文章编号:1671-6833(2003)01-0063-03

神经网络学习样本点的选取方法比较

王少波,柴艳丽,梁醒培

(郑州机械研究所,河南 郑州 450052)

摘 要:为了比较训练人工神经网络的所需样本点的选取,分别采用随机遍历法、正交设计法和均匀设计方法产生样本点,用于训练神经网络.分析结果表明,在样本点个数相同情况下,均匀设计法的代表性最好,正交设计法次之,而随机遍历法较差.随机遍历法随着样本点个数的增多,同样可以提高其代表性.当函数随变量在区间内变化较小(因素水平可以取的较少)时,正交设计法也不失为一个好的选择.均匀设计法在多变量,且每个变量需要选取较多水平数的情况下,更能体现它的优越性.

关键词:神经网络;正交设计;均匀设计

中图分类号:O 157.5 **文献标识码**:A

0 引言

目前人工神经网络[1]已经广泛地被应用于工程实际,其特色在于信息的分布式存储和并行协同处理.虽然单个神经元的结构极其简单,功能有限,但大量神经元构成的网络系统所能够实现的行为却是极其丰富多彩的.和数字计算机相比,神经网络系统具有集体运算的能力和自适应的学习能力.另外,它还有很强的容错性和鲁棒性,善于联想、综合和推广.神经网络用于求解多目标优化问题具有巨大的优势.

人工神经网络模型各式各样,目前已有数十种,它们是从各个角度对生物神经系统不同层次的描述和模拟.代表性的网络模型有感知器、多层映射BP 神经网络、GMDH 网络、RBF 网络、双向联想记忆(BAM)、盒中脑(BSB)、Hopfield 模型、Boltzmann 机等.运用这些网络模型可实现函数近似、数据聚集、模式分类、优化计算、概率函数估计等功能,因此人工神经网络广泛用于人工智能、自动控制、机器人、统计学等领域的信息处理中.

神经网络近似函数,处理信息的能力完全取决于网络中各种神经元之间的耦合权值.对于较大规模的网络,权值不可能一一设定,因此网络本身必须具有学习的功能,即能够从示范模式的学习中逐渐调整权值,使网络整体具有近似函数或处理信息的功能.

在神经网络学习前如何有效地选取特征样本点呢?本文以多层映射BP 神经网络为例,分别研究了三种不同的选取方法(随机遍历法、正交设计法[2]和均匀设计方法[3])对网络的训练效率和精度的影响.

1 正交设计法和均匀设计法

正交设计是最常用而且有长久历史的实验设计,是一种多因素实验设计和实验结果分析相结合的科学实验方法.用它进行实验时,是用已印好的表格——正交表科学地安排实验,使得实验尽可能地减少.

正交表是安排实验、分析实验的一种简单而容易掌握的有力工具.它是根据数理统计原理归纳出来的一种合理安排实验的表格,用 $L_n(m^k)$ 表示.表中 L 是代表"正交"之义;n 表示横行数即实验次数;m 表示每纵列中的不同字码的个数,即每个因素的水平数;k 表示纵列数,即该均匀设计表最多安排的因素数.

均匀设计是在正交设计的基础上进一步发展而成的,是只考虑实验点在实验范围内均匀散布的一种实验设计方法.均匀设计根据数论在多维数值积分中的应用原理,构造一套均匀设计表,用来进行均匀设计.

收稿日期:2002-10-18;**修订日期**:2002-12-25

基金项目:国家先进制造技术中心资助项目(AMTRC 2002-3)

作者简介:王少波(1976-),男,河南省洛宁县人,郑州机械研究所博士研究生,主要从事流固耦合结构振动分析及综合优化设计方面的研究.

均匀设计表是一种规格化的表格,是均匀设计的基本工具,用 $U_n(m^k)$ 表示,表中 U 是均匀设计表代号,其它符号和正交表一致.

2 三种选取方法比较分析

2.1 样本点的代表性

命 $F(\boldsymbol{x})$ 为 R^s 中的一个连续多元分布,n 为一个整数.我们需要在 R^s 中找出 n 个点 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 使它们对 $F(\boldsymbol{x})$ 有一个好的代表性.那么,代表性的含义又是什么呢?让我们来考虑代表性的一个度量.

