

文章编号: 2096-1618(2019)05-0495-06

基于改进 BP-Adaboost 和 HMM 混合模型的方言情感识别

冀常鹏¹, 程琳², 李锋²

(1. 辽宁工程技术大学电子与信息工程学院, 辽宁 葫芦岛 125105; 2. 辽宁工程技术大学研究生院, 辽宁 葫芦岛 125105)

摘要:针对方言情感数据库资源匮乏以及如何提高传统模型的识别准确率的问题,建立了扬泰方言情感数据库,并提出一种自适应变异粒子群优化 BP-Adaboost 神经网络和隐马尔科夫(HMM)相结合的模型。首先采用基于相关性的特征选择(CFS)提取最优的方言情感特征;之后通过 HMM 得到最优状态序列并进行时间规整,最后将特征矢量输入自适应变异粒子群优化的 BP-Adaboost 神经网络进行情感识别。在扬泰方言情感数据库上的情感识别率达到了 86.18%,结果表明,该模型取得了较好的情感识别效果。

关键词:电子与通信工程;方言情感识别;HMM;BP 神经网络;CFS

中图分类号:TP391.4

文献标志码:A

doi:10.16836/j.cnki.jcuit.2019.05.010

0 引言

近年来,语音情感识别已经成为人工智能的一个研究热点。现今,方言情感识别作为语音情感识别的一个重要分支,已经逐渐受到人们的重视。开展方言情感识别研究是一项很有意义的工作。科大讯飞研发了深度全序列卷积神经网络(deep fully convolutional neural network,DFCNN)语音识别框架,使方言识别准确率显著提升。2017年,科大讯飞公布了《1024计划》,1024万条方言将被收集并打造中国方言库。科大讯飞在方言识别上有着突出的成就,但方言情感识别的研究在国内还处于起步阶段,情感数据库是情感识别研究的第一步,是决定识别率高低的一个重要因素。在方言情感识别中,方言情感数据库的缺乏导致了方言情感识别研究无法开展,所以文中参考 CASIA 汉语语音数据库,依据国际标准语音库的制作标准建立了一个扬泰方言情感数据库。

2018年7月18日,中国国家863计划中,多方言的高表现力情感语音交互系统建立了融合语音韵律信息和内容信息的情感识别模型,对愉悦、悲伤等情感状态的识别准确率超过85%,能够支撑较高性能的情感分析并应用于医学抑郁症患者诊治、人员情绪监控等多种场合。系统中情感特征的选择较传统模型有很大的突破,但该系统情感类别有限并且情感识别率还有待提高。如 Schuller B 等^[1]采用 HMM 进行语音情

感识别,HMM 有着极强建模能力,但 HMM 的分类决策能力较弱。Jiao C 等^[2]采用神经网络进行情感识别,神经网络的并行处理能力较强,但神经网络的建模能力较弱以及容易陷入局部最优。针对这些问题,学者们提出了混合模型^[3-6],将 HMM 对动态时间序列的极强建模能力和神经网络的较强并行处理能力相结合,取长补短,有效提升识别率,但神经网络还有陷入局部最优的问题。Hua Y 等^[7]采用粒子群优化神经网络,成功解决陷入局部最优问题。所以文中将自适应变异粒子群优化的 BP 神经网络与 Adaboost 相结合,提出 AMPSO-BP-Adaboost 和 HMM 的混合模型(AMPSO-ABP-HMM)。

1 情感数据库的建立

国际上的情感数据库都参照严格的评价、制作和分发规范制作。此次扬泰方言情感数据库参考中科院自动化研究所建立的 CASIA 汉语语音数据库进行收集和制作^[8]。

1.1 发音人

发音人选择生活在扬泰地区,当地方言口音纯正,且日常生活均是以扬泰方言交流的本地住户。实验选取了符合要求的10个人作为扬泰方言发音人。

1.2 录音语料

扬泰方言^[9-10]的特点是保留入声。同时其声母浊

音清化,古全浊声母逢塞音、塞擦音,而且不论平仄一律读送气念清音。根据数据库中语音文本的选择标准且要具有以上的扬泰方言特色,首先句子不能有明确的语义指向,并具有较高的情感自由度;其次发音时间控制在5 s以下。根据准则选取 50 条语句作为语音文本,部分发音见表 1^[11]。

