维普资讯 http://www.cqvip.com

计算机科学 1997 Vol. 24 №. 2

基于稳健误差估计器的快速 BP 算法

刘光远　邱玉辉　虞厥邦

（西南师范大学计科系　重庆 630715）（成都电子科技大学光电系）

摘　要　标准的或许多改进的 BP 算法大都采用均方误差估计器，致使学习易于陷入局部极小、收敛速度慢和对初始权敏感等。本文针对这些缺点提出了较好的算法，仿真结果证实了核算法的正确性。

关键词　标准 BP 算法，误差估计器，稳健 BP 算法，XOR 问题

1　引言

自从 Rumelhart 等科学家于 1986 年提出并系统论述了反向传播(Backpropagation，简称 BP)算法以来，该算法已在模式识别、信号处理、自适应控制、机器人、机器学习等众多领域得到了广泛的应用。近十年来许多研究人员对该网络进行了深入的研究，其研究重点集中在诸如训练时间长、易于陷入局部极小、对初始权敏感、训练存在发散或麻痹现象等 BP 网络的不足点上，这些工作都是卓有成效的。

众所周知，标准 BP 算法所定义的误差估计器是一个二次式函数，但这并不是唯一的选择。1988 年 Baum 和 Wileze 等人提出了一个误差估计器，可以克服麻痹现象，但可能导致过调或振荡。1989 年，Falman 提出了一种折衷方案，其效果比前者好。然而，这些工作并未很好地解决 BP 网对初始权敏感及收敛速度慢等问题。针对这些方法的局限性，本文利用拉格朗日乘数法提出了一种稳健 BP 算法。

2　稳健 BP 算法

事实上，对于 BP 算法，所选用的误差(函数)估计器即使不是二次式，只要该误差估计器在目标函数等于实际输出时能达到最小值就满足要求。最优化理论中的拉格朗日乘数法给予我们启示，即只要找到一项适当的约束，求解某误差函数之极小使之能满足要求，则我们的问题即可得到解决。

(上接第 39 页)

作集合对象的基类之间存在继承关系，为操作产生的对象的基类指明了适当的位置。

参考文献

[1] Won Kim，面向对象数据库系统：诺言·现实·前景(1)，计算机科学，21(4)1994

[2] Won Kim，面向对象数据库系统：诺言·现实·前景(2)，计算机科学，21(5)1994

[3] M. Mosaffari and Y. Tanaka, ODM: An object-oriented data model, New Generation Computing, 7(1) 1989

[4] Jan Van Den Bussche and Andreas Heuer, Using SQL with object-oriented database, Info. Sys., 18(7)1993

[5] M. A. Roth, H. F. Korth and A. Siberschatz, Extended algebra and calculus for nested relational databases, ACM Trans. Database Systems, 13(4) 1988

[6] D. Van Gucht and P. C. Fischer, Multilevel nested relational structures, J. Comput. Sys. Sci, 36(1) 1988

[7] G. M. Shaw and S. B. Zdonik, An object-oriented query algebra, Proc. 2nd Int. Workshop Database Programming Language, 1989

[8] G. M. Shaw and S. B. Zdonik, A query algebra for object-oriented database, Proc. 6th Int. Conf. Data Engineering, 1990

刘光远　博士生，主要研究领域为神经网络。

维普资讯 http://www.cqvip.com

现设所研究的网络在前馈方式下包含多个联接层，且其间没有跳跃层，层数用 0 到 N 表示，每层神经元个数为 $n_i(i=0,1,\cdots N)$。对于模式 $p(p=1,2,\cdots,P)$，第 k 层的第 j 个神经元的状态为 $x_j^p(k)$，$j=1,2,\cdots,n_k$。第 $k-1$ 层的第 i 个神经元与第 k 层的第 j 个神经元的突触联接为 $w_{ij}(k)$，第 $k-1$ 层输出通过联接权 $w_{ij}(k)$ 得到第 k 层第 j 个神经元的输入为：

$$a_j^p(k)=\sum_{i=1}^{n_{k-1}} w_{ij}(k)x_i^p(k-1), \quad (1)$$

$j=1,2,\cdots,n_k;p=1,2,\cdots,P$。

因此，第 k 层神经元输出（即前向传播方程）为：

$$a_j^p(k+1)=f(a_j^p(k))=f(\sum_{i=1}^{n_{k-1}} w_{ij}(k)x_i^p(k-1)) \quad (2)$$

式中 $x_j^p(0)(j=1,2,n_0)$ 是外部输入向量，也即输入样本，$f(\cdot)$ 是非线性函数，通常取为：

$$f(x)=(1+e^{-cx})^{-1} \text{ 或 } f(x)=\tanh(cx) \quad (3)$$

设目标输出为 $d_j^p(j=1,2,\cdots,n_N;p=1,2,\cdots,P)$，网络的实际输出为 $x_j^p(N)$（j 与 p 取值同上）。我们定义一种新的误差估计器：

