文章编号：1000－8152(2002)06－0963－04

网络泛化能力与随机扩展训练集

杨慧中[1]，卢鹏飞[1]，张素贞[2]，陶振麟[2]
（1. 江南大学 通信与控制工程学院，江苏无锡 214036；2. 华东理工大学 自动化研究所，上海 200237）

摘要：针对神经网络的过拟合和泛化能力差的问题，研究了样本数据的输入输出混合概率密度函数的局部最大熵密度估计，提出了运用 Chebyshev 不等式的样本参数按类分批自校正方法，以此估计拉伸样本集，得到新的随机扩充训练集．使估计质量更高，效果更好．仿真结果证明用这种方法训练的前馈神经网络具有较好的泛化性能．

关键词：前馈神经网络；泛化能力；最大局部熵密度函数；Chebyshev 不等式
中图分类号：TP183；O211.66　　**文献标识码**：A

Generalization of networks and random expanded training sets

YANG Hui-zhong[1], LU Peng-fei[1], ZHANG Su-zhen[2], TAO Zhen-lin[2]
(1. College of Communication & Control Engineering, Southern Yangtze University, Jiangsu Wuxi 214036, China;
2. Research Institute of Automation, East China University of Science, Shanghai 200237, China)

Abstract: Aiming at the problems of over-fitting and generalization for neural networks, the locally most entropic colored Gaussian joint input-output probability density function (PDF) estimate is studied, and a new method by means of Chebyshev inequality is proposed to self-revise respectively according to every cluster. In terms of the method, a random expanded training set is obtained. The simulation results illustrate that generalization of feed forward neural networks using the expanded training sets is greatly improved.

Key words: feed forward neural networks; generalization; locally most entropic probability density function; Chebyshev inequality

1 引言(Introduction)

神经元网络的泛化能力研究是一个在神经网络的理论研究和工程应用上具有重大意义的问题．传统的避免泛化误差的方法是交叉验证(cross-validation)．将训练集分割成训练集和验证集或测试集．实际操作时将训练和测试交替进行，监督泛化误差的变化曲线，防止过拟合产生．一个可实际使用的网络通常需要成千上百次的反复训练．

许多学者认为影响网络泛化能力的主要原因是训练样本和网络自身的缺陷[1－4]．其中 Holmstrom 等人[5]提出了应用统计估计方法，改造训练样本的概念．George[6]也提出了在恰当地选择随机扩展过程的分布特性的基础上，创造一个人为的无穷序列的输入-输出训练向量的设想．对于在实际工程中获取训练样本数据较困难、样本质量也难以保证的情况下，这是一个十分令人振奋的想法．然而，由于人为序列的伪随机特性，经过有限次迭代以后产生的数据集，其统计特性往往已与原来的 George 聚类[6]有了很大差异，局部范围内甚至会变得面目全非，失去了该聚类方法的意义．为此，本文重新研究了这种基于实际输入-输出联合概率分布函数(PDF)的局部最大熵(LME)概率密度估计，由该概率密度估计拉伸新的随机训练向量，并用 Chebyshev 不等式进行方差校正，使样本参数按类分批自校正．用这种方法产生的新训练集在学习过程中，可避免过拟合和改善网络的泛化能力．仿真结果证明了这种方法对提高网络泛化能力的有效性．

2 基于局部最大熵密度的训练集扩充(Expansion of the training set based on local most entropic density)

在实际应用中我们只能得到有限数据集 $\{Z_n\}_{n=1}^N$，并且通常采用数据循环来解决数据的有限可用性．这种形式的数据循环可以使得网络训练误差最小化，但不能解决对于训练集以外的数据特性的拟合，即泛化性问题．

收稿日期：2001－11－28；收修改稿日期：2002－07－01．

George[6]提出的最大化局部熵密度原理是试图导出有限数据集 $\{Z_n\}_{n=1}^N$ 的密度函数 $f_z'(z)$ 的替代式 $\hat{f}_{\mathrm{LME}}$，由 $\hat{f}_{\mathrm{LME}}$ 产生的新的训练数据可以防止过拟合，改善网络的泛化性. 其基本思路是：假定数据向量 Z 属于聚类数为 K 的 $H_1, H_2, \cdots, H_K$，相应先验概率为 $\pi_1, \pi_2, \cdots, \pi_K$ 中的一个，再假定每个假设 $H_k(k = 1,2,\cdots,K)$ 代表在整个输入 - 输出空间 R^{m+l} 的一个子集上以概率 $\pi_k(k = 1,2,\cdots,K)$ 出现的局部输入 - 输出数据特征. 则在该空间由均值向量 $U \in R^{m+l}$，协方差矩阵 $R_{(m+l)\times(m+l)}$ 确定的具有 $(m + l)$ 阶矩的 $(m + l)$ 维高斯密度 $f_{\mathrm{ME}}(z) = N(U,R)$，其微分熵可表示为 $h(f) \triangleq -\int_{R^{m+l}} f(z)\log f(z)\mathrm{d}z$.

