前馈网络的一种超线性收敛BP学习算法*

梁久祯[1] 何新贵[2] 黄德双[2]

[1](北京航空航天大学计算机科学与工程系 北京 100083)

[2](北京系统工程研究所 北京 100101)

E-mail: liang-jz@263.net

摘要 分析传统BP算法存在的缺点,并针对这些缺点提出一种改进的BP学习算法.证明该算法在一定条件下是超线性收敛的,并且该算法能够克服传统BP算法的某些弊端,算法的计算复杂度与简单BP算法是同阶的.实验结果说明这种改进的BP算法是高效的、可行的.

关键词 前馈神经网络,BP学习算法,收敛性,超线性收敛.

中图法分类号 TP18

80年代,Rumelhart,Hinton和Williams[1]发明的一般Delta法则,为多层BP前馈神经网络的学习奠定了基础.特别是对BP算法的讨论一直是神经网络界的研究热点.BP学习算法的学习过程目的是使误差能量函数降低到给定的精度,但是由于误差函数的高维复杂性,使BP学习算法存在着难以克服的缺点,如学习过程易陷入局部极小、学习算法的收敛速度很慢、学习过程易出现震荡现象等.

本文针对误差函数超曲面的特征以及传统的BP学习算法的缺点提出了一种改进的BP学习算法,并证明了该算法在一定条件下是超线性收敛的,该算法能有效地克服传统BP算法的收敛速度慢和学习过程震荡等现象.本文最后给出了XOR问题和九点模式识别问题的实验结果.

1 基本BP算法及其存在的问题

考虑具有两层权值的BP网络,误差能量函数取$E=E(W)$,其中W为所有权构成的向量.显然,$E(W)$为一高维的超曲面,设它的沿W的某一分量W_i的剖面如图1所示.

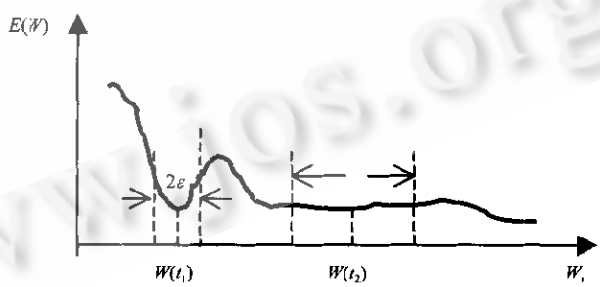

Fig. 1 Section plan of $E(W)$

图1 $E(W)$的剖面图

BP梯度算法的基本形式为

* 本文研究得到国家自然科学基金(No. 69705001)资助.作者梁久祯,1968年生,讲师,主要研究领域为神经网络优化算法,模糊知识处理,并行计算,分布式计算.何新贵,1938年生,研究员,博士生导师,主要研究领域为人工智能,模糊技术,神经网络,知识处理.黄德双,1964年生,博士后,研究员,主要研究领域为模式识别,神经网络,进化计算,模糊控制.

本文通讯联系人:梁久祯,北京 100083,北京航空航天大学计算机科学与工程系

本文1999-01-11收到原稿,1999-08-27收到修改稿

$$W(k+1)=W(k)-\alpha(k)\nabla E(W(k)), \tag{1}$$

其中$\alpha(k)$为学习速度.通常,$\alpha(k)$的取法有两种:(1) $\alpha(k)=\alpha$为一常数,只需在每一步迭代前计算梯度$\nabla E(W(k))$即可,称为简单梯度法.但由图1可知,若在$W(t_1)$的某个邻域a的取值使$\|\alpha\nabla E(W(k))\|<\varepsilon$,其中$\varepsilon$为一个正数,则在$W(t_1)$处会发生震荡现象.而当学习进入误差函数曲面的平坦区(如$W(t_2)$处)时,由于$\|\alpha(k)\nabla E(W(k))\|\to 0$,学习速度很慢,即网络又会出现"麻痹"现象.另一方面,若我们试图减小α使网络学习减小震荡,而当进入误差函数曲面的平坦区时,由于α较小,学习速度会更慢.反之,若欲加速学习速度而把α调大,在误差函数曲面陡变区网络学习震荡加剧.由此可见,对定常系数α来说,这是难以克服的缺点.(2) 优化$\alpha(k)$的最速下降法,即优化$\min\limits_{\alpha(k)}E(W(k)-\alpha(k)\nabla E(W(k)))$,等价于解如下关于$\alpha(k)$的方程组:

$$\nabla E(W(k+1))^T\nabla E(W(k))=0. \tag{2}$$

由式(2)可知,在学习过程中,前后两次的搜索方向是正交的,即所谓搜索中的"拉锯"现象,尤其是在学习的最后阶段,由于$\nabla E(W(k))\to 0$,"拉锯"现象更为剧烈,故收敛速度很慢.