表 1 数据库文本语句		
序号	扬泰方言发音	汉语含义对照
1	wu ma sang na lei	我马上拿来
2	gou hua lei wu cen	国华来完成
3	ta lei zi ci zong gao	他们支持中国
4	su lin tei piao si jie	苏联代表世界
5	cen lin xi hong mei sa	程琳喜欢美食
6	qi tong jin xing gai ga	集团进行改革
7	wu lei xu yao pang cu	我们需要帮助
8	gou jia gai ga qi yi	国家改革企业
9	bo jing zao kai ao yun	北京召开奥运
10	xiu hua dou dei da ga kai xin	笑话逗得大家开心

1.3 录音设备及语音采集

采用 TASCAM DR-05 标准配置录音笔和麦克风用于录制,48 kHz采样频率,16 bit的采样精度,单声道采样。实验地点选择在一个安静的房间。录音之前,安排发音人模拟录音,调动情绪后立即进行语音采集。录音时,发音人会分别用 5 种情感进行表达规定的 50 条语句。

1.4 数据存储及语料库标注规范

录制后保存的.wav文件用 Praat 软件进行标注,标注包括语音音段和韵律标签。每个语音文件都会用文本、情感、姓名 3 种信息进行标注整理,标注系统采用中国社会科学院语言所语音室的 C-ToBI3.0 和 SAM-PA-C 标注规范。

1.5 语音情感数据库评价规范

语音正式采集之前进行试录音,语音采集之后选择 10 个人作为裁判对录音进行情感听辨检测。可信度大于 0.75 的语料作为实验数据。最终建立的语音库中包含了 10×50×5 共 2500 条语句。每种情感语料的组成如表 2 所示,其中 1500 个为训练集,500 个为测试集。

表 2 各种情感语料组成					
情感分类	悲伤	喜悦	愤怒	平静	恐惧
语句数量	500	500	500	500	500

2 方言情感识别算法模型

2.1 自适应变异粒子群算法

粒子群算法 (particle swarm optimization, PSO) 是一种易于实现的、高精度的、快速收敛的进化算法。由于 PSO 算法结构相对简单,因此运算速度较快,但在寻优过程中容易陷入局部最优值。文中在 PSO 算法中加入了自适应变异 (adaptive mutation, AM) 过程来解决该问题^[12]。AMPSO 算法的速度、位置更新公式为

$$v_i(k+1) = \omega v_i(k) + c_1 r_1 [p_i^{best}(k) - x_i(k)] + c_2 r_2 [g^{best}(k) - x_i(k)] \tag{1}$$

$$x_i(k+1) = x_i(k) + v_i(k+1) \quad i = 1, 2, \dots, m \tag{2}$$

其中: k 为迭代次数; $v_i(k)$ 是粒子在 k 次迭代的速度; $x_i(k)$ 是粒子当前位置; 加速因子 r_1, r_2 为 $[0, 1]$ 的随机数, 实验中取 $r_1 = r_2 = 0.5$; 学习因子 c_1, c_2 为非负常数, $c_1 = 1.4, c_2 = 1.0$; $p_i^{best}(k)$ 为 i 个粒子在 k 次迭代中的最优位置; $g^{best}(k)$ 是 k 次迭代后粒子群的最佳位置。

惯性权重 ω 取值在 0.1 ~ 0.9, 实验中取 $\omega = 0.55$ 。若随迭代的进行, ω 的值线性减小, 则算法的收敛能力将会大大提高。设惯性权重的最大值 ω_{max} , 最小值 ω_{min} , t 为当前迭代次数, 迭代总次数 t_{max} , 则有:

$$\omega = \omega_{max} - t \times \frac{\omega_{max} - \omega_{min}}{t_{max}} \tag{3}$$

