用神经元网络辨识非线性系统中的网络结构选择*

鲍晓红 贾英民
(北京航空航天大学第七研究室·北京,100083)

摘要:本文定义了神经元网络的权值拟熵,在对多层前馈网训练的常规目标函数中加入权值拟熵作为约束项以改变网络的权值分布从而修定网络结构.将此方法用于一类非线性系统的神经网络辨识中可以优化网络模型输入项数和隐节点数目.

关键词:多层前馈网;权值拟熵;权值拟概率;非线性系统辨识;目标函数

1 引 言

随着社会发展,越来越多的非线性现象已引起广泛重视,建立描述非线性系统的模型是研究非线性问题的基础,有了适当的模型才能开展有关的分析、设计、预报和仿真等研究工作.

由于非线性系统的复杂性,要找到一种统一的且在计算复杂性方面实用的模型比较困难,在控制领域中较常用的为 NARMAX 模型[1],以往的辨识研究大多是针对此模型的某个子集模型.神经网络的出现,特别是多层前馈网对一般连续非线性函数的逼近特性为一般的 NARMAX 模型提供了统一的参数化表达.这样,辨识工作就转化为确定网络结构参数(层数,各层节点个数)和求取权值.目前已作的工作如[2,3]等,它们用于辨识的网络结构大都是预先选好的,即针对模型(4)并假定输入输出延迟数 n_u 和 n_y 已知,这样,网络的输入和输出节点个数就确定下来,隐层取 1～2 个,其隐节点个数的选取思想是尽量多,以保证必要的逼近精度.再通过一些优化算法极小化预报误差从而求取网络权值.但是现已知,隐节点选取过多会降低系统的外推能力,过少会使逼近精度降低,所以隐节点个数的恰当选取是前馈网在具体应用时的一个重要问题.再有,一般情况下 n_u,n_y 并不确切知道,最多只知道其上界,这又对应着如何恰当选取输入节点个数.这些问题导致了网络结构优化方面的工作.

这些工作可分为下面两种途径:

1) 网络由简单逐渐增至复杂[4,5].

· 优化理论法[4].此方法理论意义明确,但具体计算时要求逐步增加隐节点的个数且在不同隐节点下对网络重新训练并要作某种相关性检验,因而计算量大,并且此方法仅针对输入项数已知的情况.

2) 网络由复杂变为简单[6,7].

· 灵敏度法[7].此方法思路简单,但每去掉一个对输出影响最小的权值都要对网络进行重新训练,因而计算量相对大.

· 惩罚函数法[6,7].此方法思路为在常规目标函数中加入限制网络权值分布的惩罚项,在学习过程中迫使大网的多数权值趋于零或可合并,从而达到简化网络结构的目的.此方法意义很直观,且只要训练一次,计算量小,但惩罚函数的选择目前尚处于尝试阶段.

本文定义了一种新的惩罚函数来进行多层前馈网的结构优化,具体为先定义网络权值的拟概率,在此基础上定义权值的拟熵,并把它加入目标函数中,网络通过极小化此修正的目标

* 国家自然科学基金资助项目.
本文于 1996 年 1 月 2 日收到.1996 年 6 月 27 日收到修改稿.

函数来同时达到函数拟合和优化网络结构的目的.仿真证明,此方法用于非线性系统辨识中对合理选择网络输入项数和隐节点数目有指导作用.

2 网络的权值拟熵

在下文中我们把 n 个输入,k 个隐节点,m 个输出的三层前馈网记为$(n+1)-k-m$,即把阈值也并入权值同时增加一个值为 1 的输入作为输入层第 $n+1$ 个节点.

定义 1 对结构为$(n+1)-k-m$ 的三层前馈网,其权值的拟概率${}^1p_{ij}$,${}^2p_{li}$ 定义为

$$ {}^1p_{ij}=\frac{{}^1W_{ij}^2}{\sum_{i_0=1}^{k}\sum_{j_0=1}^{n+1}{}^1W_{i_0j_0}^2},\quad {}^2p_{li}=\frac{{}^2W_{li}^2}{\sum_{l_0=1}^{m}\sum_{i_0=1}^{k}{}^2W_{l_0i_0}^2}. $$

其中${}^1W_{ij}$为第一层第 j 个元到第二层第 i 个元的权值,${}^2W_{li}$ 为第二层第 i 个元到第三层第 l 个元的权值. $j=1,\cdots,n+1;i=1,\cdots,k;\ l=1,\cdots,m$.

