第25卷 第2期 力 学 进 展 Vol. 25 No. 2
1995年5月25日 ADVANCES IN MECHANICS May 25, 1995

多层前向网络研究进展及若干问题*

董 聪 郦正能 夏人伟 何庆芝
北京航空航天大学，北京 100083

提要 本文概述了多层前向网络研究的发展历史，对其中有代表性的若干成就进行了较为系统的介绍和评论，分析了当前研究工作中存在的一些问题，提出了解决这些问题的几种可行方案。在对多层前向网络的有效逼近机理进行深入剖析的基础上，提出了合理的有限规模多层前向网络应当遵循的若干构造原则。

关键词 多层前向网络；机理分析；最小二乘逼近；学习算法

1 引 言

人工神经网络（Artificial Neural Network，ANN）是由大量简单的处理单元，以某种拓扑结构广泛地相互联接而构成的复杂非线性动力学系统。它是在对以人脑为主要代表的生物神经系统的组织结构和行为特征进行研究的基础上提出的，它更侧重于对人脑某些特定功能的模拟。强调大量神经元之间的协同作用和通过学习的方法解决问题是人工神经网络的重要特征[1-5]。

诞生于本世纪40年代的人工神经网络方法，在经历了近50年的发展演变之后，其应用如今已渗透到各个领域。在智能控制，模式识别，计算机视觉，自适应滤波和信号处理，非线性系统辨识及非线性系统组合优化等领域已经并正在取得令人鼓舞的成就[6-10]。

对整个神经网络的发展和应用状况进行系统的介绍和评论不是一篇短文所能做到的。鉴于此，本文只想对人工神经网络中的一个非常重要的分支——多层前向网络的研究状况及其存在的问题进行深入的剖析。选择前向网络为研究对象，不仅因为它是目前应用最为广泛的网络之一，也因为对前向网络的映射能力和学习算法的研究相对来讲进行得较深入、彻底。因此，对前向网络研究与应用中存在的问题进行深入剖析，不仅有利于前向网络本身理论与应用研究工作的展开和深入，而且从中获取的经验和启示也可为其它类型神经网络的理论与应用研究工作提供有价值的参考。

* 中国博士后基金资助项目

2 多层前向网络研究概况

多层前向网络（Multilayer Feedforward Neural Network，MFNN）是在对简单感知机模型的能力及其缺陷进行研究的基础上提出的，它是具有明显层次结构的网络模型．网络由输入层、输出层和若干个隐层组成，网络之间通过神经元（节点）顺序单向联接（图1）．每一层神经元只接受前一层神经元的输入，并在节点上进行复合（线性叠加）和畸变（非线性映射）．通过复合反映不同神经元之间的耦合作用和耦合强度（由相对权值来表征），通过畸变改变输入信息的结构和性态．

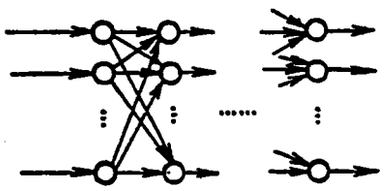

图1　多层前向网络

多层前向网络的研究已有很长的历史．虽然Amari，Brysop，Werbos和Parker都对多层前向网络学习算法的形成做出过重要贡献[11-14]，但真正使人们重新认识到多层前向网络可能具有重大潜在应用价值的是Rumelhart等人的杰出工作[15]．在《Parallel Distributed Processing》一书中，他们不仅从认识科学的角度出发，分析了隐单元训练作为知识内部表达的重要作用，还提出了误差反向传播算法（Back-Propagation，BP）以解决隐单元的训练问题．他们采用BP模型进行了大量的数值仿真实验，成功地解决了XOR问题，对称性检验问题和字称问题等．他们的研究工作向人们展示了多层前向网络用于描述和处理复杂非线性问题的巨大潜力．

多层前向网络以其优良的适应能力在许多领域得到了广泛的应用．目前，多层前向网络已经成为人工神经网络中最重要、最富有代表性的模型之一[16]．

对多层前向网络的研究主要沿着3个方向在迅速地发展和展开．①关于理想网络映射能力的数学证明；②关于有限规模网络逼近机理的研究；③学习算法．

3 理想网络映射能力的数学证明

人们对多层前向网络映射能力的认识首先来源于大量的仿真实验和工程实践中的感性认识．在应用多层前向网络解决模式识别问题的研究中，Lippmann通过大量的仿真实验得出结论[17,18]：采用3层网络可以形成若干个复杂的决策域；而采用4层网络则可以形成任意复杂的决策域．Wieland和Leighton则通过一个具体的例子说明3层网络具有把指定空间分解为若干个子空间的能力[19]．信号分类和图像分割实际上是一种广义的数学映射．因此，大量的仿真结果很自然地给人们一种启示：多层前向网络可能具有实现任意复杂非线性映射的能力．

