第16卷第5期 红外与毫米波学报 Vol. 16, No. 5
1997年10月 J. Infrared Millim. Waves October, 1997

一种采用基于知识的优化过程的BP算法研究*

朱江海 戚飞虎
(上海交通大学计算机科学与工程系,上海,200030)

摘要 针对BP算法的固有缺点,提出一种实用有效的改进算法.此算法在每次得到的搜索方向上,都进行一维优化,从而解决了BP算法须人工由经验选取固定步长而带来的弊病;针对误差函数高度非线性的特征,改进算法采用基于知识的处理过程,全面利用每一步计算所得到的误差函数值和导数进行误差函数曲面地形判断,指导搜索计算,使算法具有极好的收敛稳定性,同时大大提高了收敛速度.

关键词 BP算法,神经网络,学习,一维优化,梯度.

引言

多层前馈神经网络结构简单而功能强大,它的常用学习算法是BP算法.BP网络已被广泛应用于各个领域,如用于分类、函数逼近[1]、模式识别[2]、字符识别[3]、以及用于控制系统[4]等.但是,BP算法有许多固有的缺点[5],如训练时需要给出隐层节点个数,学习速度慢,有可能收敛于局部极小点[12]等.文献[6]给出了对网络剪枝的方法,文献[7]对BP网络隐层节点个数的确定方法进行了研究,文献[8]对激励函数进行修改以改善BP算法的收敛稳定性,文献[2]则先构造一个决策树,而后导出一个对应的多层网络,而使训练过程较快.

本文给出一个采用基于知识的优化过程的BP算法,它能够有效地提高算法的收敛速度和收敛稳定性,而又不需要改变激励函数和目标函数的形式,也不需要用复杂的方法事先构造一个决策树,因而有更好的通用性.

传统的BP算法如下:设给定了 P 个样本 (x_p, y_p), $p=1,2,\ldots P$. E_p 为第 p 个样本的误差函数,BP算法的权值迭代公式为

$$\omega(k+1)=\omega(k)+\Delta\omega(k), \tag{1}$$

$$\Delta\omega(k)=-\eta\frac{\partial E}{\partial\omega(k)}. \tag{2}$$

其中 η 为步长, $\partial E/\partial\omega(k)=\sum_p \partial E_p/\partial\omega(k)$, $k=0,1,2\ldots$ 为迭代次数.

在传统的BP算法中,学习的关键问题在于步长 η 的选择,它对收敛速度和收敛结果有很大影响. η 过小,会在误差曲面平坦区域妨碍迭代合理步进; η 过大,迭代可能会在峡谷两

* 国防预研基金(编号96J2.4.2)及国家自然科学基金(编号69572026)资助项目
稿件收到日期1996-08-12,修改稿收到日期1997-06-27

边跳跃，产生振荡. 由于BP算法中迭代步长是固定的，很难选取一个能够自始至终适合迭代运算步长，使得BP算法在实际使用中效果不佳. 文献[9]证明在BP算法中，使用一个常数的学习率，将不能保证算法在一般意义下均能收敛. 于是，人们想到使步长随迭代过程而变小[10]，即步长在迭代刚刚开始时取得较大，以使误差较快下降；随着迭代次数增加，步长逐渐减小. 这些方法虽然有一定的效果，但仍有很大的盲目性，起始步长和步长减小速度仍然要靠经验来选取. 文献[11]将步长优化引入了BP算法中，但只使用了二次函数逼近，而且没有考虑误差函数多峰的可能性，因此在实际使用中计算效果不好. 本文提出了基于知识的搜索算法. 在每步的迭代过程中，它根据在此之前已经得到的点和当前点的误差函数的值和导数，对误差函数曲面的地形做出判断，然后继续搜索. 由于基于知识方法的引入，可以很好地解决在各种复杂地形下搜索时遇到的困难，使算法有很高的效率.

1 采用基于知识的优化过程的BP算法

针对BP算法缺陷，本文提出一种改进算法. 在轮过整批样本后，得到本次前进方向为

$$S=\frac{\partial E}{\partial \omega(k)}=-\sum_{p}\frac{\partial E_p}{\partial \omega(k)}. \tag{3}$$

然后，在S方向上做一维优化，找出此方向上的一维极小点

$$\Delta\omega(k)=\eta_{opt}S. \tag{4}$$

这样就解决了传统BP网络中步长过大过小产生的问题. 在误差曲面的平坦地带，迭代可以大步前进；在狭谷地带，迭代将会小步搜索. 各种情况下均可得到最优迭代效果.

