第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)03—0202—03 文献标识码：A 中图分类号：TP183

BP 网络模型中输出神经元过早饱和的机理研究

胡上尉，刘琼荪，孙海雷
(重庆大学数理学院，重庆 400044)

摘 要：在数学理论上分析了 BP 神经网络输出神经元出现过早饱和现象的内在机制，提出了相关定理并加以证明。阐释了动量项对出现该现象起着十分重要作用，分析了其它一些文献提出的改进算法与该文所提出的定理之间的联系。因此，该定理对改进 BP 算法在理论上能够提供很好的指导。

关键词：前馈神经网络；BP 算法；过早饱和；内在机制

Analysis of Premature Saturation of the Output Units in Backpropagation Network

HU Shangwei, LIU Qiongsun, SUN Hailei
(College of Mathematics and Physics, Chongqing University, Chongqing 400044)

【Abstract】 The mechanism that gives rise to the phenomenon of premature saturation of the output units of BP network from mathematics is described. The theorem for the occurrence of premature saturation is presented and proved. It is concluded that the momentum term plays the leading role in the occurrence of the phenomenon. In addition, the connection between some improved algorithm proposed by some researchers and the conditions of the theorem in this paper is analysed. Therefore, it can provide a good guidance in theory for improving BP algorithm.

【Key words】 Feedforward neural networks; BP algorithm; Premature saturation; Mechanism

1986 年Rumelhart 和McCelland领导的科学家小组在《Parallel Distributed Processing》一书中，对具有非线性连续转移函数的多层前馈网络的误差反向传播(简称BP)算法进行了详尽的分析，实现了Minsky关于多层网络的设想[2]。BP万能逼近定理：含一个隐层的 3 层BP网络，只要隐节点数足够多，就能以任意精度逼近有界区域上的任意连续函数。这使BP网络成为目前应用最为广泛的一种神经网络，在很多的领域中都得到广泛的应用，主要有模式识别、控制工程、优化计算、信号处理和经济预测等。但是BP网络也存在着收敛速度慢、容易陷入局部极小[1~6]等缺陷和不足。这在很大程度上限制了BP网络的进一步应用。因此许多在该方向上的研究者对它提出了不少好的改进方法[4~6]。

BP网络收敛速度慢除了基于梯度下降来调整权值之外，在很大程度上是由于输出神经元出现过早饱和，导致学习过程出现“平台”或“瘫痪”现象。许多研究者也意识到了这一现象，因此提出了很多改进方法，主要通过引入自适应学习因子、动量因子、陡度因子等[1,2,6]。

1 BP 网络模型的描述

BP算法的基本思想是[2]：学习过程由信号的正向传播与误差的反向传播两个过程组成。正向传播时，输入样本从输入层传入，经各隐层逐层处理后，传向输出层。若输出层的实际输出与期望输出不符，则转入误差的反向传播阶段。它主要通过梯度下降法来修改权值，使得总误差函数达到最小。

设网络共 L 层，N_l 为第 l 层的神经元个数，其中 l= 1,2,…,L。输入的样本总数为 P。第 p 个样本输入向量为 $X_p=(x_{p1},x_{p2},..,x_{pi},..,x_{pn})^T$，其期望输出向量为 $d_p=(d_{p1},d_{p2},...,d_{pk},...,d_{pN_L})^T$，实际输出向量为 $o_p^{(l)}=(o_{p1}^{(l)},o_{p2}^{(l)},...,o_{pk}^{(l)},...,o_{pNl}^{(l)})^T$，p=1,2,…,P。定义总误差函数如下：

$$E=\sum_{j=1}^{N_L}E_j \tag{1}$$

其中 E_j 为第 j 个输出神经元的误差。

$$E_j=\frac{1}{2}\sum_{p=1}^{P}(d_{pj}-o_{pj}^{(L)})^2 \tag{2}$$

第 p 个样本的第 l 层的第 j 个神经元输出为

$$o_{pj}^{(l)}=f(net_{pj}^{(l)})=\frac{1}{1+e^{-net_{pj}^{(l)}}}$$

其中，$f(\cdot)$为激励函数，$net_{pj}^{(l)}=\sum_{i=1}^{N_{l-1}}w_{ij}^{(l)}o_{pj}^{(l-1)}$ 第 k 步迭代后权值更新按以下规则进行：

$$w_{k+1}=w_k+\Delta w_k$$
$$\Delta w_k=-\eta\nabla E(w_k)+\alpha\Delta w_{k-1} \tag{3}$$