$$E_R(w_{ij},x_j^p(N))=\sum_{p=1}^{P}\sum_{j=1}^{n_N}\varphi(r_j^p) \quad (4)$$

式中 $r_j^p=d_j^p-x_j^p(N)$，如果 $\varphi(r_j^p)=(r_j^p)^2$，这就是传统的均方误差估计器。

根据稳健统计（Robust Statistic）理论，有稳健解的函数能抑制系统的过失误差，对于初始权值不会存在敏感性问题。为此，我们提供两种稳健性误差估计器：

（1）逻辑斯谛误差估计器

$$E_{RL}(w_{ij},x_j^p(N))=\frac{\alpha}{\beta}\sum_{p=1}^{P}\sum_{j=1}^{n_k}\log[\cosh(\alpha r_j^p)]$$

$$\alpha,\beta>0 \quad (5)$$

（2）柯西误差估计器

$$E_{RC}(w_{ij},x_j^p(N))=\sum_{p=1}^{P}\sum_{j=1}^{n_N}\ln(1+(r_j^p)^2] \quad (6)$$

利用拉格朗日乘数法我们可以得到稳健 BP 算法。先看对于单个模式 p 的拉格朗日函数（局部拉格朗日函数）有：

$$L^p(w_{ij},x_j^p,b_j^p)=E_R^p(w_{ij},x_j^p(N))+\sum_{k=1}^{N}\sum_{j=1}^{n_k}b_j^p(k)\cdot$$

$$[x_j^p(k)-f(a_j^p(k))] \quad (7)$$

对于全部样本模式，必有全局拉格朗日函数为：

$$L=\sum_{p=1}^{P}L^p(w_{ij},x_j^p,b_j^p) \quad (8)$$

如前所述，L 的两项和中之第一项正好是输出误差，它可为(5)、(6)或(5)+(6)三种表达式中的任一种，第二项是对每层的 N 求和，可解释为约束项，其中 $b_j^p(k)$ 为拉格朗日乘数。如果约束满足，则所有约束项为零（这可由(2)式得到解释），因此可得：

$$x_j(k)=f(\sum_{i=1}^{n_{k-1}} w_{ij}(k)x_i(k+1))$$

$$k=1,2,\cdots,N;j=1,2,\cdots,n_k \quad (9)$$

(9)式与(2)式实质上同义，可见约束项正好定义了网络的前向动力学特性，亦即 $x_j(k-1)$、$w_{ij}(k)$、$x_j(k)$ 之间的关系。

下面我们将寻找序列 $w_{ij}(1),w_{ij}(2),\cdots,w_{ij}(N)$ 使得满足约束的目标函数极小化，容易证明：

$$\nabla L(w_{ij},x_j,b_j)=0 \quad (10)$$

成立。方程(10)完全描述了网络的动力学特性，它由以下三个子方程组成：

$$\frac{\partial L(w_{ij},x_j,b_j)}{\partial b_j(k)}=0, \quad j=1,2,\cdots,N_0 \quad (11)$$

$$\frac{\partial L(w_{ij},x_j,b_j)}{\partial x_j(k)}=0, \quad j=1,2,\cdots,N_0 \quad (12)$$

$$\frac{\partial L(w_{ij},x_j,b_j)}{\partial w_{ij}}=0, \quad j=1,2,\cdots,N_0 \quad (13)$$

对于某个具体模式中，由方程(11)有

$$\frac{\partial L(w_{ij},x_j,b_j)}{\partial b_j(k)}=\frac{\partial L^p(w_{ij},x_j^p,b_j^p)}{\partial b_j^p(k)}$$

$$k=1,2,\cdots,N;p=1,2,\cdots,P。 \quad (14)$$

由于对任意局部拉格朗日函数，上式并不依赖于 b_j^p，因此得

$$x_j^p(k)=f(\sum_{i=1}^{n_{k-1}} w_{ij}(k)x_i^p(k-1))$$

$$k=1,2,\cdots,N;p=1,2,\cdots,P \quad (15)$$

此即前向动力学方程(2)或(9)。

由方程(12)，对最后一层（k=N）有：

$$\frac{\partial L^p(w_{ij},x_j^p,b_j^p)}{\partial x_j^p(N)}=0 \text{ 得}$$

$$b_j^p(N)=-\frac{\partial E_R(w_{ij},x_j^p(N))}{\partial x_j^p(N)} \quad (16)$$

而对 $k\neq N$ 的那些网络层则有

$$b_j^p(N)=\sum_{i=1}^{n_k} w_{ij}(k+1)b_i^p(k+1)f'(a_i^p(k+1)) \quad (17)$$

令 $y_j^p(k)=b_j^p(k)f'(a_j^p(k))$ (18)

