doi:10.3772/j.issn.1002-0470.2017.11 – 12.002

基于卷积神经网络学习的语音情感特征降维方法研究①

薄洪健② 马 琳 孔祥浩 李海峰③

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 针对语音信号认知中需要对语音情感快速精准的解析问题,提出了一种基于卷积神经网络(CNN)学习的特征降维方法。在原始语音情感数据提取大量特征的基础上,通过对不同维度特征进行归正获得其相应的特征矩阵。应用 CNN 对特征矩阵进行学习,对收敛后的 CNN 网络全连接层的权值进行分析,根据网络学习特性定义基于 CNN 的特征筛选准则(FR-CNN),即通过对比每类特征激活权值的不同,计算选择出最有利于分类的特征,得到降维高效的语音情感认知特征集 F。在中国科学院自动化研究所提供的多模态情感数据库 CHEAVD 上,提取全部 8 类情感数据进行了实验测试,使用全体特征集构建的 CNN 分类器的类平均识别错误率相比基线减少了 2.1%,而本文方法得到的降维后特征集 F 通过相同的 CNN 分类器的类平均错误率相比基线减少了 9.4%。在对大量特征进行降维筛选的基础上,仅使用原特征集 15% 的特征,不仅有效增加了分类器的收敛速度,还使得识别错误率有所减小,同时在构筑实际语音情感识别系统时能够减少系统的复杂程度。本研究综合了数据的不同类型的特征信息,采用 CNN 网络学习特性进行特征二次优选与降维,为语音情感的特征提取问题提供了一个新的思路。

关键词 模式识别,语音情感,卷积神经网络(CNN),特征优选准则,特征降维

0 引 言

研究对情感语音信号的快速精准的解析方法,高效地实现从其中提取有效情感特征,是实现计算机听觉认知的重要基础问题。由于不同情感状态下的语音信号呈现的是一种非线性、时变信号,情感信息便包含在这一复杂信号中。因此,提取高效低维语音情感特征是进行情感计算的重中之重。

在 20 世纪 80 年代中期出现了语音情感识别相关研究,并开创了情感分类中使用声学特征的先河[1,2]。1999 年 Moriyama 等人[3]提出了情感和语音之间的线性关联模型,并将此模型运用到实际领域,能够方便采集用户的语音信息并建造出了识别用户情感的图像采集系统,这是语音情感识别在电子商务中的初步应用。2000 年 Cowie 等人[4]开发的 FEELTRACE 情感标注系统,能够为语音情感数据进行标准化的标注。2007 年 Grimm 等人[5,6]在自发语音情感识别的研究中,利用三维情感描述模型对维度情感识别问题进行回归预测。2010 年慕尼黑工业大学的 Eyben、Schuller 等人[7]实现了对语音情感数据进行特征提取的开放式的工具包 openSMILE,能够批量提取常用语音情感特征,得到学术界的广泛认可[8,9]。2015 年 Shahzadi 等人[10]使用非线性动力学特征,并使用遗传算法和支持向量机结合的方法对语音情感进行识别。2016 年 Trigeor-

① 国家自然科学基金(61671187),深圳市基础研究(JCYJ20150929143955341,JCYJ20150625142543470)和语言语音教育部-微软重点实验室开放基金(HIT. KLOF. 2015OXX,HIT. KLOF. 2016OXX)资助项目。

② 男,1984 年生,博士生;研究方向:脑机接口,智能信息处理;E-mail: bohongjian@hit.edu.cn

③ 通信作者,E-mail: lihaifeng@hit.edu.cn

(收稿日期:2017-06-27)

gis 等人[11]使用卷积神经网络(convolutional neural network, CNN)与长短时记忆神经网络进行结合,处理上下文感知的语音情感识别问题。陶建华等[12]以语音的韵律和声学特征为因素对情感语音的合成问题进行了探究。蒋丹宁等人[13]利用概率神经网络和隐马尔可夫模型对声学参数统计特性和时序特性进行处理,通过特征融合提高了识别率。韩文静等人[14]将全局统计特征与基于语段的时序特征相结合,有效解决了基于不同时长的情感特征不能够有效表达情感信息的问题。随着深度学习在各领域内掀起热潮,也有研究者将包括 CNN 在内深度学习方法运用到语音情感识别中,如 Trigeorgis[11]、Badshah[15]等人。

在大多数研究工作中,情感特征都使用混合多种技术提取的声学特征,虽能有效提高情感语音识别效果,但同时大量的冗余信息导致计算复杂度提高,对识别率会有影响。对所提取特征进行有效性分析,从中筛选更具有代表性的特征,能够提高情感语音分析精度。因此,需要对语音情感信号进行降维从而选择有效的特征,但国内外对语音情感信号的降维方面的研究较少,有增量流形学习[16]和双向二维加权[17]等方法也是从数学角度来进行降维,而忽略了对于特征本身有效性的研究。

