函数、MV-Softmax损失函数在识别任务中更具有优越性。

3.3.2 SITW数据集下的实验对比

上述单一数据集下的实验结果仅反映损失函数在单一数据集上的性能,为进一步验证损失函数在不同数据集上的性能,综合体现其泛化性能,在SITW数据集下采用不同损失函数开展了声纹识别研究,并采用相同的指标进行实验,实验结果如表3所示。

(1) D-F-Softmax损失函数相较于F-Softmax损失函数,EER下降1.7%,minDCF在P值为0.1、0.01、0.001分别降低6.1%、1.8%、4.7%,D-AAM-Softmax损失函数相较于AAM-Softmax,EER下降1%,minDCF在P值为0.1、0.01、0.001分别降低2.7%、1.4%、2.9%,进一步验证了权重指示函数$d(p_t)$的有效性。

(2) DV-AAM-Softmax损失函数相较于边界损失函数、挖掘损失函数和MV-Softmax损失

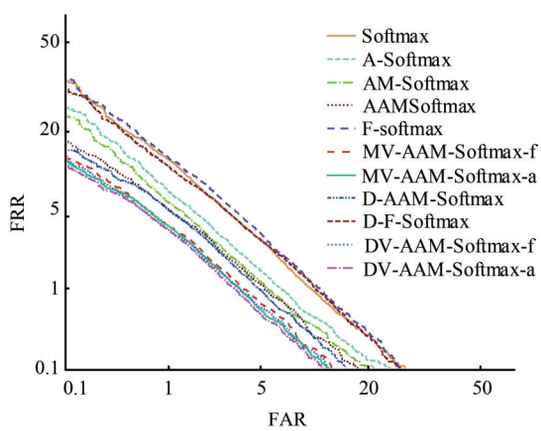

图4 Voxceleb1数据集上不同损失函数的DET曲线对比

Fig. 4 Comparison of DET curves of different loss functions on Voxceleb1 data set

表3 SITW测试集上不同损失函数的性能比较

Table 3 Performance comparison of different loss functions on SITW test sets

损失函数	EER/%	minDCF(P=0.1)	minDCF(P=0.01)	minDCF(P=0.001)
Softmax	6.29	0.296	0.546	0.752
A-Softmax	5.21	0.256	0.510	0.706
AM-Softmax	4.73	0.226	0.433	0.644
AAM-Softmax	3.91	0.179	0.348	0.551
F-Softmax	6.35	0.312	0.553	0.763
MV-AAM-Softmax-f	3.78	0.168	0.334	0.542
MV-AAM-Softmax-a	3.75	0.165	0.323	0.538
D-AAM-Softmax	3.87	0.174	0.343	0.535
D-F-Softmax	6.24	0.293	0.543	0.727
DV-AAM-Softmax-f	3.73	0.165	0.320	0.521
DV-AAM-Softmax-a	3.70	0.162	0.316	0.515

函数,都取得了最佳性能。DV-AAM-Softmax损失函数相较于边界损失函数AAM-Softmax,EER下降5.4%,minDCF在P值为0.1、0.01、0.001分别降低9.5%、9.2%、6.5%,DV-AAM-Softmax损失函数相较于挖掘损失函数F-Softmax,EER下降41.7%,minDCF在P值为0.1、0.01、0.001分别降低48.1%、42.9%、32.5%,DV-AAM-Softmax损失函数相较于MV-AAM-Softmax损失函数,EER下降1.3%,minDCF在P值为0.1、0.01、0.001分别降低1.8%、2.2%、4.3%,进一步验证了DV-AAM-Softmax损失函数具有良好的分类性能。

(3) DV-AAM-Softmax损失函数自适应形式相较于固定形式,EER下降0.8%,minDCF在P值为0.1、0.01、0.001分别降低1.8%、1.3%、1.2%,进一步表明自适应式(14)优于固定式(13)。

由图5也可看出,DV-Softmax损失函数优于现有损失函数,进一步验证了该损失函数具有良好的泛化性能。

综合上述实验可知,通过将样本进一步区分注意、特征挖掘和边缘最大损失的优点继承到一个公式中,DV-Softmax损失函数在声纹识别中显示了其可靠的性能,更有利于声纹的分类。

