文章编号: 1001-0920(2002)01-0065-04

基于快速神经网络算法的非特定人语音识别

田 岚[1], 陆小珊[1], 白树忠[2]
(1. 山东大学 信息科学与工程学院, 山东 济南 250061; 2. 山东大学 电气工程学院, 山东 济南 250061)

摘 要: 提出一种用于语音识别的改进的快速神经网络算法, 即动态不等步长的误差分段学习算法。将步长看作误差和网络节点输出的函数, 对各权值按不同步长进行动态调整, 并将其应用于一个基于前馈神经网络模型的非特定人语音识别系统。实验表明, 该算法比传统 BP 算法在训练速度上可提高十几倍, 训练出的语音识别网络系统具有较高的识别率。

关键词: 非特定人语音识别; 神经网络; 学习算法
中图分类号: TP 183 文献标识码: A

Speaker-independent speech recognition based on a fast NN algorithm

TIAN Lan[1], *LU Xiao-shan*[1], *BAI Shu-zhong*[2]
(1. College of Information Science & Engineering, Shandong University, Ji'nan 250061, China; 2. College of Electric Engineering, Shandong University, Ji'nan 250061, China)

Abstract: An improved learning algorithm —— dynamic different step error segmenting algorithm is presented, in which the step is regarded as the function of the error and the output function of network node, and weight is regulated dynamically by different step. By adopting the fast NN algorithm, a speaker-independent speech recognition system based on a BP NN is set up. The experiment shows that the new algorithm is over 10 times faster than the traditional BP algorithm and the resulting neural network has better performance and spreading ability.

Key words: speaker-independent speech recognition; neural network; learning algorithm

1 引 言

神经网络(NN)是一种与传统模式识别完全不同的分布式并行信息处理系统, 由于它具有自适应、自组织及联想等反映人脑加工、存储和搜索信息的某些特征, 使其特别适合于语音识别这类感知问题。

基于神经网络的语音识别与传统的语音识别有着本质的差异。传统的语音识别通常采用模板匹配或概率模型的方法, 其参考模型空间是训练样本集经某种处理后所得的模板集或概率模型, 识别时, 以待识别单元与哪个模板最接近或哪个模型产生的可能性最大作为系统的输出; 而基于神经网络的语音识别系统是对词表的总体建立模型, 该模型的参数集(或权值)通过自学习而获得, 这个参数集是整个词表的语音特征在系统中的映射, 参数子集与词之间没有对应关系。因此, 基于神经网络的语音识别系统对知识的存储是分布的[1]。

收稿日期: 2001-02-03; 修回日期: 2001-04-23

作者简介: 田岚(1965—), 女, 山东济南人, 副教授, 硕士, 从事语音处理及应用等研究; 陆小珊(1965—), 男, 江苏南京人, 讲师, 从事通信及语音处理等研究。

多层结构的前馈神经网络模型已广泛应用于模式识别、联想存储等领域。为使神经网络能作为系统的重要部件或系统本身参与应用, 对网络进行训练或学习是至关重要的。学习算法是训练前馈神经网络连接权值的常用算法。本文首先分析传统 BP 算法的原理, 进而针对其权值调整和收敛速度慢的缺陷, 提出一种动态不等步长的误差分段学习快速算法, 从而有效地缩短了训练时间。在此基础上, 将其应用于一个基于神经网络的非特定人语音识别系统, 取得了满意的实验结果。

2 算法原理及其改进

多层结构的神经网络由输入层(N 个神经元)、输出层(M 个神经元) 和隐含层构成, 随着层数的增加, 网络的模式分类功能逐渐增强。根据 Rumelhart 的结论, 一个具有 Sigmoid 非线性响应函数的三层网络, 能以任意给定精度逼近任何给定的连续函数。BP 学习算法是多层前馈神经网络所使用的监控式学习算法。传统方法按照最小均方误差准则, 使用梯度搜索技术, 以期最小化网络的实际输出和期望输出的均方差, 网络的学习过程是一种误差边传播边修正权值的过程。若采用 Sigmoid 型转移函数

$$f(x) = 1/(1 + e^{-x})$$
$$f'(x) = f(x)[1 - f(x)]$$

输入层、隐含层及输出层的权矩阵分别为 W_{it}, W_{th} 和 W_{hj}, 其学习方法如下:

误差函数为

$$E = \frac{1}{2}\sum_{j=1}^{M}(T_j - y_j)^2 \tag{1}$$

其中, T_j 和 y_j 分别是期望输出和网络实际输出。对输出层权值的调节

$$W_{hj}(n+1) = W_{hj}(n) + \Delta W_{hj} \tag{2}$$

由于 $y_j = f(\sum W_{hj} y_h) = f(s_j)$, y_h 为隐含层节点的输出, 则输出层权值调节量

$$\Delta W_{hj} = -\eta \frac{\partial E}{\partial W_{hj}} = -\eta \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial W_{hj}} = \eta(T_j - y_j) f'(s_j) y_h = \eta \delta_j y_h \tag{3}$$

其中

$$\delta_j = (T_j - y_j) f'(s_j) = (T_j - y_j) y_j (1 - y_j)$$

为输出层节点的误差。同理可得隐含层权值调节量

$$\Delta W_{th} = \eta \delta_h y_t \tag{4}$$

其中

$$\delta_h = \left(\sum_j \delta_j W_{hj}\right) f'(s_h) = \left(\sum_j \delta_j W_{hj}\right) y_h (1 - y_h)$$

为隐含层节点的误差。输入层权值调节量

$$\Delta W_{it} = \eta \delta_t x_i \tag{5}$$

其中

$$\delta_t = \left(\sum_h \delta_h W_{th}\right) f'(s_t) = \left(\sum_h \delta_h W_{th}\right) y_t (1 - y_t)$$

为输入层节点误差。

由此可见, BP 网络的学习是一个优化逼近的过程, 逼近某些给定的输入与输出模式的映射, 实现一个连续的函数映射 $F: X \in R^N \rightarrow Y \in R^M$, 因此 F 具有高度的非线性。

在神经网络的实际应用中, 人们往往关心以下两方面问题: 1) 网络的学习速度及学习算法的运行效率; 2) 网络的推广性能, 即训练后的网络对测试集的正确响应率。而传统的BP 算法在这些方面存在一些缺陷, 主要表现在:

1) 各权值采用相同的步长, 不适合窄长峡谷型的误差曲面, 梯度最陡下降法会使误差在两壁间跳来跳去, 权值得不到充分的训练, 收敛很慢;

2) 出现“过早饱和”现象, 即当某一神经元输出接近S 型函数上、下饱和值(0或1) 时, 梯度变得很小, 很难摆脱局部极小状态, 造成有关权值修正量近似为零, 而其它权值的调整有可能出现“过学习”现象, 使网络的响应率变差。

研究发现, 权值调整的单一步长不能兼顾前期训练和后期训练的要求, 是导致“过早饱和”的一个重要原因。若将步长看作误差和网络节点输出函数值的函数, 对各权值进行动态调整, 必将加速处于上、下饱和状态的神经元梯度方向的调节, 使其迅速脱离饱和状态。同时, 训练过程中由于误差曲面存在许多平坦区, 而且对不同的映射其平坦区的位置和范围各不相同。因此, 若开始就按误差精度进行训练, 势必造成误差在这些区域产生振荡, 使得收敛变慢。为了加快收敛速度, 我们改用选取多级误差的训练方法, 加速全局最小点的搜索, 以很小的额外代价换取网络收敛速度的加快和逼近精度的提高。这样, 由此训练出的网络将具有更好的推广性能[2~4]。

2.1 动态不等步长权值调整方法

设 P 为训练样本总数, 定义网络的目标函数

$$\begin{cases} E^{(p)}(W) = \dfrac{1}{2}\sum_{j=1}^{M}(T_j - y_j)^2 \\ E_{总} = \sum_{i=1}^{P} E^{(i)}(W) \end{cases} \tag{6}$$

权值调节为

$$W_{i,j,k}(n+1) = W_{i,j,k}(n) + \eta\delta_{j,k}(n)y_{i,k-1}(n) + \alpha\delta_{j,k}(n-1)y_{i,k-1}(n-1) \tag{7}$$