在每个假设 $H_k(k = 1,2,\cdots,K)$ 下实施最大微分熵约束，则 Z 的无条件概率分布函数是该高斯混合分布函数

$$f_{\mathrm{LME}} = \sum_{k=1}^{K} \pi_k N(U_k, R_k), \tag{1}$$

称其为 LME 密度函数.

应用某些最小失真聚类方法分割训练集为 K 个子集 $C_k(k = 1,2,\cdots,K)$. 令 N_k 是在子集 C_k 中数据向量 $Z(X,Y)$ 的个数，式(1)中参数 U_k, R_k 和 $\pi_k(k = 1,2,\cdots,K)$ 的无偏估计就能容易地以样本平均的形式得到.

$$\hat{U}_k = \frac{1}{N_k}\sum_{z_n \in C_k} z_n, \tag{2}$$

$$\hat{R}_k(\sigma_{\mathrm{M}}, \sigma_{\mathrm{DL},k}) = \begin{cases} \sigma_{\mathrm{M}}^2[\dfrac{1}{N_k - 1}\displaystyle\sum_{z_n \in C_k}(z_n - \hat{U}_k)(z_n - \\ \hat{U}_k)^{\mathrm{T}} + \sigma_{\mathrm{DL},K}^2 I], \ N_k > 1, \\ \sigma_{\mathrm{M}}^2, \sigma_{\mathrm{DL},K}^2 I, \ N_k = 1, \end{cases} \tag{3}$$

$$\hat{\pi}_k = \frac{N_k}{N}. \tag{4}$$

为使 $\hat{R}_K$ 满足可逆的条件，已对式(3)进行了广义化处理，并且在 $\sigma_{\mathrm{DL},k}^2 = 0$ 时，类 C_k 中至少需要 $(m + l)$ 个数据样本，以此保证协方差估计矩阵 $\hat{R}_k$ 以概率 1 是可逆的.

然后根据 PDF，从类 $k \in \{1,2,\cdots,K\}$ 中按式(4)给出的每个类的相关频率 $\hat{\pi}_k(k = 1,2,\cdots,K)$ 获取数据 $Z^{(i)}, i = 1,2,\cdots$，这里 $\hat{U}_k$ 和 $\hat{R}_k$ 分别由式(2)和(3)给出. 由此可产生新训练样本集，即

$$Z^{(i)} = \hat{U}_k + \hat{L}_k s^{(i)}, \tag{5}$$

$s^{(i)}$ 是一个由 $N(0,I)$ 引出的独立同分布向量序列，$\hat{L}_k$ 是由分解类 C_k 的协方差估计矩阵 $\hat{R}_k$ 而得到的下三角阵，即

$$\hat{R}_k = \hat{L}_k\hat{L}_k^{\mathrm{T}}. \tag{6}$$

式(3)中的形状参数 σ_M^2 决定聚类半径. George[6]采用了一种最大似然交叉验证的方法来确定 σ_M^2，并将 σ_M^2 代入式(3)再重新估计协方差矩阵 $\hat{R}$.

聚类数 $K \in \{1,2,\cdots,N\}$ 的确定原则是使 LME 密度函数具有最小微分熵 $h(f_{\mathrm{LME}})$，即

$$\hat{h}(f_{\mathrm{LME}}) = -(1/J)\sum_{j=1}^{J}\log f_{\mathrm{LME}}(z_j) \tag{7}$$

取最小. 这里运用 Monte-Carlo 的样本平均计算法，通过 f_{LME} 取出了 J 个数据点，然后搜索 K，使 $\hat{h}(f_{\mathrm{LME}})$ 最小.