假设粒子数目为 n , 第 i 个粒子的适应度为 f_i , 粒子当前平均适应度为 f_{avg} , 粒子群的群体适应度方差为 σ^2 , 则有:

$$\sigma^2 = \sum_{i=1}^n \left[\frac{f_i - f_{avg}}{f} \right]^2 \tag{4}$$

其中 f 是归一化定标因子, 作用是限制 σ^2 的大小。 f 随算法的变化而变化, 那么

$$f = \begin{cases} \max \{ |f_i - f_{avg}| \} & \max \{ |f_i - f_{avg}| \} > 1 \\ 1 & \text{其他} \end{cases} \tag{5}$$

σ^2 反映粒子的收敛度, 当 $\sigma^2 = 0$ 时并不能说明粒子找到全局最优位置, 而是算法陷入早熟。此时 g^{best} 需要通过变异操作跳出局部区域, 寻找全局最优解。将 g^{best} 按照一定概率 p_m 变异。计算公式如下:

$$p_m = \begin{cases} k & \sigma^2 < \sigma_d^2 \\ 0 & \text{其他} \end{cases} \tag{6}$$

其中 k 取 $[0.1, 0.3]$ 的任意值, σ^2 的取值一般情况下都是远小于其最大值。采用随机扰动的方法对 g^{best} 进行变异操作, 设 g_k^{best} 为 g^{best} 的第 k 维取值, η 是

服从 $Gauss(0,1)$ 分布的随机变量,则

$$g_k^{best} = g_k^{best} \times (1 + 0.5\eta) \tag{7}$$

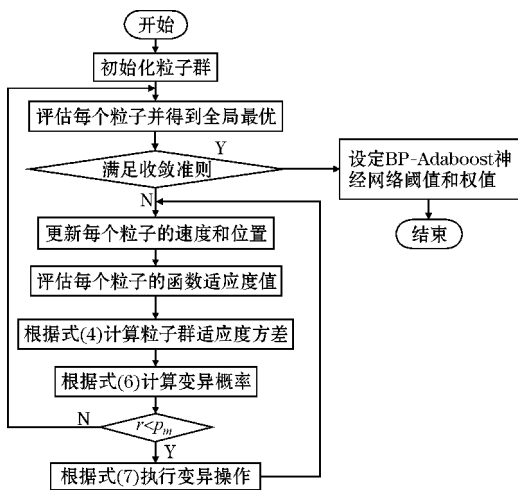


图 1 自适应变异粒子群算法流程图

2.2 BP-Adaboost 算法模型

Adaboost 算法是根据在线分配算法提出的,能够对弱分类器的错误进行适应性调整的一种迭代算法。它的核心思想是训练不同的弱分类器,并将弱分类器集合构成强分类器^[13]。BP-Adaboost 模型和具体步骤如图 2 所示。

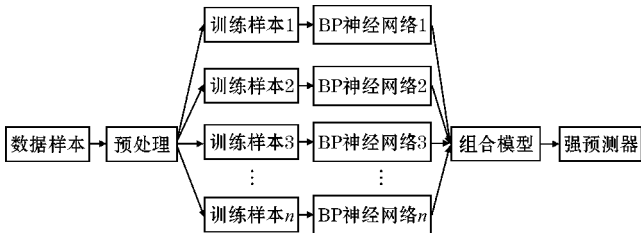


图 2 BP-Adaboost 算法结构

步骤 1:初始化和数据的获取。设 m 组训练数据,则训练数据的权重分布为 $D_i(i) = 1/m$,网络初始的权重和阈值由自适变异粒子群算法优化获得。

步骤 2:训练弱预测器。训练第 t 个预测器时得到预测序列的预测误差和 e_t 以及预测序列权重 a_t :

$$e_t = \sum_{i=1}^m D_i(i); g(t) \neq y \tag{8}$$

$$a_t = \frac{1}{2} \ln \left(\frac{1 - e_t}{e_t} \right) \tag{9}$$

步骤 3:根据公式(8)、(9)调整测试数据的权重值:

$$D_{t+1}(i) = \frac{D_t(i)}{B_t} e^{-a_t y g_t(x_i)} \tag{10}$$

步骤 4:生成强预测器函数:

$$h(x) = \text{sgn} \left[\sum_{t=1}^T a_t f(g_t, a_t) \right] \tag{11}$$

2.3 HMM 算法模型

马尔可夫模型(hidden markov model, HMM)有较

强的学习和建模能力。文中采用 Viterbi 算法求解最可能的隐状态序列,算法步骤如下:

