一种神经网络快速修剪算法

乔俊飞,李 森,刘 江
(北京工业大学电子信息与控制工程学院,北京 100124)

摘 要: 通过分析非线性函数各输入参数对函数值的影响,发现当输入参数间无相互作用时,表征其灵敏度的傅立叶振幅主要集中在基频上.基于该发现,提出一种基于扩展傅立叶振幅灵敏度方法的神经网络隐含层神经元快速修剪算法.其实质是通过计算神经网络隐含层各神经元输出的指定基频上的傅立叶振幅,求取隐含层神经元对神经网络输出的灵敏度.根据各神经元的灵敏度,按照一定的准则削减冗余神经元,获得紧凑的神经网络结构.将提出的神经网络结构修剪算法用于污水水质参数化学需氧量(COD)的软测量过程中,实验结果与扩展傅里叶振幅灵敏度算法相比,在修剪效果相同的情况下,其运行时间得到明显减小.

关键词: 灵敏度; 傅立叶振幅; 神经网络修剪算法
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2010) 04-0830-05

A Fast Pruning Algorithm for Neural Network

QIAO Jun-fei, LI Miao, LIU Jiang
(*College of Electronic and Control Engineering, Beijing University of Technology, Beijing* 100124, *China*)

Abstract: By analyzing the impact of the input factors on the output value in nonlinear function, it is suggested that when the input factors are independent, the Fourier Amplitudes which showing sensitivity values are relied mainly on the fundamental frequency. As a result, a fast pruning algorithm for the hidden neurons in the neural network is proposed based on the Fourier amplitude sensitivity test method. In essence, the Fourier amplitudes on the assigned frequencies of the hidden layer outputs are computed. Then the sensitivity of each hidden neuron to the neural network output is obtained. Finally, the redundant hidden neurons are pruned according to their sensitivity values to obtain a network with compact structure. The propose method is used in the soft measurement for Chemical Oxygen Demand(COD), which is a quality parameter of waste water. The experimental result shows that our proposed method is much faster than the Fourier amplitude sensitivity test method. The remaining neurons are the same after pruning for the two methods.

Key words: sensitivity; fourier amplitude; neural network pruning

1 引言

神经网络的结构设计确定问题一直是业界非常关注的一个研究话题.如果神经网络过于复杂,就会导致计算效率低,泛化能力差;如果神经网络过于简单,其信息处理能力又十分有限.因此根据任务获取合适的神经网络结构是亟待解决的问题.对于有冗余神经元的神经网络,近年来学者们提出了许多修剪算法,其中比较著名的有影响因子(Impact Factor)算法、最优脑损伤算法(OBD)[1]、最优脑外科算法(OBS)[2]、最优细胞损伤算法(OCD)[3]、方差无效修剪(Variance Nullity Pruning VNP)[4]等.这些算法的共同缺陷是修剪往往是在训练误差达到最小才开始,并且修剪后需要重新训练网络.此外,对于不同初始规模的网络,修剪后的剩余神经元个数往往不同.

扩展傅立叶振幅灵敏度算法(EFAST)算法[5]可以用来估计一个系统的输入参数对系统输出值的影响.Philippe等人[6]将其成功运用于简化神经网络结构,通过计算每个隐含层神经元输出值的灵敏度来确定每个隐含层神经元对输出的贡献大小,直接剔除冗余神经元来对神经网络结构进行修剪.这种算法可以克服上述算法的不足,但是其缺点是:如果网络的初始规模较大,其迭代次数可达到百万、千万次以上,运行时间大幅度增加.本文提出了一种快速EFAST算法,该算法在计算灵敏度时只计算基频上的傅立叶振幅,而忽略了非基频上的傅立叶振幅影响.快速EFAST算法对神经网络的修剪效果与EFAST算法完全一致时,其运行时间大大减少.