2 超线性BP学习算法

考虑BP学习算法梯度下降法的如下形式:

$$W(k+1)=W(k)-\alpha(W(k))\nabla E(W(k)), \tag{3}$$

其中$\alpha(W(k))=\mathrm{diag}(\alpha_1,\alpha_2,\ldots,\alpha_N)$,$\alpha(W(k))$取法满足:(1) 当$E(W)$陡变时,$\|\alpha(W)\|$较小;(2) 当$E(W)$缓变时,$\|\alpha(W)\|$较大.我们取

$$\alpha(W(k))=\beta\cdot(E(W(k)))^{1/r}\cdot\mathrm{diag}\left(e^{-\left|\frac{\partial E(W(k))}{\partial W_1}\right|},e^{-\left|\frac{\partial E(W(k))}{\partial W_2}\right|},\ldots,e^{-\left|\frac{\partial E(W(k))}{\partial W_N}\right|}\right)\|\nabla E(W(k))\|^{-1}, \tag{4}$$

其中$\beta,r>0$为常数,则我们改进的BP算法迭代式为

$$W(k+1)=W(k)-\beta\cdot(E(W(k)))^{1/r}\cdot\mathrm{diag}\left(e^{-\left|\frac{\partial E(W(k))}{\partial W_1}\right|},e^{-\left|\frac{\partial E(W(k))}{\partial W_2}\right|},\ldots,e^{-\left|\frac{\partial E(W(k))}{\partial W_N}\right|}\right)G(W(k)), \tag{5}$$

其中$G(W(k))$为$E(W(k))$的单位梯度向量.

定义1. 若W^*为$W(k)$的极限点,定义误差,如$e(W(k))=\|W(k)-W^*\|$.设存在$\lim\limits_{k\to\infty}\frac{e(W(k+1))}{e(W(k))}=\gamma$,可知$\gamma\in[0,1]$.若$0<\gamma<1$,则称序列$\{W(k)\}$以收敛比为$\gamma$进行线性收敛;若$\gamma=0$,则称$\{W(k)\}$超线性收敛;若$\gamma=1$,则称$\{W(k)\}$次线性收敛.

引理1. 设$\Omega\subset R^n$为有界闭集,$W\in\Omega$,$C^{(2)}(\Omega)$为定义在Ω上的二阶连续函数空间,若$E(W)\in C^{(2)}(\Omega)$,则$\nabla E(W)$在Ω上有界.

引理2[3]. 设$W(k)\to W^*$,记$h(k)=W(k)-W^*$,$\delta(k)=W(k+1)-W(k)=h(k+1)-h(k)$,且$W(k),W^*,h(k),\delta(k)\in\Omega\subset R^n$,则$\{W(k)\}$超线性收敛等价于$h(k+1)=o(\|\delta(k)\|)$.

定理1. 当$r>2$时,由式(5)产生的迭代序列$\{W(k)\}$是收敛的,且对第2层权值序列来说,其收敛速度是超线性的.

证明:易证$E(W(k+1))<E(W(k))$,即$\{E(W(k))\}$单调下降.另一方面,$E(W(k))$有下界,故极限存在.并且$E(W(k))$关于W是一致连续的,于是,必存在W^*,使得$\lim\limits_{k\to\infty}W(k)=W^*$.即由式(5)产生的迭代序列$\{W(k)\}$是收敛的.