由感性认识上升为理性认识是人类认识发展的必然趋势．从多层前向网络的特定结构出发，采用严格的数学方法分析理想网络的映射能力成为随后研究工作的一个主要焦点．

Funahashi[20]，Hornik[21,22]和Cybenko[23]等人对理想网络的映射能力进行了严格的数学证明，其主要成果可概括为以下3条定理：

定理1　(Funahashi)[20]：设$h(x)$为有界单调递增连续函数，K为R^n的紧致子集（有界闭子集），$f(X)=f(x_1,x_2,\cdots,x_n)$为$K$上的实值连续函数，则对任意$\varepsilon>0$，存在整

数N和实常数α_{ij}，θ_j，β_j($i=1,2,\cdots,n$；$j=1,2,\cdots,N$)，使

$$\hat{f}(x_1,x_2,\cdots,x_n)=\sum_{j=1}^{N}\beta_j h\left(\sum_{i=1}^{n}\alpha_{ij}x_i-\theta_j\right) \tag{1}$$

满足

$$\max_{X\in K}|\hat{f}(x_1,x_2,\cdots x_n)-f(x_1,x_2,\cdots,x_n)|<\varepsilon \tag{2}$$

即对$\varepsilon<0$，存在一个3层前向网络，其输入节点和隐节点分别为n和N个，网络的输入-输出关系为$\hat{f}(x_1,x_2,\cdots,x_n)$，隐层的节点函数为$h(\cdot)$，输入层和输出层的节点函数为线性的，使得(2)式成立。

Funahashi通过证明一个能够近似表达任意给定连续映射的3层网络的存在定理，解决了多层前向网络的映射能力分析问题。该定理可以直接推广到多输出的k层($k\geqslant 3$)网络中去。由于n维线性空间V^n上不同范数之间存在等价性，因此，上述定理也可以很自然地推广到任意范数意义下的逼近问题中去。

需要说明的是，Hornik等人的研究方法和Funahashi是不同的，他们不是去研究一个具体网络的特性，而是去研究具有某种共同属性的网络集合的特性。他们首先定义一个多层前向网络集合，再对隐节点的特性赋予某种假设条件，然后证明该集合在$L^p(R^n,\mu)$空间和$C(K)$空间中的稠密性，从而说明了多层前向网络集合在L^p范数意义下，对任意L^p下有定义的给定映射的近似表达能力。

考虑一个含有n个输入、1个输出、N个隐节点的单隐层网络集合：

$$\Omega_n^{(N)}(h)=\left\{\hat{f}:R^n\to R^1\,|\,\hat{f}(X)=\sum_{j=1}^{N}\beta_j h\left(\sum_{i=1}^{n}\alpha_{ij}x_i-\theta_j\right)\right\}$$

其中，$h(\cdot)$为隐节点函数，α_{ij}和β_j分别为输入层到隐层和隐层到输出层的联接权值，θ_j为隐节点的阈值。

当隐节点的个数N是任意大的数时，该集合可描述为：

$$\Omega_n(h)=\bigcup_{N=1}^{\infty}\Omega_n^{(N)}(h)$$

对此集合，有如下定理：

定理2 (Hornik)[22]：若$h(\cdot)$是非常值且有界，则对所有R^n上的有限测度μ，$\Omega_n(h)$在$L^p(R^n,\mu)$中稠。

定理3 (Hornik)[22]：若$h(\cdot)$非常值且有界连续，则对R^n中的所有紧致子集K，$\Omega_n(h)$在$C(K)$中稠。

由于$L^p(R^n,\mu)$收敛意味着在L^p范数的条件下按测度μ收敛，因此定理2意味着，只要$h(\cdot)$非常值且有界，则R^n中的所有可测函数都能够用$\Omega_n(h)$按测度μ逼近。

前面已经说过，Hornik等人的证明是针对一类网络集合而不是针对一个具体的网络。在使用上，我们只可能使用一个具体的网络，不管它是理想网络还是有限规模网络，而不可能去使用一个网络集合。因此，Hornik定理中对隐节点函数特性的放宽条件并不适用于具体网络。这一点应当引起应用研究人员的注意，