其中

$$\eta = \overline{E^{(p)}(W)} + \epsilon[2y_j(n) - 1]^4 \tag{8}$$

$W_{i,j,k}$ 为第 $k-1$ 层中第 i 个神经元与第 k 层中第 j 个神经元之间的权值, $y_{i,k}$ 为第 k 层中第 i 个节点的输出值, $\delta_{j,k}$ 为第 k 层中第 j 个节点的误差, ϵ 初始值取 10～16。可见, η 随误差的减小而减小; 当 $y_j(n)$ 接近 0 或 1 时, 增加步长, 以平衡权值的调整随节点 j 趋近饱和而减小, 使训练加速。

设 $\Delta E_{总} = E_{总}(n) - E_{总}(n-1)$, 如果 $\Delta E_{总} > 0$, 则 $\epsilon = \epsilon/2$; 如果 $\Delta E_{总} < 0$, 则

$$\epsilon = \begin{cases} \epsilon, & \epsilon = \text{初始值} \\ 2\epsilon, & \epsilon < \text{初始值} \end{cases} \tag{9}$$

$$\delta_{j,k} = \begin{cases} (T_j - y_j)F'(s_{j,k}), & k \text{ 为输出层} \\ F'(s_{j,k})\sum_{l=1}^{M}(\delta_{l,k}W_{j,l,k}), & k \text{ 为隐含层} \end{cases} \tag{10}$$

即与节点相连的低一级权值按同一步长调整。如果节点处于 0 或 1 的极端状态, 则与之相连的权值得到充分加强, 使该节点状态沿梯度方向快速改变。

2.2　误差分段训练方法

前馈网络的训练过程是在误差曲面上寻找全局最小点的过程, 而误差曲面通常存在很多极小点。若开始训练时各权值随机赋予一些小数, 那么加入不同的训练样本后, 误差梯度有可能趋向局部极小点, 造成开始训练时误差的振荡, 经过几个振荡周期(一般 4～5 个) 后, 调整方向趋于一个全局最小点。如果开始就按要求的误差精度进行训练, 每个样本训练到满足要求则需要较长时间, 有时还会出现 "过学习", 即训练过程随着网络目标函数的进一步递减, 使得网络的推广性变差。当样本较多时, 这个过程将会更长。

误差分段训练正是基于这一现象提出的。首先在较大的误差上进行训练, 以加速寻找全局最小点; 然后逐级降低误差, 直到满足要求为止。一般误差分为 3～4 级训练较好。对于前馈网络的模式分类, 设输出大于 1－ERROR 为一类, 小于 ERROR 为另一类, ERROR 为给定的误差精度。采用等比误差序进行训练, 初始误差为 0.4(一般取 0.3～0.4), 分 N 段进行训练。首先求出误差放大倍数 $A = 0.4/\text{ERROR}$, 则等比为 $B = 1/A^{(N-1)/2}$。各段误差分别为: $0.4, 0.4B, 0.4B^2, \cdots, \text{ERROR}$。

将以上方法结合起来, 便是本文提出的改进学习算法。具体过程是先设置较大的初始误差, 对每个样本按当前误差进行训练, 并按动态不等步长方法调整权值。当所有样本在当前误差训练好后, 再按误差逐级减小的分段方法重复上述训练过程, 直到满足误差精度为止。

3　实验结果

为了研究改进算法的性能, 我们在联想 586 微机上以 10 个字音的 60 个频谱为输入矢量, 采用 60–20–10 的前馈网络与传统 BP 算法做了对比实验, 结果如表 1 所示。

表 1　算法性能比较

误差	BP 算法		改进算法	
	训练时间 min	权值调整次	训练时间 min	权值调整次
0.4	4.5	11 012	0.6	995
0.1	16	38 678	1	1 390
0.01	—	—	3.4	5 018

可以看出, 改进算法比传统的 BP 算法在训练速度上提高了至少一个数量级, 当误差精度提高时, 效果更加显著。从训练过程权值调整次数上看, 改进算法也比传统算法大为减少, 同时训练出的网络有较好的推广性能。

4　语音识别系统

基于前馈神经网络模型及上述改进的 BP 学习算法, 我们建立了一个非特定人语音识别系统, 其结构如图 1 所示。该系统选用 0～9 共 10 个字音为识别字表, 每个字发音不超过 1 s。语音通过多媒体声卡输入, 16 位 A/D 转换, 11 kHz 频率采样及做高频提升预处理。