则(17)式变为：

$$y_j^p(k)=\sum_{i=1}^{n_k} w_{ij}(k)y_i^p(k+1)f'(a_j^p(k)) \quad (19)$$

利用(18)式，可将(16)式变换成 $y_j^p(N)$ 的计算式，相应的各种误差估计器的 $y_j^p(N)$ 的计算公式列于表 1 中。

维普资讯 http://www.cqvip.com

表 1　计算 $y_j^p(N)$的公式

均　方	$y_{pj}^p(N)=2f'(a_j^p(N))(d_j^p-x_j^p(N))$非稳健
柯　西	$y_{pj}^p(N)=\frac{2(d_j^p-x_j^p(N))}{1+(d_j^p-x_j^p(N))^2}f'(a_j^p(N))$
逻辑斯谛	$y_{pj}^p(N)=\beta\tanh(\alpha(d_j^p-x_j^p(N)))f'(a_j^p(N))$
复　合	$y_{pj}^p(N)=my_{pj}^p(N)+(1-m)y_{pj}^p(N)\quad m\in[0,1]$

又由(13)式得：

$$\sum_{p=1}^{P}\sum_{j=1}^{n_k}b_j^p f'(a_j^p(k))x_j^p(k-1)=0 \tag{20}$$

再利用公式(18)即知

$$\sum_{p=1}^{P}\sum_{j=1}^{n_k}y_j^p(k)x_j^p(k-1)=0 \tag{21}$$

显然，上式中包含的 w 是 L 的不动点。我们仍采用梯度下降法寻找目标函数关于 w 的极小点：

$$\Delta w_{ij}(k)=-\lambda\frac{\partial L(w_{ij},x_j,b_j)}{\partial w_{ij}(k)} \tag{22}$$

即　$$\Delta w_{ij}(k)=\lambda\sum_{p=1}^{P}y_i^p(k)x_j^p(k-1) \tag{23}$$

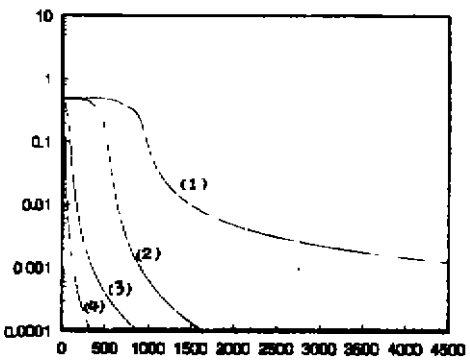
式中 λ 为调节阶距。从而

$$w_{ij}(k)\leftarrow w_{ij}(k)+\Delta w_{ij}(k) \tag{24}$$

算法的实现过程同标准 BP 算法一样，限于篇幅，此略。

3　仿真实验

我们采用 2-2-1 三层前向网络分别对表 1 中的四种误差估计器进行了计算机仿真实验。其结果示于图 1 和表 2 中。所取控制参数为：最大总体误差 $E_{max}=10^{-4}$，m＝0.5，λ＝0.8。为便于观察曲线的收敛处情形，图 1 中的纵坐标采用对数标度(误差用对数标度)。



(1)均方　(2)柯西　(3)逻辑斯谛　(4)复合

图 1　四种误差估计器的仿真结果

表 2

误差估计器	迭代次数
①均方(不带冲量项)	35453
②柯　西	1648
③逻辑斯谛	884
④复合(m＝0.5)	333

可以看出，我们提出的算法其结果明显优于传统的标准 BP 算法。在各进行的 20 次实验中，新算法对随机设定的初始权并不表现出敏感性。特别值得一提的是，当两种稳健误差估计器结合时，其收敛速度令人满意。

4　结束语

本文中，利用拉格朗日乘数法结合稳健误差估计器详细描述了稳健 BP 算法。与传统方法相比较，我们的方法考虑了网络前向传播方程的不动点约束，并把它结合在代价函数中，因此方法的数学理论是严谨的。所进行的仿真实验表明：我们的算法收敛速度快且对初始权不敏感。事实上，该算法还具有对函数良好的逼近能力，其结果将另文报道。

参考文献

[1]D. Rumelhart et al. ,Learning Representation by Back-propagating Errors,Nature,1986,323:533～536

[2]D. Rumelhart, J. McCleand, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge:Bradford Books,MIT Press,1986,(1-2)

[3]E. Baum,F. Wilczek,Supervised Learning of Probability Distribution by Neural Networks,In: Anderson D. ed. Proc. of the 1987 IEEE Conf. on Neural Information Processing Systems—— Natural and Synthetic, New York:American Institute of Physics,1988:52～61

[4]P. J. Huber,Robust Statistic,Wiley,New York,1981

[5]M. Athams and P. L. Falb,Optimal Control Theory, McGram Hill,1966

[6]D. Hush,J. Salas,Improving the Learning Rate of Back-propagation with the Gradient Rease Algorithm,Proc. of the IEEE Inter. Conf. on Neural Networks,1988,(I),441～447