第23卷 第1期
1997年1月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 23, No. 1
January, 1997

对多层前向神经网络研究的几点看法

阎平凡

(清华大学自动化系 北京 100084)

摘 要 从不同的领域对多层前向网络的作用本质作了分析,对泛化能力、模型选择、有限样本量等主要问题做了定性讨论;对当前前向网络研究中的一些问题提出了看法.

关键词 神经网络,多层感知器,统计建模.

1 引言

文献[1]对当前国内人工神经网络(ANN)研究的现状作了分析,作者对该文指出的一些问题也有同感. 有关联想记忆网络,文[2]已发表了一些很好的看法,这里着重讨论关于多层前向网络(简记为 MFN)研究中的问题. 国内在 MFN 方面的研究工作很多,其中不少是很好的,但也存在一些问题,一是有些理论研究不够深入,甚至没有抓住关键问题;二是有些应用没有说清楚为什么 ANN 是较合理的选择,使人感到其应用有一定盲目性. 为此,提出以下应注意的问题.

1)从多方面深入理解 MFN 作用的本质;

2)对 MFN 的主要要求是什么;

3)用好 MFN 的关键问题及主要困难;

4)对一些笼统的提法或似是而非的问题,应加以分析.

2 对 MFN 作用的理解

要理解 MFN 的作用不能只局限于对 ANN 的研究,应搞清楚其与相关领域的关系和差异.

2.1 函数逼近

用一些简单函数的复合去逼近一个未知函数 f,工程上常用的有两种类型. 一是谱方法,二是有限元法. 谱方法把 f 分解成不同频率的谐波成份,基函数的正交性显而易见. 这种分解在频域的分辨率几乎是无限的(单频的谐波在频域是一个 δ 函数),有利于频域分析. 但各基函数在时(空)域是全局性的,时(空)域分辨率很低. 有限元法则是在空(时)域把 f 分成一些小区域,每一小区可用简单的函数表示,其空(时)域分辨率高而频域分

1)国家攀登计划认知科学(神经网络)重大关键项目资助

收稿日期 1995-08-07

辨率低.

从函数逼近角度可把 MFN 看作由许多简单函数组成的复合函数,或者看作函数的综合过程. 文献[3]给出一个广义网络逼近定理.

以下各类网络在由 d 维欧氏空间紧集上的连续函数构成的空间上,在下述意义下是稠密的,即对任一上述函数 f^* 都存在一个收敛于 f^* 的网络函数序列 f_n.

所列网络包括了不同结构和不同节点函数的网络. 可以说任何一种复合函数都对应一种 MFN,反之任一 MFN 也都可写成复合函数的形式,可把网络看作逼近算法的图(Graph)示形式. MFN 表示的复合函数一般不能硬性归属于前面说的两种极端情况. 基于记忆的映射网络如:CMAC[4],资源可重置网络[5]和 RBF 网络应属于有限元类,多数以指数函数为节点函数的网络,则看作谱分解更合理.

许多工程问题,如动态系统建模,信号的滤波及变换等,都是一个函数到另一函数的映射[6],文[7]对 MFN 的逼近能力作了扩展,证明它不仅可逼近 R^n 上的函数,且可逼近函数簇(泛函),文[8]给出了更一般的多层(网络)泛函的概念,可把 MFN 看作无限维函数空间上的算子.

2.2 回归与分类

由随机样本组 (x_i, y_i) $i=1,\cdots,n$ 寻找 y 与 x 间关系的最佳(在均方误差意义下)函数 $y=f(x)$,是统计学中的一个古老分支——回归分析的研究内容,可以说任一参数或非参数回归模型都对应一个 MFN,任一 MFN 也可写出其对应的回归模型. 文[9]在从统计观点研究 ANN 的学习时,把希望输出与观测输入的关系写成 $d=g(x)+\varepsilon, g(x)=\mathbf{E}[d|x]$. 这就是回归模型. 文[10]从非线性数据分析的观点讨论了 MFN 的作用,证明当用作分类器时,是按后验概率划分的,对此,文[11]有更详细的讨论.

