文章编号：1004-4353(2020)01-0049-06

基于特征降维及参数优化的语音情感识别

俞颖[1,2]， 黄凤华[1,2]， 刘永芬[3]
（1.阳光学院 空间数据挖掘与应用福建省高校工程研究中心；2.阳光学院 人工智能学院；
3.福建农林大学 金山学院：福建 福州 350001）

摘要：针对传统BP神经网络在语音情感识别过程中存在的计算量偏大和容易陷入局部最优解的问题，提出了一种基于特征降维及参数优化的情感识别改进方法.首先提取情感语料库的高维度联合特征，利用快速主成份分析法(Fast_PAC)进行特征降维以达到降低问题复杂性的目的；然后引入遗传算法对BP神经网络进行参数优化以避免限入局部最优问题；最后构建语音情感识别分类器，并利用CASIA汉语语料库及柏林德语语料库进行情感识别验证.验证结果表明，与传统的支持向量机(SVM)方法、传统的主成份分析法(PCA算法)结合SVM模型识别方法相比，本文方法能有效地降低语音情感的特征维度，且具有运算量少和识别精度高的优点.

关键词：快速主成份分析法；遗传算法；BP神经网络；语音情感识别
中图分类号：TP301.6 **文献标志码**：A

Speech emotion recognition based on feature dimension reduction and parameter optimization

YU Ying[1,2]， HUANG Fenghua[1,2]， LIU Yongfen[3]
（1. *Spatial Data Mining and Application Research Center of Fujian Province, Yango University*；
2. *Artificial Intelligence College, Yango University*；
3. *Jinshan College, Fujian Agriculture and Forestry University: Fuzhou* 350001, *China*）

Abstract: The traditional BP neural network has been existing some burning questions in the process of speech emotion recognition, especially the high computational and local optimum trending. Against these shortcomings, we present a novel method of emotion recognition based on feature dimension reduction and parameter optimization. The recognition method is divided into three stages. In the first stage, it extracts the high-dimensional joint features of the speech emotion database. This is, in fact, aimingto reduce the complexity of the problem which is carried out by the fast principal component analysis (Fast_PAC) method. In the second stage, genetic algorithm is used to optimize the parameters of BP neural network to avoid the local optimum problem. Finally, we construct a speech emotion recognition classifier, and take the experiments on the CASIA Chinese corpus and Berlin German corpus for emotion recognition verification. The experiments show that the proposed method can effectively reduce the feature dimension of speech emotion comparing with other competitive methods, such as the traditional support vector machine (SVM) method and the traditional PCA combined with SVM model recognition method. Furthermore, it demonstrates the advantages of less computation and higher recognition accuracy.

Keywords: Fast_PAC; genetic algorithm; BP neural network; speech emotion recognition

收稿日期：2019-12-26 **作者简介**：俞颖(1984—)，女，讲师，研究方向为数据挖掘、模式识别.
基金项目：福建省教育厅中青年教师科研项目(JAT190977)；福建省自然科学基金资助项目(2019J01088)

0 引言

语音是人际交流的重要媒介.语音信号中不仅包含所要传递的语义信息,还包含丰富的情感信息,因此如何使计算机从语音信号中自动识别出说话人的情感状态及其变化,是实现自然人机交互技术的关键前提.目前,语音情感识别存在两大难点:一是如何寻找有效的语音情感特征,二是如何构造合适的语音情感识别模型[1].研究[2-3]显示,单一特征情感识别的效果并不理想,因此学者们更多的是采用多特征联合的方法来识别语音情感;但采用多特征联合的方法易使情感特征的维数偏高,进而增加计算的复杂度.近年来,支持向量机(SVM)和人工神经网络(ANN)模型被广泛应用于语音情感识别.例如:文献[4]通过构造多个SVM分类器进行语音感情识别,该方法虽然提高了语音情感识别率,但因所构造的SVM分类器较多使得识别过程较为复杂;文献[5]提出了一种将传统的主成份分析法(PCA算法)和SVM分类器相结合的语音情感识别方法,该方法可有效降低语音情感识别的计算量,但传统的PCA算法在降维过程中需要较高的时间耗费;文献[6]采用改进遗传算法优化BP神经网络来进行语音情感识别,该方法的语音情感识别率较高,但识别过程中所用的特征维数较高,增加了语音情感识别的计算量.基于上述研究,本文利用快速主成份分析法(Fast_PCA算法)和优化后的BP神经网络提出一种新的语音情感识别方法,并通过实验验证本文方法的有效性.