针对误差函数高度非线性的特征，改进算法在一维优化的实现中主要考虑以下两个方面：(1) 增加安全措施，保证每步都得到一个改进解，而不致于使迭代变得发散；(2) 采用基于知识的处理方法，充分利用误差函数值与导数的信息对误差函数曲面地形进行判断，指导搜索计算，逼近极小点.

1.1 基于知识的一维优化方法

一维优化过程[13~15]首先要进行区间确定，即做初步粗略搜索，确定最优值存在的区间.

1.1.1 区间确定

误差函数用f来表示，以ga、gb分别代表$f'(a)$、$f'(b)$在S方向上的分量. 在搜索方向S上，找出使$ga<0$且$gb>0$的区间，则最优值必然存在于(a,b)区间之内. 区间确定过程典型的做法是采用逐渐扩展的模式. 它的第$k+1$个试验点b由以下递推公式确定：

$$x_{k+1}=x_k+2^k\Delta,\quad k=0,1,2,3,\ldots \tag{5}$$

这里$a=x_0$是搜索起始点，Δ是所选的适当大小步长参数. 当$gb<0$时，区间不断扩展，直到$gb>0$为止. 采用逐渐扩展的方法后，可以使迭代点迅速到达误差曲线谷底，大大加快收敛速度. 区间确定的过程如图1所示，最后得到区间确定结果为$a=x_0$，$b=x_{k+1}$.

考虑到误差函数的高度非线性，须对上述区间确定方法做改进. 从起始点a点，搜索方向S后，按步长扩展，得b点. 考虑a、b点的函数值和导数值组合：函数值有$fb<fa$或$fb>fa$两种情况，导数值$ga<0$，而gb有小于或大于零两种情况，共4种组合情况，如图2所示.

图 2(a)当 $ga<0$ 且 $gb>0$，认为函数在(a,b)上为单峰，且最优解就在(a,b)区间上，区间确定完毕；图 2(b)当 $gb<0$ 且 $fb<fa$，则将 a 点前移到 b 点，并将步长扩大，继续做区间确定；图 2(c)当 $gb<0$ 且 $fb>fa$，则函数在(a,b)上必为多峰. 仍以 a 点为起始点，将步长缩短，继续进行区间确定.

以上改进算法中，图 2(a)与原方法相同；(b)比原方法向前推进了一步，使最后的区间确定结果为 $a=x_k, b=x_{k-1}$，与原来结果 $a=x_0, b=x_{k+1}$相比较，区间大大缩小了；情况(c)的情

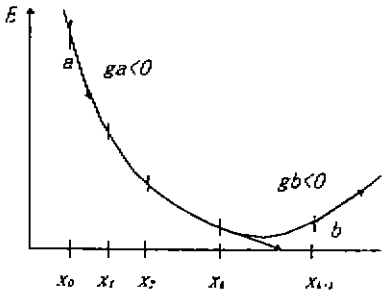

图 1　区间确定过程

Fig. 1 The process of interval determination

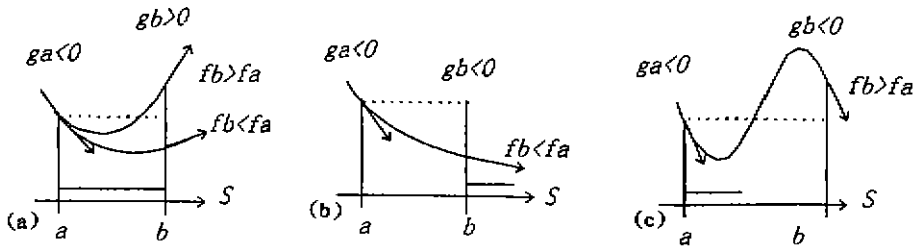

图 2　改进的区间确定方法

Fig. 2 Interval determination by the proposed method

况下若用原方法会在下一步时扩大步长，而改进后为缩小区间，这样就更适应多峰情况了. 为安全起见，步长扩大和步长缩短的比为无理数($\pi:2$). 并且限制区间确定的重复执行次数，当大于一定次数时，跳出本次一维优化，转入下次的搜索方向计算.

经区间确定后，就可以用较精确的方法，如三次插值法或二等分法求得最优解估计值.

1.1.2　三次插值法

当函数足够光滑，并且在区间上单峰、连续时，则可以用一个足够高阶的多项式来预测最优点的位置. 三次插值法是最常用和最有效的多项式近似法. 如图 3 所示，为求出 Q 值，可进行如下运算：

$$z=3(fa-fb)/(b-a)+ga+gb, \tag{6}$$

$$w^2=z^2+ga+gb, \tag{7}$$

$$Q=1-(gb+w+z)/(gb-ga+2w). \tag{8}$$

其中 fa、fb 分别代表 $f(a)$、$f(b)$，ga、gb 分别代表 $f'(a)$、$f'(b)$在 S 方向上的分量. 可以证明 $0\leqslant Q\leqslant 1$. 这就保证了求得的最优点估计值在$[a,b]$区间内.