- (1) 初始化: $a_0(1) = 1, a_0(j) = 0, (j \neq 1)$;
- (2) 递推公式: $a_t(j) = \max_i a_{t-1}(i) x_{ij} y_{ij}(z_t), (t = 1, 2, \dots, T; i, j = 1, 2, \dots, N)$;
- (3) 最终输出: $P_{\max}(S, \frac{Z}{\lambda}) = a_T(N)$ 。

2.4 HMM 算法模型

AMPSO-ABP-HMM 情感识别混合模型如图 3 所示。

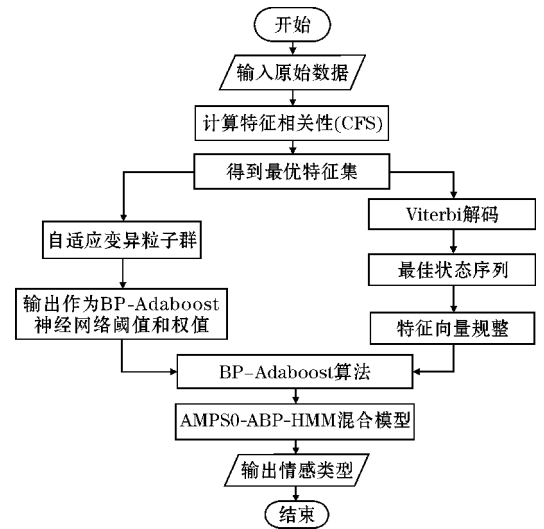


图 3 AMPSO-ABP-HMM 情感识别混合模型

3 仿真实验与结果分析

3.1 特征提取

首先对 2500 条语句进行预处理、分帧和加窗,并提取音质特征、谱特征、韵律特征 3 种特征结合作为情感识别的特征集^[14],包括共振峰特征、振幅能量、短时能量、基音频率、Mel 频率倒谱系数(MFCC)5 种特征及其衍生参数共 72 维特征参数,如表 3 所示。一个高效的特征集是情感识别率高低的关键,由于语音特征值维数很大,其中会包含很多情感值没有贡献或者贡献小的特征值,所以采用基于相关性的特征选择(CFS)提取最优特征^[15],对 2500 个数据的 72 维特征进行特征提取,采用相关性测度去寻找特征之间相关性较低,但与种类标记高度相关的特征构成的特征子集,可以达到去除冗余属性和类无关属性的目的,最后提取反映方言情感的主要原始数据特性 30 维。

表 3 提取的情感特征

特征项	具体特征	特征个数
时间构造	短时平均过零率、无声部分时间比例	2
振幅构造	短时平均能量、短时能量变化范围、短时能量平均变化率、短时能量标准差、短时能量正导数平均值、负导数绝对值平均值、正负导数绝对值均值之差、短时能量短时平均振幅、短时振幅平均变化率、短时最大振幅、短时能量最后十帧导数均值、短时最后十帧二次倒数均值	12
基频构造	基频平均值、最大值、最小值、基频平均变化率、基频动态范围、基频标准差、基频的 1/3 与 1/4 分位点、基频正导数平均值、负导数绝对值平均值、正负导数绝对值均值之差、基频最后十帧导数均值、基频最后十帧二次导数均值	12
共振峰构造	第一共振峰频率、平均值、动态范围、平均变化率、均方差、1/3 与 1/4 分位点、第一共振峰正导数平均值、负导数绝对值平均值、正负导数绝对值均值之差	10
MFCC 系数	12 阶的 MFCC 系数、一阶差分和二阶差分 MFCC 系数	36