Funahashi 和 Hornik等人的研究成果向人们展示了多层前向所网络蕴涵的具大潜能，为网络的实际应用奠定了坚实可靠的理论基础．然而，这些研究成果并未能向人们展示应当如何去构造一个高效的网络；应当如何根据具体问题选择合适的节点函数；多隐层网络的跨层联接会对网络的实际逼近能力产生怎样的影响；增加隐层数目和增加隐节点数目在改进逼近效果方面存在什么本质差异，各自的作用机制是什么；对于有限规模的离散点集，怎样的采样结构才是合理的；神经网络逼近与传统逼近方法的主要区别在哪里；……．数学家们只对抽象的存在问题感兴趣．因此 White 曾经写道："我们的研究结果表明：形成多层前向网络一般逼近能力的本质原因在于，具有足够多的处理单元的各层之间的相互作用，各层上处理单元特性的特定选择是次要的……"．作为一个数学家，White 是对的．但应用研究人员却必须考虑如何去构造一个现实可行的网络以解决手头的具体问题．他会敏感地意识到节点特性的特定选择和网络的联接结构形式会对逼近效果产生巨大的影响；他会发现增加隐层不仅能够改善网络对已有离散点集的逼近效果，而且能够改进网络的预测能力……．

从大量的仿真实验中人们获得了宝贵的经验，他们尝试着将这些经验知识升华为一种构造性的而不是存在性的指导规则，他们想了解神经网络神奇逼近能力产生的内在机制是什么，采用什么样的措施可以强化这种能力．

在有限规模的网络结构，有限数量的离散采样集合的特定条件下研究网络逼近能力的内在机制及其强化措施，构成了多层前向网络机理研究的主体内容[16]．

4 有限规模多层前向网络的逼近机理

考虑一个3层前向网络，其输入节点，输出节点和隐节点分别为 n，m 和 N 个，则网络可描述为

$$Y_k=\sum_{j=1}^{N}\beta_{jk}h_j\left(\sum_{i=1}^{n}\alpha_{ij}x_i+\alpha_{0j}\right)+\beta_{0k}\quad k=1,2,\cdots,m \tag{3}$$

其中，x_i 和 Y_k 分别是网络的输入向量 X 和输出向量 Y 的第 i 个和第 k 个分量，$X=(x_1,x_2,\cdots x_n)^T$，$X\in R^n$，$Y=(Y_1,Y_2,\cdots Y_m)$，$Y\in R^m$；α_{ij} 和 β_{jk} 分别为第 i 个输入节点到第 j 个隐节点和第 j 个隐节点到第 k 个输出节点的联接权系数，α_{0j} 和 β_{0k} 分别为第 j 个隐节点和第 k 个输出节点的阈值（就信号处理或图象分割而言）或平移参数（就连续函数映射而言），$h_j(\cdot)$ 为第 j 个隐节点的节点函数．

令
$$A=\begin{bmatrix}\alpha_{01} & \alpha_{02} & \cdots & \alpha_{0N}\\ \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1N}\\ \cdots & \cdots & \cdots & \cdots\\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nN}\end{bmatrix}\qquad B=\begin{bmatrix}\beta_{01} & \beta_{02} & \cdots & \beta_{0m}\\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m}\\ \cdots & \cdots & \cdots & \cdots\\ \beta_{N1} & \beta_{N2} & \cdots & \beta_{Nm}\end{bmatrix}$$

以及

$$\alpha_j=(\alpha_{0j},\alpha_{1j},\cdots,\alpha_{nj})^T,\quad \beta_k=(\beta_{0k},\cdots,\beta_{Nk})^T,\quad G(a_j,X)=h_j(X^Ta_j)$$

其中，A 由输入层到隐层的所有可调的网络参数组成，$A\in R^{(n+1)\times N}$，B 由隐层到输出层的所有可调的网络参数组成，$B\in R^{(N+1)\times m}$，α_j 和 β_k 分别为矩阵 A 和 B 的第 j 列和第 k 列，

$G(\alpha_j, X)$ 表示第 j 个隐节点的输出，$j=1,2,\cdots,N$；$k=1,2,\cdots,m$。则网络（3）可写成

$$Y_k=[1 \quad G(\alpha_1,X)\cdots G(\alpha_N,X)]\beta_k \qquad k=1,2,\cdots,m \tag{4}$$

为描述方便，网络（4）简记为：

$$Y=F(A,B,X)$$

其中，F：$A\times B\times X\subset R^{(n+1)\times N}\times R^{(N+1)\times m}\times R^n\to R^m$。

按照上述定义和记号，考虑如下一类离散点集上的最小二乘逼近问题：

设有未知映射 $y=f(X,\Theta)$，$X\in[a,b]^n\subset R^n$，$y\in R^m$，$\Theta\in R^L$，f：$X\times\Theta\subset R^n\times R^L\to R^m$；已知在离散点集 $X_d=\{X^1,X^2,\cdots,X^p\}\subset[a,b]^n$ 上 $f(\cdot,\Theta)$ 所成的 p 个向量，$\{(X^i,y^i)\mid y^i=f(X^i,\Theta),\ X^i\in X_d,\ i=1,2,\cdots p\}$，对网络（4）