定义 2 定义网络的权值拟熵

$$ E={}^1E+{}^2E. \tag{1} $$

其中 $$ {}^1E=-\sum_{i_1=1}^{k}\sum_{j_1=1}^{n+1}{}^1p_{i_1j_1}\ln{}^1p_{i_1j_1},\quad {}^2E=-\sum_{l_1=1}^{m}\sum_{i_1=1}^{k}{}^2p_{l_1i_1}\ln{}^2p_{l_1i_1}. $$

下面,我们对熵函数作一下简单分析.回顾信息论中所定义的熵函数,设 $p_i=\frac{n_i}{n_0}$,$\sum_{i=1}^{q}n_i=n_0(n_i\geqslant 0,n_0>0,q$ 为某正整数),则熵函数为

$$ H(p)=H(p_1,\cdots,p_q)\triangleq-\sum_{i=1}^{q}p_i\ln p_i. \tag{2} $$

定理 1[8] 熵函数是非负的.即 $H(p)=H(p_1,\cdots,p_q)\geqslant 0$,当且仅当对某 i,$p_i=1$,其余的 $p_k=0(k=1,\cdots,q$,且 $k\neq i)$ 时等号成立.

定理 2[8] 当 p_i 为等值$\frac{1}{q}$ 时,熵函数(2) 取最大值.

则可知对于函数(2) 其值域范围是

$$ 0=H(0,\cdots,0,1,0,\cdots,0)\leqslant H(p_1,\cdots,p_q)\leqslant H\left(\frac{1}{q},\cdots,\frac{1}{q}\right)=\ln q. $$

且上下界是可以达到的.

推论 1 对于我们定义的网络权值拟熵(1),当 $m=1$ 时,有

$$ 0\leqslant E\leqslant\ln((n+1)k)+\ln k. \tag{3} $$

且当两层都只有一个权值不为零其余值都为零时,E 达到最小值,而当在不计符号的情况下各层权值分别均匀分布时 E 达到最大值.

由此拟熵函数的性质可以看出,当权值分布很集中时,函数值小,反之当权值分布很均匀时,函数值大.我们利用这个性质把它作为惩罚函数来控制前馈网训练时的权值分布.

3 非线性系统辨识

一类广泛的离散时间非线性系统可用NARMAX 模型表示,我们在此考虑如下SISO 系统

$$ y(t)=f(y(t-1),\cdots,y(t-n_y),u(t),\cdots,u(t-n_u))+\varepsilon(t). \tag{4} $$

其中 $u(t)$,$y(t)$ 分别是输入输出,n_u,n_y 分别是输入输出的最大延迟,$\varepsilon(t)$ 是零均值白噪声,$f(\cdot)$ 是连续非线性函数.

'对此模型,我们采用带输入延迟的多层前馈网作为辨识模型.具体辨识过程用串 - 并辨识(如图 1)

网络初始结构定为$(n_y+n_u+1)-k-1$,其中 k 为一较大正整数.

对于此动态结构的辨识，传统的目标函数一般取

$$J_c = \frac{1}{2N}\sum_{q=1}^{N}(y(q) - \hat{y}(q))^2. \tag{5}$$

在此我们为了同时获得网络的简化结构定义修正的目标函数

$$J = J_c + \lambda E. \tag{6}$$

其中 E 为(1)中所定义的权值拟熵函数，$\lambda > 0$ 为可调系数. 其意义在于由于 E 有极值性质(3)，在学习过程中通过调整网络权值极小化(6)使网络在进行输入输出数值拟合的同时，尽量使权值分布集中化，从而出现大量的零或可合并的等值小量而达到精简网络结构的目的.

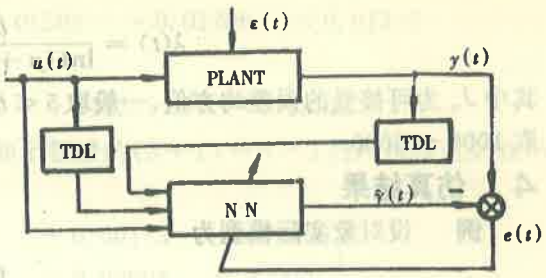

图 1　神经网络辨识模型

下面我们给出对于修正的目标函数(6)的学习算法. 为突出重点，本文选用最简单的 BP 算法. 即令 Θ 为网络中所有可调参数构成的向量，有

$$\Theta(t+1) = \Theta(t) - \eta\nabla J. \tag{7}$$

其中 η 为迭代步长.