定义1 命 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 为 R^s 中的 n 个点,则函数

$$F_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} I\{\boldsymbol{x}_i \leqslant \boldsymbol{x}\}$$

称为 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 的经验分布,其中所有不等式表示不等式两边的矢量的各分量均分别满足该不等式,$I\{A\}$ 为 A 的指标函数,即

$$I\{A\}=\begin{cases}1,(\text{若 } A \text{ 成立})\\ 0,(\text{若 } A \text{ 不成立})\end{cases}$$

定义2 命 $F(\boldsymbol{x})$ 为 R^s 上的一个cdf 及 $P=\{k_k, k=1,\cdots,n\}$ 为 R^s 上的一个点集,则

$$D_F(n,P)=\sup_{x\in R^s}|F_n(\boldsymbol{x})-F(\boldsymbol{x})|$$

称为 P 关于 $F(\boldsymbol{x})$ 的一偏差,此处 $F_n(\boldsymbol{x})$ 为 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 的经验分布.显然 F 一偏差是 P 关于 $F(\boldsymbol{x})$ 的代表性的度量.F 一偏差越小,其代表性越好.

当 $F(\boldsymbol{x})$ 为 $C^s=[0,1]^s$ 上的均匀分布时,即 $U(C^s)$ 则偏差定义如下:命 $P=\{\boldsymbol{x}_k, k=1,\cdots,n\}$ 为 C^s 上的一个点集,对于 $\gamma\in C^s$,命 $N(\gamma,P)$ 表示 P 中满足 $\boldsymbol{x}_k\leqslant\gamma$ 的点数,则

$$D(n,P)=\sup_{x\in C^s}\left|\frac{N(\gamma,P)}{n}-v(|0,\gamma|)\right|$$

称为 P 的偏差,其中,$v(|0,\gamma|)=\gamma_1,\cdots,\gamma_n$ 表示矩形 $[0,\gamma]$ 的体积[4].

训练神经网络的样本点,在整个可行区域内符合均匀分布,显然可以用 $D(n,P)$ 的大小评判各种样本点的代表性.在2.2的算例中,把区间 $[-1,1]$,映射到 $[0,1]$.取 $\gamma=0.5$,计算三种方法所产生样本点的偏差.

随机遍历法:0.155;正交设计法:0.115;均匀设计方法:0.035.

显然,均匀设计法产生的样本点的代表性最好,正交设计法次之,随机遍历法则最差.

2.2 算例结果及分析

本文构造了一个三层的BP 神经网络来近似三个变量的复杂函数

$$f(x_1,x_2,x_3)=\frac{(x_1+x_1x_2+x_2^2+x_2x_3+x_3^2)}{3},$$

式中:$x_1,x_2,x_3\in[-1,1]$.

根据Kolmogorov 定理设计的神经网络结构如图1所示,第一层(即输入层)有3个处理单元,中间层有7个处理单元,第三层(即输出层)有1个处理单元.中间层神经元的传递函数为S 状曲线,输出层神经元的传递函数为线性函数.

例子是在MATLAB 6.1[5]环境下实现的,为了比较三种不同方法产生的样本点对神经网络训练结果的影响,保证神经网络训练过程参数设定一致,并要求产生相同个数的样本点25个.

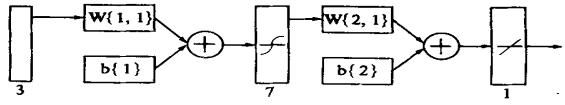

图1 三层BP 神经网络

Fig.1 Three layer BP neural net

(1) 随机遍历法生成样本点:

```
ar =rand(3,25)    %产生(3×25)的"0—1均匀分布"随机矩阵ar;
x =ar *2—1        %转变到[—1 1]区间,作为样本点
for i =1:25
f(i) =(x(1,i) +x(1,i) *x(2,i) +x(2,i) *x(2,i) +x(2,i) *x(3,i) +x(3,i) *x(3,i))/3   %计算函数值end
```

(2) 正交设计法生成样本点:选用 $L_{25}(5^6)$ 正交表,取前3列分别对应 x_1,x_2,x_3,"5"表示 x_1,x_2,x_3 分别取5个水平的值(本例分别取 $-1,-0.5,0,0.5,1$),这样根据正交表可以产生25个样本点:首先将所用正交表存入矩阵ar,然后作变化:$x=(ar-1)/2-1$,其它步骤类似随机遍历法.