3.2 粒子群算法性能分析

实验针对粒子群是否加入自适应变异过程的差异进行分析。在实验过程中设置粒子群算法加速部分最大迭代次数为 100,见图 4。每次迭代过程中记录其粒子适应度方差 σ^2 ,可以看出,AMPSO-BP 收敛速度比 PSO-BP 快很多,并且跳出了局部最优后在第 37 次迭代时寻到最优解。图 5 显示了这两种算法的训练误差,可以看出:AMPSO-BP 算法训练 42 次时达到

0.04098 的最小误差;而 PSO-BP 算法训练 42 次误差值较大为 0.1516,最终在 73 次时达到 0.03789 的最小误差。AMPSO-BP 算法不但降低 PSO 算法陷入局部最优的概率,而且提升了神经网络训练的速度和精度。

3.3 实验结果分析

实验在扬泰方言情感语音库上提出自适应变异粒子群优化的 BP-Adaboost 和 HMM 混合模型进行悲伤、喜悦、愤怒、平静、恐惧 5 种情绪模式识别,识别率如表 4 所示。

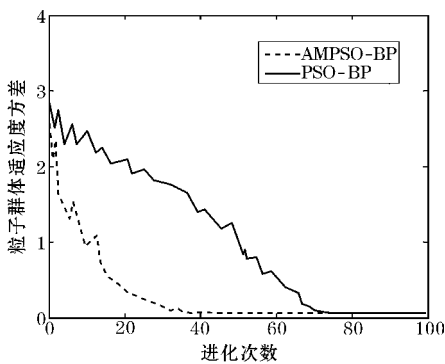


图 4 适应度方差比较曲线

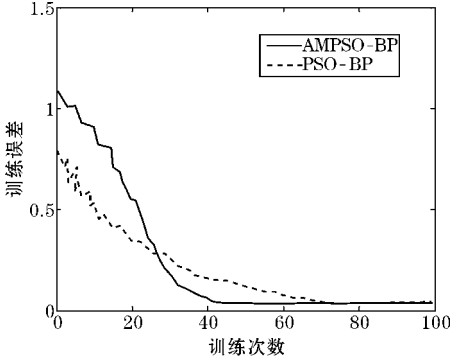


图 5 训练误差比较曲线

表 4 5 种情感识别率

情感	情感识别率/%				
	愤怒	恐惧	喜悦	平静	悲伤
愤怒	87.47	3.55	2.46	3.35	3.17
恐惧	3.32	86.15	3.44	4.06	3.03
喜悦	2.52	2.25	88.50	3.24	3.49
平静	4.40	4.26	4.11	82.84	4.39
悲伤	2.29	3.79	1.49	6.51	85.92

通过 BP、SVM、HMM 以及文中提出的混合模型,分别对扬泰方言情感数据库的 5 种情感识别的识别率进行对比,验证提出的 AMPSO-ABP-HMM 混合模型在

方言情感识别中的效果。如表 5 所示,模型的 5 种情感平均识别率达到了 86.18%,相较 BP 提高了 11.71%,相较 SVM 提高了 9.38%,相较 HMM 提高了

8.17%, 相较 BP-Adaboost 提高了9.29%, 该混合模型对方言数据库的识别率大大提升且稳定性好, 对于相似特征的错误识别率较低, 达到了良好的效果。

表5 情感识别率比较

模型	情感正确识别率/%					平均识别率/%
	愤怒	恐惧	喜悦	平静	悲伤	
BP	75.33	74.42	75.87	73.56	73.15	74.47
SVM	77.15	75.24	79.07	76.59	75.93	76.80
HMM	80.93	75.05	79.32	76.47	78.26	78.01
ABP	79.68	76.42	78.21	74.95	75.19	76.89
文中模型	87.47	86.15	88.50	82.84	85.92	86.18

4 结束语

方言作为经过多年历史变迁而留下的中国传统文化需要被推广以及保护, 目前国内对于方言情感识别的研究少之又少, 文中通过建立扬泰方言情感数据库, 并提出了一种隐马尔科夫模型和自适应变异粒子群优化的 BP-Adaboost 神经网络相结合的模型, 实验中通过该模型对悲伤、喜悦、愤怒、平静、恐惧 5 种情绪的平均识别率达到了86.18%, 该模型的识别率以及训练速度都达到了良好的效果, 未来语音情感识别的发展会更加迅速的, 对于方言的语音情感识别会做更进一步的研究。

参考文献:

[1] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition[C]. International Conference on Multimedia & Expo. IEEE, 2003.