3 样本参数的按类分批自校正(Self-revision of sample parameter in terms of every cluster)

George[6]提出交叉验证确定 σ_M^2，以此确定聚类半径，其运用最小熵原理确定样本数据聚类数的本质是将原样本数据精细分类，然后以类为单位迭代产生符合该类统计特性的数据，扩充训练样本集. 由此产生足够多的统计特性符合要求的各类数据作网络训练样本，对改善网络的过拟合与提高泛化能力无疑是十分重要的.

然而在迭代、变换过程中，我们能采用的随机数只是伪随机的，即使是计算机函数库中直接调用的随机数，经检验其统计特性也并不理想，由此而得到的高斯白噪声序列也并非真实，均值与方差的显著水平都很低，经有限次迭代变换，数据的实际统计特性与理论上的特性往往相距甚远(高斯分布时的显著水平远低于 0.01). 我们对文献[6]中的例 1 产生的若干类数据重新作了统计分析，发现扩展后普遍有了变化，有些类的样本特性甚至已面目全非. 按扩展数据聚类，最佳聚类数已不是 $K = 11$. 这就与文中提出的随机扩展样本数据的初衷大相径庭了. 样本数据的变化，对网络的训练显然会带来不利影响.

本文利用 Chebyshev 不等式进行方差校验，以此进行按类分批自校正，解决样本参数的变化问题，取得了较好的效果.

设随机变量 ζ 的 PDF 为 $p(x)$，$\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$，其有限均值 μ 和方差 σ^2 为

$$\begin{aligned} \mu &= \int_{-\infty}^{+\infty} xp(x)\mathrm{d}x, \\ \sigma^2 &= \int_{-\infty}^{+\infty}(x - \mu)^2 p(x)\mathrm{d}x. \end{aligned} \tag{8}$$

令 ε 为一个特定常数,则 $|\zeta-\mu|\geqslant\varepsilon$ 的概率为

$$P(|\zeta-\mu|\geqslant\varepsilon)=\int_R p(x)\mathrm{d}x. \tag{9}$$

式中,R 表示满足不等式 $|\zeta-\mu|\geqslant\varepsilon$ 的所有 ζ 值组成的集合.显然,

$$\varepsilon^2\int_R p(x)\mathrm{d}x\leqslant\int_R(\zeta-\mu)^2p(x)\mathrm{d}x\leqslant\sigma^2. \tag{10}$$

由式(9)和(10)得

$$P\{|\zeta-\mu|\geqslant\varepsilon\}\leqslant\frac{\sigma^2}{\varepsilon^2}. \tag{11}$$

式(11)就是著名的 Chebyshev 不等式.我们称用式(11)进行样本参数按类分批自校正的方法为 Chebyshev 自校正方法.其原理为:

用文献[6]的方法确定聚类样本数据有 K 类,第 k 类的均值与方差分别为 $U_k,\sigma^2_{Mk},k=1,2,\cdots,K$.按类各自扩展第一批数据 $\{Z^{k_1}_n\},k=1,2,\cdots,K$,其母体记为 Z^{k_1},均值为 U_{k_1}.由

$$E\{Z^{k_1}-(U_{k_1}-U_k)\}=U_k,$$

易知 $\{Z^{k_1}_n\}$ 的均值校正方法.

Chebyshev 不等式利用随机变量 ζ 的数学期望与方差对 ζ 的概率分布进行估计,它给出了落在区间 $|\zeta<\mu|<\varepsilon$ 外概率的上限 $\frac{\sigma^2}{\varepsilon^2}$,一旦越过了这个上限,我们有理由相信数学期望或方差或是两者都发生了变化.上面已对数学期望进行了校正,现研究对方差的校正方法.设

$$P\{|Z^{k_1}-U_{k_1}|\geqslant\varepsilon_{k_1}\}=p_{k_1}>\frac{\sigma^2_{Mk}}{\varepsilon^2_{k_1}},$$

由 $U_{k_1}=U_k$ 可知,第 k 类第一批扩展数据的方差已不再是 σ^2_{Mk}.令 $p_{k_1}=\alpha\frac{\sigma^2_{Mk}}{\varepsilon^2_{k_1}}$.鉴于 $\frac{\sigma^2_{Mk}}{\varepsilon^2_{k_1}}$ 是概率的上限,我们可以有充足的理由认为 Z^{k_1} 的方差已由 σ^2_{Mk} 至少增大为 $\alpha\sigma^2_{Mk}$.取 Z^{k_1} 的方差为下限值 $\alpha\sigma^2_{Mk}$,则 $\left\{\frac{Z_{k_1}}{\sqrt{\alpha}}\right\}$ 可以满足 Chebyshev 不等式:

$$P\left\{\left|\frac{Z_{k_1}}{\sqrt{\alpha}}-\frac{U_{k_1}}{\sqrt{\alpha}}\right|\geqslant\varepsilon_{k_1}\right\}=$$

$$P\{|Z^{k_1}-U_{k_1}|\geqslant\sqrt{\alpha}\varepsilon_{k_1}\}\leqslant$$

$$\frac{\alpha\sigma^2_{Mk}}{\alpha\varepsilon^2}=\frac{\sigma^2_{Mk}}{\varepsilon^2}.$$