1.1.3　二等分法

二等分法是一种区间减缩法，当函数值和一阶导数值可求时，它在区间减缩法中效率最

高.区间确定后,得 $ga<0$ 且 $gb>0$,取 $x=(a+b)/2$,计算 x 点函数值和导数值,若 $f'(x)>0$,保留(a,x),删去(x,b);若 $f'(x)<0$,则保留(x,b),删去(a,x),如图 4 和 5 中阴影所示.

二等分法不仅可以用于单峰的情况,也可用于多峰情况.因此,如果三次插值法失败,即得到的 x 比 a、b 还大,则可判定区间上为多峰情况,转用二等分法来做最优化.

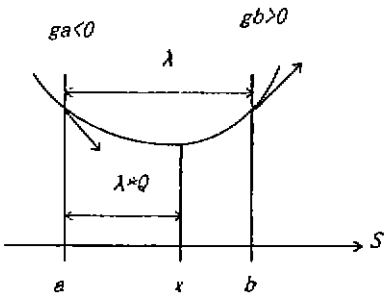

图 3 三次插值法
Fig. 3 Cubic interpolation

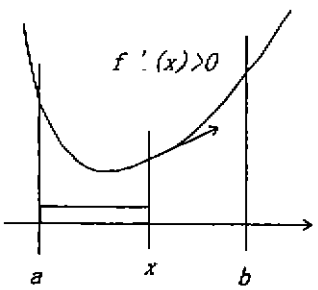

图 4 二等分法($f'(x)>0$)
Fig. 4 Bisection algorithm

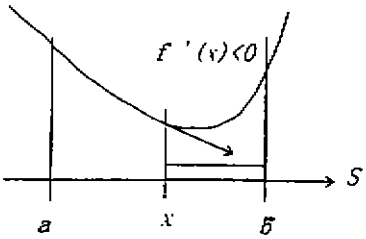

图 5 二等分法($f'(x)<0$)
Fig. 5 Bisection algorithm

针对高度非线性,须对二等分法做改进.原二等分法中每步计算中只利用了导数信息,而没有利用误差值的信息.改进后不仅利用导数信息,而且利用误差值信息对误差函数曲面地形进行判断,以减少误差函数计算次数,并且提高收敛性能.

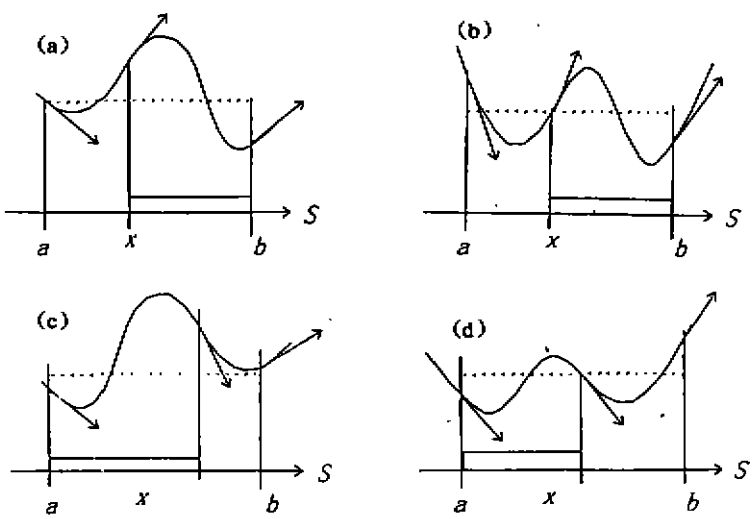

图 6 改进的二等分法
Fig. 6 Bisection algorithm by the proposed method

在区间确定后,已得到 $ga<0$ 且 $gb>0$.考虑 fa 与 fb 比较,有 2 种情况:fa 比 fb 大或 fb 比 fa 大;$f(x)$与 fa、fb 的比较,有 3 种情况:$f(x)$比 fa、fb 都小,$f(x)$在 fa、fb 之间或 $f(x)$比 fa、fb 都大;$f'(x)$有 2 种情况:大于 0 或小于 0,组合起来共有 12 种情况.考虑多峰

可能性时，其中的 4 种组合情况不应使用原来的二等分法，即简单地在 $f'(x)>0$ 时保留$[a, x]$，而 $f'(x)<0$ 时保留$[x, b]$. 改进的二等分算法如图 6 所示，图中阴影的部分为应当保留的区间. 上述 4 种情况做改进二等分法后，将不再保持 $ga<0$ 且 $gb>0$ 的特性，继续做改进二等分法时，要考虑由此而出现的所有的可能性.