[2] Jiao C, Wang W. Studying on emotion recognition model based on BP network in E-Learning[C]. IEEE International Conference on Software Engineering & Service Sciences, 2010.

[3] Li H, Artieres T, Gallinari P. Data driven design of an ANN/HMM system for on-line unconstrained handwritten character recognition[C]. Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on. IEEE, 2002.

[4] Lin X, Li Y, Dai H, et al. Application of speech recognition system based on algebra algorithm and HMM[J]. Computer Engineering & Design, 2010, 31(24): 5324-5327.

[5] Lv G, Hu S, Lu X. Speech emotion recognition based on dynamic models[C]. International Conference on Audio. IEEE, 2015.

[6] Longfei Li, Yong Zhao, Dongmei Jiang, et al. Hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition [P], 2013.

[7] Hua Y, Chengwei H, Yun J, et al. Speech Emotion Recognition Based on Particle Swarm Optimizer Neural Network [J]. Journal of Data Acquisition and Processing, 2011, 26(1): 57-62.

[8] 李爱军, 王天庆, 殷治纲. 863 语音识别语音语料库 RASC863——四大方言普通话语音库[C]. 全国人机语音通讯学术会议, 2003: 41-44.

[9] 李子煜, 汪鑫, 张优优, 等. 卷积神经网络在语言识别中的应用——以江苏省方言分类为例[J]. 科技传播, 2018, 10(7): 95-97.

[10] 章婷, 朱晓农, 朱璘. 江淮官话通泰片声调类型[J]. 南京师范大学文学院学报, 2015(4): 149-156.

[11] 张策, 韦鹏程, 陆晓燕, 等. 重庆方言语音识别系统的设计与实现[J]. 计算机测量与控制, 2018(1): 256-259.

[12] Cervantes A, Galvan I M, Isasi P. AMPSO: A New Particle Swarm Method for Nearest Neighborhood Classification [J]. IEEE TRANSACTIONS ON CYBERNETICS, 2009, 39(5): 1082-1091.

[13] Bhalla J S, Aggarwal A. Using Adaboost Algorithm along with Artificial neural networks for efficient human emotion recognition from speech [C]. 2013 International Conference on Control, Automation, Robotics and Embedded Systems (CARE). IEEE, 2013.

[14] 赵力,黄程韦.实用语音情感识别中的若干关键技术[J].数据采集与处理,2014,29(2):157-170.

情感识别系统设计[J].化工自动化及仪表,2018,45(3):205-211.

[15] 侯一民,陈帅旗,周慧琼.基于 GA-CFS 的语音

Dialect Emotion Recognition based on Improved BP-Adaboost and HMM Hybrid Model

JI Changpeng¹, CHENG Lin², LI Feng²

(1. School of Electronics and Information Engineering, Liaoning Technical University, Huludao 125105, China;2. Institute of Graduate, Liaoning Technical University, Huludao 125105, China)

Abstract: Aiming at the shortage of dialect emotion database resources and how to improve the recognition accuracy of traditional models, the yangtai dialect emotion database was established, and a model combining the BP-Adaboost neural network optimized by adaptive mutation particle swarm and Hidden Markov Mode (HMM) was proposed. Firstly, correlativity based feature selection (CFS) was used to extract the optimal dialect emotional features. Firstly, using the feature selection (CFS) based on correlation to extract the optimal dialect emotional features. After that, the optimal state sequence was obtained by HMM, and the time was structured. Finally, the feature vector was input into the bp-adaboost neural network optimized by adaptive mutation particle swarm optimization for emotion recognition. . The emotion recognition rate of yangtai dialect emotion database reached 86.18%. The results showed that the model has a good effect on emotion recognition.

Keywords: electronics and communication engineering; dialect emotion recognition; HMM; BP neural network; CFS