由此可用 $\frac{Z_{k_1}}{\sqrt{\alpha}}$ 对 $\{Z^{k_1}_n\}$ 数据集进行方差自校正.各批数据的拉伸均按此法处理.

顺便指出,我们采用 Chebyshev 不等式对数据参数进行自校正,是不想轻易否定 George 的拉伸方法,故而依据方差的下限值 $\alpha\sigma^2_{Mk}$ 进行校正.而且,这样处理的计算量也较小.事实上,直接采用 Z^{k_1} 的方差估计值 $\hat{\sigma}^2_{Mk}$ 进行校正,也存在着显著水平或置信区间的选取问题,实际效果并没有 Chebyshev 自校正方法好.

4　仿真结果(Simulation results)

例 1　以聚丙烯腈聚合反应中聚合物总数的神经网络模型训练为例,由现场采集的数据和人工分析值组成 143 组样本数据,用间隔法将样本集分为训练子集和测试子集,两个子集分别有 72 组数据和 71 组数据.由这些实际样本数据集,采用 George[6] 提出的最大似然交叉验证方法求得 σ^2_M 为 5.705,聚类数 $K=6$.再分别用 George[6] 方法和 Chebyshev 自校正的方法(取,$\varepsilon_{k_1}=0.01,\frac{\sigma^2_{Mk}}{\varepsilon_{k_1}}=0.01,k=1,2,\cdots,6$),各自产生 200 组随机扩充数据集,选择一个 4—15—1 的神经网络结构,在同等初始值(w_0)条件下,分别将实际训练样本集、由 George 方法及 Chebyshev 自校正方法产生的扩充数据集用标准 BP 算法进行学习.训练中,除开始时的 100 次迭代运算用实际样本训练集外,以后全部用随机扩充数据集.用相同的测试样本集分别测试各方法的泛化误差.训练和泛化的误差如图 1 所示.从图 1 可看出,用实际训练集学习的泛化误差最大,最后达到 4.2012 左右,George 算法的泛化误差其次,在 3.3969 左右,Chebyshev 自校正算法则泛化误差最小,仅为 1.3984 左右.上述结果是在综合了多次运行结果后得出的.

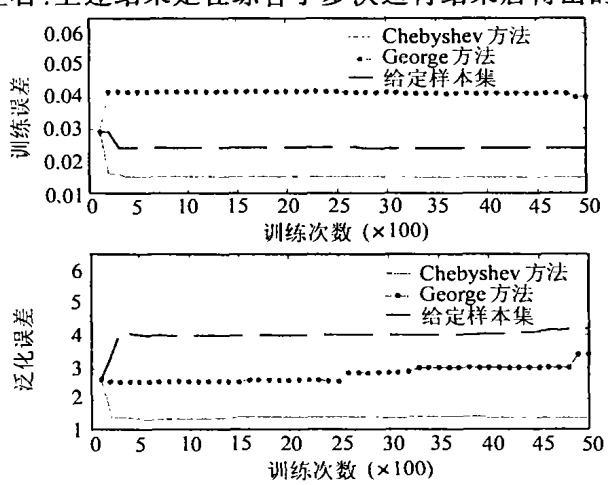

图 1　例 1 的训练均方误差与泛化均方误差

Fig. 1　The squared errors of training and generalization with Example 1

例 2 设一动态系统的模型为:

$$y(k) = 10.8 - 0.5\exp(-y(k-1)^2)\cdot y(k-1) - [0.3 + 0.9\exp(-y(k-1)^2)]\cdot y(k-2) + u(k)+0.2u(k-2)+0.1u(k-1)\cdot u(k-2)+e(k).$$

