文章编号：1000－8152(2001)02－0257－03

一种估计人工神经网络泛化误差的新方法*

李 杰 韩正之
(上海交通大学智能工程研究所·上海，200030)

摘要：神经网络的结构学习就是要确定网络的拓扑，使之有较好的泛化能力．本文考虑了确定性前向网络，而其训练集合是随机点集的结构学习问题．文章定义了一种新的结构学习目标函数，给出了它与目前常用的目标函数比较的优越性，讨论了相关的学习算法，还给出了一个例子说明这种学习的效果．

关键词：人工神经网络；泛化误差；结构学习；随机点集
文献标识码：A

A New Method to Estimate the Generalization Error of Artificial Neural Network

LI Jie and HAN Zhengzhi
(Institute of Intelligence Engineering, Shanghai Jiaotong University · Shanghai, 200030, P.R.China)

Abstract: The constructional learning is used to determine the architecture of neural network such that the network holds a satisfactory generalization. This paper considers the constructional learning in the case where the training set is randomly chosen from an input-output space. A new objective function of constructional learning is presented. It is illustrated the reason why this objective function is superior to other functions. The learning algorithm for this objective function is also analyzed. Finally, a simulation example is given to show the efficiency of the method presented in this paper.

Key words: aritificial neural network; generalization; constructional learning; stochastic set

1 引言(Introduction)

考察一个神经网络的精度主要有两个指标：一致性(university)和泛化能力(generalization)．一致性是指人工神经网络在样本上的逼近能力．泛化能力描述了人工神经网络在整个输入和输出空间上的逼近能力[1~8]．在应用神经网络建模的时候，泛化能力应该是一个更为重要的指标，近年来受到较多的重视．

决定泛化能力主要有两个因素：一是网络的结构，一是网络的训练程度．已经知道，过分地追求在样本点上的逼近会产生过度训练(over-trainning)，导致网络的泛化能力的减弱[9,10]．因而提高泛化能力主要应该从网络结构着手．结构学习主要有结构删除法(pruning)[8]和结构增长法(growing)[10]．这两者都是通过定义一个反映网络泛化能力的目标函数，然后结构学习就转化为优化这个函数．已经提出的一些结构学习方法都没有将样本点集作为随机集合来考虑，这样做一方面与实际不相符合，另一方面又影响网络的泛化能力．本文在考虑了样本点集的随机性的基础上，提出了一种新的与泛化能力有关的目标函数，并且讨论了这种目标函数的优越性及相应的学习算法．

2 神经网络的结构学习(The structural learning of neural networks)

1) 问题的描述．

设神经网络要逼近的函数为 $g(x): X \to Y_1, X \subset \mathbb{R}^n, Y_1 \subset \mathbb{R}^m$．$g(x)$ 的输出受到可加噪声的干扰，即量测为 $y = y_1 + n = g(x) + n$，n 满足正态分布 $N(0,\sigma^2)$．设 $x \in X$ 为随机变量，其密度函数为 $\rho(x)$．样本 $(x,y) \in X \times Y$，其密度为 $p(x,y)$，根据条件概率公式得：$p(x,y) = \rho(x) \cdot p(y \mid x)$[9,11]．因为 $y = g(x) + n$，所以

$$p(y \mid x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}[y - g(x)]^2\right\}, \tag{1}$$

于是

* 基金项目：本文的部分研究工作得到国家自然科学基金(69874025)的资助．
收稿日期：1998－12－28；收修改稿日期：2000－01－10．

$$p(x,y)=\rho(x)\cdot\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}[y-g(x)]^2\},\tag{2}$$

神经网络的学习样本集 $D=\{(x_i,y_i),i=1,2,\cdots,m\}$ 就是按 $p(x,y)$ 在 $X\times Y$ 中取得的.

神经网络的泛化能力一般由泛化误差来度量.用 $f_w(x)$ 表示神经网络确定的函数,其参数 w 为网络权值,它的泛化误差 E_g 的定义为[4,7]:

$$E_g[f_w(x)]=\int p(x,y)[y-f_w(x)]^2\mathrm{d}y\mathrm{d}x.\tag{3}$$

2) 神经网络的结构学习.

这里讨论只有一个隐含层的前向神经网络.神经网络的结构学习就是决定最优的隐含层神经元个数 n 的过程,通常将它转化为求取结构泛化误差的极小值.

