基于一类非线性特性的 FNN 训练算法*

王 凌　郑大钟
(清华大学自动化系　北京　100084)

摘　要　针对BP 算法收敛缓慢和易于陷入局部极小的缺点, 将基于一类非线性特性的动量项引入 BP 算法的梯度搜索, 提出前向神经网络(FNN)的一种通用且简单的全局训练算法(IBPM 算法)。结合升温策略, 算法在优化精度和训练速度方面有较大的改善。典型算例的仿真验证了算法的有效性。

关键词　BP 算法, 动量, 非线性, IBPM 算法
分类号　TP 18

Training Algorithm for FNN Based on a Class of Nonlinear Property

Wang Ling, Zheng Dazhong
(Tsinghua University)

Abstract　To solve the slow convergence of BP algorithm and avoid getting stick in local minima, a general and simple training algorithm for feed-forward neural networks(FNN), named IBPM, which combines a class of nonlinear property of momentum term with the gradient descent of BP algorithm, is presented. Using "temperature-raising" strategy, convergence rate and training speed of the algorithm are greatly improved. Simulation results verify the efficiency of such algorithm, and some conclusions to choose parameters are provided.

Key words　BP algorithm, momentum, nonlinear, IBPM algorithm

1 引　　言

前向神经网络(FNN)因其优良的逼近能力和算法的简单性而得到广泛应用。误差反传算法(BP)[1]解决了 FNN 的训练问题, 对神经网络的发展起了很大作用, 但存在收敛速度慢, 易陷入局部极小, 网络推广性能差等缺点。目前的改进方法主要有改变学习率, 加动量项, 修改激励函数, 采用合适的训练模式(如逐一式、批处理、跳跃式), 引入全局优化技术(如模拟退火和遗传算法)[2]等。这些方法或是性能改善能力有限, 或是计算量和存储量增加巨大而应用不良。

本文在BP 梯度搜索中引入一类非线性特性的动量项, 提出 FNN 的一种简单且通用的全局训练算法(IBPM)。利用非线性特性, 训练过程在权空间能遍历局部极小, 并具有突跳特性(不同于模拟退火基于概率分布的突跳机制[2])。结合升温策略(即自适应控制非线性强度系数), 算法在优化精度和训练速度方面有较大的改善。仿真结果表明, 算法在确保优化精度的同时大大加快了训练速度。

2 IBPM 算法及其训练机制

BP 算法本质上是利用梯度信息来调整权值。在误差曲面较平坦处, 导数值较小使权值调整较小, 从而收敛缓慢; 在曲率较大处, 导数值较大使权值调整较大, 但会出现跃冲极小点现象而难以收敛。由于目标函数高维曲面存在多极小点, 且不能保证目标函数在权空间的正定性, 单一梯度下降法难免陷入局部极小。基于物理学中"系统下一运行状态同时取决

* 国家自然科学基金项目(69684001)和国家攀登计划基金项目
1998-09-18 收稿, 1998-11-23 修回

于当前状态的能量和动量"的思想,本文在BP梯度搜索中引入一类非线性特性的动量项(非线性函数$g(\cdot)$如图1),构成一种新的权值迭代式(BPM算法,如式(1)),使权值下一步的调整量同时依赖于当前目标函数相对于权值的导数和当前权值变化量的非线性作用。

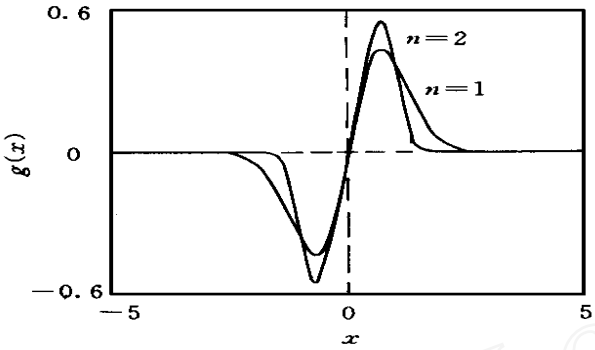


图1 函数 $g(x)=x\exp(-x^{2n})$

$$\begin{cases} w_{i,j}(k+1)=w_{i,j}(k)-\eta\partial E/\partial w_{i,j}(k)+\\ \qquad p(k)g[w_{i,j}(k)-w_{i,j}(k-1)]\\ g(x)=x\exp(-x^{2n}),\quad n=1,2,\cdots \end{cases} \tag{1}$$

其中,w 为权值,E 为目标函数,η为学习率,$p(k)$ 为非线性项强度系数。对于传统BP算法,$g(x)=0$;对于传统的带动量项BP(BPM)算法[1],$g(x)=x$。