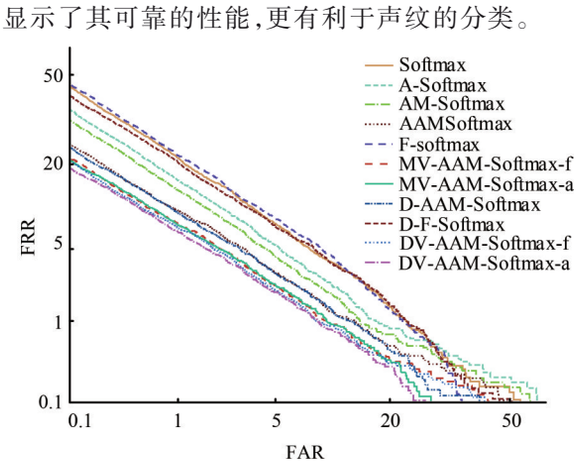

图 5　SITW 数据集上不同损失函数的 DET 曲线对比

Fig. 5　Comparison of DET curves of different loss functions on SITW data set

4　结束语

针对声纹识别领域现有损失函数方法未能有效区分样本的重要度,提出了一种集合样本分类权重自适应函数和分类器之间权重函数的DV-Softmax损失函数。首先,基于目标检测领域内的挖掘损失函数,在其样本分类为简单样本和硬样本的基础上,提出模糊样本的概念,进而提出权重自适应函数$d(p_t)$,有效地进行样本权重分类。其次,结合人脸识别领域内的MV-Softmax损失函数,在其基础上加入模糊样本的概念,进一步改善分类器之间的关系,并具有一定的语义指导进行分类。最后,基于Voxceleb1和SITW数据集开展了声纹识别研究,实验结果表明,该损失函数能有效提升声纹识别的性能。

参考文献:

[1] Ranjan R, Castillo C D, Chellappa R. L2-constrained softmax loss for discriminative face verification[J]. Arxiv Preprint, 2017, 3: No. 170309507.

[2] Liu W, Wen Y, Yu Z, et al. Sphereface: deep hypersphere embedding for face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hololulu, USA, 2017: 212-220.

[3] Wang F, Cheng J, Liu W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.

[4] Deng J, Guo J, Xue N, et al. Arcface: additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4690-4699.

[5] Thienpondt J, Desplanques B, Demuynck K. Cross-lingual speaker verification with domain-balanced hard prototype mining and language-dependent score normalization[J]. Arxiv Preprint, 2020, 7: No. 200707689.

[6] Li X, Wang W, Wu L J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21002-21012.

[7] Ma C, Sun H, Zhu J, et al. Normalized maximal margin loss for open-set image classification[J]. IEEE Access, 2021, 9: 54276-54285.

[8] Lee J, Wang Y, Cho S. Angular margin-mining softmax loss for face recognition[J]. IEEE Access, 2022, 10: 43071-43080.

[9] Boutros F, Damer N, Kirchbuchner F, et al. Elasticface: elastic margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 1578-1587.

[10] Wang X, Zhang S, Wang S, et al. Mis-classified vector guided softmax loss for face recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020, 34(7): 12241-12248.

[11] Nagrani A, Chung J S, Zisserman A, et al. Voxceleb: a large-scale speaker identification dataset[J]. Arxiv Preprint, 2017, 6: No. 170608612.

[12] Mclaren M, Ferrer L, Castan D, et al. The speakers in the wild (SITW) speaker recognition database[C]// Proceedings of the Interspeech, San Francisco, USA, 2016: 818-822.

[13] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 761-769.

[14] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2980-2988.

[15] Desplanques B, Thienpondt J, Demuynck K, et al. Ecapa-tdnn: emphasized channel attention, propagation and aggregation in tdnn based speaker verification [C] //Interspeech, Shanghai, China, 2020: 3830-3834.

[16] Shen H, Yang Y, Sun G, et al. Improving fairness in speaker verification via group-adapted fusion network[C] //ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022: 7077-7081.