收稿日期:2008-06-17;修回日期:2009-11-04
基金项目:国家863高技术研究发展计划(No.2009AA04Z155,No.2007AA04Z160);国家自然科学基金(No.60873034,No.60674066);教育部博士点基金(No.200800050004);北京市自然科学基金(No.4092010)

2 EFAST 算法及其应用

2.1 EFAST 算法

设给定系统的数学模型为:

$$y = f(Z_1, Z_2, \cdots, Z_p) \tag{1}$$

式中:$Z_1, Z_2, \cdots, Z_P$ 是作用于输出 Y 的 p 个输入参数.在这 p 个输入参数中,第 h 个输入参数用 Z_h 表示,则 Z_h 以频率 w_h 在$[a_h, b_h]$内振荡[6],即:

$$Z_h(s) = \frac{a_h + b_h}{2} + \frac{b_h - a_h}{\pi}\arcsin(\sin(w_h s)) \tag{2}$$

给定函数 $f(s)$的傅立叶形式可以表示为:

$$f(s) = \sum_{j=-\infty}^{+\infty} (A_j\cos(js) + B_j\sin(js)) \tag{3}$$

其中傅立叶系数 A_j 和 B_j 表示为:

$$A_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)\cos(js)\mathrm{d}s \tag{4}$$

$$B_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)\sin(js)\mathrm{d}s \tag{5}$$

输出 Y 的方差可以用下式表示[4]:

$$\begin{aligned} Var(Y) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f^2(s)\mathrm{d}s - [E(Y)]^2 \\ &\approx \sum_{J=-\infty}^{+\infty} (A_J^2 + B_J^2) - (A_0^2 + B_0^2) \\ &\approx 2\sum_{J=1}^{+\infty} (A_J^2 + B_J^2) \end{aligned} \tag{6}$$

输入参数 Z_h 对方差 Y 的影响为:

$$D_h = Var_{Z_h}[E(y \mid Z_h)] = 2\sum_{k=1}^{+\infty} (A_{kw_h}^2 + B_{kw_h}^2) \tag{7}$$

其中,kw_h 表示 Z_h 的第 k 次谐波对应的频率.

Z_h 在 $k=1$ 时傅立叶振幅的值最大,随着 k 增加,傅立叶振幅也随之变小.给定一个常数 M,使得傅立叶振幅在 Mw_h 后小到可以忽略不计.M 叫做干扰因子,一般可以取 4 或 6.如果有 p 个输入参数,$w_{\max}$表示所有输入参数基频的最大值,那么计算傅立叶常数选取的最高频率至少等于 $Mw_{\max}$.因此式(6)可以写成:

$$Var(Y) = 2\sum_{w=1}^{Mw_h} (A_w^2 + B_w^2) \tag{8}$$

Z_h 的一阶灵敏度表达式可表示为:

$$S_h = \frac{Var_{Z_h}[E(y \mid Z_h)]}{Var(Y)} = \frac{\sum_{k=1}^{+\infty} (A_{kw_h}^2 + B_{kw_h}^2)}{Var(Y)} \tag{9}$$

S_h 表示 Z_h 独立于其他输入参数对系统输出的贡献.

计算某个输入参数 Z_h 的灵敏度时,当 w_h 与其余参数的基频或 N 次谐波($N \leqslant M$)对应的频率相近时,就会产生频率混叠,不能完全把 Z_h 的作用和其他输入参数区分开来.为解决这个问题,Saltelli 等人[5]提出解决的办法是,令

$$w_h = 2M\max(w_{\sim h}) \tag{10}$$

其中,$\max(w_{\sim h})$是除 Z_h 之外的其他输入参数基频的最大值.$[0, M\max(w_{\sim h})]$是除 Z_h 外的其他所有输入参数对输出的作用区域对应的频率范围,$[M\max(w_{\sim h}), Mw_h]$为 Z_h 的作用区域对应的频率范围.Z_h 在频率为 $N\max(w_h)$,$(N=1,2,\cdots,M)$附近的傅立叶振幅较大.

关于 Z_h 的总灵敏度用下式计算:

$$ST_h = \frac{\sum_{w=M\max(w_{\sim h})+1}^{Mw_h} (A_w^2 + B_w^2)}{\sum_{w=1}^{Mw_h} (A_w^2 + B_w^2)} \tag{11}$$

ST_h 包括 Z_h 独立于其他输入参数对输出的作用和 Z_h 与其他输入参数的协同作用,因此称为总灵敏度.

归一化后第 h 个输入参数的灵敏度为:

$$ST_h = \frac{ST_h}{\sum_{l=1}^{p} ST_i} \tag{12}$$

2.2 基于 EFAST 算法的非线性函数分析

用 EFAST 算法对非线性函数的输入输出之间的关系进行定量分析.