在网络中　　$\hat{y} = \sum_{i=1}^{k} {}^2W_{1i}O_i, \quad O_i = \mathrm{sig}\Big(\sum_{j=1}^{n+1} {}^1W_{ij}x_j\Big).$

其中向量 $x = [u(t),\cdots,u(t-n_u),y(t-1),\cdots,y(t-n_y),1]^{\mathrm{T}}$，$\mathrm{sig}(z) = \dfrac{1}{1+\exp(-z)}$.

令 $e(q) = y(q) - \hat{y}(q)$，在输入每组辨识信号(N 个)之后统一调整权值如下：

$${}^1W_{ij}(t+1) = {}^1W_{ij}(t) + \frac{\eta}{N}\sum_{q=1}^{N}\frac{\partial\hat{y}(q)}{\partial {}^1W_{ij}}e(q) - \eta\lambda\frac{\partial E}{\partial {}^1W_{ij}}.$$

其中　　$\dfrac{\partial\hat{y}}{\partial {}^1W_{ij}} = {}^2W_{1i}O_i(1-O_i)x_j, \quad \dfrac{\partial E}{\partial {}^1W_{ij}} = \dfrac{\partial^1 E}{\partial {}^1W_{ij}} = -\sum_{i_1,j_1}(\ln {}^1p_{i_1j_1} + 1)\dfrac{\partial {}^1p_{i_1j_1}}{\partial {}^1W_{ij}},$

$$\frac{\partial {}^1p_{i_1j_1}}{\partial {}^1W_{ij}} = \begin{cases} 2\dfrac{{}^1W_{ij}\sum_{i_0,j_0}{}^1W_{i_0j_0}^2 - {}^1W_{ij}^3}{(\sum_{i_0,j_0}{}^1W_{i_0j_0}^2)^2}, & i_1 = i, j_1 = j, \\ -2\dfrac{{}^1W_{i_1,j_1}^2\,{}^1W_{ij}}{(\sum_{i_0,j_0}{}^1W_{i_0j_0}^2)^2}, & \text{其它.} \end{cases}$$

同样　　$${}^2W_{1i}(t+1) = {}^2W_{1i}(t) + \frac{\eta}{N}\sum_{q=1}^{N}\frac{\partial\hat{y}(q)}{\partial {}^2W_{1i}}e(q) - \eta\lambda\frac{\partial E}{\partial {}^2W_{1i}}.$$

其中　　$\dfrac{\partial\hat{y}}{\partial {}^2W_{1i}} = O_i, \quad \dfrac{\partial E}{\partial {}^2W_{1i}} = \dfrac{\partial^2 E}{\partial {}^2W_{1i}} = -\sum_{i_1}(\ln {}^2p_{1i_1} + 1)\dfrac{\partial {}^2p_{1i_1}}{\partial {}^2W_{1i}},$

$$\frac{\partial {}^2p_{1i_1}}{\partial {}^2W_{1i}} = \begin{cases} 2\dfrac{{}^2W_{1i}\sum_{i_0}{}^2W_{1i_0}^2 - {}^2W_{1i}^3}{(\sum_{i_0}{}^2W_{1i_0}^2)^2}, & i_1 = i, \\ -2\dfrac{{}^2W_{1i_1}^2\,{}^2W_{1i}}{(\sum_{i_0}{}^2W_{1i_0}^2)^2}, & \text{其它.} \end{cases}$$

(6)中 λ 的选取很重要，过小会不起作用，过大会引起网络误差增加. 因此考虑到既要使熵函数发挥作用，又要对拟合误差影响小，我们选用一种随学习步数衰减的系数使其初值很大，逐渐衰减至不影响误差的程度.

$$\lambda(t)=\frac{bJ_m a^T}{\ln((n+1)k)+\ln k}\cdot a^{-t}. \tag{8}$$

其中J_m为可接受的误差均方值，一般取$5<b<10$，$1<a<1.01$，T为终止步数，根据情况可取1000～5000.

4 仿真结果

例 设对象实际模型为

$$y(t)=\frac{0.6}{1+\exp(-0.5u(t)+0.4y(t-1)+0.1)}+\varepsilon(t).$$

在模型未知的情况下用神经网络进行建模仿真研究，其中$\varepsilon(t)=0.01\mathrm{rand}[-1,1]$.

显然，此系统用$(2+1)-1-1$的网络，权值取${}^1W=[-0.5,0.4,0.1]$，${}^2W=[0.6]$即可实现系统动态.

