神经网络的学习误差函数及泛化能力*

李 杰　韩正之
(上海交通大学智能工程研究所　200030)

摘　要　用于训练神经网络的样本点集不可避免地会受到噪声污染。利用神经网络的概率描述,通过研究 K—L 信息距离和神经网络泛化能力的关系,构造一个新的神经网络学习误差函数。泛化能力分析和仿真结果表明了该学习误差函数的合理性。

关键词　神经网络,泛化能力,学习误差函数,概率表示,K—L 信息距离
分类号　TP 18

The Learning Error Function of Neural Network and Its Generalization

Li Jie, Han Zhengzhi
(Shanghai Jiaotong University)

Abstract　The training data is inevitably contaminated with noise. At first the neural network is represented with a probability relation. Then the K_L information distance is adopted to be a learning error function, which keeps the consistence with the generalization. Finally the effectiveness of the new learning functions is proved through generalization analysis and the simulation example.

Key words　neural network, generalization, learning error function, probability representation, K_L information distance

1 引　　言

进入 80 年代中期,神经网络的研究和应用开始活跃起来,其中研究的重点在于神经网络的非线性逼近能力和学习能力[1,2]。神经网络的学习就是根据样本点集得到一个学习误差函数,然后优化这个函数的过程。学习误差函数一般取为神经网络在样本点集上的偏差 $E_m(w)$。但由于各种原因,样本点集中不可避免地会受到噪声的污染,在优化 $E_m(w)$ 的同时,神经网络的学习也会受到噪声的干扰,由此产生过度学习等现象,严重影响了神经网络的泛化能力[3,4]。

针对噪声的干扰,本文利用神经网络的概率描述,通过研究 K—L 信息距离和神经网络泛化能力的关系,构造一个新的神经网络学习误差函数。最后,用一个简单的仿真实例来说明改进的学习误差函数的有效性。

2 神经网络的概率描述和 K—L 信息距离

用来训练神经网络的学习样本点集合为 $D=\{(x_i,y_i),i=1,2,\cdots,m\}$。由于系统的输出受到满足正态分布 $N(0,\sigma_1^2)$ 的可加性噪声 n^1 的污染,因此每个样本点的输出为 $y_i=g(x_i)+n_i$,其中 n_i 为噪声的一个简单子样。因为不能把噪声 n_i 从输出 y_i 中分离出去,为此在神经网络 $f(x,w)$ 的输出上,也加上一个满足正态分布 $N(0,\sigma^2)$ 的噪声 n。然后优化网络权值 w 和噪声方差 σ^2,通过 $f(x,w)+n$ 来逼近实际输出 $y=g(x)+n^1$,使得神经网络 $f(x,w)$ 能够逼近真实函数 $g(x)$(如图 1)。

当神经网络 $f(x,w)$ 输出上的噪声 n 的方差为 σ^2 时,将其概率记为 $p_{w,\sigma}(x,y)$,表示此时随机变量 (x,y) 在输入输出空间 $X\times Y$ 中的密度函数[5]。其表达式为

* 国家自然科学基金项目(69874025)
1998 - 08 - 31 收稿,1999 - 01 - 18 修回

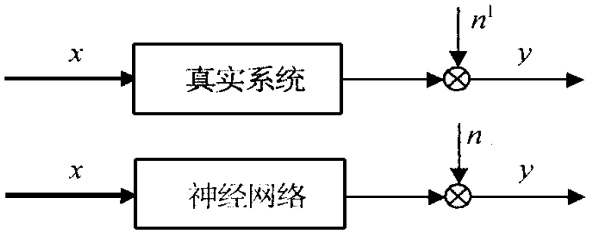

图1 用加噪声的神经网络逼近受噪声污染的真实系统

$$P_{w,\sigma}(x,y) \equiv p(x)p(y|x,w,\sigma) = p(x)\left\{\frac{1}{\sqrt{2\pi\sigma}}\exp\left[\frac{1}{2\sigma^2}(y - f(w,x))^2\right]\right\} \quad (1)$$