令

$$T=\begin{bmatrix} y_1^1 & y_2^1 & \cdots & y_m^1 \\ y_1^2 & y_2^2 & \cdots & y_m^2 \\ \cdots & \cdots & \cdots & \cdots \\ y_1^p & y_2^p & \cdots & y_m^p \end{bmatrix} \qquad O=\begin{bmatrix} Y_1^1 & Y_2^1 & \cdots & Y_m^1 \\ Y_1^2 & Y_2^2 & \cdots & Y_m^2 \\ \cdots & \cdots & \cdots & \cdots \\ Y_1^p & Y_2^p & \cdots & Y_m^p \end{bmatrix}$$

其中，y_j^i 表示 y^i 的第 j 个分量，$Y_j^i=[1 \quad G(\alpha_1,X^i)\cdots G(\alpha_N,X^i)]\beta_j$，$i=1,2,\cdots,p$；$j=1,2,\cdots,m$。要求按下述指标

$$\|T-\hat{O}\|_F^2=\min_{A,B}\|T-O\|_F^2 \tag{5}$$

确定网络的最优参数 $\hat{A}$，$\hat{B}$，使 $F(\hat{A},\hat{B},\cdot)$ 在（5）的意义下作为对 $f(\cdot\Theta)$ 的近似描述。其中，$\|\cdot\|_F$ 表示 Frobenius 范数，简称 F 范数。

根据式（4）可将式（5）重新写作

$$\|T-\hat{Q}\|_F^2=\min_{A,B}\|T-G_AB\|_F^2 \tag{6}$$

其中 $G_A\in R^{p\times(N+1)}$。

$$G_A=\begin{bmatrix} 1 & G(\alpha_1,X^1) & \cdots & G(\alpha_N,X^1) \\ 1 & G(\alpha_1,X^2) & \cdots & G(\alpha_N,X^2) \\ \cdots & \cdots & \cdots & \cdots \\ 1 & G(\alpha_1,X^p) & \cdots & G(\alpha_N,X^p) \end{bmatrix}$$

如果 $p<\{[(n+1)N+(N+1)m]/m\}$，$T=G_AB$ 可能有无穷多组解，不存在最小二乘逼近问题。此时的网络是一种广义插值网络。当 $p\geqslant\{[(n+1)N+(N+1)m]/m\}$，且 $\mathrm{rank}(G_A)=N+1$ 时，式（6）有唯一的最小二乘解，下面讨论这种情况下的网络逼近机理。

对于每个固定的 A，由矩阵论知[24]

$$\min_B\|T-G_AB\|_F^2=\|T-G_AG_A^+T\|_F^2 \tag{7}$$

其中，$G_A^+T=B$，G_A^+ 是 G_A 的 Penrose 广义逆。因此，

$$\|T-\hat{O}\|_F^2=\min_A\|T-G_AG_A^+T\|_F^2 \tag{8}$$

令 t^k 表示 T 的第 k 列，则有

$$\|T-\hat{O}\|_F^2=\min_A\left\{\sum_{k=1}^m\|t^k-G_AG_A^+t^k\|_F^2\right\} \tag{9}$$

$E = G_A G_A^+$ 是沿 G_A 的零空间 $N(G_A)$ 向 G_A 的值域空间 $R(G_A)$ 进行投影的正交投影矩阵.

由于 $R(G_A) \oplus N(G_A) = R^p$，因此任意 $t^k \in R^p$ 均可唯一地分解为

$$t^k = t_1^k + t_2^k, \quad t_1^k \in R(G_A), \quad t_2^k \in N(G_A) \tag{10}$$

其中 t_1^k 是 t^k 沿着 $N(G_A)$ 到 $R(G_A)$ 的投影.

结合（9）、（10）两式可得

$$\| T - \hat{O} \|_F^2 = \min_A \left\{ \sum_{k=1}^{m} \| t_2^k \|_F^2 \right\} \tag{11}$$

其中

$$t_2^k = \begin{cases} 0 & t^k \in R(G_A) \\ t_2^k & t^k \notin R(G_A) \cup N(G_A) \\ t^k & t^k \in N(G_A) \end{cases}$$

对比（7）、（8）两式是有益的. 在传统的逼近方法中，由于 G_A 仅为 X 的函数，当采样数据已定时，传统逼近方法只能实现如式（7）所示的一次优化过程；而在神经网络逼近中，由于 G_A 同时也是 A 的函数，因此允许实现如式（8）所示的二次优化过程. 将式（8）改写为以下形式：

$$\min_A \| T - G_A G_A^+ T \|_F^2 = \min_A \| (I - G_A G_A^+) T \|_F^2 \tag{12}$$