但若事先对系统了解甚少，我们会用一较大的网来辨识系统，如取$n_y=2$，$n_u=1$，网络结构为$(3+1)-7-1$，辨识输入信号用六级PRBS序列，作为对比，我们先用BP算法对目标函数(5)训练2000次，取$N=6$，迭代步长$\eta=0.3$.

引入误差指数EI来表征模型拟合精度$\mathrm{EI}=\sqrt{\sum_{q=1}^{N}e^2(q)/\sum_{q=1}^{N}y^2(q)}$，则得到

$${}^1W=\begin{bmatrix}-0.65288 & 0.61440 & 0.89569 & 0.23845\\ 0.05769 & 0.25768 & 0.91565 & -0.16215\\ -0.01763 & 0.22616 & 0.43350 & -0.20967\\ -0.05365 & 0.32974 & 0.14337 & -0.35629\\ 0.62542 & 0.05827 & 0.60620 & 0.26826\\ 0.76574 & 0.11964 & -0.01501 & 0.61274\\ 0.77355 & -0.95197 & -0.43393 & -0.47097\end{bmatrix},$$

$${}^2W=[-0.19159\quad 0.84516\quad -0.56950\quad -0.12323\quad 0.02891\quad 0.51531\quad -0.30466],$$

$J_c=0.00007$，${}^1E=2.72056$，${}^2E=1.31300$，$\mathrm{EI}=0.03789$.

可见权值分布非常无规律，看不出哪个权值或哪个节点是可以去掉的，我们把此网叫网Ⅰ.若对它加信号

$$u(t)=\sin\frac{2k\pi}{25}. \tag{9}$$

并取$N=25$，则得到此时的$J_c=0.00055$，$\mathrm{EI}=0.11966$，可见网Ⅰ的外推能力不强.

下面，我们用本文的惩罚函数法重作上例.$\lambda(t)$中的参数取$J_m=0.0002$，$b=10$，$a=1.003$，$T=2000$.则在其它条件完全相同时，进行训练后有

$${}^1W=\begin{bmatrix}-0.00190 & -0.00013 & -0.00004 & 0.00000\\ 0.33374 & 0.01047 & -0.00173 & -0.02568\\ -0.00190 & -0.00013 & -0.00004 & 0.00000\\ -0.00190 & -0.00013 & -0.00004 & 0.00000\\ -0.00190 & -0.00013 & -0.00004 & 0.00000\\ -0.00190 & -0.00013 & -0.00004 & 0.00000\\ -0.00134 & -2.73232 & -0.00001 & 0.00007\end{bmatrix},$$

$$^2W = [-0.01206 \quad 0.62985 \quad -0.01206 \quad -0.01206 \quad -0.01206 \quad -0.01206 \quad -0.00979],$$

$$J_c = 0.00007, \quad ^1E = 0.07774, \quad ^2E = 0.01620, \quad \mathrm{EI} = 0.03823.$$

可见，权值中明显存在可合并的项，说明用如下权值的(3＋1)－3－1的网可以达到相同的效果.

$$^1W = \begin{bmatrix} 0.33374 & 0.01047 & -0.00173 & -0.02568 \\ -0.00190 & -0.00013 & -0.00004 & 0.00000 \\ -0.00134 & -2.73232 & -0.00001 & 0.00007 \end{bmatrix},$$

$$^2W = [0.62985 \quad -0.06030 \quad -0.00979].$$

对于简化了的网，再加信号(9)得到$J_c = 0.00011$，$\mathrm{EI} = 0.05359$. 可以看出此简化网的外推能力要比网 Ⅰ 强. 观察到输入层第 3 个节点的权值偏小，可考虑把此节点去掉，此时加信号(9)得 $J_c = 0.00011$，$\mathrm{EI} = 0.05367$，可知网络可进一步减化为(2＋1)－3－1(即取 $n_y = 1$)而基本保持原外推特性. 我们把此简化网叫网 Ⅱ.

仿真发现 $\lambda(t)$ 中参数 a，b 的选取对结果影响很大，一般规律为增大 a 对输入层节点选取影响大，增加 b 对隐节点选取影响大. 这还要结合具体问题慎重选取.

目前惩罚函数法中较常用的惩罚函数及相应的目标函数为[7]：

$$P_1 = \sum_{i,j} |W_{ij}|, \quad J_1 = J_c + \lambda_1 P_1. \tag{10}$$

其中 W_{ij} 代表网络中所有权值.