因为 CNN 具有容错性强、自适应和自学习能力强的特点,能够隐式地从训练数据中对有效的特征模式进行学习,而且网络的权值对特征的有效程度具有一定的指示作用。因此,本文提出了基于 CNN 学习的语音情感降维方法,针对语音情感分类识别中的特征降维问题,对原始语音信号提取大量特征构成全体向量集 F-all(m 维),对 F-all 中的向量进行长度归正得到特征矩阵集,将特征矩阵集输入到 CNN 中进行学习并提取 CNN 全连接层权值矩阵,对权值进行分析并制定筛选准则得到最优特征集 F(m' 维)。

1 卷积神经网络技术原理

卷积神经网络(CNN)是一种多层的神经网络,它能够将原始特征空间直接作为神经网络的输入,在网络的内部进行对特征的提取和优化,能够有效解决特征维度高的分类问题。CNN 中存在局部感知域和权值共享两个特性。局部感知域是指两层之间的神经元的连接是非全连接的,权值共享是指在同一层中某些神经元之间的连接的权重是共享的。这两个机制极大地减少了权值的数量,使网络更加容易收敛。

CNN 中可以定义一组卷积核函数,将每个卷积核与上层数据进行卷积计算,经过激活函数,加上偏置值,便形成了这一层特征图的局部表达。不同的卷积核函数运算,得到该层不同的特征图。

一般情况下可用下面的公式来表示卷积层进行卷积运算的过程:

$$M_j^L = f\left(\sum_{i \in N^{L-1}} M_i^{L-1} \times C_j^L + b_j^L\right) \tag{1}$$

其中,M_j^L 表示第 L 个卷积层中的第 j 个特征图,M^{L-1} 是上一层的特征图集合即这层的输入,C_j^L 为第 L 个卷积层中的第 j 个卷积核,× 表示卷积运算,b 为偏置值。$f(x)$ 为激活函数,通常使用的有 Sigmoid、ReLU 等。其目标函数为整个数据集 D 中损失函数(Loss):

$$Loss = \frac{1}{|D|}\sum_{i}^{|D|} f_L(X^{(i)}) \tag{2}$$

平均值,其中,$f_L(x)$ 计算的是单个样本 $X^{(i)}$ 的 Loss 函数值,计算的方法通常是样本估计值和预测值的欧氏距离的平方的均值。

优化采用的是随机梯度下降,根据计算负梯度 $\nabla L(W)$ 和上一次的权重值 W_t 的线性组合来更新 W,迭代公式如下:

$$W_{t+1} = W_t - \alpha Loss(W_t) \tag{3}$$

其中,α 是学习率,用来调整学习的速度。

2 基于 CNN 的特征筛选准则

2.1 不同信号长度归正方法

有时根据语音情感信号长短的不同,会提取到不同长度的特征,即不同长度的信号帧长不同。对于一个固定的系统来说,则需要把不同长度的特征归正到同一帧长。

设 N 个信号的帧长分别为 $T_1, T_2, T_3, \cdots, T_N$,并

对这 N 个信号的帧长求平均值为 $\bar{T}$, 即每个样本的帧长最终归正为 $\bar{T}$, 则对任意一个信号 i,归正公式如下:

$$T_i'(j) = T_i(\lfloor j \cdot \frac{T_i}{\bar{T}} \rfloor), j = 0,1,\cdots,\bar{T}-1 \quad (4)$$

即新信号 T_i' 的第 j 帧取自原信号 T_i 的第 $\lfloor j \cdot \frac{T_i}{\bar{T}} \rfloor$ 帧。

归正后,便可获得一个维度为 $\boldsymbol{N} \times \bar{\boldsymbol{T}}$ 的特征矩阵,矩阵的每一行代表一种类型的特征,每一列代表一帧。

2.2 CNN 参数设置

本文的分析方法对原始数据提取不同特征并进行长度归正之后,输入到 CNN 网络中进行学习,着重分析其全连接层的权值演变情况,定义特征矩阵进行降维的准则。本文给出了基于 CNN 的语音情感特征降维框架(如图 1 所示),所使用的卷积神经网络结构如图 2 所示。整个 CNN 网络结构由 6 层组成,具体参数如下。

(1) 第一层(Data):输入层。输入为不同维度特征归正后的特征矩阵。

(2) 第二层(C1):卷积层,该层设置 16 个卷积核,卷积核的大小为 1×50,水平卷积步长为 3,垂直卷积步长为 1,使用 ReLU 激活函数。

(3) 第三层(C2):卷积层,主要是对第二层的特征图进行进一步的卷积,提取相应的特征。该层也设置为 16 个卷积核,卷积核的大小为 1×20,水平卷积步长为 3,垂直卷积步长为 1,使用 ReLU 激活函数。