1 语音情感特征提取及Fast_PAC降维

1.1 语音情感特征参数的提取

常用的语音情感特征主要包含韵律学特征、基于谱的特征和音质特征.语音处理的特征参数通常是以帧为单位提取的,但由于单帧信号所含的信息量较少,因此用于情感识别的特征参数多采用连续多帧的提取特征值,然后通过计算这些特征值的统计量来组合情感识别的特征参数.基于中文与西方语种在语音和情感表达上存在的差异[5],本文将中西方语种语音信号中的基音频率、短时能量、短时幅值、短时平均过零率、共振峰、语音持续时间及梅尔频率倒谱系数(MFCC)作为原始的语音情感特征,并通过计算这7类原始语音的情感特征值及其一阶差分、二阶差分的统计值(统计值主要包括最大值、最小值、均值、中值、标准差、方差等)来获取语音信号的高维度联合特征.

1) 基音频率.基音频率(简称基频)是指发浊音时声带产生的周期性的振动频率,它能够反映声道的特征.一般来说,男性的基频较低,女性的基频较高,且不同情感状态下基频的大小不同[7].

2) 短时能量.短时能量是指每帧信号的短时平均能量,它反映的是语音的能量或语音振幅随时间缓慢变化的规律[8].设 $x(l)$ 为语音时域信号,N 为每帧的长度,$w(m)$ 为窗函数,$x_n(m)$ 为加窗分帧处理后的第 n 帧语音信号.定义 $x_n(m)=w(m)x(n+m)$,则短时能量谱 E_n 的计算公式[5] 为

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

3) 短时幅值.短时幅值也是度量语音信号能量大小的一个指标,它与短时能量的区别在于计算时无论取何采样值,都不会因为对语音信号值取二次方而造成分帧之间的能量值有较大差异.短时幅值 M_n 的计算公式[5] 为

$$M_n=\sum_{m=0}^{N-1}w(m)\,|x(n+m)|. \tag{2}$$

4) 短时平均过零率.短时平均过零率是指每帧语音信号在零值上下所波动的次数.浊音具有较低的过零率,清音具有较高的过零率,利用短时平均过零率可以从背景噪声中找出语音信号并判断出语音的起点和终点[9].

5) 共振峰.共振峰是声源通过声道时产生的一组共振频率.当人处在不同的神经紧张程度下,声道发生形变,共振频率也发生改变[10].本文利用线性预测法提取语音信号中的共振峰频率,并计算第1至

第 3 共振峰的相关统计特性. 计算所得的相关统计特性作为语音信号的特征参数.

6) 语音持续时间. 语音持续时间是指情感发音的持续时间. 因欢快、愤怒和惊奇的发音长度相对较短，而悲伤的语音持续时间相对较长，因此可以利用语音的时间构造来进行情感区分.

7) MFCC 系数. MFCC 系数反映的是人的感知能力与语音信号的频率之间存在的特定关系. MFCC 系数的计算以 Mel 频率为基准，其计算表达式[11] 为

$$\mathrm{mel}(f)=2\,595\times\log_{10}(1+f/700), \tag{3}$$

其中 f 是语音频率.

1.2　Fast_PCA 算法的特征参数降维

研究表明，利用 PCA 算法中的线性变换可将高维空间中的样本数据投影到低维空间中，从而达到特征降维的目的[12]；但传统的 PCA 算法在特征降维过程中需要对样本的协方差矩阵进行本征值和本征向量的求解，计算量较大. 快速主成份分析法[13] 是 PCA 算法的一种改进方法，该方法在特征降维过程中能够通过求解低维度的协方差转置矩阵的本征向量值及本征值来代替求解高维度协方差矩阵本征向量值及本征值，因此可实现语音情感特征的高效降维.

