DOI:10.16262/j.cnki.1000-8217.1997.01.006

前向网络全局最优化问题研究

董 聪

(清华大学土木工程系，北京 100084)

[**摘要**] 简要回顾前向网络研究的历史，对其中若干经典成就做简要的介绍和评论。在对多层前向网络全局最优化问题所涉及的几个主要方面进行深入剖析的基础上，给出了全局最优化算法应具备的基本条件和一种算法格式，对所给算法格式的收敛性做了理论证明。本文指出，将数论和多元非线性回归的有关方法和成果引入前向网络研究领域是一个值得注意的发展趋势。

[**关键词**] 前向网络，全局最优化，拓扑，非线性回归，数论

1 引言

用一元函数的复合表示及逼近多元函数是数学中的一个重要问题。1900 年，Hilbert 猜想：多元连续函数不能用一元连续函数的复合表示。Hilbert 的猜想在 1957 年被著名数学家 Arnoid 和 Kolmogorov 证伪。Kolmogorov 证明：任一定义在 $[0,1]^n$ 上的连续函数可用一些一元连续函数的复合表示。Kolmogorov 的工作初步奠定了多层前向网络映射能力数学证明的基础[1]。

1965 年，Nilsson 提出了含有隐节点的多层网络的构想[2]。1969 年，Minsky 和 Papert 对感知机在处理非线性问题方面的局限性做了系统的分析，对多层网络是否存在有效的学习方法持怀疑态度[3]。1986 年，Rumelhart，Hinton 和 Williams 发现十余年前由 Werbos 发明的误差反传算法（Back Propagation error，BP）能够有效地解决多层网络中隐节点的学习问题[4]，证明 Minsky 对多层网络的看法是不正确的，这一发现在一定程度上促成了人工神经网络研究热潮的再度兴起[5]。Lippmann[6,7]通过仿真研究发现：三层网络可以形成若干个复杂的决策域，而四层网络则可以形成任意复杂的决策域。大量的仿真结果给人们一种启示：多层前向网络可能具有实现任意复杂非线性映射的潜力。

Funahashi[8]，Hornik[9,10]和陈天平[11]对理想网络的映射能力做了系统的证明，结论是：在一个相当宽的范围内，多层前向网络具有以任意精度逼近定义在紧致子集上的任意非线性连续函数的能力。在 Funahashi 的证明中，隐节点函数限定为有界单调递增连续函数，Hornik 和陈天平则发现，有界性是必要的，单调递增的限制条件并非必要。与此相应，本文作者给出了 BP 算法的广义描述。广义 BP 算法和网络权值的具体修改方式不发生直接关联，不依赖于神经元节点函数和系统误差函数的具体描述，只要求它们一阶可导即可。本文作者还证明，基

国家自然科学基金、航空科学基金和 863 计划资助课题.

本文于 1996 年 5 月 9 日收到.

于广义BP算法的任何全样本前向网络权值修改方式都是收敛的。Shatz 对发育中的大脑进行了深入的研究[12]，他发现在大脑发育过程中，轴突有选择地收缩和长出新的分枝，新的分枝通过竞争的方式选择正确的靶位，并通过一定的机制消除选址的错误。本文作者将 Shatz 的发现以计算机模拟程序的方式再现出来，提出了多层前向网络中跨层连接的概念及网络拓扑结构简化的通用算法。文献[13]对网络的泛化机制和改进措施进行了系统的分析，指出最简网络拓扑结构不仅有利于硬件实现，也有利于网络泛化功能的改善。由于与前向网络相关的其它一些问题已相继得以解决，前向网络的全局最优逼近和全局最优拓扑构造这一长期以来一直未能有效解决的难题，迅速成为研究的热点。本文将系统介绍近来在该领域取得的一些进展，以期对前向网络理论与应用研究的深入与发展有所帮助。