(4) 第四层(C3):卷积层,16 个卷积核,卷积核的大小为 1×10,水平卷积步长为 2,垂直卷积步长为 1,使用 ReLU 激活函数。

(5) 第五层(FC1):全连接层,神经元的个数为 1000 个。

(6) 第六层(FC2):全连接层,共有 8 个神经元,代表了八分类问题。

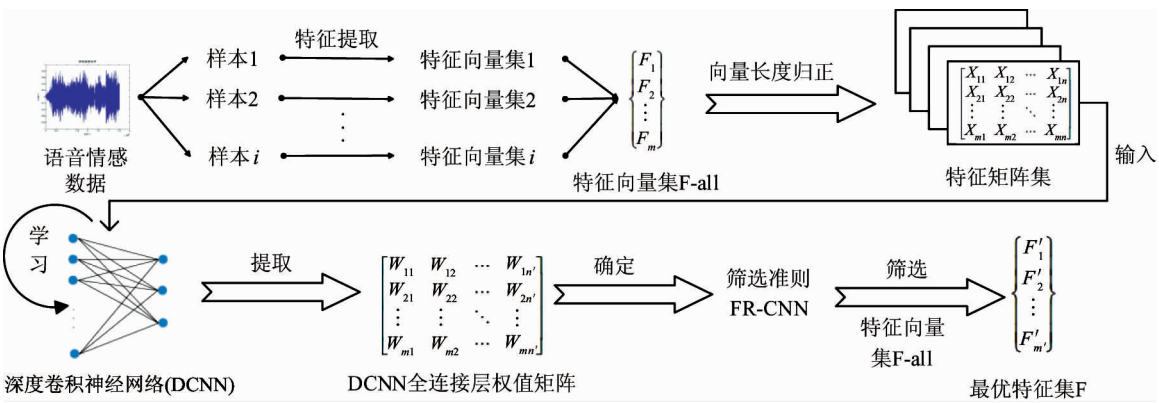

图 1 基于 CNN 的语音情感特征降维框架图

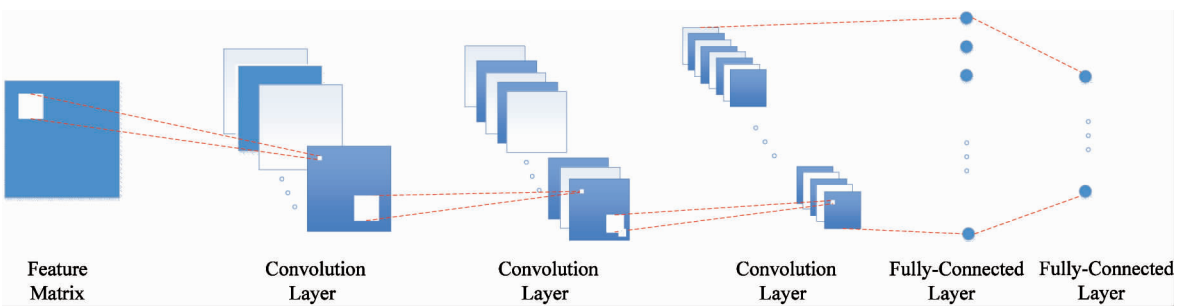

图 2 CNN 网络结构图

网络中各层的具体参数由实验择优确定。

CNN 在学习过程中,有效的特征会对特征分类有积极的效果,从而使与其相关的网络连接权值增大,从而达到分类识别的目的。因此,经过学习后的 CNN 的全连接层权值就包含着对于特征本身的评价信息,这些信息可以用来对特征进行筛选降维。

2.3 基于 CNN 权值矩阵特性的特征筛选准则

从语音情感信号提取的特征矩阵中确定最有效的特征是特征降维的关键。因此,本文提出基于 CNN 的权值矩阵特性的特征筛选准则 FR-CNN。

主要思路为,将特征矩阵输送到 CNN 中进行学习,进而提取其全连接层的权值,通过其权值大小的分布来判断哪种特征对分类更加有效果,然后对比不同类别激活权值的不同,确定对分类最为有效的数种特征,从而达到降维的目的。

具体步骤如下:

首先求得有益加权权值矩阵。设 CNN 中全连接层接受的输入特征图数量为 s, 特征图的维度为 $n \times l$, 且有 t 层全连接层,每一层全连接层分别用 $\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_t$ 表示,且节点数目为 $k_1, k_2, \cdots, k_t$。第 t 层全连接层,有 k_t 个节点,代表 CNN 处理的是 k_t 分类问题。

当 $t=1$ 时,即只有 1 个全连接层,即 $\boldsymbol{W}_1$ (维度为 $[s \times n \times l, k_t]$)可作为有益加权权值矩阵。