设 $\boldsymbol{D}$ 是构成语音情感特征向量的样本矩阵，$\boldsymbol{D}\in\mathbf{R}^{n\times m}$，其中 n 为语音样本数量，m 为语音样本特征维数. 设 mA 为样本均值，k 为降维的维数. 则 Fast_PCA 算法降维的具体步骤可描述为：

Step 1　将 $\boldsymbol{D}$ 矩阵中的每个样本减去 mA，得到中心化样本矩阵 $\boldsymbol{Z}_{n\times m}$.

Step 2　计算协方差转置矩阵 $\boldsymbol{T}$，$\boldsymbol{T}=\boldsymbol{Z}\times\boldsymbol{Z}^{\mathrm{T}}$.

Step 3　计算协方差转置矩阵 $\boldsymbol{T}$ 的最大 k 个特征值和特征向量 $\boldsymbol{V}_1$.

Step 4　对特征向量 $\boldsymbol{V}_1$ 左乘 $\boldsymbol{Z}^{\mathrm{T}}$，得到协方差矩阵的特征向量 $\boldsymbol{V}$，$\boldsymbol{V}=\boldsymbol{Z}^{\mathrm{T}}\times\boldsymbol{V}_1$.

Step 5　对 $\boldsymbol{V}$ 进行归一化处理.

Step 6　计算 $\boldsymbol{Z}\times\boldsymbol{V}$，将特征向量线性降维到 k 维空间.

2　BP 神经网络参数优化

2.1　BP 神经网络原理

BP 神经网络具有较强的非线性映射能力，其能够通过学习自适应地更新神经网络的权值来逼近求解问题的最优解，因而被广泛应用于图像分类、语音识别等领域. BP 神经网络属于多层前馈神经网络，包含 1 个输入层、多个隐含层和 1 个输出层，层与层之间采用全连接方式，其最大的优点是可以通过训练样本反向传播调节网络的权值和阈值来实现网络的误差平方和最小的目的[14]. 3 层 BP 神经网络结构如图 1 所示. 图 1 中，$x_1,x_2,\cdots,x_n$ 为 BP 神经网络的输入信号值，$y_1,y_2,\cdots,y_m$ 为 BP 神经网络的输出信号值. 尽管 BP 神经网络具有很强的自学习和自适应能力，但其仍存在一些不足之处，如网络的权值及阈值是随机初始化的，网络的收敛速度较慢，当网络中存在多个极小值时问题的解容易陷入局部最优解.

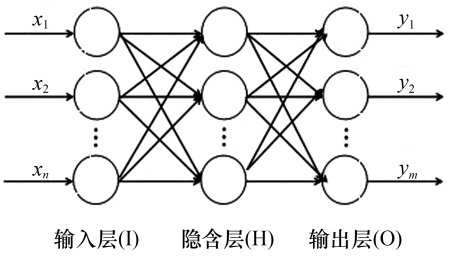

图 1　3 层 BP 神经网络的结构

2.2　遗传算法优化 BP 神经网络

为了克服 BP 神经网络自身存在的缺陷，本文采用遗传算法对 BP 神经网络的权值和阈值进行全局优化搜索，通过训练、搭建语音情感分析 BP 网络模型来提高语音情感识别的精度. 利用遗传算法优化 BP 神经网络的具体步骤如下：

Step 1　初始化 BP 神经网络，确定网络的输入层、隐含层及输出层，产生网络的初始权值和阈值.

Step 2　设置遗传算法的进化迭代次数、初始化种群规模、选择交叉概率和变异概率等参数.

Step 3 随机产生一个种群,并进行染色体编码;计算 BP 网络误差,确定染色体的适应度值.

Step 4 对种群进行遗传迭代,根据个体适应度选择染色体并进行交叉和变异,由此产生一个新的种群.

Step 5 计算新种群的适应度,并更新该种群的染色体.

Step 6 判断是否满足退出条件,如果是则可获得最优 BP 神经网络的权值和阀值,转 Step 7;否则转 Step 4,继续迭代.