2 全局最优逼近算法

前向网络的全局最优主要有两个方面：一是网络对离散点集的全局最优逼近；二是网络的全局最优拓扑构造。影响第一个方面的主要因素有：网络结构，网络权值的初始设定，网络的学习算法，局部极小点及其逃逸算法（随机游动算法），“过拟合”及其解决方法等。影响第二个方面的主要因素有：网络的构造法则，网络拓扑的学习算法等。网络结构依赖于网络所要逼近的具体对象，由于完成同样功能要求的网络结构可以多种多样，因此从网络的功能要求出发来决定网络结构的努力是不现实的。现实的作法是，建立合理的网络构造法则，通过学习的方式来决定网络结构。网络权值初始设定的好坏不仅依赖于网络的拓扑结构，也依赖于网络所要逼近的对象的具体性质和误差曲面的具体形态，这是一个学习过程结束之后才可以判定的问题。在实际应用中，研究这一问题没有太多的意义。网络的学习算法是构成网络全局最优逼近的一个非常关键的因素，目前的研究大多集中在网络权值的学习上，对网络拓扑结构（隐层数，隐节点数和节点函数类型等）的学习研究得很少。权值的学习又多集中在学习速率方面，对学习算法的收敛性和收敛点特性的研究做得比较少。局部极小点问题是非线性优化算法长期以来存在的尚未妥善解决的一个实质性问题。从理论上讲，只有随机优化算法有能力从根本上解决任意非线性系统的全局最优化问题。由于受计算条件的制约和问题本身的复杂性，长期以来，随机优化算法的研究多表现为零散的仿真实验和启发式的构想，真正系统化的理论研究做得很少。在人工神经网络研究领域，目前用得最多的两种随机优化算法是 Holland[16]提出的遗传算法（Genetic Algorithm，GA）和 Metropolis[17]提出的模拟退火算法。从原理上讲，遗传算法及其变种只是借用了一些生物学中的名词，如种群（population）、交叉（crossover）和变异（mutation）等，并不和生物进化学说（如达尔文的自然选择，孟德尔的遗传学说）直接关联。从算法结构上讲，遗传算法缺乏一种实现全局最优化的内在机制，正因为如此，才使得遗传算法的变种层出不穷[18]。迄今为止，没有人从理论上证明遗传算法具有实现全局最优化的能力。模拟退火算法的情况有些不同，自从 1983 年 Kirkpatrick 首次采用非平稳 Markov 链对模拟退火算法的全局优化能力进行证明以来，不断有人采用类似方法证明：模拟退火算法具有以概率 1 逼近全局最优解的能力。但已有的仿真结果令人失望[19]，因此一系列的改进算法相继出台。理论证明和仿真结果之间的巨大反差使人有理由怀疑：理论证明过程和网络的实际学习过程之间存在定性（本质）的差异。由于这一问题比较复杂，作者将另文详述。

避免局部极小点的随机算法的关键是对应的逃逸算法，目前这方面的研究做得不够深入。逃逸能力和逃逸概率是两个不同的概念。如果实际的逃逸概率很小，且不说伪随机数发生器有一个和计算机位数相关联的循环周期，即使无此周期的存在，实际计算中也很难逃出局部极小点。这方面深入的研究工作目前几乎没有做。“过拟合”现象是网络隐节点过多的必然结果，它的出现直接影响了网络的泛化能力。避免“过拟合”的关键是选择合适的隐节点函数和合理的网络拓扑结构。目前，由于作者等人的工作，网络拓扑简化问题已基本解决。节点函数类型的学习和网络拓扑结构的学习虽然有了一些初步的研究成果[1,5]，但还有许多问题有待更深入的探索。网络的构造法则原则上有两种方式：一种是删减法则，给出的前向网络拓扑结构简化的通用算法；另一种是扩张法则，如文献[1]中提出的增殖变异法就属此类。从原则上讲，删减法则虽有利于硬件实现，并在一定程度上可以缓解网络的“过拟合”现象，但由于局部极值点的存在，因此不可能从根本上解决网络的全局最优拓扑构造问题，解决这一问题必须采用扩张式的网络构造法则。文献[1]对扩张式的网络拓扑构造法则的神经生理学基础进行了比较充分的论证。

