文章编号:1674-7070(2012)05-0456-05

基于 BP 网络的汉语普通话声调识别

李仕强[1]　王水平[2,3]

摘要

研究了一种常用的模式分类器——BP 神经网络,分析了 BP 网络的训练及识别过程,提取了能体现声调特性的特征数据组成分类特征向量,设计了具有一个隐含层的 3 层前馈网络作为分类器,对普通话声调样本库做了分类识别实验,分析了不同隐含层节点数的识别实验结果. 实验结果表明,提取的音频特征基本有效,分类效果良好,具有一定的应用价值.

关键词

BP 网络;音频特征分析;声调识别

中图分类号 TN912.34

文献标志码 A

0　引言

随着计算机技术、网络技术和通讯技术的不断发展,音频、图像和视频等多媒体数据约占互联网信息高速公路上所传送数据的 70%,其中声音媒体是除视觉媒体外最重要的媒体形式[1],各行各业对声音媒体的使用也越来越广泛. 我国汉语普通话的使用率非常高,汉语是一种有调语言,汉语普通话的声调分为 4 种:阴平(一声)、阳平(二声)、上声(三声)及去声(四声)[2]. 许多学者利用声调与基因频率之间的因果关系,提出了隐马尔科夫模型、神经网络、决策树和支持向量机等算法进行声调识别. 其中,神经网络是公认的识别效果较好的且应用比较广泛的一类非线性分类器. 本文将重点分析 BP 网络的训练及识别过程,提取汉语的声调特征,采用具有一个隐含层的 3 层前馈网络作为分类器,对声调样本进行分类实验. 由于 BP 算法存在学习速度和收敛速度慢的问题,本文在计算权值调整量时引入了动量项 α,减少了学习过程的振荡,促使 BP 网络更快地收敛,实际应用效果明显.

1　BP 神经网络模型与算法

人工神经网络(Artificial Neural Network)是模拟生物神经网络进行信息处理的一种数学模型,由多个非常简单的处理单元彼此按照某种方式相互连接而成,是依靠系统状态对外部输入信息的动态响应来处理信息的[3]. 人工神经网络最大的特点是可以通过改变连接参数来调整系统,使其适应于复杂环境,从而可以实现类似于人类大脑的学习、归纳和分类等功能. 在实际应用中,超过 80% 的神经网络采用 BP 网络及其各种变化形式,BP 网络体现了人工神经网络中最精华的部分.

1.1　BP 神经网络模型

反向传播神经网络(Back-Propagation Neural Network,BP 网络)是对非线性可微分函数进行权值训练的多层网络,BP 网络是一种具有 3 层或 3 层以上的神经网络,包括输入层、隐含层(中间层)和输出层[4]. 上下相邻层的神经元之间全连接,而同一层的神经元之间无连接. 图 1 给出了包含一个隐含层的 BP 前向神经网络的模型结构,它由一个输入层、一个隐含层和一个输出层组成.

收稿日期 2012-02-09

资助项目 2010 年江苏省政府留学奖学金项目;江苏省教育科学"十二五"规划课题(C-C/2011/01/42);南京信息工程大学教学建设与改革工程项目(11JY056);南京信息工程大学高等教育调研及政策研究课题(2012GJ002);南京信息工程大学实验室开放项目(12KF034,12KF028)

作者简介

李仕强,男,硕士,实验师,主要研究方向为信息安全与智能计算. booyee@nuist.edu.cn

1　南京信息工程大学　网络信息中心,南京,210044

2　南京信息工程大学　江苏省网络监控工程中心,南京,210044

3　南京信息工程大学　计算机与软件学院,南京,210044

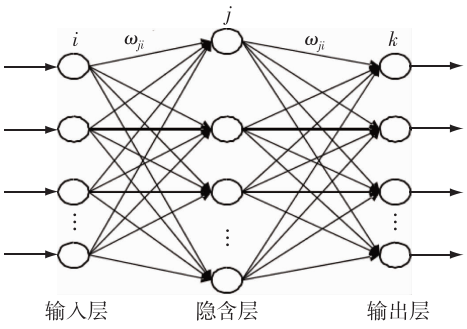

图 1　BP 神经网络结构

Fig. 1　BP neural network structure

1.2　BP 网络学习算法

BP 网络的学习过程包含工作信号的正向传播和误差信号的反向传播两部分. 本文采用了 Back-Propagation 学习算法. 图 2 为本文采用的神经网络学习流程,通过训练样本的学习,从大量输入样本中发现和抽取内在的特征,并将其作为分类的依据.