给定一个输入参数间无相互作用的非线性函数:

$$f(x) = \frac{\tan x}{x^2} + \frac{\mathrm{e}^x}{x^3} + x\log_2 x + x\sin x \tag{13}$$

令 $Z_1(x) = \tan x/x^2$, $Z_2(x) = \mathrm{e}^x/x^3$, $Z_3(x) = x\log_2 x$, $Z_4(x) = x\sin x$.当 $x \in [1,10]$时,$f(x)$、$Z_1(x)$、$Z_2(x)$、$Z_3(x)$和 $Z_4(x)$的变化如图 1 所示.

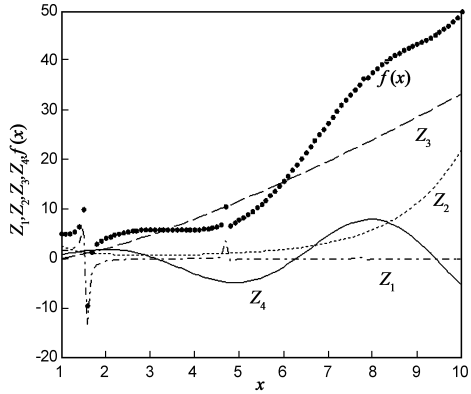

图1 $f(x)$、$Z_1(x)$、$Z_2(x)$、$Z_3(x)$和$Z_4(x)$随x的变化

不难看出:函数 $f(x)$的各个输入参数的关系比较复杂,很难直观看出输入参数对函数输出的贡献.用 EFAST 方法分析,则很容易看出各输入参数对输出的影响.由 EFAST 算法可以得到该非线性函数的输入 $Z_1(x)$、$Z_2(x)$、$Z_3(x)$和 $Z_4(x)$相应的灵敏度为:$S_1 = 17.82\%$,$S_2 = 22.44\%$,$S_3 = 50.88\%$,$S_4 = 8.86\%$.通过分析可以的得到 $Z_4(x)$对整个函数输出的影响是最小的.

3 一种基于快速 EFAST 方法的神经网络结构修剪算法

3.1 快速 EFAST 算法

将一个前神经网络的输出看作为一个非线性函数,那么其上一层中的各神经元输出值可以看作为该非线性函数的输入参数.这样,就可利用 EFAST 方法分析各神经元对神经网络输出的贡献.对贡献很小的神经元,可以当作冗余神经元,将其删除.直接将 EFAST 方法用于修剪前向神经网络的最大问题是计算量大,特别是对于规模较大的神经网络这种方法难以在线实现.经过分析,我们发现前向神经网络中各层神经元之间无相互作用.输入参数间若无相互作用,那么其傅立叶振幅主要集中在基频上,于是可以得到快速 EFAST 算法,进而降低神经网络修剪计算的复杂度.仍然以式(13)给出的非线性函数为例,取 w_h 为 32,$w_{\sim h}$ 为 1 到 4,通过 EFAST 算法得出的 Z_1, Z_2, Z_3, Z_4 在各个频率下的傅立叶振幅如图 2 所示.

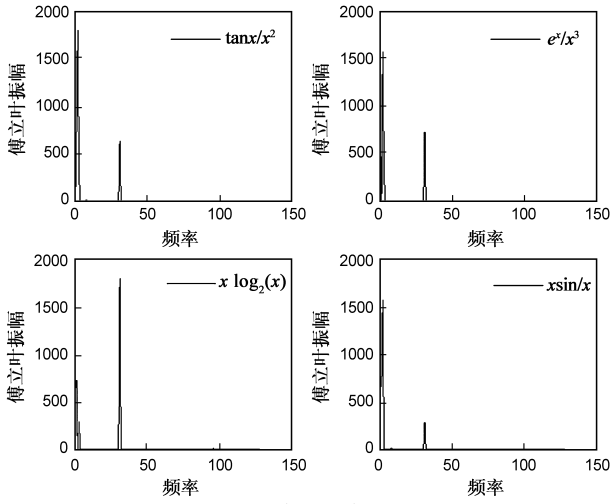

图2 不同频率下的傅立叶振幅

可以看出:kw_h($k = 1,2,3,\cdots$)上的傅立叶振幅体现了该输入参数 Z_h 对函数的独立作用,kw_h 附近频率上的傅立叶振幅体现了 Z_h 与其他输入参数之间的相互作用.对于函数(13),各个输入参数的作用是独立的,从图 2 中容易看出傅立叶振幅的值几乎全部集中在基频上.