1.2 采用基于知识的优化过程 BP 算法的实现

在实现中，我们将上述各种情况下的前后迭代点处的误差函数值和导数的信息以及每种组合情况所代表的地形下相应的搜索策略一起存在知识库中，这样，知识库就包括了所有在计算中可能遇到的地形情况. 在每次迭代计算后，将前后迭代点处的误差函数的值和导数信息输入知识库，就可以得到对当前地形的判断，然后指导搜索的进行.

2 实验及结果分析

以 XOR 问题为例，取三层前馈神经网，输入层、隐层和输出层的神经元个数为 2、2、1，初始权值取(−0.1，0.1)间的随机值，激励函数为单极 Sigmoid 函数，迭代控制精度为 10^{-4}. 表 1 所示为两种算法在不同步长下随机设置权值初值后的 10 组运行结果的平均值. BP 法和改进算法解 XOR 问题的误差收敛曲线如图 7 和 8 所示.

表 1　XOR 问题

Table 1 XOR Problem

	BP 算法					改进算法				
η	0.1	0.3	0.6	0.8	0.9	0.1	0.3	0.6	0.8	0.9
平均次数	31209.4	32164.4	32726.0	32523.4	28508.6	43.0	53.0	28.4	48.8	42.4
平均时间	30.59s	31.01s	32.23s	31.9s	27.79s	1.65s	2.00s	1.45s	1.78s	1.51s
平均误差	0.00130	0.00040	0.00013	0.0001	0.0001	0.0001	0.001	0.0001	0.0001	0.0001
控制精度	未达到	未达到	未达到	达到	达到	达到	达到	达到	达到	达到

由表 1 可见，BP 算法对步长取值很敏感，当 $\eta=0.9$ 时收敛较快，平均为 28,508 步，时间为 27.79s；而 η 取 0.1 时，迭代超过 32,000 步(时间 30s 以上)，仍未收敛到 10^{-4}. 改进算法对起始步长不敏感，在 $\eta=0.1, 0.3, 0.6, 0.8, 0.9$ 时，平均为 28.4～53.6 步，时间 1.45～2.00s. 改进算法与 BP 算法比较，在迭代次数上少于 1/672，时间上少于 1/18.

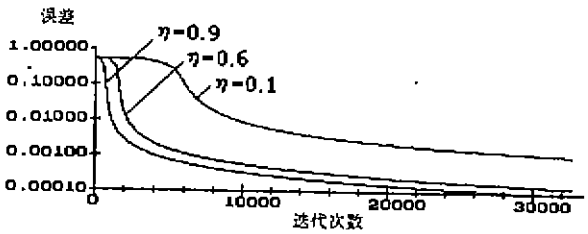

图 7　BP 求解 XOR 问题收敛曲线

Fig. 7 Convergence curve of XOR problem solved by BP algorithm

从图7可以看到，BP算法在取 $\eta=0.1,0.6,0.9$ 时，收敛速度不同. 小步长 $\eta=0.1$ 时，误差函数收敛很慢，5000～6000步后才开始较快下降；而步长较大，取 $\eta=0.9$ 时，误差函数在几百步后便开始较快下降，但在1000步后下降速度就变慢了. 这是因为迭代计算已达谷底，而步长过大，使得迭代在谷底两边来回跳跃，产生振荡，不能很快收敛. 若控制精度取 10^{-2} 或 10^{-3}，则BP法可在数百步或千余步收敛. 由此可见，BP算法越靠近极小值点，收敛速度越慢，尤其在控制精度取较小值的时候. 正是由于BP算法中步长固定，使得迭代中收敛速度快和迭代稳定不振荡的要求不能同时兼顾.

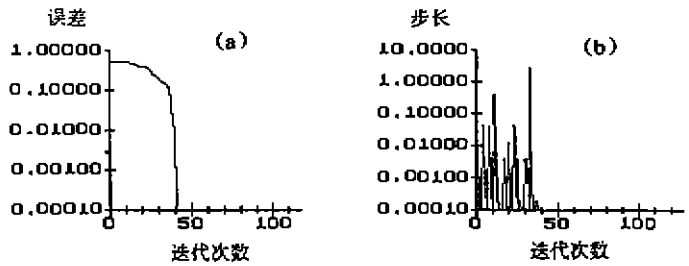

图8 改进算法求解XOR问题
(a) 收敛曲线($\eta=0.1,0.6,0.9$) (b) 步长变化曲线
Fig. 8 XOR problem solved by the proposed algorithm
(a) convergency curve (b) variation curve of searching steps

图8(a)表示的是改进算法的收敛曲线. 改进算法在取 $\eta=0.1,0.6,0.9$ 时，收敛曲线基本相同. 误差函数自始至终都能以很快的速度下降，收敛速度快和迭代稳定的要求得以兼顾. 控制精度越高，改进算法的优势越大. 图8(b)给出步长的变化曲线. 迭代过程中步长变化达3～4个数量级，与BP算法的固定步长有明显的区别.