回归分析中有一类型,其形式与 MFN 更接近,即投影寻踪回归(Projction Purswit Regression,PPR)[12],其输入 x 与输入 y 间关系可表示为

$$u_k = \sum_{j=1}^{p} w_{kj}\, x_i - \theta_k = \boldsymbol{w}_k^T \boldsymbol{X} - \theta_k, \quad k = 1,2,\cdots,h,$$

$$y_i = \sum_{k=1}^{m} \beta_{ik} f_k(u_k) = \sum_{k=1}^{m} \beta_{ik} \sigma(u_k), \quad i = 1,2,\cdots,m.$$

其大致过程是,第一步,先把高维的输入 x 投影到低维空间 u,第二步,由 u 的非线性函数 σ 的线性组合逼近 y,要学习的参数有三个:β_{ik},平滑函数 σ 的形式和输入层数值 w_k,文[13,14]对 PPR 与 MFN 的学习作了比较,文[15]从 PPR 的观点讨论了 Cascade-Correlation 学习,文[16]介绍了 PPR 网络的实现方法.

3)正规化理论(Regularization Theory)

T. Poggio 等在文[17]中把监督学习与逼近联系起来,并用正规化方法处理它们,说明了正规化方法可等价于一个只含一个隐层的 MFN,称之为正规化网络. 进一步又提出广义正规化网络,它包含了很广一类逼近网络,文[18]是对正规化理论与 ANN 的全面总结. MFN 可看作某一正规化方法的实现,由于 MFN 学习本身就是一个反问题[1,19],它与正规化理论的联系是显而易见的.

3　对 MFN 的基本要求和主要困难

在有限数据(样本)上学习后有推广(泛化)能力,是对 MFN 的基本要求(有关泛化问题作者已有讨论[20]).除此,网络的容错性、学习速度等有时也很重要,相对其它性能而言,泛化能力是主要的.泛化能力与网络结构(规模)、给定样本数量与问题本身的复杂程度有关.问题的复杂程度无法控制,实际可存在两类问题:

1)给定网络结构,为达到好的推广能力需要多少训练样本.

2)训练样本量已定,确定合理的网络结构以保证好的推广能力.

第二类问题实际上遇到的最多,即在样本量已定时,如何根据对问题的先验知识(如果有的话)来设计网络结构,下一步是用 BP 法学习参数(权和阈值).前者是结构学习,后者是参数学习,目前大部分研究是针对参数学习的,应看到更有决定意义的是结构设计,目前对此问题尚没有理想的方法,文[21]对此作了初步探讨,下一节将从统计建模观点讨论一下基本原则.

4　统计建模与最简单原则

建模是一种对未知世界的逼近方法,对模型的三个基本要求是:Flexibility(适用性),Dimensionality(计算复杂性)和 Interpretability(透明性).Flexibility 是模型对不同问题提供准确逼近的能力,不准确(模型与真实规律的偏离)表现为估计结果的偏置(Bias).Dimensionality 是据有限样本确定模型中的参数时产生的不稳定性,表现为估计结果的方差(Variance),所谓"维数灾难"是指为把方差限制到要求的范围内所需数据量随维数(自变量数)的迅速增长(如指数增长).Flexibility 和 Dimensionality 是一对矛盾,对此可简单说明如下[22]:

给定数据(x_i,y_i)后,在均方误差意义下 y 对 x 的回归 $E[y|x]=\int yP_{y|x}(x,y)\mathrm{d}y$ 是最好的模型.因为

$$E[(y-\hat{f}(x)^2)|x]=E[(y-E[y|x])^2|x]+[\hat{f}(x)-E[y|x]]^2$$
$$\geqslant E[(y-E[y|x])^2|x].$$

可用后一项,即 $\hat{f}(x)$与 $E[y|x]$的接近程度作为 $\hat{f}(x)$好坏的度量,经过推导可得

$$E[(\hat{f}(x)-E[y|x])^2]=E[((\hat{f}(x)-E[\hat{f}(x)])+(E[\hat{f}(x)]-E[y|x]))^2]$$
$$=E[(\hat{f}(x)-E[\hat{f}(x)])^2]+E[(E[\hat{f}(x)]-E[y|x])^2]$$
$$+zE[(\hat{f}(x)-E[\hat{f}(x)])(E[\hat{f}(x)]-E[y|x])]$$
$$=(E[\hat{f}(x)]-E[y|x])^2 \qquad \text{偏置}$$
$$+E[(\hat{f}(x)-E[\hat{f}(x)])^2] \qquad \text{方差}$$

结合 MFN 看,若网络过于简单,则不足以反映未知规律 $y=f(x)$ 的复杂性,因而无法准确逼近 $f(x)$(偏置大).反之若结构过于复杂,则要由给定数据学习的未知参数太多,使学习结果不稳定(方差大).