对于输入参数之间不存在相互作用的函数,为了减少算法迭代次数,可以只取基频上的傅立叶振幅进行叠加计算灵敏度.则 Z_h 的总灵敏度可以用下式计算:

$$ST_h = \frac{A_{w_h}^2 + B_{w_h}^2}{\sum_{w=1}^{\max w_{\sim h}} (A_w^2 + B_w^2) + (A_{w_h}^2 + B_{w_h}^2)} \tag{14}$$

根据式(14)计算出式(13)给出的函数的灵敏度为:$S_1 = 17.76\%$,$S_2 = 22.40\%$,$S_3 = 51.04\%$,$S_4 = 8.80\%$.与 EFAST 算法计算出的灵敏度相比,其最大相对误差不超过 1%,可以满足神经网络结构修剪的需要.

3.2 快速 EFAST 算法修剪神经网络

一个输入层有 l 个神经元,隐含层有 p 个神经元的单输出三层前向神经网络,隐含层第 j 个神经元的输出可以表示为:

$$Z_j = \sigma\Big(\sum_{i=1}^{l} w_{i,j}^1 x_i + b_j\Big) \tag{15}$$

其中,x_i 为神经网络的第 i 个输入,$w_{i,j}^1$ 为第 i 个输入到第 j 个隐层神经元的输入层权值,b_j 为第 j 个隐层神经元的阈值,σ 为 sigmoid 函数,Z_j 为隐含层第 j 个神经元的输出.

隐含层神经元输出和神经网络输出的关系为:

$$Y = \sum_{j=1}^{p} w_j^2 Z_j \tag{16}$$

其中,w_j^2 为输出层权值,Y 为神经网络的输出.将 p 个隐神经元的输出作为快速 EFAST 算法的 p 个输入参数,将神经网络的输出作为快速 EFAST 算法的输出,通过计算输入参数 $Z_1, Z_2, \cdots, Z_P$ 对于输出 Y 的灵敏度,然后按照一定的准则修剪灵敏度小的神经元.

由该前向神经网络的结构可知,其各隐含层神经元的输出 $Z_1, Z_2, \cdots, Z_P$ 之间没有相互作用,傅立叶振幅值主要集中在基频上,因此,可以采用式(14)计算其灵敏度.

神经网络结构修剪算法具体步骤如下:

(1)训练一个有冗余神经元的神经网络到 n 步.根据实验结果,发现如果 n 选的过小(< 100 步),则采集的信息量不够充足,计算出的灵敏度不够准确,因此只要 n 不过小,其选取没有特别严格的限制.

(2)记下训练过程中每个隐神经元输出的变化范围$[a_h, b_h]$.

(3)计算神经元的总灵敏度:为每个隐神经元的输出指定各自的基频.设初始神经元个数为 p,如果要计算第 h 个隐神经元的灵敏度,那么可以把每个隐神经元的频率指定为$[1,2,3,\cdots,h-1,w_h,h+1,\cdots,p]$,每个输入参数根据式(2)定义的曲线振荡,再根据式(14)计算总灵敏度 ST_h.

(4)重复步骤(3),直到计算完 p 个隐神经元的总灵敏度.

(5)根据式(12)计算每个神经元的灵敏度 S_n($n = 1,2,3,\cdots,p$),删除 S_n 小于 5%[6]的隐含层神经元.

(6)继续训练神经网络到一定步数,重复步骤(2)~(5),直到新神经网络中所有隐含层神经元的灵敏度均小于 5% 停止修剪.

(7)训练神经网络直到偏差达到给定值.

3.3 快速 EFAST 算法修剪神经网络迭代次数分析

EFAST 算法的迭代次数约为 $p^2 M w_h$(p 表示隐神经元的个数).可见,迭代次数随着隐神经元个数增加而成倍增加.当隐神经元数目较多时,原算法迭代次数是很庞大的,运行时间也会较长.新的算法迭代次数约为 $p^2 w_{\sim h}$.当 M 取 4,$w_h = 8$ $w_{\sim h}$,原算法的迭代次数是新算法的 32 倍.可见,隐神经元个数越多,就越能显出快速 EFAST 算法的优越性.