3 结论

本方法较成功地解决了BP算法须由经验选取固定步长而带来的弊病，大大提高了算法收敛速度. 针对误差函数的高度非线性特征，本方法采用基于知识的处理过程，尽可能多地全面利用误差函数值和导数的信息，使收敛速度和稳定性得到了极大提高. 同时，作为知识库输入所需要的误差值和导数本来就是传统BP算法计算的一部分，故改进算法的计算量没有明显的增加，需要额外计算的部分仅仅是查询知识库而得到对下步迭代的指导. 该算法已成功地应用于字符识别神经网络的训练，取得了很好的效果. 实践证明，该算法有良好的收敛速度和收敛稳定性.

该改进算法也可以与冲量法或共轭梯度法相结合使用. 后者在搜索中不断对前进方向 S 进行修正，然后改进算法在 S 方向上进行基于知识的一维优化. 两者结合，可望使整个算法效率得到更大的提高.

REFERENCES

1 RUCK D W, ROGERS S K. *IEEE Trans. Neural Networks*, 1990, **1**: 296
2 Brent R P. *IEEE Trans. Neural Networks*, 1991, **2**: 346
3 Fukushima K, Wake N. *IEEE Trans. Neural Networks*, 1991, **2**: 355
4 Antsaklis P J. *IEEE. Neural Networks*, 1990, 1: 242
5 鲍立威. **模式识别与人工智能**(BAO L W. *Patteern Recognition and Artificial Intelligence*), 1995, **8**: 1
6 Karnin E D. *IEEE Tran. Neural Networks*, 1990, **1**: 239
7 Huang S C, Huang Y F. *IEEE Tran. Neural Networks*, 1991, **2**: 47
8 Fukumi M, Omatu S. *IEEE Tran. Neural Networks*, 1991, **2**: 535
9 Kuan C M, Hornik K. *IEEE Tran. Neural Networks*, 1991, **2**: 484
10 李艳斌，戚飞虎，等. **无线电工程**(LI Y B, QI F H, *et al. Radio Engineering*), 1995, **25**: 6
11 孙杳如. **微型计算机**(SUN Y R. *Micro Computer*), 1995, **15**: 38
12 孙德保，高　超，等. **工程最优化——方法与应用**，北京：北京航天大学出版社(SUN D B, GAO C, *et al. Optimization of Engineering—Method and Application*. Beijing: Beijing Space Aeronautical Univ. Press), 1990
13 孙　兵. **工程最优化——方法与应用**，北京：北京航天大学出版社(SUN B. *Optimization of Engineering—Method and Application*. Beijing: Beijing Srace Aeronautical Univ. Press), 1990
14 卜英勇. **优化设计方法**，南京：中南工业大学出版社(BO Y Y. *Optimized Design Method*. Nanjing: Zhongnan Univ. of Technology Press), 1986
15 万耀青. **最优化计算方法——常用程序汇编**，北京：工人出版社(WAN Y Q. *Optimized Calculation Method—Commonly Used Programs*, Beijing: Worker's Press.), 1983

A NEW BACKPROPAGATION ALGORITHM WITH KNOWLEDGE-BASED OPTIMIZING PROCESS*

ZHU Jian-Hai　QI Fei-Hu

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shainghai 200030, China)

Abstract　An effective algorithm to deal with the shortcomings of backpropagation was proposed. The proposed algorithm makes one-dimensional optimization after getting searching directions. By using this method, the shortcomings of fixing learning rate with an empirical value in traditional BP can be overcome, and the rate of convergence can be improved obviously. According to the highly nonlinear feature of the error function, the algorithm makes use of both the value and derivative of error function to predict the shape of surface, and conducts the searching process knowledge-basedly. This method makes the algorithm converge stably and speedily.

Key words　backpropagation algorithm, artificial neural network, learning, one-dimensional optimization, gradient.

* The project supported by the Foundation of Preliminary Research in National Defence and the National Natural Science Foundation of China
Received 1996-08-12, revised 1997-06-27