鉴于$g(0)=0$,BPM迭代式保留了BP原先的不动点。由于函数$g(\cdot)$的非线性特性,$\Delta w(k)$较大时,即系统远离不动点(局部极小),$g[\Delta w(k)]$的作用较小,系统几乎以梯度下降方式趋于不动点;$\Delta w(k)$较小时,即系统已进入不动点的小邻域,系统由$g[\Delta w(k)]$得到较大的驱动,能够爬越能量波峰以克服局部极小而进入其它能量低谷,并最终趋于全局极小。因此,将基于这类非线性特性的动量项引入BP梯度搜索,权值具有快速和平稳的变化过程,增强了克服局部极小的能力,尤其能克服神经元激励函数饱和区和目标函数平坦区的影响,可大幅度改善学习速率和收敛性能。

为避免不合适非线性作用引起随机振荡使算法难以收敛,可通过自适应改变非线性作用项的强度得到较强的综合性能。本文算法保留了BP简单易实施的优点,仅增加少量计算和存储,算法本身的调节参数较一些改进BP算法和全局优化算法要少。

现以一维函数$E(x)$的优化问题为例来分析非线性项强度的影响。假定单一梯度法按$x(k+1)=f[x(k)]$迭代,令$y(k+1)=x(k+1)-x(k)$,则迭代式可转化为式(2)。若原迭代式的不动点x^*是稳定的,即满足$A=\left|\mathrm{d}f/\mathrm{d}x\right|_{x=x^*}<1$,则式(2)在不动点$(x^*,0)$处的Jacobian矩阵和特征方程为式(3)。

$$\begin{cases} x(k+1)=f[x(k)]+pg[y(k)]\\ y(k+1)=x(k+1)-x(k) \end{cases} \tag{2}$$

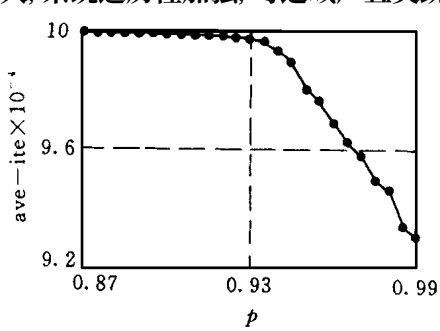
$$\begin{cases} J(x^*,0)=\begin{bmatrix} A & P\\ A-1 & P \end{bmatrix}\\ \lambda^2-(A+P)\lambda+p=0 \end{cases} \tag{3}$$

当$0\le p<1$时,两特征值的模恒小于1,$(x^*,0)$仍稳定;当$p=1$时,特征值为模等于1的一对共轭复根;当$p>1$时,特征值为模大于1的一对共轭复根或模大于1的一对实根,$(x^*,0)$将不稳定。因此,$p\ge 1$时系统将引入复杂的动态(分岔、混沌、周期解或发散)。尽管混沌可用来进行全局搜索[3],但神经网络的训练是高维且复杂的优化问题,$p\ge 1$时训练过程易分散,故本文仅讨论$0\le p<1$的情况。

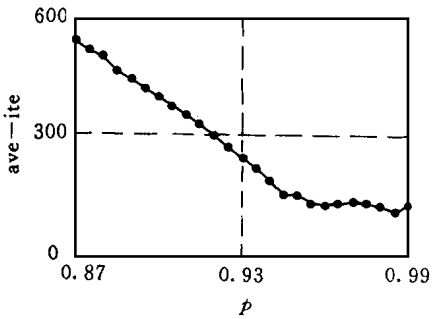
3 非线性强度系数影响的仿真研究

异或问题(XOR)是研究神经网络训练算法性能的典型算例[1],其优化曲面不规则且存在多局部极小,很难进行全局优化。采用{(0,0,0),(0,1,1),(1,0,1),(1,1,0)}样本,研究非线性强度系数p对BPM算法平均实际误差、平均迭代步数、局部极小个数的影响。每组参数均随机运行100次,2 000步未达到精度(10^{-3})则认为是局部极小。采用2-3-1网络结构(隐层和输入层分别采用$y=1/[1+\exp(-\mu x)]$和线形激励函数,并附加阈值单元。取$\mu=1$,$\eta=0.8$,$g(x)=x\exp(-x^2)$,初始权值取(-1.0,1.0)内随机值,仿真结果分别如图2,图3,图4所示。

由仿真结果可见:在研究范围内,强度系数大则算法精度高,迭代步数少,但局部极小个数有增加趋势。考察权值迭代方法,加大强度系数可使非线性项作用增大,系统遍历性加强,可达域广且突跳性强,