Step 7 更新 BP 神经网络的权值和阈值,生成优化 BP 神经网络模型.

3 算法流程

本文提出的语音情感识别改进方法的具体工作流程如图 2 所示,具体操作步骤为:

1)对语音情感语料库进行预处理.首先通过分析语音情感语料库的特征为语料库中的语音数据添加识别标签,然后对语音数据进行特征提取、特征联合以及归一化处理.

2)建立训练集 D_1 和测试集 D_2.首先利用 Fast_PCA 算法计算语音特征参数的主成份分量并分析其对语音特征的贡献度,然后通过确定有效的特征维数将语音特征集划分为训练集 D_1 和测试集 D_2.

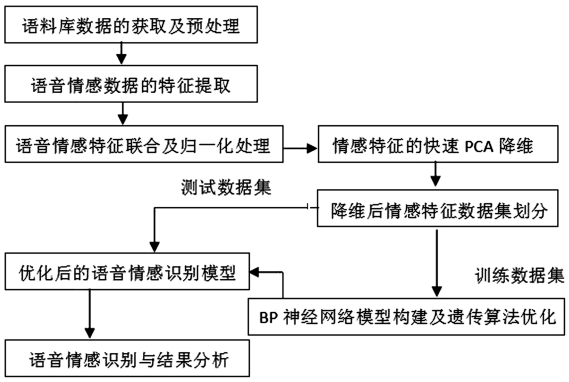

图 2 改进的语音情感识别方法的流程图

3)建立语音情感识别模型.首先利用训练集 D_1 对 BP 神经网络进行训练,并采用遗传算法对网络模型的参数进行优化;然后利用迭代动态调节神经网络权值(阈值)获得最优的语音情感识别模型.

4)分析语音情感的识别性能.首先利用测试集 D_2 对建立的最优语音情感识别模型进行验证,然后计算情感识别精度并进行精度分析.

4 算法的实验验证

4.1 实验环境及数据

算法在 MatlabR2014a 环境下实现编程,计算机的配置为:Intel(R) i5-4570R,8G 内存,Windows7.为了更好地进行语音情感识别效果对比,分别选择国内具有代表性的中科院自动化所模式识别实验室提供的 CASIA 汉语情感语料库[15]和柏林工业大学提供的德语情感语料库[16]进行语音情感识别验证.这两个情感语料均在无噪声环境下获取,采样率为 16 kHz,采样精度为 16 bit.情感语料库的基本信息如表 1 所示.

表 1 语料库信息表

语料库	情感种类	情感类别名称	情感语料数量
CASIA 汉语情感语料库	6 类	angry(愤怒)、fear(恐惧)、happy(快乐)、sad(悲伤)、surprise(惊讶)、neutral(中立)	1 200
柏林德语情感语料库	7 类	fear(恐惧)、disgust(厌恶)、happy(快乐)、boredom(无聊)、neutral(中立)、sadness(悲伤)、anger(愤怒)	535

4.2 实验及结果分析

在 CASIA 汉语情感语料库、柏林德语情感语料库中提取每句语料的 7 类原始语音情感特征(基音频率、短时能量、短时幅值、短时平均过零率、共振峰、语音持续时间及 MFCC),然后计算这 7 类原始特

征的特征值及其一阶差分、二阶差分的统计值. 根据计算所得结果，将其组合成 186 维的语音情感联合特征，用以表示每句语料的情感信息.

为了验证本文所提出的语音情感识别改进方法对语音情感特征的降维效果，采用 Fast_PCA 算法分别对 2 个语料库的特征参数进行降维处理. 图 3 和图 4 为 2 个语料库降维后的前 10 维主成份分量对原始语料信息的贡献比例，表 2 为 2 个语料库在不同降维处理时所耗费的时间及对原始语料信息的贡献比例. 从图 3 和图 4 可以看出，第 1 维到第 10 维对原始语料信息的贡献比例呈逐渐降低的趋势，其中第 1 维对原始语料信息的贡献比例分别为 36.87％和 28.29％，第 2 维对原始语料信息的贡献比例均为 15％左右，第 10 维对原始语料信息的贡献比例均低于 5％以下. 这表明经过 Fast_PCA 算法特征降维后，对原始语料信息的贡献程度起主要作用的主成份分量集中在低维区. 从表 2 可以看出，增加维数时降维时间虽呈增加趋势，但 CASIA 汉语情感语料库和柏林德语情感语料库的特征降维时间分别均低于 0.1 s 和 0.2 s；当语料情感特征维度降维至 35 维时，其对原始语料信息的累计贡献比例已经超过 95％. 上述结果表明，采用 Fast_PCA 算法的降维效果较好.