式中 $p(x)$ 为输入变量 x 在输入空间上的概率密度函数。

记 $p_{w^1,\sigma_1}(x,y)$ 为真实系统 $y = g(x) + n^1$ 的概率描述，其中 n^1 的方差 σ_1 未知，神经网络的概率为 $p_{w,\sigma}(x,y)$。则它和 $p_{w^1,\sigma_1}(x,y)$ 的K—L信息距离定义为[6]

$$I(p_{w^1,\sigma_1}(x,y), p_{w,\sigma}(x,y)) = \int p_{w^1,\sigma_1}(x,y)\ln\frac{p_{w^1,\sigma_1}(x,y)}{p_{w,\sigma}(x,y)}\mathrm{d}y\,\mathrm{d}x \equiv J(w,\sigma) - H_{w^1} \quad (2)$$

其中

$$J(w,\sigma) = \int p_{w^1,\sigma_1}(x,y)\frac{1}{2\sigma^2}(y - f_w(x))^2\mathrm{d}y\,\mathrm{d}x + \ln\sigma$$

而 $H_{w^1} = J(w^1,\sigma_1)$ 与权值 w 和 σ 无关。可以证明，K—L信息距离恒为正值，当且仅当这两个神经网络的权值相同，即 $w = w^1$；同时输出上的噪声也相同，即 $\sigma = \sigma_1$ 时，K—L信息距离为零。此时，神经网络 $f(x,w)$ 便可精确实现真实系统 $g(x)$。

K—L距离给出了两个概率描述之间符合程度的一个度量，可用作神经网络学习的误差函数，因此神经网络的学习过程转化为如下的最优化问题

$$\min_{w,\sigma}\{I(p_{w^1,\sigma_1}(x,y), p_{w,\sigma}(x,y)), w \in W, \sigma \geq 0\} \quad (3)$$

由于 H_{w^1} 与权值 w 和 σ 无关，所以神经网络的学习过程又可化为

$$\min_{w,\sigma}\{J(w,\sigma), w \in W, \sigma \geq 0\} \quad (4)$$

即函数 $J(w,\sigma)$ 为神经网络的学习误差函数。

3 学习误差函数的实现

神经网络的优劣，具体体现在它的泛化误差 $E_g(w)$ 上，泛化误差越小，泛化能力越强，即

$$E_g(w) = \int p_{w^1,\sigma_1}(x,y)(y - f_w(x))^2\mathrm{d}y\,\mathrm{d}x \quad (5)$$

式中 y 为系统的观测输出(含噪声)。而学习误差函数 $J(w,\sigma)$ 可化为

$$J(w,\sigma) = \int p_{w^1,\sigma_1}(x,y)\frac{1}{2\sigma^2}(y - f_w(x))^2\mathrm{d}y\,\mathrm{d}x + \ln\sigma = E_g(w)/2\sigma^2 + \ln\sigma \quad (6)$$

由优化理论知，当达到最优值时，$\sigma \equiv E_g(w)$，则

$$\min_{w,\sigma}J(w,\sigma) = \frac{1}{2}\min_{w}\ln E_g(w) + \frac{1}{2} \quad (7)$$

上式说明函数 $J(w,\sigma)$ 与泛化误差 $E_g(w)$ 之间存在一致性。因此，以 $J(w,\sigma)$ 为学习误差函数时，不会出现过度学习的情况。但在学习误差函数 $J(w,\sigma)$ 的表达式(7)中，$E_g(w)$ 和噪声 n 的方差 σ^2 是未知的，必须利用样本点集合 D，分别构造两个函数作为 $E_g(w)$ 和 σ^2 的逼近。

首先，利用样本点上的偏差

$$E_m(w) = \frac{1}{m}\sum_{x_i \in D}(y_i - f_w(x_i))^2$$

代替式中的 $E_g(w)$。因为是用 $f(x,w) + n$ 来逼近样本点集 D，因此

$$n_i = y_i - f(w,x_i), \quad (x_i,y_i) \in D$$

是分布满足 $N(0,\sigma^2)$ 的噪声 n 的简单子样，于是有

$$E_2(w) = \frac{1}{m}\sum_{i=1}^{m}\left\{[y_i - f_w(x_i)] - \frac{1}{m}\sum_{j=1}^{m}[y_j - f_w(x_j)]\right\}^2 \approx \sigma^2 \quad (8)$$