由于 $(I - G_A G_A^+)$ 是沿 G_A 的值域空间 $R(G_A)$ 向其零空间 $N(G_A)$ 进行投影的投影矩阵，因此式（12）的几何意义是：通过调整 G_A 的结构来构造一个投影矩阵 $(I - G_A G_A^+)$，使其在 F 范数的意义下尽可能地与 T 正交. 以 G_A^j 表示矩阵 G_A 的第 j 列，则从式（9）可以清楚地看出：网络在式（5）的约束下进行的离散点集上的最小二乘逼近，其实质就是由隐层上的各节点函数组成了一组基函数 $h_1(\cdot), h_2(\cdot), \cdots, h_N(\cdot)$，它们在有限点基 X_d 上形成了一组基向量 $G_A^1, G_A^2, \cdots, G_A^{N+1}$，通过调整基向量的内部结构实现对指定向量集合 $\{t^k\}$ 的最佳逼近. 在输入向量已定的条件下，由于 $\{a_{ij}\}$ 是无级可调的，基向量的结构柔性主要取决于隐层节点的节点函数特性. 因此我们得出结论：与理想网络不同，有限规模多层前向网络中隐层节点函数特性的特定选择是构成网络有效逼近能力的最关键的因素.

增加最后一个隐层的隐节点个数，将扩展值域空间 $R(G_A)$ 的维数，因此也就扩大了网络的有效逼近区域，使网络能够适应更多采样点的变化. 但隐节点个数 N 的上限应当满足限制条件 $N_{\max} \leqslant \min\{L, p-1\}$（其中 L 为参数矢量 Θ 的维数），否则会出现过分逼近现象，从而降低了网络的预测能力；增加隐层数目将显著改善基向量的变结构能力，因此将提高对采样点集的逼近精度，同时也增强了网络的预测能力. 就实用目的来讲，4 层网络已经具有合适的结构柔性.

仅就对连续函数的逼近来讲，阈值条件的引入不仅不必要而且会产生意想不到的不良后果. 从几何上讲，阈值的作用相当于在 $(n+1)$ 维超曲面上打孔. 因此，阈值的引入必然会破坏原始函数的连续性条件. 在对连续函数的逼近中，我们将原先的阈值参数当作平移参数处理. 从理论上讲，只需在输出层引入平移矢量就可满足对连续非齐次函数的逼近要求. 采用神经网络进行信号处理和图象分类等目的时，阈值条件的引入是十分必要的. 阈值是形成分离界面的基础.

关于神经网络逼近和传统逼近方法的优劣比较是令人感兴趣的问题．对这个问题的回答要根据具体问题来定．在传统的逼近方法中，基变量是固定的；而在神经网络逼近中，由于引入了可调参数$\{a_{ij}\}$，基变量的结构特性是可在一定范围内变化的．也就是说，传统逼近方法的突出特点是在固定基底的条件下完成逼近过程，而神经网络方法则是在可变基底的条件下完成逼近过程．

如果要完成如下非线性映射的逼近

$$f(x)=\sum_{i=0}^{n}c_i x^{q_i}\qquad c_i\in R,\qquad x\in[a,b]\subset R$$

当$q_i=i$（$i=0,1,\cdots,n$）时，采用任何方法，包括采用神经网络方法都不可能比采用以$\{1,x,\cdots,x^n\}$为基底的传统逼近方法做得更好；同样，当$0<q_i<1$（$i=0,1,\cdots,n$）时可以指望，采用Sigmoid函数作为隐节点函数的神经网络方法会比传统逼近方法做得好些．

对上述问题的分析使我们意识到：在有限规模网络的限制条件下，对于给定的非线性映射，必然存在一组最佳的基函数组合，以其为基底可获得对给定映射的最佳逼近．

对于给定的问题，如果已知最佳基函数的结构，采用传统方法和采用神经网络方法没有本质区别；如果对最佳基函数的结构一无所知，传统方法很难给出一个满意的解答，而神经网络方法却可能给出一个满意的解答．

现实的问题往往是介于以上两种极限状态之间：即对于最佳基函数的结构，我们可能有一些粗略的知识，但这种知识不够完整、精确．在这种情况下，互补型混合网络的使用被证明是一种好的选择[16]．

互补型混合网络由两部分组成：在同一隐层中，有N_1个隐节点的节点函数是根据我们已有的关于被描述对象的知识预先选择的，它们被看成是网络知识结构中的经验成份；有N_2个隐节点的节点函数是普通的Sigmoid型函数，它们被用来对网络中原有的知识进行修改和补充．由于$N_1+N_2=N$，因此网络的规模是预先固定的．这样做有两个好处：一则便于算法的实现和并行化；二则便于对网络的自适应学习能力进行检验．N_1和N_2的具体取值由一个与逼近误差发生关联的学习监督机构自动地进行调节，以确保网络对于被描述对象有一个比较合理的知识结构．