(3) 均匀设计法生成样本点:选用 $U_{25}(25^9)$ 均匀表,取1,5,8列分别对应 x_1,x_2,x_3,"25"表示 x_1,x_2,x_3 分别取25个水平的值,这样根据均匀表可以产生25个样本点:首先将所用均匀表存入矩阵ar,然后作变化:$x=(ar-1)/12-1$,其它步骤类似随机遍历法.

以上述三种样本点分别训练三个结构相同BP 网络1,网络2,网络3.在区间 $[-1,1]$ 随机取10个点,来测试三个网络的映射效果.其结果如表1所示.

表 1　网络映射效果测试数据

Tab.1　Data for testing mapping effect of nets

序号	(x_1,x_2,x_3)	$f(x_1,x_2,x_3)$	网络 1	网络 2	网络 3
1	(0.9003,−0.5377,0.2137)	0.2120	0.11882	0.29145	0.26467
2	(−0.0280,0.7826,0.5242)	0.4158	0.37355	0.46238	0.50114
3	(−0.0871,−0.9630,0.6428)	0.2394	−0.10061	−0.25977	0.23887
4	(−0.1106,0.2309,0.5839)	0.1310	−0.19046	0.17976	0.15455
5	(0.8436,0.4764,−0.6475)	0.5278	0.72012	0.33497	0.46930
6	(−0.1886,0.8709,0.8338)	0.6090	0.92633	0.67022	0.64853
7	(−0.1795,0.7873,−0.8842)	0.1283	−0.051377	0.12661	0.15243
8	(−0.2943,0.6263,−0.9803)	0.0869	−0.19303	0.066389	0.11591
9	(−0.7222,−0.5945,−0.6026)	0.2606	0.52852	0.35298	0.25044
10	(0.2076,−0.4556,−0.6024)	0.3193	0.3543	0.63919	0.32101

图 2 表示出三种网络的映射值与目标函数值的关系.网络 1 的映射平均误差达到 110.6%,网络 2 的映射平均误差为 50.1%,网络 3 的映射平均误差仅为 13.7%.显然在三种形成样本点的方法中,均匀设计法效果最好,正交设计法次之,而随机遍历法最差.随机遍历法在自变量区间内随机选取样本点,显然很盲目,在样本点比较少的情况下(比如本文例子 25 个)有很大的局限性,并不能很有效地代表整个函数区间.

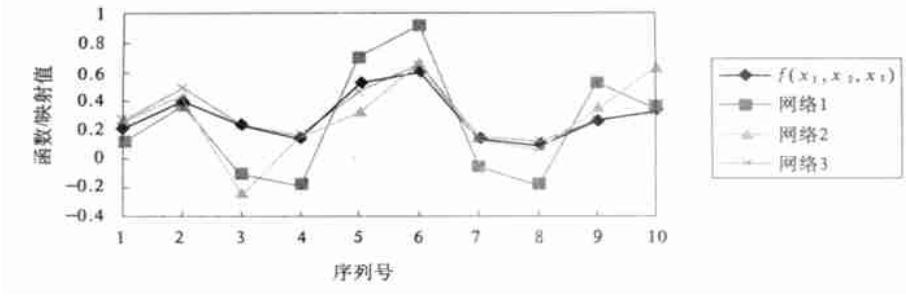

图 2　网络映射对比图

Fig.2　Contrast for net mapping

正交设计法考虑各因素的综合可比性,必须是全面实验,而每个因素的水平必定有重复,这样一来实验点在实验范围内就不可能充分地均匀分散,实验点的数目就不能过少,本文例子三个变量局限于 25 个样本点,就不得不牺牲每个变量的水平数,只能取 5 个水平,这样也影响样本点的代表性,不过其效果比随机遍历法明显要好.