从本质上讲，网络的全局最优逼近是问题的核心，其它问题都可以看作是为实现这一目的而诱发的附带问题，至少从工程应用的角度看是这样。

全局最优点不一定是极值点，在网络权值取值区间为紧致子集的情况下，全局最优点就可能是边界点。当隐节点函数为有界连续可导函数时，不论所要逼近的原始函数特性如何，其网络描述形式皆为连续可导函数，因此，网络的全局最优逼近点必为边界点或误差面的极小值点。在边界点上，误差面的梯度矢量中未达边界的各维的梯度分量为零，已达边界的各维的梯度分量通常非零，寻优过程被强制中止。

网络的逼近能力依赖于特定的网络结构。组成前向网络结构的四要素是：网络的隐层数，各隐层的节点数，各层间的连接关系和隐节点的特性。数学家们对理想网络的杰出研究成果使人们普遍产生了一种误解，似乎对于多层前向网络而言，网络的隐层数和各层间的连接关系二者与隐节点的函数特性是无关紧要的，起作用的只是隐层中的节点个数。Funahashi证明[8]，当隐节点函数为有界单调递增连续函数时，三层前向网络具有以任意精度逼近定义在紧致子集上的任意非线性连续函数的能力。Hornik[9,10]和陈天平进一步证明[11]，隐节点函数的有界性是必要的，单调递增条件并非必要。需要指出的是：理想网络至少有三个基本特性：第一是有无穷多个采样点；第二是有一个合理的采样结构；第三是原则上允许有任意多个隐节点。如果说采样结构可以采取一些合理的作法，比如采用数论中的完全佳格点集作为采样点[14,15]来加以改善的话，那么，进行无穷多次采样和采用任意多个隐节点在实用中则难以达到。实际问题必然是在有限个离散采样点集的条件下，采用有限数目隐节点的网络对所给对象进行逼近。研究工作表明，用于函数逼近时，多层前向网络中，各隐层的隐节点数目不应多于采样点数，否则，必有冗余的隐节点可以归并；各隐层的隐节点数目也不应等于采样点数，否则，网络将成为插值网络。插值网络对已学样本的系统误差为零，也就是说具有简单而良好的直接记忆功能，但不具有容错性和保证泛化能力的机制，而后者是函数逼近所必需的。综上所述不难得出以下结论：

(1) 定义采样点数为N，则用于函数逼近时，多层前向网络各隐层节点数的上限为$N-1$；

(2) 用于函数逼近时，多层前向网络在极小值点和边界点上的系统误差非零；

(3) 改善网络逼近精度的方式有四种：增加隐节点数，增加隐层数，采用合适的层间连接和合适的隐节点函数；

(4) 多层前向网络中，逆信息流方向，各隐层所含的隐节点数以递增为宜；

(5) 网络对离散点集的最优逼近能力依赖于网络结构。

由于网络对离散点集的最优逼近能力依赖于网络结构，而网络结构的四要素中，除各隐层的节点个数存在上限外、各层间的连接关系、隐层数目和隐节点函数类型均无上限条件。因此，严格意义上的网络对于离散点集的最优逼近问题或者不存在，或者很难求解。只有在限定隐层数目和限定隐节点函数类型的条件下，网络对离散点集的最优逼近解才存在并唯一(最小系统误差唯一，相应的权矢量不一定唯一)。已经证明，任何全样本前向网络训练算法均是收敛的。有些文献由于作者对非线性规化方法的收敛条件缺乏了解而得不到收敛解，其实，这和算法本身无直接关系。对于这类情况本文不予讨论。

在算法结构正确的前提下，网络的逼近误差必单调下降，也就是说，网络必收敛于极小值点(或边界点，以下统称为极值点)。如果极小值点附近的误差曲面比较平缓，网络结构的收敛速率会比较低，此时可调整系统误差函数以加速收敛。

前文已经说过，在限定隐层数目和限定隐节点函数类型的条件下，网络对离散点集的最优逼近解存在且唯一。现在，我们讨论最简单的一种类型，即三层前向网络，隐层的节点函数类型已确定，仅隐节点数目未定，在此特定条件下的网络全局最优逼近问题。由于这一问题涉及到网络结构的演化和局部极值点的逃逸算法这些网络全局最优化问题中最根本的问题，因此，如果这一问题能够解决，那么更一般的网络优化问题也就能迎刃而解。