在大多数结构学习中,网络结构 $F_n=\{f_n(w)\}$ 的泛化误差 E_{F_n} 定义为:在这个映射族中所有神经网络映射 $f_n(w,x)$ 的泛化误差 $E_g[f_n(w,x)]$ 的下确界[6,9],$E_{F_n}=\inf\limits_{w}\{E_g[f_n(w,x)]\}$.事实上,由于(3)式含有未知函数,因而无法得到 E_{F_n}.通常的应用中总利用 μ-LMS 算法得到的 w_D,把 w_D 当作最优权值 w^*,将映射 $f_n(w_D)$ 的泛化能力当作结构 F_n 的泛化能力来进行结构学习[7,8].

3 一种新的泛化指标(A new criterion of generalization)

为了消除样本点集的随机性对决定结构泛化误差的影响,定义网络结构 F_n 的泛化误差为一个数学期望,即:

$$E_{F_n}=\mathrm{E}_w[E_g[f_n(w_D,x)]]=\int p_n(w_D)\cdot E_g[f_n(w_D,x)]\mathrm{d}w_D,\tag{4}$$

其中的 $\mathrm{E}_w[\cdot]$ 为相对于权值求期望值.

为了计算 E_{F_n},首先求取 $p_n(w_D)$.(x_i,y_i) 为样本点,其输出含有分布为 $N(0,\sigma^2)$ 的随机噪声 n,所以 (x_i,y_i) 相对于权值 w 的条件分布为正态分布,即其密度为

$$p(x_i,y_i\mid w)=\frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(f_n(w,x_i)-y_i)^2],\tag{5}$$

同理,记 $\underline{x}=[x_i]_{m\times1}$,$\underline{y}=[y_i]_{m\times1}$,样本点集 $D=\{(x_i,y_i),i=1,2,\cdots,m\}$ 相对于权值 w 的密度函数为

$$p(\underline{x},\underline{y}\mid w)=\prod_{i=1}^{m}p(x_i,y_i\mid w).\tag{6}$$

假设初始权值的分布为 $\rho^{(0)}(w)$,则权值相对于样本点集为 $D=(\underline{x},\underline{y})$ 的条件分布:

$$\rho(w\mid\underline{x},\underline{y})=\frac{p(w,\underline{x},\underline{y})}{p(\underline{x},\underline{y})}=\frac{p(\underline{x},\underline{y}\mid w)*\rho^{(0)}(w)}{p(\underline{x},\underline{y})}=\frac{\rho^{(0)}(w)*\prod\limits_{i=1}^{n}p(x_i,y_i\mid w)}{\int_w\rho^{(0)}(w)*\prod\limits_{i=1}^{n}p(x_i,y_i\mid w)\mathrm{d}w}.\tag{7}$$

(7)式反映了为了适应样本点集 $D=(\underline{x},\underline{y})$,权值 w_D 的分布情况.在计算上式时,一般认为初值的绝对值应小些,因此,定义每个初始权值的分布为 $N(0,k^2)$,k^2 为一个根据实际情况选取的常数.

神经网络映射 $f_n(w,x)$ 的泛化误差 $E_g[f_n(w,x)]$ 的计算有许多成熟的方法[4,5]:如 Akaike 信息判据法(AIC)、最终预测误差法(FPE),Cross-Validation 等.这里采用 FPE 方法,利用样本点集 D 上的学习误差函数 $E_D(w)$,得到神经网络 $f_n(w,x)$ 的泛化误差为:

$$E_g[f_n(w,x)]=\frac{1}{N}*\frac{N+d}{N-d}E_D(w),\tag{8}$$

式中的 N,d 分别为学习样本点的数目,权值的维数,对于一个网络结构是固定不变的.把上式代入结构泛化误差估计的定义式得:

$$E_{F_n}=\frac{1}{N}*\frac{N+d}{N-d}\int_{w_D}p_n(w_D)E_D(w_D)\mathrm{d}w,\tag{9}$$

上式中所有变量都是已知的或可计算的,从而,网络结构的泛化误差 E_{F_n} 是可以计算的.