网络训练的目的主要是确定网络的各连接权值,在训练过程中不断地调整连接权值和阀值,直至总误差函数满足预设的精度要求,则结束训练,形成分类器.

基本 BP 训练算法的实质是最速下降的静态寻优方法,在修正权系数时不考虑先前积累的经验,仅按照该时刻的负梯度方向进行修正,因而收敛速度缓慢. 为解决该问题,本文在每个权值调整量上加入了一个动量项 α,用来提高收敛性能,即采用式(1)来计算各连接权值的调整结果.

$$\Delta\omega_{jk}(n) = \gamma(1-\alpha)D(k) + \alpha D(k-1), \quad (1)$$

其中,$D(k)$代表 k 时刻的负梯度,$D(k-1)$为 $k-1$ 时刻的负梯度,γ 为学习速率,动量项 α 的取值范围为[0,1]. 该动量项起到了阻尼效果,可以减少学习过程的振荡,进而提高收敛速度.

2　预处理及特征提取

本文的处理对象是人类发音器官发出的语音信号. 人类的语音同其他声音相比有其固有的特性,而语音信号处理的最本质的目标就是研究语音信号的基本特征,比如周期、频率、能量等.

语音信号的预处理也可以称作是前端处理,其目的是在对语音信号提取具体特征之前,使待处理的语音信号更能满足实际需求. 对语音的预处理工作主要包括预加重、分帧和加窗等. 预加重是指对语音的高频部分进行加重处理,增加语音的高频分辨率,一般采用一个一阶 FIR 高通数字滤波器来实现预加重. 语音是一个时变信号,但可以认为在一小段时间(一般 10 ~ 30 ms)内具有短时平稳性,因此可以对语音信号分成各个小段来处理,这就是所谓的分帧处理. 为了保证帧与帧之间能平滑过渡,保持语音连续性,一般采用交叠分帧的方法,并且在每一帧的数据上加上一个窗函数,本文采用汉明(Hamming)窗.

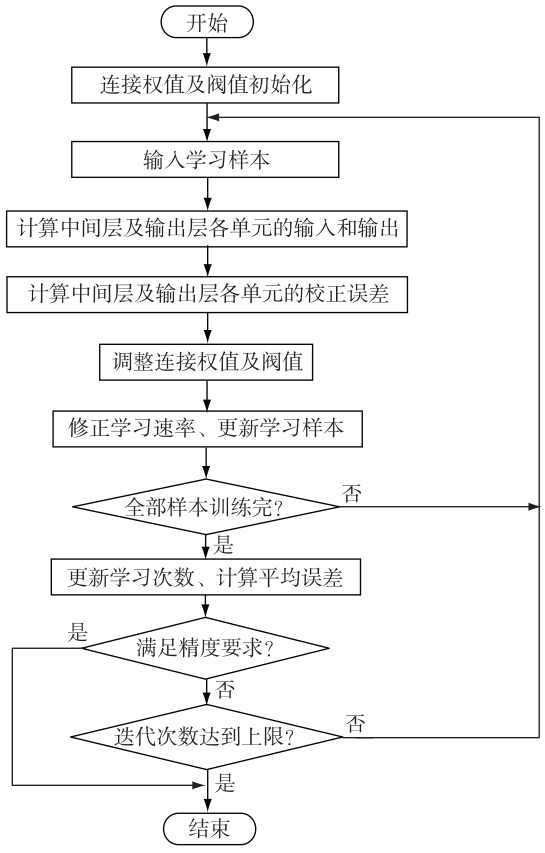

图 2　3 层 BP 神经网络学习流程

Fig. 2　Learning process of 3 – layer BP neural network

图 3 为普通话音节“shu”的 4 种声调语音的语谱. 从语音信号的语谱可以看出,汉语的 4 种声调所对应的基频曲线形成了 4 种不同的变化趋势.