当 $t > 1$ 时,为求得最有效的权值矩阵,对每一层,我们以下一层权值为权重对该层权值进行加权,即有如下公式:

$$\boldsymbol{W} = \boldsymbol{W}_1 \times \boldsymbol{W}_2 \times \cdots \times \boldsymbol{W}_t \tag{5}$$

$\boldsymbol{W}_k$ 的维度为 $[s \times n \times l, k_t]$, 即 $\boldsymbol{W}$ 作为有益加权权值矩阵。$\boldsymbol{W}$ 的第 h 列表示对分类为 h 有益的加权权值。

求得有益加权权值矩阵后,对其中任意一列 $\boldsymbol{W}_h$, 维度为 $[s \times n \times l, 1]$, 维度为可以进行重排列得到 $\boldsymbol{W}_h'$:

$$[s \times n \times l, 1] \rightarrow [s, n, l] \tag{6}$$

此时,矩阵 $\boldsymbol{W}_h'$ 可以看成是 s 幅 $n \times l$ 权值图像。对每一幅图像都进行取绝对值操作,即求得一幅图中所有值对平均值的偏差程度,这样会使有效的权值包括正相关权值和负相关权值变为较大值,而无效的权值会趋近于 0。即对第 r 幅图像进行如下操作:

$$\boldsymbol{W}_h''(t,i,j) = \left| \boldsymbol{W}_h'(t,i,j) - \frac{1}{n \times l} \sum_{i=1}^{n} \sum_{j=1}^{l} \boldsymbol{W}_h'(t,i,j) \right| \tag{7}$$

然后求得所有图像的平均,得到

$$\boldsymbol{T}_h(i,j) = \frac{1}{s} \sum_{t=1}^{s} \boldsymbol{W}_h''(t,i,j) \tag{8}$$

对矩阵 $\boldsymbol{T}_h$ 求得每一行的标准差,i 的取值范围是 $(1 \leqslant i \leqslant n)$:

$$\overline{\boldsymbol{T}_h}(i) = \frac{1}{l} \sum_{j=1}^{l} T(i,j) \tag{9}$$

$\overline{\boldsymbol{T}_h}$ 为一个列向量,表示特征矩阵中每一行(即为一个特征)对 h 分类的有效程度,有效程度越大说明这一行的特征对分类为 h 越有益。

同理对所有分类都可以求得 $\overline{\boldsymbol{T}_1}, \cdots, \overline{\boldsymbol{T}_{t_k}}$。进而可以求得特征矩阵第 i 行(即为一个特征)对所有分类有益的程度 $\bar{X}(i)$:

$$\bar{X}(i) = \frac{\sum_{j1=1}^{t_k-1} \sum_{j2=j1+1}^{t_k} | \bar{X}_{j1} - \bar{X}_{j2} |}{\frac{1}{2}t_k^2 - \frac{1}{2}t_k} \tag{10}$$

为了求出最有效的特征,对所有行求偏差,得到列向量 $\boldsymbol{P}$, 公式为

$$\boldsymbol{P}(i) = \left| \frac{\bar{X}(i) - \frac{1}{n} \sum_{i=1}^{n} \bar{X}(i)}{\frac{1}{n} \sum_{i=1}^{n} \bar{X}(i)} \right| \tag{11}$$

可以对 $\boldsymbol{P}$ 进行排序,选取 m' 行,使得每一行都有满足 $\boldsymbol{P}(i) \geqslant \varphi$($\varphi$ 为阈值,本研究认为偏差标准值的0.5即为有效特征,因此取 $\varphi = 0.5$)。对应选择特征矩阵中最有效的 m' 行做为特征集 F。

因此,本文对于特征的筛选准则是首先将全部的特征矩阵输送到一个 CNN 中进行识别,得到收敛后的 CNN 网络权值,运用特征筛选算法,即可从 CNN 网络权值的分布计算得到最优的特征集。

具体如算法 1 所示。

算法 1　基于 CNN 的权值矩阵特性的特征筛选算法

Input:特征矩阵,CNN 全连接层权值
Output:最优特征集 F

1) 根据全连接层权值矩阵求有益加权权值矩阵:

$$\boldsymbol{W} = \boldsymbol{W}_1 \times \boldsymbol{W}_2 \times \cdots \times \boldsymbol{W}_t$$

for $h = 1$ to k_t do

对 $\boldsymbol{W}$ 的一列 $\boldsymbol{W}_h$,维度为 $[s \times n \times l, 1]$, 进行重排列得到 $\boldsymbol{W}_h'$, 维度为 $[s, n, l]$

3) 对 $\boldsymbol{W}_h'(h = 1, \cdots, k_t)$, 进行如下操作:

for $t = 1$ to s do
for $i = 1$ to n do
for $j = 1$ to l do

$$\boldsymbol{W}_h''(t, i, j) = \left| \boldsymbol{W}_h'(t, i, j) - \frac{1}{n \times l} \sum_{i=1}^{n} \sum_{j=1}^{l} \boldsymbol{W}_h'(t, i, j) \right|$$