将 $E_m(w)$，$E_2(w)$ 分别代入(6)式，则学习误差函数 $J(w,\sigma)$ 化为

$$J(w) = \frac{E_m(w)}{2E_2(w)} + \frac{1}{2}\ln E_2(w) \quad (9)$$

此时，神经网络的学习过程便转化为以 $J(w)$ 目标函数的无约束极小化问题，而对 $J(w)$ 的优化有许多成熟的方法。学习完成后，根据式(7)，神经网络的泛化误差 $E_g(w)$ 为

$$E_g(w) = \exp(2\min\{J(w,\sigma) - 1\}) \approx \exp(2J(w) - 1) \quad (10)$$

4 仿真结果

我们设计了一个简单例子，用于说明BP算法的缺点和改进后的学习效果。其中的真实系统为1—10—1结构的神经网络，它的权值和神经元的偏差向量分别是随机产生的，输出上的可加性噪声的

方差为0.2。学习用的神经网络为1—20—1结构, 学习样本点为区间[-5, 5]上的100个点。由于神经网络结构的确定一直缺乏理论指导, 在实际应用中往往采用实验法来选取网络结构。

神经网络的学习采用BP算法。当学习误差函数为$E_m(w)$时, 学习结果如图2所示。图中实线为真实系统, 虚线为学习后的神经网络输入输出关系。显然, 神经网络在过于拟合样本点中的噪声使得整体逼近效果并不理想, 即泛化能力不好。

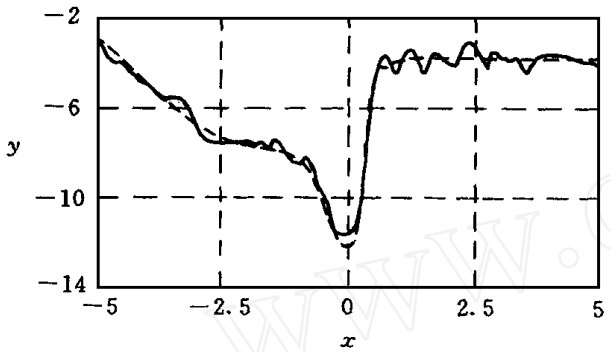

图2　学习误差函数为$E_m(w)$时过度学习的结果

利用学习误差函数式(9), 得到的仿真结果如图3所示。利用式(10)得到神经网络的泛化误差估计为0.212, 非常接近真实系统的噪声的方差0.2, 这说明逼近效果很好。

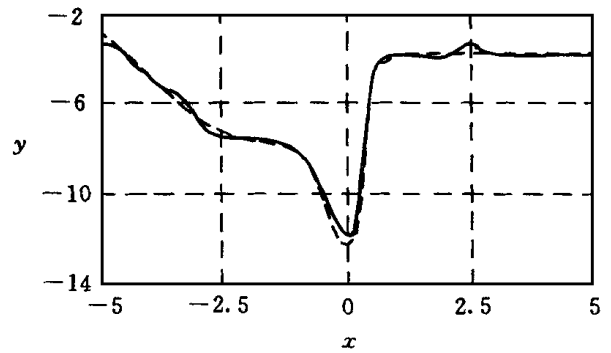

图3　利用学习误差函数(9)得到的学习结果

5　结　　论

本文从可加性随机干扰的观点出发, 研究了神经网络学习误差函数的选取问题。通过在神经网络的输出上也加上一个正态分布的噪声, 以二者的和来逼近包含噪声的学习样本点。这样, 神经网络的输出便可当作以权值为参数的概率分布来处理, 这属于传统概率统计范畴。文中证明了K—L信息距离与泛化能力的一致性, 说明用它作为神经网络的学习误差函数是合理的。

在本文的讨论中, 都是假定神经网络的结构已经确定, 且该网络结构可以实现真实系统, 然后对网络结构进行权值学习。这一不足有待于今后进一步研究和改进。

参 考 文 献

1　Hornik K. Approximation capabilities of multilayer feedforward network. Neural Networks, 1991, 4: 251_257

2　Cybenko G. Approximation by superposition of a Sigmoid function. Signals and Systems, 1989, 2(4): 303_314

3　Levin E, N Tishby, S A Solla. A statistical approach to learning and generalization in layered neural networks. Proc IEEE, 1990, 78(10): 1568_1574

4　董聪, 刘西拉. 广义BP算法及网络容错和泛化能力的研究. 控制与决策, 1998, 13(3): 120—124

5　陈开明. 概率论与数理统计. 上海: 上海科学技术出版社, 1989

6　L荣. 系统辨识——使用者的理论. 上海: 华东师范大学出版社, 1990

作 者 简 介

李　杰　男, 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　男, 1947年生。现为上海交通大学智能工程研究所所长, 教授, 博士生导师。研究方向为非线性控制, 神经网络, 远程教育。