我们把网络所拥有的关于被描述对象的知识划分为两个不同的层次：①表层知识．即网络关于被描述对象个体特征的描述和再现．其学习过程表现为网络对其联接权系数矩阵取值的特定选择；②深层知识，或称为系统知识．即网络关于被描述对象所属集合的整体特征的抽象和概括．其学习过程表现为网络对节点函数类型及其相对比例的特定选择．传统的神经网络研究强调了对于表层知识的学习和掌握，对于深层知识的抽象与概括，并未给予应有的重视，而后者恰恰是神经网络方法优于传统方法的本质所在．

有限规模多层前向网络存在最佳基底的思想使我们得出结论：对于多层网络，应当建立各层节点和输出节点间的直连通道，以便一旦最佳基函数生成，便迅速传输出去．实际使用中，不少人都尝试过这种作法，其效果均十分良好．

接下来我们研究一下与学习算法有关的一些问题．

5 前向网络的学习算法

误差的反向传播算法（BP）是神经网络学习算法中最有影响的算法之一．它的提出和完善不但解决了 Minsky 和 Papert 在《Perceptrons》一书中设想的多层网络的学习问题，而且极大地推动了多层网络理论与应用研究的深入和发展[10,25]．

BP 算法本质上是一种梯度投影法，收敛速率慢是这类算法的主要问题．对传统梯度投影法的修正和改进目前主要采用两种方式：①修正投影方向．如共轭梯度法，Davidon-Fletcher-Powell 法等[26]．②修正增量步长（即神经网络研究中的所谓学习率）．如 Jacobs 的冲量法[27]．刘俊民等人的自适应学习率方法等[10,28]．修正投影方向必然伴随着大量的中间运算．因此，这类方法虽然可以显著减少学习次数，却未必能够显著减少学习时间．从实用的角度讲，建立一种合理的自适应学习率修正方法更为简单可行．

本文不打算介绍原始BP算法，对此算法感兴趣的读者不妨参阅文献[10,25]．本文只介绍由 Chen 和 Billings 等人提出的并行递推预测误差算法及对这些算法的若干修正[3,,16]．

考虑一个多层前向网络，其输入节点和输出节点分别为 n 和 m 个．设网络中所有可调的参数（权值和阈值）共有 n_Θ 个，它们构成向量 Θ．网络可描述为

$$y(t,\Theta)=F(X(t),\Theta) \tag{13}$$

其中，$X(t)$ 为 n 维的网络输入向量，$y(t,\Theta)$ 为 m 维的网络输出向量，$F(\cdot,\cdot)$ 为向量形式的非线性映射，t 为参量．

给定一组样本序列 $\{(X(t),\ d(t))\mid X(t)\in R^n,\ d(t)\in R^m,\ t=1,2,\cdots,N;\ N$ 为有限数$\}$．记误差序列 $\varepsilon(t,\Theta)$ 和代价函数 $J(\Theta)$ 分别为

$$\varepsilon(t,\Theta)=d(t)-y(t,\Theta)\qquad t=1,2,\cdots,N \tag{14}$$

$$J(\Theta)=\frac{1}{2N}\sum_{t=1}^{N}\varepsilon^T(t,\Theta)\ \varepsilon(t,\Theta) \tag{15}$$

则在最小范数意义下确定网络的最优参数 $\hat{\Theta}$ 的过程可描述为：

$$J(\hat{\Theta})=\min_{\Theta}\frac{1}{2N}\sum_{t=1}^{N}\varepsilon^T(t,\Theta)\ \varepsilon(t,\Theta) \tag{16}$$

记网络式（13）关于其参数的梯度为 $G(t,\Theta)$，它是 $n_\Theta\times m$ 维的矩阵，并记

$$H(\Theta)=\frac{1}{N}\sum_{t=1}^{N}G(t,\Theta)\ G^T(t,\Theta) \tag{17}$$

则按递推预测误差算法，确定 $\hat{\Theta}$ 的递推过程可描述为

$$\begin{cases}\Delta(k)=r_m\Delta(k-1)+r_g G[k,\hat{\Theta}(k-1)]\ \varepsilon[k,\hat{\Theta}(k-1)]\\ P(k)=\dfrac{1}{\lambda}\{P(k-1)-P(k-1)G(k,\hat{\Theta}(k-1))\ [\lambda I+G^T(k,\hat{\Theta}(k-1))\\ \qquad P(k-1)G(k,\hat{\Theta}(k-1))]^{-1}\ G^T(k,\hat{\Theta}(k-1))P(k-1)\}\\ \hat{\Theta}(k)=\hat{\Theta}(k-1)+P(k)\Delta(k)\end{cases} \tag{18}$$