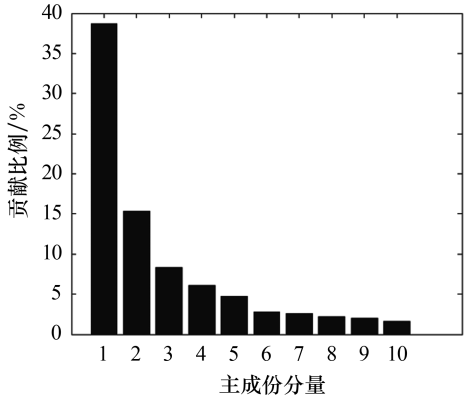

图 3　降维后 CASIA 汉语情感语料库的前 10 维主成份分量对原始语料信息的贡献比例

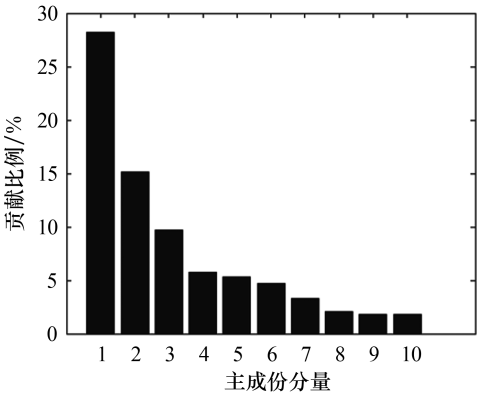

图 4　降维后柏林德语情感语料库的前 10 维主成份分量对原始语料信息的贡献比例

表 2　不同维数的降维处理时间及对原始语料信息的贡献比例

降维维数	CASIA 汉语情感语料库		柏林德语情感语料库	
	降维时间/s	贡献比例/％	降维时间/s	贡献比例/％
1 维	0.031	36.87	0.124	28.19
5 维	0.046	73.19	0.127	64.24
15 维	0.049	87.45	0.132	84.77
25 维	0.050	93.72	0.136	92.23
35 维	0.062	98.20	0.141	96.31
45 维	0.067	98.38	0.141	98.40
55 维	0.068	99.01	0.175	98.77
65 维	0.076	99.05	0.158	99.89

为了进一步验证本文方法对语音情感识别的有效性，将本文方法与传统的无特征降维的 SVM 情感识别方法(SVM)、文献[4]方法(PCA＋多级 SVM)及文献[5]方法(PCA＋SVM)进行对比. BP 神经网络和遗传算法的相关参数设置如下：BP 神经网络的输出层采用二进制进行识别，CASIA 汉语情感语料库识别网络的输出层节点数为 6，柏林德语情感语料库识别网络的输出层节点数为 7；BP 神经网络的最大迭代次数为 2 000，学习率为 0.01，目标精度为 0.001；遗传算法的初始种群规模为 30，交叉概率为 0.3.

表 3 为基于 CASIA 汉语情感语料库(其中 50％用于训练集，50％用于测试集)的不同方法的语音

情感识别效果.表4为基于柏林德语情感语料库(其中70%用于训练集,30%用于测试集)的不同方法的语音情感识别效果.由表3和表4可以看出,本文方法的语音情感平均识别率显著优于其他3种方法(SVM、PCA+SVM和PCA+多级SVM).