3) 对 $\boldsymbol{W}_h''(h = 1, \cdots, k_t)$, 对第一个维度求平均,得到:

$$\boldsymbol{T}_h(i, j) = \frac{1}{s} \sum_{t=1}^{s} \boldsymbol{W}_h''(t, i, j)$$

4) 对矩阵 $\boldsymbol{T}_h(h = 1, \cdots, k_t)$, 求得每一行的偏差程度, i 的取值范围是 $(1 \leqslant i \leqslant n)$:

$$\overline{\boldsymbol{T}_h}(i) = \frac{1}{l} \sum_{j=1}^{l} \boldsymbol{T}(i, j)$$

$$\bar{X}(i) = \frac{\sum_{j1=1}^{t_k - 1} \sum_{j2=j1+1}^{t_k} | \bar{T}_{j1} - \bar{T}_{j2} |}{\frac{1}{2} t_k^2 - \frac{1}{2} t_k}$$

$$\boldsymbol{P}(i) = \left| \frac{\bar{X}(i) - \frac{1}{n} \sum_{i=1}^{n} \bar{X}(i)}{\frac{1}{n} \sum_{i=1}^{n} \bar{X}(i)} \right|$$

5) 对 $\boldsymbol{P}$ 进行排序,选取 m' 行,满足 $\boldsymbol{P}(i) \geq \varphi$
并对应选择特征矩阵中 m' 行做为特征集 F

end

3　实验及结果分析

3.1　实验数据集

本文所使用的语音情感数据库是中国科学院自动化研究所提供的多模态情感数据库 CHEAVD (CASIA Chinese Emotional Audio-Visual DATABASE)[18],该数据库的数据来源于影视剧中截取的音视频片段,每个音视频片段标注为常见情感中的一种,属于离散情感数据库。

该数据库有两个特点,一是音视频长度是不固定的,有的非常短,不到 1s,有的长达 40s,本文通过对文件长度的统计,观察到大部分文件的长度主要介于 2 ~5s 之间。二是每种情感的样本个数分布不均匀,其中最多的"neutral"情感共有 815 个样本,而最少的"disgust"情感只有 50 个样本。

本实验使用 CHEAVED 中全部音频数据进行情感识别,分为 8 类情感,分别是 happy、neutral、sad、worried、surprise、angry、disgust、anxious,表 1 为每类情感的具体数目。

3.2　语音情感特征提取

为了能够筛选出更加有效的特征来对情感语音进行识别,不能仅仅使用单纯的 MFCC 特征,因而需要加入新的情感特征。本实验共选择 130 维的情感特征,其中包括 65 维的帧特征和 65 维的统计特征。65 维的帧特征包括 12 维 MFCC 特征,平滑的基音曲线,局部频率微扰,谐噪比,局部振幅微扰等,如表 2所示。而 65 维的全局特征是对 65 维帧特征的平均值。这些特征均通过开源软件 OpenSmile 方便提取到。

表 1 每类情感数目

情感 \ 数据集	训练集	测试集
Happy	307	38
Neutral	725	90
Sad	256	31
Worried	93	11
Surprise	67	8
Angry	399	49
Disgust	45	5
Anxious	89	77

表 2 情感语音特征

特征名称	维度
F0final(平滑的基频)	1
voicingFinalUnclipped(浊音概率)	1
jitterLocal(局部抖动)	1
jitterDDP(帧间抖动)	1
shimmerLocal(局部微扰)	1
logHNR(谐噪比)	1
pcm _ RMSenergy(均方能量)	1
Zcr(过零率)	1
audspec _ lengthL1norm	1
audspecRasta _ lengthL1norm	1
MFCC(梅尔频率倒谱系数)	14
pcm _ fftMag	15
audSpec _ Rfilt	26
总计	65

3.3 特征长度归正

CHEAVED 数据库的一个特点就是文件长度是不固定的,文件长度主要介于 2 ~5s 之间,因此每一个样本的帧的数目是不一样的,但 CNN 网络需要维度一致的输入,因此需要对帧的数目进行归正,也即是对特征的长度进行归正。本文计算了所有样本的平均值,为了使数值具有适应性,本实验选取了比平均值稍大的一个数——700,即把所有的样本的帧数都归正到 700 帧,也即所有的帧特征的长度都归正为 700。而对于统计特征,每一个样本是同一数值,也将其扩展为 700 维。最后得到的每个样本的特征矩阵的大小为 130 ×700。