14 世纪的法国修道士 William Ockham 提出过一个最简单(最节省)原则:"与已知事

实符合(一致)的理论中的最简单者就是最好的理论”,后人称此原则是“Occam's Razor”. 文[23]从 Bazes 观点对此作了一个定量说明. 这一思想与爱因斯坦的一句名言“Any theory must be as simple as possible, but not simpler”的精神是一致的. 对建模问题说就是与给定数据(例子,或训练样本)一致的最简单者是最好的选择. 所以通常的建模准则总包含两项,一是模型输出与给定数据一致(用对数似然函数表示,当数据为高斯分布时就是均方误差),二是对模型复杂性(参数多少)的约束项(此项的作用可理解为适当引入偏置以减少方差). 如信息判据(AIC, NIC)[24],正规化[25],最小描述长度(MDL)[26,27]都如此. 模型有泛化能力是基于它表示的规律具有某种程度的平滑性(简单性),正规化理论,回归模型也都是以不同形式对 $f(x)$ 的平滑性加以约束.

5 几点看法

对于为什么采用 ANN 以及用好 ANN 的关键问题,有一些过于笼统甚至似是而非的提法,应当分析. 例如:

1)“ANN 的构造和作用机制更接近人脑,所以对于模式识别,对环境自适应等问题用 ANN 比其它方法好”。应当看到现有的 ANN 模型只是人脑的一种表面的(非本质的)粗略近似,其规模和机制的复杂程度离人脑还有很大差距,现有 ANN 仍是一种工程上的人造模型. 用现有的 ANN 模型去解决上述一类问题并不注定比传统方法好,传统方法中的一些基本困难(如模型选择、变量选择、有限样本问题)也没有消失.

2)“ANN 本质上具有学习和自适应能力,只要有例子,ANN 就能通过学习解决问题”. 并非只有 ANN 才有学习能力,例如有较长历史的模式识别与可训练机器理论[28,29]也是通过学习来解决问题的,也并非只要有例子就一定能学到其中隐含的规律,哪一类问题是可以学习的,对那些可以学习的问题其学习的复杂性如何,正是计算学习理论研究的问题[30—32],一些研究者不理解上述问题的重要性,以为学习速度是主要矛盾.

3)“ANN 可以模拟任何非线性关系,且不需有关系统的模型知识”. 并非只有 ANN 具有逼近任何非线性规律的能力,同时有能力的方法不一定就是好方法,还要看它的计算复杂性是否可以接受(有效性). 非参数回归可以模拟任何非线性规律,很早对它的一致性(即随着样本数 n 的增加误差一致下降,且 $n\to\infty$ 时 $e\to o$)已给出了理论证明,但至今非参数回归用的不多(有一些被称为非参数回归的方法实际上是一种半参数法),这是由于所需样本数随问题规模急剧增长(Dimensionality 不好),文[33]证明非参数回归的收敛速度正比于 n^{-r},其中 $r=(p-m)/(2p+d)$,p 代表未知规律 f 的平滑程度,m 是要估计的 f 的导数次数,d 为输入变量数,粗略看可以认为逼近误差按 $n^{-(p/d)}$ 下降.

对 MFN 尚没有明确的规律,有人对逻辑型网络(奇偶校验网络)作了试验,把逻辑函数看作谓词,输入变量维数是谓词的阶数,结果发现学习时间随输入维数指数上升[34].

4)“传统的模式识别和回归分析要事先抽取特征,ANN 可自动压缩变量维数”,MFN 的隐层可理解为对输入变量做了变换,但这代替不了开始的特征(变量)选择,特别是当样本量有限时提高网络性能的一个重要措施是压缩原始变量的维数,这一工作可用 ANN[35]也可以用其它方法,有时还要依靠专家知识.

此外还有以下问题值得注意:

1)有限样本量问题　样本量有限是统计模式识别和回归分析的主要困难,如果样本量不受限制(也有足够的计算资源来处理这些样本),则不管模型多复杂也不会出现过拟合现象,统计建模中的 Dimensionality 问题也不存在了. 最近文[36]对此问题作了分析,结合 ANN 的讨论可见文[37,38]. 没有其它先验知识时,解决问题的信息全靠从训练样本得到. 能达到的最好效果决定于样本的数量和质量(如果认为样本来源可靠则只决定于样本的数量). 希望所选网络尽可能把隐含在样本中有关问题的信息都提取出来,这就要求网络与样本量"匹配"[38].