4 仿真实验

化学需氧量(Chemical Oxygen Demand COD)是污水处理过程中的重要水质参数,它是指废水中能被氧化的物质在被化学氧化剂氧化时所需要的氧量,以毫克/升为单位.然而 COD 的在线测量困难,目前多用软测量的方法.神经网络具有良好的非线性逼近能力,且具有自适应的功能,目前的 COD 软测量也多采用神经网络模型.现在分别采用基于 EFAST 算法和快速 EFAST 算法的自组织神经网络对 COD 进行预测.图 3 给出了 COD 神经网络预测模型,其输入分别为混合液悬浮固体(MLSS)、油、pH、氰化物(TCN),输出为污水处理过程中出水的 COD.

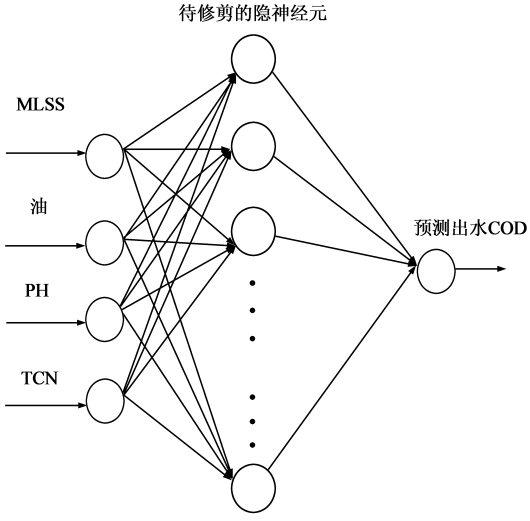

图3 化学需氧量软测量神经网络模型

实验数据来自某污水处理厂水质分析日报表,将其分为训练样本 70 组,测试样本为 30 组.神经网络的初始模型分别取 4-16-1 和 4-32-1 两种形式,用 EFAST 算法和本文提出的快速 EFAST 算法对这两种神经网络进行修剪.两种方法采用同样的修剪准则,即将隐含层中灵敏度小于 5%的神经元删掉.图 4 给出基于 EFAST 算法和本文提出的快速 EFAST 算法对该神经网络中隐含层各神经元灵敏度的计算结果.图 4 中 S_n 为 EFAST 算法得出的第 n 个隐神经元的灵敏度(用浅灰色直方图表示),Sf_n 为本文提出的快速 EFAST 算法得出的第 n 个隐神经元的灵敏度(用深灰色直方图表示).

由图 4 可知,本文提出的快速 EFAST 算法计算出的灵敏度值与 EFAST 算法计算出的灵敏度值相差很小(不超过 0.05%),两种算法对神经网络修剪后剩下的神经元是完全相同的.此外,虽然神经网络的初始结构不同,但修剪后的神经网络可以保持稳定的结构.

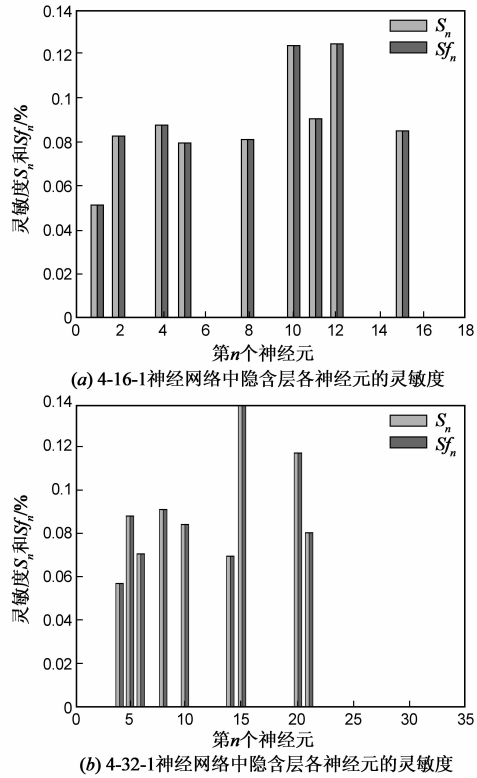

图4 EFAST算法和快速EFAST算法计算出的各神经元灵敏度

根据污水处理过程的实际需求,设定 COD 的预测误差不得超过 1%,采用本文提出的快速 EFAST 算法和 EFAST 算法分别对结构为 4-16-1 和 4-32-1 的前向神经网络进行修剪,由灵敏度实验数据可知,两种方法修剪后的神经网络结构均为 4-9-1.而两种修剪方法的运行时间见表 1.不难看出本文提出的方法较 EFAST 方法具有明显速度优势.