表3 不同方法对CASIA汉语情感语料库中的语音数据进行情感识别的结果 %

方法	情感特征维数	各类语音的情感识别率						平均识别率
		angry	fear	happy	sad	surprise	neutral	
SVM	68维	69.0	66.0	57.0	47.0	60.0	55.0	59.0
PCA+SVM	28维	73.5	64.5	61.5	66.5	70.5	76.0	68.8
本文方法	25维	84.9	76.0	73.6	83.3	73.0	81.5	78.7

表4 不同方法对柏林德语情感语料库中的语音数据进行情感识别的结果 %

方法	情感特征维数	各类语音的情感识别率							平均识别率
		fear	disgust	happy	boredom	neutral	sadness	anger	
PCA+多级SVM	30维	71.0	60.9	52.1	55.6	44.3	69.3	81.9	63.7
本文方法	20维	73.6	66.6	71.4	77.4	68.9	81.8	75.7	73.6

5 结束语

研究表明,与传统的SVM情感识别方法、PCA+SVM方法及PCA+多级SVM方法相比,本文提出的基于Fast_PCA算法的快速降维及遗传算法参数优化的BP神经网络语音情感识别方法,不仅能够以较低的时间代价实现特征维数降维,有效克服局部最优问题,而且情感识别的平均精度显著优于上述3种方法,因此本文方法具有很好的实用价值.本文所采用的语音情感语料库均是在无噪声条件下提取的,而在实际中语音信号的提取往往会受到背景噪声的影响,因此今后我们将进一步研究噪声环境下的语音情感识别算法.

参考文献:

[1] 王富,孙林慧,苏敏,等.基于参数寻优决策树SVM的语音情感识别[J].计算机技术与发展,2018,28(7):63-65.

[2] ZHU J C, LIU Z L. Analysis of hybrid feature research based on extraction LPCC and MFCC[C]//Tenth International Conference on Computational Intelligence and Security. Kunming: IEEE, 2014:732-735.

[3] 李高玲,帖云,齐林.基于随机森林分类优化的多特征语音情感识别[J].微电子学与计算机,2019,36(1):70-73.

[4] 任浩,叶亮,李月,等.基于多级SVM分类的语音情感识别算法[J].计算机应用研究,2017,34(6):1682-1684.

[5] 蒋海华,胡斌.基于PCA和SVM的普通话语音情感识别[J].计算机科学,2015,42(11):270-272.

[6] 陈闯,Ryad Chellali,邢尹.改进遗传算法优化BP神经网络的语音情感识别[J].计算机应用研究,2019,36(2):344-345.

[7] 徐照松,元建.基于BP神经网络的语音情感识别研究[J].软件导刊,2014,13(4):11-12.

[8] 崔星星,苏智剑.一种新呼吸音信号特征提取方法与应用[J].中国医学物理学杂志,2018,35(2):214-218.

[9] 刘晨轩,蓝贤桂.语音信号短时分析算法研究与实现[J].价值工程,2012,12:191-192.

[10] 李强,刘晓峰,贺静.基于语音特征的情感分类[J].小型微型计算机系统,2016,37(2):385-387.

[11] 周萍,沈昊,郑凯鹏.基于MFCC与GFCC混合特征参数的说话人识别[J].应用科学学报,2019,37(1):24-32.

[12] 廖周宇,王钰婷,谢晓兰,等.基于粒子群优化的支持向量机人脸识别[J].计算机工程,2017,43(12):248-250.

[13] MITTAL N, WALIA E. Face recognition using improved fast PCA algorithm[C]// Proceedings of the 2008 Congress on Image and Signal Processing. Sanya: IEEE Computer Society, 2008:554-558.

[14] 杨怡涵,柳炳祥.一种基于遗传算法优化BP神经网络的陶瓷原料分类方法[J].陶瓷学报,2018,39(3):340-342.

[15] Institute of Automation, Chinese Academy of Sciences. CASIA Mandarin emotional corpus[DB/OL]. [2019-10-12]. http://www.chineseldc.org/resource_info.php? rid=76Casis.

[16] BURKHARDT F, PAESCHKE A, ROLFES M, et al. Adaabase of German emotional speech[C]//Proc of Interspeech. Lisbon: ISCA, 2005:1517-1520.