2 非线性强度系数对平均实际误差的影响



3　非线性强度系数对平均迭代步数的影响

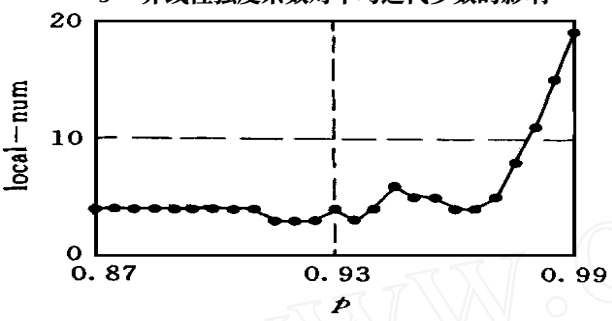


图 4　非线性强度系数对局部极小个数的影响

有利于全局搜索, 能提高精度并加速收敛, 但作用过强会使系统在某些局部极小之间振荡。为使算法有较好的综合性能, 强度系数应自适应选取。

4　带升温策略的 IBPM 算法及其仿真研究

鉴于强度系数对算法性能的影响规律, 设计“升温”策略

$$p(k) = \min\{0.988, p(k-1) \times \ln[e + \lambda(1 - p(k-1))]\}$$
$$p(0) = 0.9,\quad k = 1, 2, \cdots$$

训练初始阶段, 权值变化幅度较大, 较小的强度系数驱动系统在适当突跳性引导下较平滑地遍历局部极小, 逐步进入性能较好区域。随着训练的进行, 搜索可能陷入某些局部极小或平坦区, 以致权值变化较小, 但强度系数的增大使非线性项作用增大, 系统穿越平坦区和跳跃局部极小的能力加强, 能进一步趋于高精度解, 从而克服 BP 达到有限精度后难以全局收敛的缺点。

令 $\lambda = 0.05$, BPM 惯性系数取 0.7, 采用 2 - 3 - 1 网络结构, $g(x) = x\exp(-x^4)$, 其他参数同前。以 XOR 为例, 结合升温策略, IBPM, BP 和 BPM 各做 100 次随机仿真, 算法性能比较列于表 1。

可见, 结合升温策略的 IBPM 性能较 BP 和 BPM 有较大改善(高精度要求下训练速度提高了 1 000 余倍), 而 BP 和 BPM 很难达到 6 以上的精度指数。IBPM 在快速实现高精度优化的同时, 迭代步数随精度要求提高的增幅不大, 2 - 3 - 1 结构的迭代步数 s 与精度指数 n (对应于精度要求 10^{-n}) 的关系近似为 $s \propto e^{0.38n}$, 且 100 次随机实验中 96 次达到精度要求。若采用 2 - 4 - 1 结构, 迭代步数与精度指数的关系近似为 $s \propto e^{0.51n}$, 每次随机实验均能达到精度要求。

5　算法与网络参数的选择

以 XOR (精度指数 5) 来考察算法和网络参数对优化性能的影响。基准参数取 2 - 3 - 1 网络, 初始权值为 (-1.0, 1.0) 内随机值, $p(0) = 0.9$, $\mu = 1$, $\eta = 0.8$, $\lambda = 0.05$, $g(x) = x\exp(-x^4)$, 每组参数均作 100 次随机仿真。主要结论归纳如下:

1) 初始强度系数 $p(0)$ 的影响(表 2): $p(0)$ 增大使迭代式中动量项作用加大, 梯度项作用减小, 穿越误差曲面平坦区和克服局部极小的能力加强, 能提高收敛速度且降低误差, 但局部极小个数有增长趋势。$p(0)$ 过小使系统处于“低温”阶段过长, 影响收敛速度。相反, 动量项作用过强使系统以一种畸变的梯度下降, 易造成小幅振荡而难以收敛。为避免早熟收敛, $p(0)$ 不能选得过大; 为加快收敛速度, $p(0)$ 不能选得过小。

表 1　算法性能比较

精度要求	IBPM 算法		BP 算法		BPM 算法	
	平均误差	平均步数	平均误差	平均步数	平均误差	平均步数
1.0e-1	8.904 887e-2	79	9.936 103e-2	1 052	9.785 501e-2	837
1.0e-2	8.783 803e-3	90	9.986 716e-3	1 477	9.956 307e-3	958
1.0e-3	9.143 957e-4	105	9.998 319e-4	4 117	9.994 630e-4	1 855
1.0e-4	9.415 675e-5	128	9.999 833e-5	31 238	9.999 419e-5	9 885
1.0e-5	9.554 921e-6	164	9.999 982e-6	278 614	9.999 941e-6	85 969
1.0e-6	9.602 23e-7	225		∞		∞
1.0e-7	9.711 64e-8	293		∞		∞