4 例子(Example)

假设被逼近函数为:$g(x)=\cos(x)$,$x\in[-5,5]$,其中的输出可加性噪声为正态分布,方差为 0.2.学习样本随机产生,数目为 100 个.然后,利用结构增长方法,从隐层神经元数目为 2 的结构开始学习,直到神经元数目为 20 为止.其中神经网络的泛化误差是利用公式(9)得到的.其中隐层节点数目为 6 时,泛化误差最小,即对于这个逼近问题,网络的最优结构是含有 6 个隐层节点,然后进行参数学习.其相对应的结果如图 1 ~ 图 3 所示:

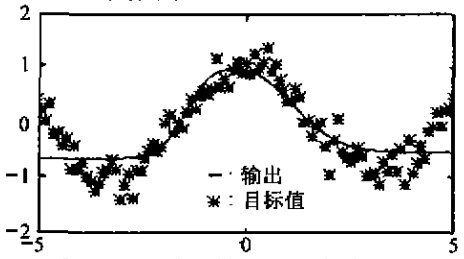

图 1 2个隐层神经元时的逼近效果

Fig. 1 The approximation of the output of the network with 2 neurons in hiden layer

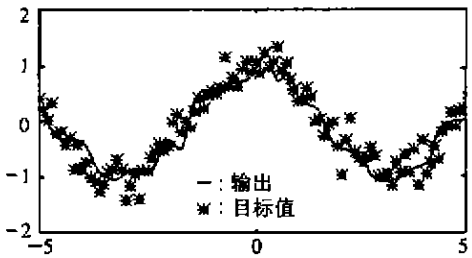

图 2 20个隐层神经元时的逼近效果

Fig. 2 The approximation of the output of the network with 20 neurons in hiden layer

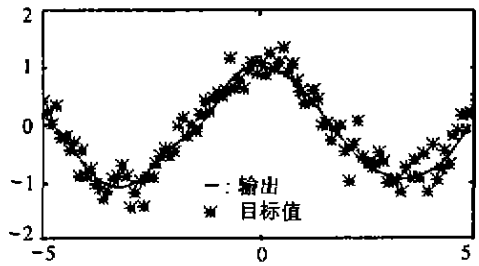

图 3 6个隐层神经元时的逼近效果

Fig. 3 The approximation of the output of the network with 6 neurons in hiden layer

5 结束语(Conclusion)

本文考虑了一类确定的网络结构,但是由于样本点集的随机性,通过学习得到的权值也是随机的.因此将一个结构的泛化误差定义为学习样本点集后泛化误差的期望.采用本文的定义,需要知道权值的分布.这个权值分布的计算可能比较复杂,不过在大多数情况下,可设它满足正态分布.利用这个分布,可以得到网络的结构泛化误差,进而可以确定一个最优结构,这个结构的选取可以采用删除法或增长法.

参考文献(References)

[1] Cybenko G. Approximation by superposition of a sigmoidal function [J]. Mathematical Control Signals Systems, 1989,(2):303 – 314

[2] Hornik K and Stinchcombe M. Multilayer feedforward networks are universal approxiamtors [J]. Neural Networks, 1989,(2):359 – 366

[3] Hassoun M H. Foundation of Artificial Neural Network [M]. Cambridge: MIT Press,1995

[4] Dong Cong and Liu Xila. A study on generalized BP algorithm of neural network and its fault tolerance and generalization capability [J]. Control and Decision, 1998,13(2):120 – 124(in Chinese)

[5] Murata N, Yoshizawa S and Amari S. Network information criterion – determining the number of hidden units for an artificial neural network model [J]. IEEE Trans. on Neural Networks, 1994,5(6):865 – 871

[6] Kwok Tin-Yau and Yeung Dit-Yan. Constructive algorithms for structure learning in feedforward neural networks for regression problem [J]. IEEE Trans. on Neural Networks,1997,8(3):630 – 645

[7] Chung F L and Lee T. Network-growth approach to design of feedforward neural networks [J]. IEE Proc.-Control Theory Appl., 1995,142(5):572 – 581

[8] Russell Reed. Pruning algorithms – a survey [J]. IEEE Tran. on Neural Networks,1993,4(5):239 – 242

[9] Amari S and Murata N. Asymptotic statistical theory of overtraining and cross-validation [J]. IEEE Trans. on Neural Networks,1997,8(5):985 – 996

[10] Doering Axel, Galicki Miroslaw and Witte H. Structure optimization of neural networks with the A*-algorithm [J]. IEEE Trans. on Neural Networks,1997,8(6):1434 – 1445

[11] Rogn L. System Identification – The Theory for Users [M]. Shanghai: East China Normal University Press,1990(in Chinese)

本文作者简介

李 杰 1971年生.上海交通大学控制理论与应用专业博士生.研究方向为人工智能,神经网络及其在系统辨识中的应用.

韩正之 1947年生.教授,上海交通大学控制理论与应用专业博士生导师,智能工程研究所所长.研究领域为:非线性控制,神经控制,远程教育.