语音信号基音频率的估计在语音信号处理应用中具有十分重要的作用,估计的方法有很多,最基本的有基于短时自相关和基于短时平均幅度差函数的方法. 本文采用短时平均幅度差函数(AMDF)来计算音频的基音轨迹,平均幅度差函数只需加法、减法和取绝对值等计算,算法简单,在基音频率检测中使用得相当普遍[5]. 短时平均幅度差函数定义如下:

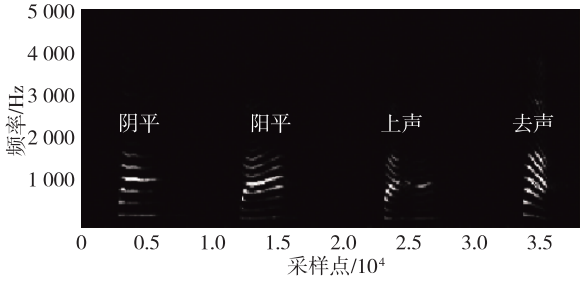

图 3　音节“shu”的 4 种声调音频语谱

Fig. 3　Audio spectrogram of 4 tone of the syllable “shu”

$$A_{\mathrm{DMF}}(k) = \sum_{m=-\infty}^{+\infty} | x(n+m)w(m) - x(n+m-k)w(m-k) | . \tag{2}$$

图 4 为音节“shu”的 4 种不同声调的基音轮廓曲线.

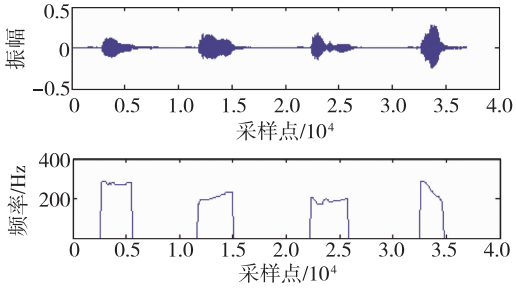

图 4　音节“shu”的波形图及基频轮廓曲线

Fig. 4　Waveform and pitch contour of the syllable “shu”

根据图 4 可以看出,基频的变化规律对于声调的区分度很高. 因此,本文在提取基频特征参数 $F_0 = \{f_0^1, f_0^2, f_0^3, \cdots, f_0^N\}$ 的基础上,进一步计算了差分基频参数 $\Delta F_0 = \{\Delta f_0^1, \Delta f_0^2, \Delta f_0^3, \cdots, \Delta f_0^{N-1}\}$. 另外,本文还计算了各帧的能量 $E = \{e_1, e_2, e_3, \cdots, e_N\}$, $e_i = \sum_{n=0}^{M-1} | x(n) |^2$ 及能量差分 $\Delta E = \{\Delta e_1, \Delta e_2, \Delta e_3, \cdots, \Delta e_{N-1}\}$. 每帧音频的特征向量包括四个参数 $\{f_0^i, \Delta f_0^i, e_i, \Delta e_i\}$,其中 $1 \leqslant i \leqslant N-1$,整个音节的特征参数可以用特征矩阵 $\boldsymbol{T}$ 表示:

$$\boldsymbol{T} = \begin{bmatrix} f_0^1 & \Delta f_0^1 & e_1 & \Delta e_1 \\ f_0^2 & \Delta f_0^2 & e_2 & \Delta e_2 \\ \vdots & \vdots & \vdots & \vdots \\ f_0^{N-1} & \Delta f_0^{N-1} & e_{N-1} & \Delta e_{N-1} \end{bmatrix}. \tag{3}$$

为了对特征向量的维数进行缩减,本文对矩阵中的 4 列参数分别采用最小二乘法做曲线拟合,各取 4 个拟合参数,总共 16 个特征参数组成如下特征矩阵.

$$\boldsymbol{C} = \begin{bmatrix} c_1^1 & c_1^2 & c_1^3 & c_1^4 \\ c_2^1 & c_2^2 & c_2^3 & c_2^4 \\ c_3^1 & c_3^2 & c_3^3 & c_3^4 \\ c_4^1 & c_4^2 & c_4^3 & c_4^4 \end{bmatrix}. \tag{4}$$