2)变量选择问题　前面已指出,为保证估计结果的稳定性,所需样本量大致按问题规模(输入变量维数)指数上升,可见当样本量有限时,保证推广能力的最有效方法是降低输入向量的维数,只保留那些重要的变量. 特征选择的具体方法在模式识别和回归分析中都有,这里不再讨论.

3)先验知识的运用　如果能利用先验知识,则有可能限制模型的复杂性,对解决问题有很大帮助,特别是样本量少而问题又较复杂时,应设法尽量利用先验知识. 对此文[39]作了讨论,许多人主张把传统人工智能技术与 ANN 结合起来,在一定程度上也是由于传统 AI 更易于利用人类的先验知识以降低 ANN 学习的负担.

6　讨论

提出以上看法不是想说明 ANN 的许多能力别的方法也能做到,从而限制 ANN 的应用与研究,也不能要求事先把为什么用 ANN 完全搞清楚才能用它,因为未经过实践是不能搞清楚的,只是为了更深入有效的研究 ANN 及其应用提供一些看法,供大家深入思考. MFN 和一些传统方法的等效性说明原来方法的一些基本问题仍然存在,它们的研究结果对研究 ANN 是很有用的,出现上述问题的一个原因是在 ANN 的发展初期,国外一些文章较多强调了 ANN 与传统方法的不同,而对使用中的一些困难,如结构设计(建模问题)、样本量、变量选择等很少讨论. 有一些应用文章(不少是权威性的工作)对上述问题的处理不够合理,例如有些应用中,模型中的未知参数(权系数个数)甚至比样本数还多. 对这种情况最近国外也有所讨论[40],应引起足够注意. 当然 MFN 确有自己的特点,作者认为主要是:

1)构造上更接近生物神经系统. 信息的表示与处理是分布在联接权上,尽管离真实脑模型还有很大距离,其功能也没有达到所期望的那样强,但总是朝这方向前进了一步.

2)有很大的灵活性,同一形式的 MFN,可用作函数逼近、回归、正规化等等,而且更形象化. 在某种意义上也可以灵活地引入不同性质的约束,也可方便地引入信息反馈与上下文信息,加一些局部反馈又可模拟时间序列或动态系统等等. 促进了上述各方法的融合.

3)本质上的并行处理及结构上的规则性,对硬件实现有利.

参　考　文　献

[1]　张承福. 对当前神经网络研究的几点看法. 力学进展,1994,**26**(2):181—186.

[2]　张承福等. 联想记忆神经网络中的若干问题. 自动化学报,1994,**20**(5):513—521.

[3] Barron A R. Statistical properties of artificial neural networks. Proc. 28th Conference on decision and control, 1989, 280—285.

[4] Albus. A new approach to manipulator control, the cerebellar model articulation controller (CMAC). *J. on Dynamical Systems Measure and control*, 1975, 220—227.

[5] Platt J. A resource-allocating network for function interpolation. *Neural computation*, 1991, **3**: 213—225.

[6] Barrett J F. The use of functional in the analysis of nonlinear physical system. *J. Electro and control*, 1963, **15**: **567—615**.

[7] Chen Tianping *et al*. Approximation of continuous functions by NN with application to dynamic systems. *IEEE Trans. NN*, 1993, **4**(6): 910—918.

[8] Modha D S. *et al*. Multilayer functionals, in Mathamatical Approch to NN, Taylor J G ed. Amsterdam Elsevier Science Publishers, 1993. 235—260.

[9] White H. Artificial neural networks, approximation and learning theory. Cambirdge, MA. Blackwell.

[10] Asoh H. Nonlinear data analysis and MLP. Proc. of IJCNN-90, 1990, Ⅱ 411— Ⅱ 415.

[11] Gish H. Apobabilistic approach to the understanding and training of NN classifier. Proc. ICASSP'90, 1990, 1361—1364.

[12] 成平等. 投影寻踪——一类新的统计方法. 应用概率统计, 1986, **3**: 267—276.

[13] Maechler M, *et al*. Projection pursuit learning networks for regression. Proc. Tools for AI, IEEE Press, 1990, 350—358.