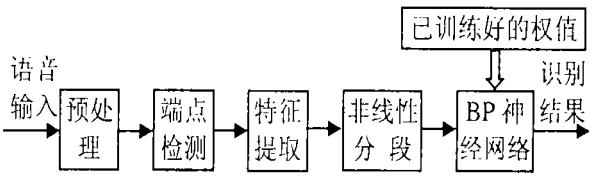

图 1　语音识别系统结构

语音端点检测: 以原始语音信号和一阶差分信号的短时能量以及能零积为参数, 进行多门限端点检测, 检测精度可达 1 帧。

特征参数提取: 每帧取 256 点, 帧间重叠 128 点, 加矩形窗, 作 256 点快速 Walsh 变换(FWT), 按下式逐帧计算功率谱特征参数

$$p(k) = \sum_{j=1}^{D} [X^2_{CAL}(2Dk + j) + X^2_{SAL}(2Dk + j)]$$
$$k = 0, 1, \cdots, N/2D - 2 \qquad (11)$$

当取 $D = 8$ 时, 一帧语音信号可求出 15 个特征值。

非线性分段: 为减少数据的冗余度, 将相关性较强的特征帧合并, 采用非线性时间校准方法, 即将每个发音非线性分成 8 段(即 8 帧)。对应汉语语音的声母段 2～3 帧, 声 / 韵过渡段 2～3 帧, 韵母段 2～4 帧。这样, 每个发音共提取 $15 \times 8 = 120$ 谱值, 进而送入神经网络进行训练和识别。

神经网络: 采用三层 120-30-10 前馈网络, 网络的输入节点为 $8 \times 15 = 120$, 输出节点为 10。对于隐含层节点的选取, 考虑到有利于语音内部区别特征的提取, 避免隐层节点数过少而导致网络对输入模式识别能力变差, 同时兼顾系统的规模不宜过大, 系统选为 30 点。训练算法采用本文提出的动态不等步长误差分段学习算法。

实验中对 5 个发音人(2 男 3 女) 的语音进行了 10 个字音(0 ～ 9) 的识别实验。每个音每人用普通话发 20 遍, 共计 1 000 次发音, 其中以每人每个发音的前 10 次作训练样本, 后 10 次发音作测试样本。识别实验结果如表2所示, 每个音的识别时间小于 0.3 s。

表 2 10 个字音识别结果

数字音	0	1	2	3	4	5	6	7	8	9
识别率 / %	96	96	100	98	96	100	96	96	98	92
总识别率 / %	96.8									

实验中发现, 随着训练样本的增加, 识别率会进一步提高。例如, 对第一位发音者增加训练样本, 对较易混淆的音采用更多样本训练, 然后以每个音发音 20 遍进行识别, 结果在 200 次识别中, "0, 6, 9" 的正识均为 196 次, 误识 4 次, 识别率达到 98%。

5 结 论

利用神经网络进行语音识别是一种很好的方法, 特别是对非特定人语音的识别。该方法是在大量的学习中, 从输入模式中发现和抽取语音特征并进行分类, 减少了传统识别方式中人为提取加工知识的过程, 通过自学习建立规则更容易获得性能优良的识别系统。采用动态不等步长的误差分段训练方法, 不仅可以大大提高网络的学习速度, 减少运算次数, 而且训练的网络具有良好的推广性能, 基于该算法建立的非特定人语音识别系统取得了较高的识别率。

在识别时, 神经网络可以进行 0 或 1 的并行邻近搜索, 搜索时间与存贮项数目无关, 识别响应是实时的。所以对小字表的非特定人语音识别系统而言, 神经网络法具有良好的应用前景。

参考文献(References):

[1] J B Hampshire, A H Waibel. A novel objective function for improved phoneme recognition using time-delay neural networks[J]. IEEE Trans on Neural Network, 1990, 1(2): 216-228.

[2] 焦李成. 神经网络系统理论[M]. 西安: 西安电子科技大学出版社, 1992.

[3] M K Weir. A method of self-determination of adaptive learning rates in back propagation[J]. IEEE Trans on Neural Network, 1991, 4(3): 371-379.

[4] P Burrascano, P Lucci. Smoothing back propagation cost function by delta constraining[A]. Proc of Int Joint Conf on Neural Network[C]. San Diego, 1990. 75-78.

下 期 要 目