表2 初始强度系数对算法性能的影响

$p(0)$	平均步数	平均误差 × e-6	极小个数
0.8	256	9.832 436	5
0.85	214	9.750 663	5
0.9	164	9.554 921	4
0.92	159	9.572 273	8
0.95	146	9.321 678	14

2) 学习率η的影响(表3): η过小导致目标函数曲面平坦区的权修正量太小, 以致收敛缓慢; η过大则使曲面曲率较大处权修正量较大而出现跃冲现象, 引起振荡以致算法难以收敛。由于BPM中没有改变BP搜索的主框架, 可采用BP动态选择学习率等改进方法[1]。

表3 学习率η对算法性能的影响

η	平均步数	平均误差 × e-6	极小个数
0.1	387	9.701 493	7
0.7	171	9.614 802	5
0.8	164	9.554 921	4
0.9	164	9.621 696	5
0.99	151	9.605 638	7

3) 升温因子λ的影响(表4): λ加大即加快升温, 能提高收敛速度和降低实际误差, 但局部极小个数增多。若升温过快, 则权值变化轨道不太平滑, 动量项作用易引入不利的振荡而增加收敛到局部极小; 若升温过慢, 则系统处于"低温"阶段过长, 不利于优化的快速性。

表4 升温因子λ对算法性能的影响

λ	平均步数	平均误差 × e-6	极小个数
0.02	785	9.964 519	3
0.04	192	9.723 045	4
0.06	151	9.488 599	6
0.08	151	9.395 658	8

4) 初始权选择范围的影响(表5): 选择范围决定初始点的分布, 范围适当有利于收敛速度的影响。若范围过大将增加优化曲面中局部极小个数, 神经元总输入易进入激励函数饱和区, 从而影响算法性能。相反, 范围过小将导致权值的差异较小, 势必影响收敛速度。

表5 初始权选择范围对算法性能的影响

范 围	平均步数	平均误差 × e-6	极小个数
(-0.1, 0.1)	508	9.371 699	1
(-0.5, 0.5)	186	9.408 937	3
(-1.2, 1.2)	161	9.539 729	5
(-2, 2)	182	9.768 926	10

5) 激励参数μ的影响(表6): 若神经元总输入x进入激励函数$y = 1/[1 + \exp(-\mu x)]$的非线性饱和区, 且实际输出与期望输出不一致, 则激励函数较小的导数值将使权值变化幅度较小, 产生"平台"现象而导致收敛缓慢, μ增加将使饱和区函数导数值增加, 有利于收敛速度的提高, 但神经元总输入脱离饱和区的范围减小, 致使局部极小个数增多。

表6 激励参数μ对算法性能的影响

μ	平均步数	平均误差 × e-6	极小个数
0.8	206	9.573 426	4
1	164	9.554 921	4
1.5	139	9.612 677	10
2	123	9.610 901	17

6) 网络结构的影响(表7): 隐节点数少于样本个数时, 通常会产生局部极小; 隐节点数多于样本个数时, 优化曲面的维数增加, 使得网络能鉴别各样本, 很少收敛到局部极小, 但计算和储存量增加, 不利于实际设计。随着隐节点数的增加, 迭代步数减少且极小个数也减少。

表7 网络结构对算法性能的影响

隐节点数	平均步数	平均误差 × e-6	极小个数
2	174	9.543 328	25
3	164	9.554 921	4
4	156	9.585 256	0
5	138	9.618 818	0

此外, 相对于BP算法, BPM仅增加了$p(0)$和λ参数, 比一些改进BP算法和全局优化算法的调节参数少, 应用简单方便, 这也是工程上受欢迎之处。

6 结 语

本文在BP算法中引入一类非线性特性的动量项, 并结合升温策略, 使其对FNN的训练具有快速性和全局性。文中通过仿真验证了算法的有效性, 并归纳了一些选择参数的规律性结论。本文算法的实际应用及其理论与参数选择尚需进一步研究和完善。

参 考 文 献

1 焦李成. 神经网络系统理论. 西安: 电子科技大学出版社, 1992

2 王凌, 郑大钟. 前向网络的两种混合学习策略. 清华大学学报, 1998, 38(9): 95—97

3 Zhou C S, Chen T L. Chaotic annealing for optimization. Physical Review E, 1997, 55(3): 2580_2587

作者简介见本刊1998年第13卷第1期第82页。