3　基于 BP 网络的普通话声调识别

3.1　BP 网络的设计

本文选用标准的 3 层 BP 网络,包含一个输入层、一个隐含层和一个输出层. 输入节点个数取决于样本数据特征提取后的特征向量的维数,共 16 个特征数据,因此 BP 网络输入层节点数为 16. 网络的输出节点数取决于分类结果,本文将音频信号分为 4 个声调,因此 BP 网络输出层节点数为 4. 在对网络进行训练时,如果网络的输入属于第 i 类(如语音是第 1 类),则在网络的输出单元中,第 i 个(如语音则是第 1 个)节点输出为 1,其余的节点输出都为 0.

隐含层节点的个数的选择对网络的性能影响很大,隐含层单元个数太少可能会导致训练不出所需要的网络,容错性较差;隐含层单元个数过多,则会导致网络规模庞大,结构过于复杂,网络训练需耗费大量的时间. 隐含层节点个数最终的取值要结合具体的实验进行,本文将隐含层节点个数预设为 5 ~ 14 个,在样本集上进行训练和测试比较.

3.2　BP 网络训练

实验数据为 12 名在校大学生(男生女生各 6 名)普通话发音中选出的 45 个单音节(a, ai, bao, bo, can, chi, du, duo, fa, fu, ge, hu, ji, jie, ke, la, ma, na, pao, pi, qi, qie, sha, shi, shu, tu, tuo, wan, wen, wu, xia, xian, xu, ya, yan, yang, yao, yi, yun, ying, yu, yuan, zi, zhi, zan),每个音节采集 4 种声调,每名学生的语音采样为一个连续音频文件,采样频率为 22.05 kHz,精度为 16 位. 因此,实验样本集包括 2 160(12 × 45 × 4)个单音节音频信号,其中 6 名学生(3 名男生和 3 名女生)的语音作为 BP 网络训练集(1 080 个),剩余的作为识别测试集(1 080 个). 训练过程如下:

1) 对音频数据进行预处理、特征提取,得到特征矩阵 $\boldsymbol{T}$;

2) 对得到的特征矩阵进行归一化处理,声调特征归一化后的特征向量为 $\boldsymbol{C}$,这样可以便于 BP 网络中权值和阈值的调节及运算;

3) 将特征向量 $\boldsymbol{C}$ 作为 BP 网络的输入矢量,输

出矢量为一维向量 $\boldsymbol{R}$,与 4 种声调分别对应,$\boldsymbol{R}$ = [1,0,0,0] 对应 tone1,$\boldsymbol{R}$ = [0,1,0,0] 对应 tone2,$\boldsymbol{R}$ = [0,0,1,0] 对应 tone3,$\boldsymbol{R}$ = [0,0,0,1] 对应 tone4;

4) 运用 BP 网络学习算法进行训练,直至收敛误差满足要求.

4 分类结果及分析

网络训练完成后,将原始样本集中剩余的样本进行识别测试. 训练及测试过程中分别取隐含层节点数为 5 ~ 14 中的每一种都进行训练和测试,结果如图 5 所示.

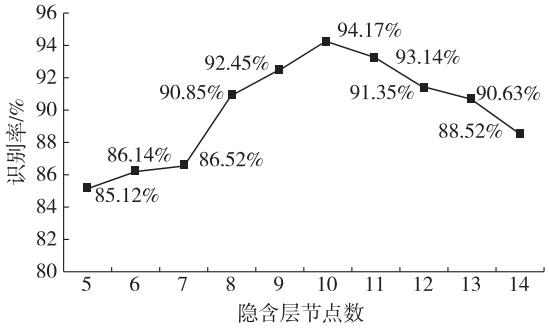

图 5 不同隐含层节点数的识别结果比较

Fig. 5 Recognition results of BP neural network with different hidden layer nodes

由图 5 可以看出,当隐含层节点数为 10 个时,总体识别率最高,达到了 94.17%. 本文还对隐含层节点数为 10 时的实验做了具体分析,具体实验数据如表 1 所列.