$\nabla E=\sum_{j=1}^{N_L}\nabla E_j$，其中 η，α 是小于 1.0 的正常数，分别为学习因子和动量因子。∇E_j 的第 i 分量为

$$\frac{\partial E_j}{\partial w_{ij}^{(l)}}=\sum_{p=1}^{P}\delta_{pj}^{(l)}o_{pi}^{(l-1)} \tag{4}$$

作者简介：胡上尉(1982－)，男，硕士生，主研方向：神经网络，遗传算法；刘琼荪，教授；孙海雷，硕士生

收稿日期：2006-03-05 **E-mail：**s.w.hu@163.com

其中 $\delta_{pj}^{(l)}$ 为第 p 个样本在第 l 层的第 j 个神经元的的信号误差。对于单极 Sigmoid 函数作为激励函数，则有

$$\delta_{pj}^{(l)}=\begin{cases}(o_{pj}^{(L)}-d_{pj})o_{pj}^{(L)}(1-o_{pj}^{(L)}); & \text{当 } l=L\text{时}\\ o_{pj}^{(L)}(1-o_{pj}^{(L)})\sum_{m=1}^{N_{l+1}}\delta_{pm}^{(l+1)}w_{jm}^{(l+1)}; & \text{当 } 1<l<L\text{时}\end{cases} \quad (5)$$

2 输出神经元过早饱和的内在原因分析

在应用标准 BP 算法训练网络的仿真实验中，在训练初始阶段网络的实际输出与期望输出相差较大，即误差较大。但多数情况下误差不会下降，从而误差曲面出现“平台”或“瘫痪”现象，其主要原因是输出神经元已经出现了过早饱和。由表达式

$$\delta_{pj}^{(L)}=(o_{pj}^{(L)}-d_{pj})o_{pj}^{(L)}(1-o_{pj}^{(L)}) \quad (6)$$

可以看出，对 $\delta_{pj}^{(L)}\to 0$ 有以下 3 种情况：

(1) $o_{pj}^{(L)}\to d_{pj}, p=1,2,..,P$，这时对应误差曲面的某个谷点；

(2) $o_{pj}^{(L)}\to 1, p=1,2,\ldots,P$；

(3) $o_{pj}^{(L)}\to 0, p=1,2,\ldots,P$。

其中第(2)、(3)种情况将使得误差信号变得很小，导致误差梯度变小，从而权值调整力度减小，且误差相对较大，最终出现“平台”现象，这不是人们所希望的。文献[1,2,6]对此提出了改进，取得较好的效果。

本文主要是研究分析其动态机制，假设实际输出与期望输出不符(即不考虑情况(1))，且只考虑 $o_{pj}^{(L)}\to 1, p=1,2,\ldots,P$ 的情况(对 $o_{pj}^{(L)}\to 0, p=1,2,\ldots,P$ 的情况可以用相同的方法加以分析)。其动态过程主要分 3 步：对每一个样本p，假设 $o_k>o_{k-1}$，且 $o_{pj}>d_{pj}$，此时 Δw_{k-1} 方向向右。当第k-1 步迭代完成时，由(3)式更新权值得到O_k，假设 $o_k>o_{\max}$ 结果如图 1 所示。

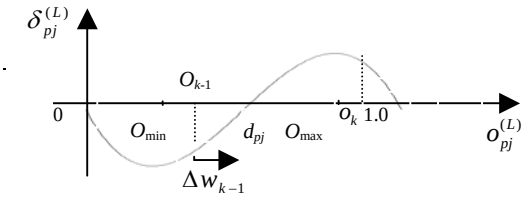

图 1 $o_k>o_{max}$

基于上面的假设，$-\eta\nabla E(w_k)$ 和 $\alpha\Delta w_{k-1}$ 是沿着互为相反的方向来修改权值，其中 $-\eta\nabla E(w_k)$ 是沿着误差函数 E_j 减少的方向进行，即方向向左；而 $\alpha\Delta w_{k-1}$ 是沿着使实际输出趋向于 1.0 的方向进行，即方向向右。假设动量项在修改权值方向上占主导地位，则当第 k 步迭代完成时得到 o_{k+1}，结果如图 2 所示。