[14] Hwang Jeng-Nang, *et al*. Regression modeling in backpropagation and projection pursuit learning. *IEEE Trans. NN*, 1994, **5**(3): 342—357.

[15] Hwang Jeng-Nang, *et al*. The cascade-correlation learning, A projection pursuit learning perspective. *IEEE Trans. NN*, , 1996, **7**(2): 278—288.

[16] Zhao Ying, *et al*. Implementing projection pursuit learning, *IEEE Trans. NN*, 362—392.

[17] Poggio T, *et al*. Networks for approximation and learning. Proc. IEEE, 1990, **78**(9): 1481—1496.

[18] Girosi F, *et al*. Regularization theory and neural network architectures. *Neural computation*, 1995, **7**: 219—296.

[19] Ogawa H. Neural network theory as inverse problem. *J. EICE*, Japan, 1990, **73**: 690—759.

[20] 阎平凡. 多层前馈网络的推广和学习问题. 中国神经网络 93 年学术会议论文集, 1993 年, 上册, 68—77.

[21] 阎平凡. 现在神经网络解决问题的困难与结构自适应问题. 中国神经网络 94 年学术大会报告, 1994, 10 月, 武汉.

[22] Gemman S. Neural networks and the Bias/variance dilemma. *Neural computation*, 1992, 4: **1—57**.

[23] Garett A J. Ockham's Razor, in maximum entropy and bayes method, Grandy W T, *et al*. Ed. Kluwer Academic Publisher, 1991. 357—364.

[24] Nobora, *et al*. NIC-Determing the number of hidden unit for ANN. *IEEE Trans. NN*, 1994, **5**(6): 865—872.

[25] Wu lizhong, *et al*. A smoothing regularizor for feedfroward and recurrent neural networkds. IEEE Trans. NN. 1996, **8**: 461—489.

[26] 阎平凡. 最小描述长度与多层前馈网络设计中的一些问题. 模式识别与人工智能, 1993, **6**(2): 143—148.

[27] Rohwer R, *et al*. Minimum description length, regularization and multimodel data. IEEE Trans. NN. 1996, **8**: 595—609.

[28] 边肇祺等. 模式识别. 北京: 清华大学出版社, 1988.

[29] Sklansky J. 著, 阎平凡等译. 模式分类器和可训练机器. 北京: 科学出版社, 1987

[30] 张鸿宾. 计算学习理论与其应用(1)(2). 计算机科学, 1992, **19**: 18—22.

[31] Anthony M, *et al*. Computational learning theory. Cambridge, England; Cambridge University Press, 1992.

[32] Anthony M, *et al*. Computational learning theory for ANN. in Mathematical approcach to NN, J G Taylor Ed. 1993, 25—62.

[33] Stone C J. Optimal global rate of convergence for nonparametric regression. *ANN. Stat*, 1982, **10**(4): 1040—1053.

[34] Tesauro G, *et al*. Scaling relationships in BP learning. *Complex system*, 1988, **2**: 39－44.

[35] Mao Jianchang, *et al*. ANN for feature extraction and multivariate data projection. *IEEE Trans. NN*, 1995, **6** (2): 296－317.

[36] Randy S J, *et al*. Small sample size effects in statistical pattern recognition, recommendation for practice. IEEE Trans. PAMI-13, 1991, 252－264.

[37] 张鸿宾. 训练多层网络的样本数问题. 自动化学报, 1993, **19**: 71－77.

[38] 阎平凡. 人工神经网络的容量, 学习与计算机复杂性问题. 电子学报, 1995, **23**(4): 63－67.

[39] Abu-Mostafa Y S. Hints. Neural computation, 1995, **7**: 639－671.

[40] Duin R. Superlearning and neural network magic. *Pattern recognition letters*, 1994, **15**: 215－217.

SOME VIEWS ON THE RESEARCH OF MULTILAYER FEEDFORWARD NEURAL NETWORKS

YAN PINGFAN

(*Dept. of Automation, Tsinghua University, Beijing* 100084)

Abstract The essential function of multilayered feedforward nerwork was analysed from the point of view of various discipline. Generalization ability, model selection and the problem of limited sample size were discussed qualitatively. Some problems in the research of multilayer feedforward network are pointed out.

Key word Neural network, multilayer perception, statistical modeling.

阎平凡　简介见本刊第 21 卷第 1 期.