由于极值点和网络结构有关，因此从局部极值点逃逸的方式至少有两种：一是改变网络结构，促使极值点发生转移。在本文的限定条件下，可通过扩充隐节点数来实现这种方式；二是采用随机游动算法，在固定网络结构的条件下，通过更改权矢量，实现由当前极小值点向更小的极值点或最小值点的转移。由于隐节点数存在上限，因此通过扩充隐节点数来实现极值点转移的方式必然是收敛的。当隐层节点函数为有界连续可导函数时，不论所要逼近的对象的具体特性如何，其网络描述均是连续可导的。因此，在权值取值区间为紧致子集的条件下，网络所表达的连续函数其系统误差的最小值点必然存在。因此，只要权值的随机游动算法可保证权值取值区间内各点在概率意义上是可达的，则必可保证网络结构以概率1实现对离散点集的全局最优逼近。

按此思路，可设计多种前向网络全局最优逼近算法。其中的一种算法——前向网络全局最优逼近算法格式如下：

(1) 根据问题的具体情况初定一个较小的网络结构。设定随机更改权矢量的最大循环次数为 R_m，可接受的系统误差为 ε，当前隐节点数为 N_h；

(2) 采用广义BP算法训练网络权矢量 $\boldsymbol{W}$；

(3) 如果 $E(\boldsymbol{W})$ 非极小值点，转到步骤2；否则，执行下一步；

(4) 如果 $E(\boldsymbol{W})<\varepsilon$，转至步骤12；否则，执行下一步；

(5) 置随机修改次数 $R=0$；

(6) 生成权矢量 $\boldsymbol{W}$ 的高斯型随机修改量 $g(R)$；

(7) 按下述方式修改权矢量：

如果 $E(\boldsymbol{W}(\boldsymbol{I}_1+g(R))^T)<E(\boldsymbol{W})$，则 $\boldsymbol{W}=\boldsymbol{W}(\boldsymbol{I}_1+g(R))^T$，转至步骤2；

如果 $E(\boldsymbol{W}(\boldsymbol{I}_1-g(R))^T)<E(\boldsymbol{W})$，则 $\boldsymbol{W}=\boldsymbol{W}(\boldsymbol{I}_1-g(R))^T$，转至步骤2；

否则，$R=R+1$，执行下一步；

(8) 如果 $R<R_m$，转至步骤6；否则，执行下一步；

(9) 如果 $N_h\geqslant N-1$，转至步骤11；否则，执行下一步；

(10) $N_h=N_h+1$，增加一个隐节点。对新增隐节点的联接权值，按节点权值修改方式修改，转至步骤2；

(11) 非正常结束，原因为以下两种情况之一：第一，当前网络为全局最优逼近网络，但系统误差达不到要求，可采用以下方式试着加以解决：(A)改变隐节点函数的类型，(B)增加隐层数目；第二，当前网络非全局最优逼近网络，可采用以下方式试着加以解决：(A)增大高斯型摄动的方差，(B)增大高斯型随机权值修改的次数；

(12) 系统误差满足要求，存储有关网络结构信息。

符号说明：E，系统误差；N，训练样本总数；I_1，各分量皆为1的列向量，维数和权矢量 $\boldsymbol{W}$ 的相同。

3　全局最优算法结构的讨论

设 K 为 R^n 的紧致子集，$X\in K$，$g(X)$ 为实值连续函数，$\hat{g}(X)$ 为 $g(X)$ 的神经网络描述，Ω_x 为 x 的广义体积，则 $\hat{g}(X)$ 的平方逼近误差为：

$$E=\frac{1}{\Omega_x}\int[g(X)-\hat{g}(X)]^2\mathrm{d}X \tag{1}$$

由于 $g(X)$ 不确知，通常是通过对 X 进行采样，由(2)式来取代(1)式

$$E=\frac{1}{N}\sum_{i=1}^{N}[g(X_i)-\hat{g}(X_i)]^2 \tag{2}$$

要使(2)式能够比较准确地刻划(1)式的基本方面，采样结构应完整而准确地反映 $g(X)$ 的两个特征：(1) $g(X)$ 所含的极值点总数；(2) $g(X)$ 各极值点的位置和大小。