3.4 计算最优特征集

为了对 130 维的特征进行降维,需要计算最优特征集,因此进行如下操作。

首先从所有情感样本的特征矩阵使用深度卷积神经网络(DCNN)进行学习,等网络收敛后提取其全连接层的权值,分别计算 8 种分类的有益加权权值,即为2.3节中的 $\boldsymbol{W}_k$。将 8 种分类的有益加权权值矩阵进行可视化,得到图 3 所示图像。可视化的方法是将所有权值归一化到[0, 255]之间,便可形成灰度图。其中,每一幅图为 130 ×29 的权值图像,即每一行为 1 种特征,共 130 种,每一种特征经过 3 层卷积后的数目为 29。因此图 3 中纵轴为 130,每类情感横轴为 29。

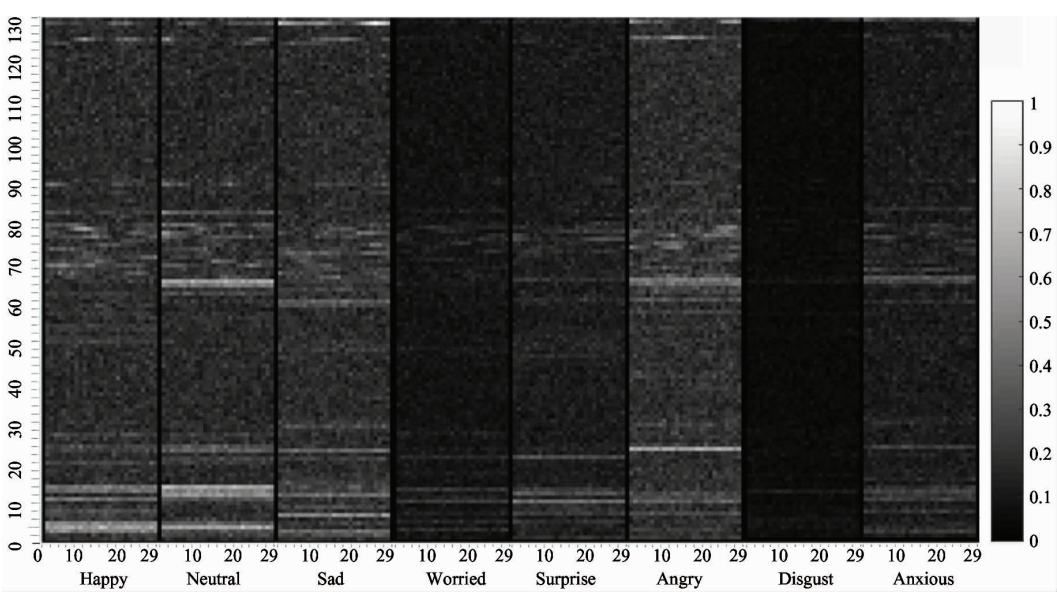

图 3 8 种分类的有益加权权值图像

由图 3 可以观察到,对于不同类别的情感,全连接层所激活的权值是有所区别的。比较亮的权值说明该处权值较大,即对应的特征对分类很有效。对于不同类之间的亮暗程度,我们推测是训练样本数目所导致的。由于在样本集中情感的数目分类不均,且 neutral 最多,因此看起来 neutral 的权值要比其它类亮得多,而 disgust 最少,因此可以看到第 7 类是最暗的。

然后运用基于 DCNN 的权值矩阵特性的行特征筛选算法计算出最优特征集。通过计算,得到了对分类有最佳效果的 20 个特征,将这些特征构成最优特征集。

3.5 最优特征集的识别结果对比实验及结果

为了验证基于本文方法 FR-CNN 筛选的特征的有效性,分别对全体特征集和最优特征集以及随机特征集(随机选择 20 个特征),再次构建 CNN 分类器对特征集进行分类识别。具体所用的分类器参数和 3.2 节所示参数相同。

为了跟 CHEAVE 数据库的基线[18]相比,本文统计了如下两个指标,MAP(macro average precision)和 ACC(accuracy),如下所示:

$$MAP = \frac{1}{s} \times \sum_{i=1}^{s} P_i \tag{12}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{13}$$

$$accuracy = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s} (TP_i + FP_i)} \tag{14}$$

具体的评价指标如准确率、召回率等如表 3 所示。

表 3 全体数据集下 3 种特征集下的评价指标

评价指标 / 情感	随机特征集准确率(%)	随机特征集召回率(%)	全体特征集准确率(%)	全体特征集召回率(%)	最优特征集准确率(%)	最优特征集召回率(%)
Happy	0	0	13.3	5.3	33.3	2.6
Neutral	38	42.2	39.6	88.9	40.4	95.6
Sad	20	3.2	40	9.7	100	3.2
Worried	0	0	0	0	0	0
Surprise	0	0	0	0	25	12.5
Angry	29.5	67.3	64	32.7	66.7	24.5
Disgust	0	0	0	0	0	0
Anxious	0	0	100	9.1	50	18.8
类平均值	10.9	14.1	32.1	18.2	39.4	19.7