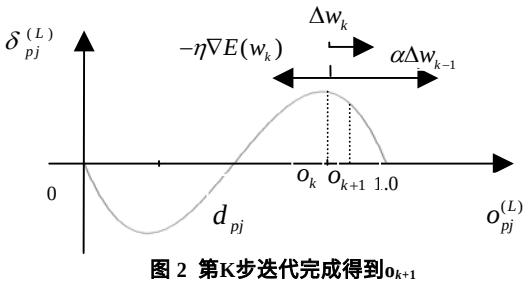

图 2 第K步迭代完成得到o_{k+1}

o_{k+1} 趋向于 1.0 而离期望输出越来越远。此时 $\delta_{pj}^{(L)}$ 变小，从而导致 $\frac{\partial E_j}{\partial w_{ij}^{(l)}}$ 也变小(如果 $o_{pi}^{(l-1)}$ 没有弥补 $\delta_{pj}^{(L)}$ 的减小)。因此，当第 k+1 步完成时得到 o_{k+2}，结果如图 3 所示(分析同上步)。

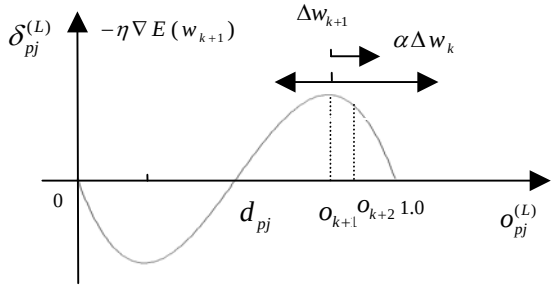

图 3 第k+1 步完成得到O_{k+2}

o_{k+2} 趋向于 1.0 而误差函数 E_j 却增加了，同时 $\frac{\partial E_j}{\partial w_{ij}^{(l)}}$ 变小了。经过多次连续的迭代，导致 $-\eta\nabla E(w_k)$ 越来越小，权值的修改几乎被 $\alpha\Delta w_{k-1}$ 所决定，即 $\Delta w_k\approx\alpha\Delta w_{k-1}$，而 $\alpha\in[0,1]$，最终导致权值几乎不会得到更新，但是有较高的误差。这说明了动量项对输出神经元出现过早饱和现象起着关键的作用，同时也表明了该输出神经元处于过早饱和状态，使误差曲面出现平坦区，导致出现收敛速度慢甚至不收敛的现象。

3 定理及证明

基于上面的输出神经元出现过早饱和内在机制的分析，本文得出如下定理，当定理中条件都满足的时候，该输出神经元将会出现过早饱和现象。

定理 设前馈神经网络共有 L 层，N_L 为第 L 层的神经元个数。当用标准 BP 算法来训练网络时，对输出层的第 j 个神经元，$j=1,2,\ldots,N_L$。如果对每一个样本 p，p=1,2,...,P，满足如下条件：

(1) $\Delta w_{k-1}\cdot\nabla E_j(w_k)>0$；

(2) $\alpha\left|\Delta w_{k-1}\cdot\nabla E_j(w_k)\right|>\eta\left|\nabla E(w_k)\cdot\nabla E_j(w_k)\right|$；