$g(X)$ 所含的极值点数反映了 $g(X)$ 的阶次和形态，是 $g(X)$ 最本质的特征。在三层网络中表现为隐层所含的节点总数和隐节点的函数类型，在多隐层网络中还体现在隐层数上；$g(X)$ 各极值点的位置和大小是其表象特征，在多层网络中表现为连接权矢量的具体取值。关于采样结构除前文已经提过可采用数论中的完全佳格点集来定义采样点外，还有许多其它好的做法，对此作者将另文详述。本文仅讨论在采样点合适的条件下如何通过学习的方式建立最优或近优网络结构。

函数逼近中，在满足精度要求的前提下，逼近函数的阶次越低越好。低阶逼近可以有效地防止“过拟合”现象的发生，从而提高了逼近函数的预测能力。反映到多层前向网络中，就是在满足精度要求的前提下，网络的隐节点数越少越好。这一思想体现在上节给出的前向网络全局最优逼近算法中，就是将增加隐节点作为外置循环来处理；随机游动算法原则上也具有搜寻极值点的能力，但效率太低，因此在我们设计的算法结构中，仅让它完成逃逸局部极值点的任务；随机游动算法中随机权值修正量可采用高斯型，也可采用其它类型，权值取值区间内各点的可达性是问题的关键，具体的随机游动方式可以任选；采用高斯型摄动时，则

其方差以大为宜。

要证明本文提出的随机优化算法可实现全局最优，只需证明以下 4 个命题成立即可

(1) BP 算法可在有限步内实现系统误差的局部极小；

(2) 当前状态非全局最小值点时，逃逸算法可在有限步内实现由当前极小值点向存在更小极值点的区域的转移；

(3) 系统误差的极小值点的数目是有限的；

(4) 通过扩充隐节点数实现极值点转移的方式可在有限步内实现。

很明显，同时满足 (1) (2) 两条的随机优化算法可单调收敛于全局最小值点。同时满足 (1) (4)条的随机优化算法可在有限步内实现全局最小化。

关于BP 算法实现局部极小的能力已有系统的证明，此处不赘述。由于隐节点数目存在上限，因此通过扩充隐节点数实现极值点转移的方式必可在有限步内实现。鉴于此，本文仅简要讨论命题 (2) (3) 成立的条件。

设网络系统逼近误差存在 m ($m \geq 1$) 个取值不同的极小值点，将其大小按降阶排列 $E_1 > E_2 > E_i \cdots > E_j \cdots > E_m$。设网络系统目前处在状态 i，定义

$$\boldsymbol{W}_{E_j} = \{\boldsymbol{W} : E_i > E(\boldsymbol{W}) \geq E_j \quad (i < j \leq m) \tag{3}$$

设权矢量 $\boldsymbol{W}$ 的维数为 r，将 $\boldsymbol{W}_{E_j}$ 所覆盖的 r 维连通域的并定义为 $\boldsymbol{W}_{E_j}$ 的广义体积 Ω_{E_j}，则 Ω_{E_j} 为 j 的不减函数。若存在 $\Omega_{E_j} > 0$，则按本文给出的算法格式，权值取值区域内各点是可达的，即

$$\mathrm{Prob}(\Omega_{E_k}) \geq \mathrm{Prob}(\Omega_{E_j}) > 0 \quad (k > j) \tag{4}$$

也就是说当前状态非全局最小值点时，逃逸算法可在有限步内实现由当前极小值点向存在更小极值点区域的转移。

对于组合优化问题，其状态空间是有限的，由于局部极小值点的数恒小于其状态数，因此按本文给出的随机优化算法，必可确保在有限步内实现组合优化问题的全局最小化。对于多层前向网络，其状态空间是无限（连续）的，但对于实际问题来讲，其网络逼近误差的局部极小值点的数目通常是有限的，因此按本文给出的算法格式，同样可保证在有限步内实现系统逼近误差的全局最小化。