可以观察到,8 类情感的分布是不均衡的,而由此训练得到的分类器由表 4 可以观察到更加倾向于数量较多的类。为了避免这种现象,分别从 8 类情感中随机每类选择 50 个构成 400 个样本的集合。再从中随机选 320 个作为训练集,80 个作为测试集,具体识别率如表 3 所示。由表 4 可以观察到,在分类平衡的数据集中,最优特征集在准确率和召回率的表现上要比全体特征集和随机特征集要好。具体结果如表 5 所示。

由以上实验可以看到,基于 CNN 的权值矩阵特性的特征筛选准则 FR-CNN 所筛选出来的最优特征集相比全体特征集以及随机特征集在识别率上有一定程度的提高。这说明本文的方法能够筛选出最佳有效的特征。

3.6 最优特征集所包含特征的分析

为最优特征集所包含的特征如下。

voicingFinalUnclipped(原始浊音概率)、logHNR_sma numeric(谐噪比)、F0final(平滑的基频)全局均值、voicingFinalUnclipped(原始浊音概率)全局均值、jitterDDP(帧间抖动)全局均值、logHNR(谐噪

比)全局均值、pcm _ fftMag _ spectralRollOff90.0(频谱 90.0% 处值)全局均值、pcm _ fftMag _ spectralFlux(频谱通量)全局均值、mfcc(第 1、4、11 个滤波器)、mfcc(第 1、2、3、4、5、8、10、11、12 个滤波器)的全局均值。

表 4 均衡数据集下 3 种特征集下的评价指标

评价指标 / 情感	随机特征集准确率(%)	随机特征集召回率(%)	全体特征集准确率(%)	全体特征集召回率(%)	最优特征集准确率(%)	最优特征集召回率(%)
Happy	30	33.3	33.3	66.7	62.5	45.5
Neutral	20	22.2	18.2	22.2	20	42.9
Sad	38.5	38.5	25	28.8	22.2	33.3
Worried	0	0	55.6	55.6	57.1	30.8
Surprise	40	15.4	33.3	40	44.4	36.4
Angry	0	0	33.3	40	43.8	77.8
Disgust	100	12.5	20	6.3	33.3	25
Anxious	17.6	66.7	66.7	3.8	57.1	36.4
类平均值	26.0	23.6	35.7	32.9	42.6	37.6

表 5 MAP 和 ACC 指标对比

方法 / 指标	MAP	ACC
Baseline	30.0%	41.9%
CNN + 全体特征集(130 维)	32.1%	41.5%
CNN + 随机特征集(20 维)	10.9%	29.6%
CNN + 最优特征集(20 维)	39.4%	42.4%

以上便是对于 8 种分类语音情感识别中最有效的 20 种特征。通过观察可以得到,mfcc 所提取的特征在这 20 维最优特征集中占了绝大多数,说明 mfcc 作为一种优秀的语音情感特征提取方法是非常有效的,这与大量研究人员使用 mfcc 提取语音特征的现状相吻合。另外如原始浊音概率、谐噪比等声学特征对于分类也相当有益。在实际应用中,可以通过提取以上的 20 种特征来实现实际的语音情感识别系统,将大量提取的 130 维特征缩减到 20 维,能够有效降低系统的复杂程度。

4 结 论

在语音情感识别中需要提取大量声学特征,特征的优化对构建实际系统中精准度的提高和复杂度的减小具有重要意义。但传统识别方法通常并没有涉及对特征的分析和优化,因此本文提出了一种基于 CNN 的权值矩阵特性的特征筛选准则 FR-CNN。我们使用 CNN 对大量特征归正后的特征矩阵进行分析,通过对于 CNN 全连接层权值特性的分析,筛选得到高效降维特征集 F。并在中国科学院自动化研究所提供的 CHEAVD 数据库上进行了验证,将 130 维特征缩减到 20 维,不仅没有损失识别率,在类平均识别率方面相比基线还提高了 9.4%,仅使用全体特征集的 15%,却取得了比全体特征集更为优秀的结果,证明了本方法的有效性。另外,还通过对最优特征集所包含特征的分析,证明了 mfcc 作为一种优秀的语音情感特征提取方法是非常有效的,原始浊音概率、谐噪比等声学特征对于分类也相当有益。本文研究针对语音情感识别中特征类型复杂的问题,利用 CNN 网络的学习特性对特征进行二次优选,在对大量特征进行降维筛选的基础上,相比使用全部特征集,使用降维特征集 F 不仅可以增加分类器的收敛速度,降低分类器的收敛难度,同时在构筑实际语音情感识别系统时,还能够减少系统的复杂程度,在保证识别率没有损失的同时,还有进一步提高系统性能的能力。在以后的工作中,可以对卷