表 1 两种算法性能比较

	EFAST	快速 EFAST
4-16-1 神经网络	58″	2″
4-32-1 神经网络	3′55″	7″

用基于本文提出的快速 EFAST 方法的结构自组织神经网络对水质参数 COD 进行预测,首先用 70 组训练样本对该神经网络进行训练,然后用于对另外 30 组测试样进行预测,COD 的预测结果与实际输出 COD 值如图 5 所示.

实线表示实测的出水 COD 值,虚线表示神经网络

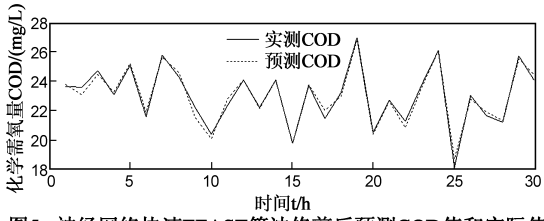

图5 神经网络快速EFAST算法修剪后预测COD值和实际值拟合曲线图

预测的出水 COD 值,由图 5 可知,基于快速 EFAST 方法的结构自组织神经网络能够有效地预测出水 COD.

5 结论

根据前向神经网络的特点,提出了一种快速 EFAST 算法修剪神经网络的算法.基于该算法的自组织神经网络具有以下特点:

(1)神经网络可以根据承担的任务,自动地修剪其自身结构.

(2)神经网络最终结构取决于承担的任务,而不依赖初始结构.即当初始神经网络的隐神经元个数不同时,用该算法修剪后的神经网络保留的神经元数目大致相同.

(3)与 EFAST 算法相比,本文给出的快速 EFAST 算法在神经网络修剪速度方面具有明显的优势,尤其针对初始规模较大的前向神经网络,修剪后得到的神经网络结构完全相同,但运行时间明显缩短.

(4)基于快速 EFAST 算法的结构自组织神经网络可以实现对 COD 的有效预测.

参考文献:

[1] Y Le Cun, J S Denker, S A Solla. Optimal brain damage[A]. Advances in Neural Information Processing Systems[C]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1990, 2: 598 – 605.

[2] B Hassibi, D G Stork. Second-order derivatives for network pruning: optimal brain surgeon[A]. Advances in Neural Information Processing Systems[C]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1993, 5: 164 – 171.

[3] T Cibas, F F Soulié, P Gallinari, S Raudys. Variable selection with neural networks[J]. Neurocomputing, 1996, 12(2 – 3): 223 – 248.

[4] M E Ricotti, E Zio. Neural network approach to sensitivity and uncertainty analysis[J]. Reliability Engineering and System Safety, 1999, 64(1): 59 – 71.

[5] Saltelli S Tarantola, K-S Chan. A quantitative model independent method for global sensitivity analysis of model output[J]. Technometrics, 1999, 41(1): 39 – 56.

[6] Philippe Lauret, Eric Fock, Thierry Alex Mara. A node pruning algorithm based on a Fourier amplitude sensitivity test method [J]. IEEE Transactions on Neural Networks, 2006, 17(2): 273 – 293.

[7] R I Cukier, C M Fortuin, K E Shuler, A G Petscheck, J H Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients [J]. Journal of Chemical Physics, 1973, 59(8): 3873 – 3878.

[8] Christian Schittenkopf, Gustavo Deco, Wilfried Brauer. Two strategies to avoid over fitting in feed forward networks[J]. Neural networks, 1997, 10(3): 505 – 516.

[9] P Engelbretch. A new pruning heuristic based on variance analysis of sensitivity information[J]. IEEE Transactions on Neural Network, 2001, 12(6): 1386 – 1399.

作者简介:

乔俊飞 男,出生于内蒙古鄂尔多斯.1998 年于东北大学获博士学位,现为北京工业大学教授、博士生导师,主要研究方向为智能控制理论及应用、神经网络分析与设计等.
E-mail: isibox@sina.com

李 淼 女,出生于天津.2006 年毕业于大连海事大学自动化系,同年进入北京工业大学自动化系攻读硕士学位,2009 年取得硕士学位,主要从事神经网络设计、污水处理智能建模与仿真的有关研究.
E-mail: limiao19831203@emails.bjut.edu.cn

刘 江 男,出生于山西.2005 年进入北京工业大学自动化系攻读硕士学位,2008 取得硕士学位,主要从事运动控制、神经网络设计的有关研究.
E-mail: liu-jiang05@emails.bjut.edu.cn