显然,用正交表安排实验,均匀性受到一定的限制,因而造成实验点的代表性不够强.均匀设计法保证样本点在区间范围内充分地均匀分散,其代表性较正交实验的样本点要强得多.

3　结论

本文用三种不同方法产生样本点,分别用于训练神经网络.分析结果表明,在样本点个数相同情况下,均匀设计法的代表性最好,正交设计法次之,而随机遍历法较差.由于本例样本点局限于 25 个,相对较少,BP 网络算法简单,致使三个网络的相对误差都比较大(随机遍历法达到 110.6%),这并不能说明随机遍历法和正交设计法不能应用于工程实际.随机遍历法随着样本点个数的增多,同样可以提高其代表性.当函数随变量在区间内变化较小(因素水平可以取的较少)时,正交设计法也不失为一个好的选择.均匀设计法在多变量,且每个变量需要选取较多水平数的情况下,更能体现它的优越性.

参考文献:

[1] 赵振宇,徐用懋.模糊理论和神经网络的基础与应用[M].北京:清华大学出版社,1997.

[2] 北京大学数学力学系概率统计组.正交设计法[M].北京:石油化工出版社,1976.

[3] 方开泰.均匀设计与均匀设计表[M].北京:科学出版社,1994.

[4] 方开泰,王　元.数论方法在统计中的应用[M].北京:科学出版社,1994.

[5] 丛　爽.面向 MATLAB 工具箱的神经网络理论与应用[M].合肥:中国科学技术大学出版社,1998.

(下转第 69 页)

call ExitThread(0)

(3) 在停止计算的菜单事件中,设置上述全局控制变量的值为"停止".另外,在上述操作线程的过程中需引用 DFMT 模块,该模块封装了 Windows 关于线程操作的函数.系统运行情况如图 1 所示.

参考文献:

[1] 周振红,杨国录,周洞汝.基于组件的水力数值模拟可视化系统[J].水科学进展, 2002,1(1):9～13.

[2] 彭国伦.Fortran 95 程序设计[M].北京:中国电力出版社,2002.

Carrying out Steering Computation in Visual Fortran

ZHOU Zhen-hong[1], ZHANG Jun-jing[1], CHEN Shi-feng[2], YANG Guo-lu[3]

(1. College of Environmental & Hydraulic Engineering, Zhengzhou University, Zhengzhou 450002, China; 2. Zhoukou Highway Bureau, Zhoukou 466000, China; 3. College of Water Resources and Hydropower, Wuhan University, Wuhan 430072, China)

Abstract: The process of numerical computation can be presented with steering computation in large-scale and nonlinear modeling, in order to develop an accurate model. In practical context of numerical simulation on the buoyant jet, a QuickWin application is developed in Visual Fortran, interface elements and graphics being designed with graphic library of QuickWin. To carry out steering computation in Visual Fortran, the key is to create multithread application, drawing and computing in a single thread, and the system responding to the event of the window in main thread. Finally, steering computation is fully implemented.

Key words: buoyant jet; numerical simulation; steering computation; Visual Fortran; QuickWin; multithread

(上接第 65 页)

Comparison of Methods to Produce Sample Points in Training ANN

WANG Shao-bo, CHAI Yan-li, LIANG Xing-pei

(Zhengzhou Research Institute of Mechanical Engineering, Zhengzhou 450052, China)

Abstract: Based on the theory of the NN in the biology, the ANN is a complicated, massive, nonlinear dynamic system that simulates the biologic neural structure. Lots of sample points are required to train the ANN. In this paper, three methods are presented to produce the sample points adopted in training the ANN. It is testified that under the same condition even design is the best method to produce sample points, followed by orthogonal design, and the third is ransacking. Ransacking will be better if the number of sample points is large. When the function changes little with the variables, fewer factor levels can be chosen and orthogonal design is a good method. Its advantages stand out for the occasion of multiple variables and many factor levels.

Key words: neural network; orthogonal design; even design