下面证明其收敛速度为超线性的.设$W(k)\in\Omega$,Ω为一有界闭集,由E的取法,$E(W(k))\in C^{(2)}(\Omega)$,由引理1可知,$\nabla E(W(k))$有界,即存在$M>0$,使$\|\nabla E(W(k))\|\leqslant M$,对任意的$W(k)\in\Omega$成立.另一方面,

$$\frac{\|h(k+1)\|}{\|\delta(k)\|}=\frac{\|W(k+1)-W^*\|}{\|W(k+1)-W(k)\|}\leqslant\frac{1}{\beta}\cdot e^{\max_i\left\{\left|\frac{\partial E(W(k))}{\partial W_i}\right|\right\}}\|W(k-1)-W^*\|(E(W(k)))^{1/r}$$
$$\leqslant\frac{1}{\beta}\cdot e^M\|W(k+1)-W^*\|(E(W(k)))^{1/r}.$$

当$r>2$时,考虑隐层至输出层为线性关系,$E(W)$为关于第2层权的二次函数,则$\|W(k+1)-W^*\|=o((E(W(k)))^{1/r})$,故上述不等式右端趋于0,即$\|h(W(k+1))\|=o(\|\delta(k)\|)$,由引理2可知,第2层权值序列是超线性收敛的. □

3 实验结果及对比情况

例 1:XOR 问题. 进行 100 次随机选取初值的重复学习,取平均值,结果见表 1.

Table 1 Comparison of learning results between super linear BP algorithm and traditional BP algorithm on XOR problem

表 1 XOR 问题超线性 BP 算法与传统 BP 算法的学习结果对比

Error precision①	Maximum learning steps②	Simple-BP algorithm③		Steepest algorithm⑥		Super-Linear-BP algorithm⑦	
		Learning steps④	Convergence ratio⑤(%)	Learning steps	Convergence ratio(%)	Learning steps	Convergence ratio (%)
0.1	2 000	1 170	66	177	80	75	73
0.01	4 000	1 754	80	499	90	328	77
0.001	8 000	2 206	89	1 221	86	548	82

①误差精度,②最大学习步数,③简单 BP 算法,④学习步,⑤收敛比,⑥最速下降算法,⑦超线性 BP 算法.

例 2:九点模式问题. 平面上 9 个点按如图 2 所示排列,分成◇和◆两类. 计算结果见表 2.

Fig. 2 Nine-Sample patterns

图 2 九点模式

Table 2 Comparison of learning results between supper linear BP algorithm and traditional BP algorithm on nine-sample patterns

表 2 九点模式问题超线性 BP 算法与传统 BP 算法的学习结果对比

Error precision①	Maximum learning steps②	Simple-BP algorithm③		Steepest algorithm⑥		Super-Linear-BP algorithm⑦	
		Learning steps④	Convergence ratio⑤(%)	Learning steps	Convergence ratio(%)	Learning steps	Convergence ratio (%)
0.5	5 000	1 505	80	367	90	46	100
0.1	10 000	3 386	25	1 835	20	694	60
0.02	20 000	10 547	15	2 398	50	1 790	30

①误差精度,②最大学习步数,③简单 BP 算法,④学习步,⑤收敛比,⑥最速下降算法,⑦超线性 BP 算法.

参考文献

1 Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation. In: Rumelhart D E, McClelland J L eds. Parallel Distributed Proceeding, Vol 1. Cambridge, MA: MIT Press, 1986

2 Xi Shao-lin. Non-Linear Optimization Method. Beijing: Higher Education Publishing House, 1992
(席少霖. 非线性最优化方法. 北京:高等教育出版社,1992)

Super-Linearly Convergent BP Learning Algorithm for Feedforward Neural Networks

LIANG Jiu-zhen[1] HE Xin-gui[2] HUANG De-shuang[2]

[1](*Department of Computer Science and Engineering Beijing University of Aeronautics and Astronautics Beijing* 100083)
[2](*Beijing Institute of System Engineering Beijing* 100101)

Abstract In this paper, some shortages of traditional BP learning algorithm are analyzed. To avoid these shortages, a modified BP learning algorithm is proposed. It is shown that this algorithm is super-linearly convergent under certain conditions. This algorithm can overcome some shortages of traditional BP learning algorithm, and has the same order of computation complexity as the traditional BP algorithm. Finally, two computing examples are given. Simulation results illustrate that this algorithm is highly effective and practicable.

Key words Feedforward neural network, BP learning algorithm, convergence, super-linear convergence.