其中 k 表示学习过程的第 k 步，r_m 和 r_g 分别为冲量因子和自适应增益，λ 为遗忘因子．

$\Delta(k)$ 是一个 n_Θ 维向量，它是代价函数 $J(\Theta)$ 的负梯度 $-\Delta J(\Theta)$ 的递推形式．$P(k)$ 则是一个 $n_\Theta \times n_\Theta$ 维矩阵，它是 $J(\Theta)$ 的近似 Hessian 阵的逆（即 $H(\Theta)$ 的逆）的递推形式．

将网络中的隐节点和输出节点按某种秩序从 1 到 p 进行编号．对第 i 个节点，设其所含的可调参数为 n_{Θ_i}，用向量 Θ_i 表示，$i=1,2,\cdots,p$，则有：

$$\Theta=\begin{bmatrix}\Theta_1\\ \vdots\\ \Theta_p\end{bmatrix},\qquad G(t,\Theta)=\begin{bmatrix}G_1(t,\Theta)\\ \vdots\\ G_p(t,\Theta)\end{bmatrix} \tag{19}$$

其中，$G_i(t,\Theta)$ 是 $n_{\Theta_i}\times m$ 阶矩阵，它是 $y(t,\Theta)$ 关于 Θ_i 的梯度，$i=1,2,\cdots,p$．

由式（19），令

$$\bar{H}(t,\Theta)=\frac{1}{N}\sum_{t=1}^{N}\begin{bmatrix}G_1(t,\Theta)G_1^T(t,\Theta) & & 0\\ & \ddots & \\ 0 & & G_p(t,\Theta)G_p^T(t,\Theta)\end{bmatrix}$$

以 $\bar{H}(t,\Theta)$ 取代 $H(t,\Theta)$，算法式（18）便可分解为 p 个子算法，即并行递推的预测误差算法：

$$\begin{cases}\Delta_i(k)=r_m\Delta_i(k-1)+r_g G_i(k,\hat{\Theta}(k-1))\varepsilon(k,\hat{\Theta}(k-1))\\ P_i(k)=\dfrac{1}{\lambda}\{P_i(k-1)-P_i(k-1)\,G_i(k,\hat{\Theta}(k-1))\,[\lambda I+G_i^T(k,\hat{\Theta}(k-1))\\ \qquad P_i(k-1)G_i(k,\hat{\Theta}(k-1))]^{-1}G_i^T(k,\hat{\Theta}(k-1))P_i(k-1)\}\\ \hat{\Theta}_i(k)=\hat{\Theta}_i(k-1)+P_i(k)\Delta_i(k)\end{cases}\qquad i=1,2,\cdots,p \tag{20}$$

对于算法（20），取 $P_i(k)\equiv I$；$\forall k(i=1,2,\cdots,p)$，则此算法即为 BP 算法．因此递推预测误差算法可看成是 BP 算法的一种变形．将 Chen 的递推预测误差算法和 Rumelhart 的惯性项修正方法[15]进行对比是有益的：Chen算法中的所谓冲量因子和自适应增益，其作用分别相当于 Rumelhart 方法中的惯性系数和学习率；Chen算法中的 $P(k)$，其作用类似于投影矩阵，目的是对梯度方向进行修正．因此，Chen算法之于BP算法，其作用和意义类似于共轭梯度法之于梯度投影法．由于 Chen算法中含有 3 个可调参数 r_m，r_g 和 λ，因此对于某些具体的对象，它可能会有比较好的收敛效率．但这种高效率的获得是以试算为代价的，因此并无指导意义．

陈大庆等人提出了带有参数变换的并行递推预测误差算法[16]．笔者认为陈大庆工作的意义在于对网络映射结构的某些调整，而不在于对 Chen 算法的改进．

建立学习率自适应算法是一件有意义的工作，这一任务可采用比较简单的方式予以实现．

令参数修正公式为

$$\Theta(k)=\Theta(k-1)+\alpha(k)\Delta(k)$$

其中，$\alpha(k)$ 为第 k 步的学习率．

设置第1步的学习率为 $\alpha(1)$，变步参数为 λ，$1<\lambda<2$，并定义 $\alpha(0)=0$，那么

$$\begin{cases} \text{如果 } J(\Theta_k)<J(\Theta_{k-1}),\ \text{则}\alpha(k+1)=\lambda\alpha(k),\ \Delta(k+1)=\Delta(k) \\ \text{如果 } J(\Theta_k)>J(\Theta_{k-1}),\ \text{则}\alpha(k+1)=\dfrac{\alpha(k-1)+\alpha(k)}{2},\ \Delta(k+1)=\Delta(k)\ (k=1,2,\cdots) \\ \text{如果 } J(\Theta_k)=J(\Theta_{k-1}),\ \text{则}\alpha(k+1)=\alpha(k),\ \text{重算 } \Delta(k+1) \end{cases}$$