(3) $\left|\nabla E_j(w_{k+1})\right|<\left|\nabla E_j(w_k)\right|$；

(4) $o_{pj}^{(L)}(w_{k+1})>o_{\max}$。

其中 $o_{\max}=\frac{1}{3}(1+d_{pj}+\sqrt{1-d_{pj}+d_{pj}^2})$

则输出层的第 j 个神经元将开始出现过早饱和现象。

证明 由 $E_j(w_{k+1})$ 一阶 Tayor 展开式可知，要使通过 $\Delta w_k=-\eta\nabla E(w_k)+\alpha\Delta w_{k-1}$ 式修改权值来减少训练误差 E_j，则必须满足

$$\Delta w_k\cdot\nabla E_j(w_k)<0 \quad (7)$$

把 $\Delta w_k=-\eta\nabla E(w_k)+\alpha\Delta w_{k-1}$ 代入式(7)可得 $\eta\nabla E(w_k)\cdot\nabla E_j(w_k)-\alpha\Delta w_{k-1}\cdot\nabla E_j(w_k)>0$。但是，上式在输出神经元开始处于过早饱和阶段并不满足。即当条件(1)满足时，训练误差将不会出现下降现象。同时由本文第 2 部分的分析可知，当动量因子在修改权值的过程中占主导地位时，即满足

$$\alpha\left|\Delta w_{k-1}\cdot\nabla E_j(w_k)\right|>\eta\left|\nabla E(w_k)\cdot\nabla E_j(w_k)\right|\cdot$$

此时 $\delta_{pj}^{(L)}$ 变小，输出层第 j 个神经元的实际输出将偏离期望输

出，导致误差增加。同时当条件 $\left|\nabla E_j(w_{k+1})\right| < \left|\nabla E_j(w_k)\right|$ 和 $o_{pj}^{(L)}(w_{k+1}) > o_{\max}$ 满足时(这是基于第 2 部分的假设)，其中 $o_{\max} = \frac{1}{3}(1 + d_{pj} + \sqrt{1 - d_{pj} + d_{pj}^2})$ (只要对式(6)令 $\frac{d\delta_{pj}^{(L)}}{do_{pj}^{L}} = 0$ 即可求得)，即误差梯度连续下降，导致 $-\eta\nabla E(w_k)$ 越来越小，权值的修改几乎被 $\alpha\Delta w_{k-1}$ 所决定，即 $\Delta w_k \approx \alpha\Delta w_{k-1}$，而 $\alpha \in [0,1]$，最终导致权值更新非常小，但是有较高的误差，导致出现“平台”现象。综上所述，当条件(1)~条件(4)都满足时,输出神经元 *j* 将开始出现过早饱和现象。(证毕)

由以上的定理可知，条件(1)、条件(2)说明了动量项在输出神经元出现过早饱和现象中起着十分重要的作用。而条件(3)、条件(4)是基于假设保证误差梯度不断变小，最终导致输出层的第 j 个神经元开始出现过早饱和现象。

4 防止出现过早饱和现象

针对上面内在机制的分析及提出的定理，为了防止输出神经元出现过早饱和现象，只要改进BP网络使得上面定理的条件不成立即可。现在有不少文献对其作出了改进，取得较好的效果[1~6]。如文献[2,4,6]提出了自适应调整学习因子和动量因子，以加快收敛速度。其中在文献[4]中提出了训练迭代步数与学习因子、动量因子的关系，同时也指出了 $\eta \neq 0$，且 α 必须小于 1.0 否则网络不会收敛。不难看出其改进主要使本文定理中的条件 2 不满足。再如文献[5]提出了调整激励函数的方法，使得

$$f(x) = \begin{cases} 0.999, & \text{当 } f(x) > 0.9999\text{时} \\ 0.0001, & \text{当 } f(x) < 0.00001\text{时} \\ f(x), & \text{其 它} \end{cases}$$

还有一些文献提出，当进入 Sigmoid 函数的饱和区域时直接修改 $o_{pk}^{(L)}(1-o_{pk}^{(L)})$ 为 $o_{pk}^{(L)}(1-o_{pk}^{(L)}) + 0.10$ 以增大误差梯度来加快收敛速度。同样，也可以看出他们的改进主要是使本文定理中的条件(4)不满足。还有其它很多的改进方法都可以归纳到本文的定理当中。

5 结束语

本文描述了 BP 网络的数学模型，通过分析输出神经元过早饱和现象的内在机制，提出并证明了出现该现象的相关定理，同时也说明了动量因子对神经元出现过早饱和现象起着十分关键的作用。此外，本文通过分析其它一些文献提出的改进算法，其内在的因素都可以归纳到本文所提出的定理当中。因此，本文对改进 BP 算法在理论上能够提供很好的指导。

参考文献

1 Hagan M T, Demuth H B, Beale M. Neural Network Design[M]. 北京: 机械工业出版社, 2002.

2 韩力群. 人工神经网络理论、设计及应用[M]. 北京: 化学工业出版社, 2002: 34-56.