表 1 隐含层节点数为 10 的 BP 网络声调识别结果

Table 1 Tone recognition results of BP neural network with 10 hidden layer nodes

样本		识别结果				正确率/%
类别	数量	tone 1	tone 2	tone 3	tone 4	
tone 1	270	250	5	4	11	92.59
tone 2	270	2	267	1	0	98.89
tone 3	270	4	3	256	7	94.81
tone 4	270	13	8	5	244	90.37
平均识别率						94.17

测试集中共有 1 080 个样本,其中正确识别 1 017 个,平均识别率为 94.17%. 实验中,第 2 声的整体识别率最高,达到 98.89%,而 tone1—tone4 和 tone4—tone1 的错误比例较高,占了错误总数的 38.1%,其原因主要在于测试集中部分样本的第 4 声调的发音较轻,与第 1 声难以区分.

参考文献

References

[1] 韩纪庆,冯涛,郑贵滨,等. 音频信息处理技术[M]. 北京:清华大学出版社,2007
HAN Jiqing, FENG Tao, ZHENG Guibing, et al. Speech signal processing technology[M]. Beijing: Tsinghua University Press, 2007

[2] 汤霖,尹俊勋,粟志昂,等. 基于两级 BP 模型的普通话声调识别系统[J]. 计算机工程与应用,2004(25):96-99
TANG Lin, YIN Junxun, SU Zhiang, et al. Mandarin tone recognition system based on two-level BP model[J]. Computer Engineering and Applications, 2004(25): 96-99

[3] 张德丰. Matlab 神经网络应用设计[M]. 北京:机械工业出版社,2009
ZHANG Defeng. Matlab neutral network application design[M]. Beijing: China Machine Press, 2009

[4] Hagan M T, Demuth H B, Beale M. 神经网络设计[M]. 北京:机械工业出版社,2002
Hagan M T, Demuth H B, Beale M. Neural network design[M]. Beijing: China Machine Press, 2002

[5] 冯康,时慧琨. 语音信号基音检测的现状及展望[J]. 微机发展,2004,14(3):95-101
FENG Kang, SHI Huikun. The current situation and prospects of pitch detection[J]. Microcomputer Development, 2004, 14(3): 95-101

[6] Emonts M, Lonsdale D. A memory-based approach to cantonese tone recognition[C]// Proceedings of the 8th European Conference on Speech Communication and Technology, 2003: 2305-2308

[7] Surendran D, Levow G A, Xu Y. Tone recognition in Mandarin using Focus[J]. Baseline, 2005, 17(4): 3301-3304

[8] Peng G, Wang W S Y. Tone recognition of continuous Cantonese speech based on support vector machines[J]. Speech Communication 2005, 45(1): 49-62

[9] 肖汉光,蔡从中. 基于 SVM 的非特定人声调识别的研究[J]. 计算机工程与应用,2009,45(9):174-176
XIAO Hanguang, CAI Congzhong. Study of speaker-independent tone recognition based on support vector machine[J]. Computer Engineering and Applications, 2009, 45(9): 174-176

[10] 王改良,武妍. 基于仿生模式识别理论的声调识别[J]. 计算机应用,2010,30(10):2709-2711
WANG Gailiang, WU Yan. Tone recognition based on biomimetic pattern recognition theory[J]. Computer Applications, 2010, 30(10): 2709-2711

Mandarin Chinese tone recognition based on back-propagation neural network

LI Shiqiang[1] WANG Shuiping[2,3]

1 Network Information Center, Nanjing University of Information Science & Technology, Nanjing 210044

2 Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044

3 School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044

Abstract The BP neural network is studied and applied to tone recognition in this paper, which is a frequently used pattern classifier. Based on analysis of training and recognition processes of BP neural network, we extract feature data reflecting the tone characteristics to build the classification eigenvector. The classifier is designed as a 3-layer BP neural network with one hidden layer, and applied to recognize the four tones of Mandarin Chinese. The recognition results by BP neural network with different hidden layer nodes are compared. The experiment result indicates that the audio feature extracted by this classifier is valid for Mandarin Chinese, and verifies its good classification performance.

Key words back-propagation neural network; audio feature extraction; tone recognition