积核的权值进行进一步分析,并得到关于特征优化的进一步的结论。

参考文献

[1] Bezooijen R V, Otto S A, Heenan T A. Recognition of vocal expressions of emotion: a three-nation study to identify universal characteristics[J]. *Journal of Cross-Cultural Psychology*, 1983, 14(4):387-406

[2] Tolkmitt F J, Scherer K R. Effect of experimentally induced stress on vocal parameters[J]. *J Exp Psychol Hum Percept Perform*, 1986, 12(3):302-313

[3] Moriyama T, Ozawa S. Emotion recognition and synthesis system on speech[C]. In: Proceedings of IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, 1999. 840-844

[4] Cowie R, Douglas-Cowie E, Savvidou S, et al. FEELTRACE: an instrument for recording perceived emotion in real time[C]. In: Proceedings of ISCA Workshop on Speech and Emotion, Belfast, Ireland, 2000. 19-24

[5] Grimm M, Kroschel K. Evaluation of natural emotions using self assessment manikins[C]. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, San Juan, Puerto Rice, 2005. 381-385

[6] Grimm M, Kroschel K, Narayanan S. Support vector regression for automatic recognition of spontaneous emotions in speech[C]. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, USA, 2007. 1085-1088

[7] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor[C]. In: Proceedings of ACM International Conference on Multimedia, Firenze, Italy, 2010. 1459-1462

[8] Schuller B, Valstar M, Eyben F, et al. AVEC 2011-the first international audio/visual emotion challenge[C]. In: Proceedings of International Conference on Affective Computing and Intelligent Interaction, Memphis, USA, 2011. 415-424

[9] Schuller B, Valstar M, Cowie R, et al. AVEC 2012: the continuous audio/visual emotion challenge - an introduction[C]. In: Proceedings of ACM International Conference on Multimodal Interaction, Santa Monica, USA, 2012. 361-362

[10] Shahzadi A, Ahmadyfard A, Harimi A, et al. Speech emotion recognition using nonlinear dynamics features[J]. *Turkish Journal of Electrical Engineering & Computer Sciences*, 2015, 23(Sup. 1): 2056-2073

[11] Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network[C]. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016. 5200-5204

[12] 陶建华, 董宏辉, 许晓颖. 情感语音合成的关键技术分析[C]. 见:全国现代语音学学术会议, 天津,中国, 2003. 520-527

[13] 蒋丹宁, 蔡莲红. 基于语音声学特征的情感信息识别[J]. 清华大学学报自然科学版, 2006, 46(1):86-89

[14] 韩文静, 李海峰, 韩纪庆. 基于长短时特征融合的语音情感识别方法[J]. 清华大学学报自然科学版, 2008(s1):98-104

[15] Badshah A M, Ahmad J, Rahim N, et al. Speech emotion recognition from spectrograms with deep convolutional neural network[C]. In: Proceedings of 2017 IEEE International Conference on Platform Technology and Service, Busan, Korea, 2017. 1-5

[16] 王海鹤, 陆捷荣, 詹永照,等. 基于增量流形学习的语音情感特征降维方法[J]. 计算机工程, 2011, 37(12):144-146

[17] 齐晓倩, 陈鸿昶, 黄海. 双向二维加权 LPP 语音特征降维算法[J]. 小型微型计算机系统, 2012, 33(7): 1588-1591

[18] Li Y, Tao J, Schuller B, et al. MEC 2016: The multimodal emotion recognition challenge of CCPR 2016[J]. *Pattern Recognition*, 2016:667-678

Research on a dimension reduction method of speech emotional feature based on convolution neural network

Bo Hongjian, Ma Lin, Kong Xianghao, Li Haifeng

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract

A feature reduction method based on convolution neural network (CNN) is proposed to solve the problem of speech emotion recognition. On the basis of extracting a large number of features of the original speech emotion data, the corresponding feature matrix is obtained by normalizing the different dimension features. The CNN is used to study the feature matrix, and the weights of the CNN network are analyzed. According to the characteristics of the network learning feature, that is, by comparing the activation weights of each class, the features that are most favorable for classification are selected by calculation, so the feature selection criterion FR-CNN is obtained. The multi-modal emotional database CHEAVD provided by the Institute of Automation of Chinese Academy of Sciences is used to test all the eight kinds of emotional data, showing that the average recognition error rate of the CNN classifier constructed with all the feature sets is reduced by 2.1% compared to the baseline results, while the average recognition error rate of the same CNN classifier constructed with dimension reduction F feature set is reduced by 9.4%. In addition, using only 15% of original feature set's features on the basis of dimensional reduction of a large number of features, can not only effectively increase the convergence speed of the classifier, but also make the recognition error rate reduced, at the same time in the actual speech emotion recognition system, the complexity of system can also be reduced. The study provides a new idea for the feature extraction of speech emotion.

Key words: pattern recognition, speech emotion, convolutional neural network (CNN), feature selection criterion, feature reduction