采用上述学习率自适应调整算法，可保证 BP 算法具有比较好的收敛效率。

6 结 束 语

从较为完整的 BP 算法的重新提出到今天不足 10 年的历史。在这短短的时间里，多层前向网络的理论研究与应用研究取得了惊人的成就。要想使多层前向网络的潜力得以充分发挥，还要做许多深入细致的研究工作。神经网络作为一种新的信息处理方式尚处在蓬勃发展时期，将神经网络方法和其它方法相结合以解决各类复杂非线性问题的尝试正在引起广泛的兴趣。

参 考 文 献

1 Brent R P. IEEE Trans. *Neural Networks*, **2**, 3 (1991) : 346—354
2 Carpenter G A, Grossberg S. *Neural Networks*, **3** (1990) : 129—152
3 Chen S, et al. *Int. J. Control*, **51**, 6 (1990) : 1215—1228
4 Chen S, et al. IEEE Trans. *Neural Networks*, **2**, 2 (1991) : 302—309
5 Gallant A R, White H, *Neural Networks*, 5 (1992) : 129—138
6 Chen S, et al. *Int. J. Control*, **51**, 6 (1990) : 1191—1214
7 Haber R, Unbehaun H. *Automatics*, **26**, 4 (1990) : 651—677
8 Roth M W. IEEE Trans. *Neural Networks*, **1**, 1 (1990) : 28—43
9 陈大庆，周凤岐. 系统仿真学报，4，2 (1992) : 57—63
10 刘俊民. 西安：西北工业大学博士学位论文 (1993, 7)
11 Amari S I. *Biological Cybernetics*, 26 (1977) : 175—185
12 Bryscn A, Ho Y C. Applied Optimal Control, New York (1969)
13 Werbos P J. Ph. D Thesis, Harward University, U. S. A (1974)
14 Parker D. Learning Logic. Stanford University, U. S. A. (1982)
15 Rumelhart D E,McClelland J L. Parallel Distributed Processing: Exploration in The Microstructures of Cognition, Vol 1 & Vol 2. Cambridge: Bradford Books/MIT Press (1986)
16 陈大庆. 西安：西北工业大学博士论文 (1993, 9)
17 Lippmann R P. Proc. 1987 IEEE Int. Conf. on Neural Networks, San Diego, 5 (1989) : 417—426
18 Lippmann R P. *Neural Communication*, **1**, 1 (1989) : 1—38
19 Wieland A, Leighton R. IEEE First Int. Conf. on Neural Networks, 3 (1987) : 385—392
20 Funahashi K I. *Neural Networks*, 2 (1989) : 183—192
21 Hornik K, et al. *Neural Networks*, 3 (1990) : 351—360
22 Hornik K. *Neural Networks*, 4 (1991) : 251—257
23 Cybenko G. *Signals and Systems*, 2 (1989) : 303—314
24 程云鹏，张凯院，徐仲. 矩阵论. 西北工业大学出版社，西安 (1987)
25 焦李成. 神经网络系统理论. 西安电子科技大学出版社，西安 (1990)
26 Haug E J, Arora J S [美]. (郁永熙，丁惠梁译). 实用最优设计. 科技出版社，北京 (1985)
27 Jacobs R A. *Neural Networks*, 1 (1988) : 295—307
28 肖日华，罗四维，丁嘉种. 中国神经网络1991学术会议论文集，南京 (1991) : 226—229

ADVANCES AND PROBLEMS IN THE STUDY OF MULTILAYER FEEDFORWARD NEURAL NETWORKS

Dong Cong Li Zhengneng Xia Renwei He Qingzhi
Beijing University of Aeronautics and Astronautics, Beijing, 100083

Abstract In this paper, ① a brief introduction is given to the history of research on multilayer feedforward neural networks; ② a systematic review is given to some important achievements in the field of multilayer feedforward networks; ③ some existing problems and tendencies in the study of multilayer feedforward networks are pointed out, and a few of feasible methods for solving those questions are suggested; ④ a systematic analysis of approximation mechanism of feedforward networks with a scale-limited discrete point set is made; ⑤ some construction rules, which govern a rational feedforward networks, are established.

Keywords *multilayer feedforward neural networks; mechanism analysis; least square approximation; learning algorithm*