其中 $e(k)$ 是均值为 0 及方差为 0.04 的高斯白噪声序列,由仿真产生 100 组输入输出数据作为训练样本,同例 1 一样,用相隔法将样本集分为训练子集和测试子集,两个子集各有 50 组数据.由这些样本数据集,采用最大似然交叉验证方法求得 σ_M^2 为 4.02,聚类数 $K = 64$. 分别用 George 方法和 Chebyshev 自校正的方法(取 $\varepsilon_{k_1} = 0.01, \frac{\sigma_{Mk}^2}{\varepsilon_{k_1}} = 0.01, k = 1,2,\cdots,64$),各自产生 200 组随机扩充数据集,选择一个 5—6—1 的神经网络结构,在同等初始值 (w_0) 条件下,分别将样本训练集,以及由上述方法产生的扩充数据集,用标准 BP 算法进行学习计算,同例 1 一样,在用随机扩充数据集学习时,除开始时的 100 次迭代运算用实际样本训练集外,以后全部用扩充数据集. 用相同的测试样本集分别测试各方法的泛化误差.其训练误差和泛化的误差如图 2 所示.

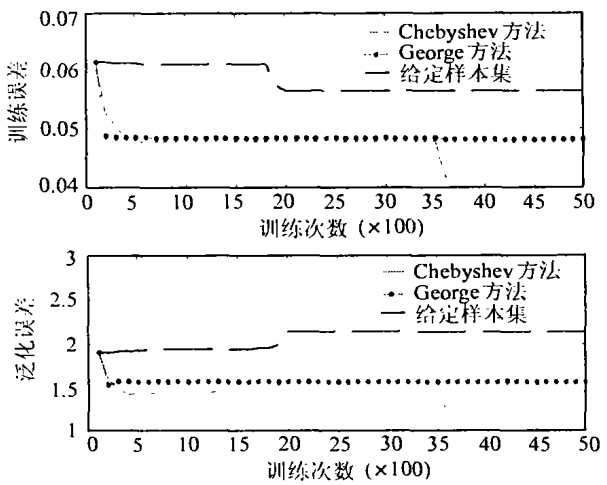

图 2 例 2 的训练均方误差与泛化均方误差

Fig. 2 The squared errors of training and generalization with Example 2

从图 2 可看出,用给定训练集训练的泛化误差最大,随着训练次数的增加,泛化误差增加, 最后稳定在 2.1351 左右; George 算法的泛化误差为 1.5654,用 Chebyshev 自校正算法的泛化误差最小,在 1.1712 左右.上述结果也是在综合了多次运行结果后得出的.

5 结论(Conclusion)

为改善前馈神经网络的泛化能力,对于给定的有限训练样本集,研究了建立局部最大熵密度有色高斯概率分布函数估计每个聚类的要求,提出了用 Chebyshev 不等式按类分批进行样本参数自校正方法改善文献[6]的样本数据统计特性.用这种方法产生的新训练集,可避免过拟合和改善前馈网络的泛化能力.仿真结果证明了这种方法的有效性.

参考文献(References)

[1] Jiang Xue-jun, Tang Huan-wen. System analysis for generalization of MFNN [J]. System Engineering Theory & Practice, 2000, (8): 36 - 40(in Chinese)

[2] Meng Jian, Qu Liang-sheng. The feed forward neural networks and application based on information optimization [J]. China Mechanical Engineering, 1997, 8(2): 54 - 57(in Chinese)

[3] Men Jian. Neural networks using entropy optimization and antenna modeling [J]. Electron Confrontation Technology, 1998, 13(6): 23 - 28(in Chinese)

[4] Lu Zi-yi, Yang Lv-xi, Wu Qiu, et al. A new approach for improving generalization ability of multilayer feedforward networks [J]. Journal of Circuits and Systems, 1997, 2(4): 7 - 12(in Chinese)

[5] Holmstrom L, Koistinen P. Using additive noise in backpropagation training [J]. IEEE Trans. Neural Networks, 1992, 3(1): 24 - 38

[6] George N K. On overfitting, generalization, and randomly expanded training sets [J]. IEEE Trans. Neural Networks, 2000, 11(5): 1050 - 1057

本文作者简介

杨慧中 1955 年生.博士.副教授.主要研究方向为过程模型化与控制. Email: yanghuizhong@163.com

卢鹏飞 1945 年生.硕士.副教授.主要研究方向为信号分析、处理、传输与控制.

张素贞 1938 年生.教授.博士生导师.主要研究方向为过程模型优化与控制.

陶振麟 1945 年生.硕士.教授.主要研究方向为计算机控制.