4 需进一步研究的问题

本文对多层前向网络的全局最优问题进行了初步的研究，给出了一种新的算法结构，同时留下了几个尚待深入研究的问题：(1) 隐层节点函数的结构化学习方法；(2) 隐层数的结构化学习方法。关于第一个方面，统计学中的投影寻踪算法值得借鉴，但仍有许多问题需要仔细研究；关于第二个方面，目前几乎没有做什么深入的研究工作，仿真实验做得比较多，机理性的分析有待加强。对于理想网络，早就知道一个隐层就够了；对于有限规模的实际网络，一个隐层有时可能不够。从定性的角度讲，增加隐层的作用早已清楚，但定量方面的结构化分析并未完成，而后者是建立关于隐层数的结构化学习算法所需要的。

对前向网络学习算法的研究，有两个领域的研究成果很值得借鉴：一个是多元非线性回归，另一个是数论。它们可能比函数逼近论更适合于处理有限规模的网络结构和有限数量的离散采样点集条件下的实际网络逼近问题。

参 考 文 献

[1] 董聪. 人工神经网络研究进展. 北京：北京航空航天大学出版社，1995.

[2] Nilsson J. Learning Machines: An Introduction to Pattern Classifying Systems. McGraw-Hill, 1965.

[3] Minsky M, Papert S. Perceptron. MIT Press, 1969.

[4] Rumelhart D E, Hinton G E, Williams R J. Learning Representation by Backpropagation Errors. Nature, 1986, **323** (6188): 533－536.

[5] 董聪，郦正能，夏人伟，何庆芝. 多层前向网络研究进展及若干问题. 力学进展，1995，**25** (2)：186－196.

[6] Lippmann R P. An Introduction to Computing with Neural Nets. IEEE ASSP Magazine, 1987, 4: 4－22.

[7] Lippmann R P. Review of Neural Networks for Speech Recognition. Neural Communication, 1989, **1** (1): 1－38.

[8] Funahashi K I. On the Approximate Realization of Continuous Mappings by Neural Networks. Neural Network, 1989, **2**: 183－192.

[9] Hornik K, Stinchcombe M, White H. Multilayer Feedforward Networks Are Universal Approximators. Neural Networks, 1989, **2**: 359－366.

[10] Honik K. Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks, 1991, **4**: 551－557.

[11] 陈天平. 神经网络及其在系统识别应用中的逼近问题. 中国科学 (A 辑)，1994，**24** (1)：1－7.

[12] Shatz C J. 发育中的大脑. 科学 (Scientific American)，1993，1：11－19.

[13] 董聪，夏人伟. 智能结构设计与控制中的若干核心技术问题. 力学进展，1996，**26** (2)：166－178.

[14] 华罗庚，王元. 数论在近似分析中的应用. 北京：科学出版社，1978.

[15] 董聪. 非线性系统可靠性分析的重要抽样法. 强度与环境，1996，**3**.

[16] Davis L. Handbook of Genetic Algorithm. New York: Van Nonstrand Reinhold, 1991.

[17] Kirkpatrick S. Optimization by Simulated Annealing: Quantitative Studies. J. Statis. Phys., 1984, **34**: 975－986.

[18] 韩祯祥，文福拴. 模拟进化优化方法及其应用——遗传算法. 计算机科学，1995，22 (2)：47－56.

[19] 胡守仁主编. 神经网络应用技术. 长沙：国防科技大学出版社，1993.

ADVANCES AND PROSPECT OF THE GLOBAL OPTIMIZATION FOR FEEDFORWARD NETWORKS

Dong Cong

(*Dept. Civil & Architectural Engineering Tsinghua University, Beijing* 100084)

Abstract In this paper, a brief review is made on the research history of feedforward networks, and some classical works are introduced and remarked. A systematic analysis is given to those main respects related to the global optimization of multilayer feedforward networks, some fundamental conditions which need be met by any algorithm of global optimization are presented, a practicable algorithm of global optimization is suggested and a theoretical proof of the reasonableness and convergence of the present algorithm is presented. The paper points out: It is a notable trend to introduce some methods and achievements of number theory and multivariate nonlinear regression into the research fields of feedforward networks.

Key words feedforward neural network, global optimization, topology, nonlinear regression, number theory