3 Haykin S. 神经网络原理[M]. 北京：机械工业出版社，2004: 109-173.

4 Dai Hengchang, MacBeth C. Effects of Learning Parameters on Learning Procedure and Performance of a BPNN[J]. Neural Networks, 1997, 10(8): 1502–1521.

5 杨安华, 彭清娥, 刘光中. BP 算法固定学习率不收敛原因分析及对策[J]. 系统工程理论与实践, 2002, (12): 22-25.

6 万小朋, 王军强, 赵美英. BP 算法改进及其在结构损伤检测中的应用[J]. 机械科学与技术, 2002, (11): 61-63.

（上接第 191 页）

Zernike 矩特征描述的是一幅图像的全局信息，而用户所关心的并非单单是图像整体的相似，而是两幅图像所表现的病理特征是否相同。

(3)利用局部病灶特征和全局 Zernike 矩特征的融合的检索效果，检索查准率比较高，效果优于前几种单一特征的检索。在返回 15 幅图像时，相关图像的平均排序值为 4.5，已达到理想值。

表 1 基于图像底层特征提取检索结果

检索特征	纹理		形状	
	15	30	15	30
Average-r	6.91	6.89	7.17	8.52
平均查准率	0.72	0.69	0.59	0.49

表 2 基于本文方法检索结果分析

检索特征	Zernike 矩		基于感兴趣区域特征提取		特征融合	
	15	30	15	30	15	30
Average-r	6.29	6.52	5.12	5.12	4.5	4.71
平均查准率	0.75	0.689	0.895	0.833	0.925	0.867

5 结束语

实验结果表明，采用本文提出的局部和全局特征融合查询有效地提高了检索结果的准确性。将该方法运行于日后开发的医学图像检索系统，协助医生对 CT、MRI 等常用的医学病理图像进行分析诊断，具有一定的实用价值。

致谢 课题得到内蒙古自治区医院放射科、介入放射科、内窥镜室的协助，特此感谢。

参考文献

1 Swain M J, Ballard D H. Color Indexing[J]. International Journal of Computer Vision, 1991, 7(1): 11-32.

2 Unser M. Texture Classification and Segmentation Using Wavelet Frames[J]. IEEE Transactions on Image Processing, 1995, 4(11): 1549-1560.

3 Flusser J, Sak T. Pattern Recognition by Affine Moment Invariants[J]. Pattern Recognition, 1993, 26(1): 167-174.

4 邵 虹, 崔文成, 张继武, 等. 低级特征和语义特征相结合的医学图像检索方法[J]. 中国图象图形学报(A), 2004, 9(2): 220-224.

5 Moghaddam B, Biermann H, Margaritis D. Defining Image Content with Multiple Region of Interest[C]//Proc. of IEEE Workshop on Content Based Access of Image and Video Libraries, 1999.

6 Gonzalez R C, Woods R E. 数字图像处理[M]. 第 2 版. 北京: 电子工业出版社, 2004-06.

7 余 鹏, 封举富. 基于高斯混合模型的纹理图像分割[J]. 中国图象图形学报, 2005, 10(3): 281-285.

8 Haralick R M. Statistical and Structural Approaches to Texture[J]. Proceedings of the IEEE, 1999, 67(5): 786-804.

9 Ortega M. Supporting Similarity Queries in MARS[C]//Proc. of ACM Multimedia, 1997: 403-413.

10 何晖光, 田 捷, 赵明昌, 等. 基于分割的三维医学图像表面重建算法[J]. 软件学报, 2002, 13(2): 219-226.

11 Teague M R. Image Analysis via the General Theory of Moments[J]. Opt. Soc. Am., 1980, 70 (8): 920-930.