我们用 J_1 作为目标函数对上例用 BP 法学习，取 $\lambda_1 = 0.001$，学习 5000 步得到结果如下：

$$^1W = \begin{bmatrix} 0.00002 & -0.00012 & 0.00015 & -0.00024 \\ 0.00013 & -0.00019 & -0.00007 & 0.00005 \\ 0.00021 & 0.00028 & -0.00013 & -0.00010 \\ -0.00029 & 0.00012 & -0.00028 & -0.00027 \\ -0.00011 & -0.00015 & 0.00003 & 0.00013 \\ 0.33830 & -0.00034 & -0.00013 & -0.00038 \\ 0.00007 & -0.00010 & 0.00028 & -0.00025 \end{bmatrix},$$

$$^2W = [0.00052 \quad 0.00014 \quad 0.00046 \quad 0.00009 \quad 0.00010 \quad 0.55925 \quad 0.00048],$$

$$J_c = 0.00013, \quad ^1E = 0.0011, \quad ^2E = 0.00062, \quad \mathrm{EI} = 0.05446.$$

可见此方法也能使权值集中分布，但拟合精度低一些，更重要的问题是我们仍无法确定该去掉哪些节点和哪些权值. 如果按权值大小决定删除项，可得到如下简化结构：

$$^1W = [0.33830 \quad -0.00038], \quad ^2W = [0.55925].$$

把此网叫网 Ⅲ，对它用信号(9)来检验，得到 $J_c = 0.00018$，$\mathrm{EI} = 0.06790$，可见网 Ⅲ 的外推能力比网 Ⅰ 强，但比网 Ⅱ 的外推精度低一些.

由此可以看出，网络过大(如网 Ⅰ)会引起较大的外推误差，而过小(如网 Ⅲ)会使拟合误差和外推误差都较大，而用本文的惩罚函数法，只要 λ 选取合适，虽然不能保证使网络结构一定最佳，但可以使网络从拟合误差，外推误差，计算复杂性等方面综合考虑起来是较优的.

5　结　论

本文定义了网络的权值拟熵并把它加入目标函数作为惩罚项，在一类非线性系统的神经网络辨识中可使网络在权值学习的过程中自动简化结构. 分析和仿真表明，此方法物理意义明

确，简单可行，是同类思路算法中较优的算法. 在应用中惩罚系数λ的选择很重要，若不当反而会使性能恶化，本文虽给出一种选择思路，但其中仍存在待调整的参数，对它们的恰当选取有待于今后理论和实际工作中进一步论证和摸索.

致谢 本文写作过程中与北京航空航天大学第七研究室周彤老师多次讨论，受到很大启发，作者在此深表感谢.

参 考 文 献

1 Leontaritis, I. J. and Billings, S. A.. Input-output parametric models for nonlinear systems. Part I : Deterministic nonlinear systems. Int. J. Control, 1985 41(2):303－328

2 Chen, S. and Billings, S. A.. Nonlinear system identification using neural networks. Int. J. Control, 1990, 51(6): 1191－1214

3 Narendra, K. S. and Parthasarathy, K.. Identification and control of dynamical systems using neural networks. IEEE Trans. Neural Networks, 1990, 1(1):4－27

4 Wang, Z. and Massimo, C. D.. A procedure for determining the topology of multilayer feedforward neural networks. Neural Networks. 1994, 7(2):291－300

5 Nabhan, T. M. and Zomaya, A. Y.. Toward generating neural networks structure for function approximation. Neural Networks, 1994, 7(1):89－99

6 Kamimura, R.. Internal representation with minimum entrophy in recurrent neural networks: Minimizing entropy through inhibitory connections. Network, 1993, 4:423－440

7 Reed, R.. Pruning algorithms——A survey. IEEE Trans. Neural Networks. 1993, 4(5):740－747

8 周荫清主编. 信息论基础. 北京：北京航空航天大学出版社，1993

Network Structure Selection in Neural Network Based Nonlinear System Identification

BAO Xiaohong and JIA Yingmin

(The Seventh Research Division, Beijing University of Aeronautics and Astronautics · Beijing, 100083, PRC)

Abstract: In this paper we define the pseudo-entropy of network weights. By adding it into the normal objective function we can obtain a rational network structure during training. Put this method into the neural network based nonlinear system identification we can acquire a proper number of input and hidden neurons.

Key words: multilayer neural networks; pseudo-entropy of weights; pseudoprobability of weights; nonlinear system identification; objective function

本文作者简介

鲍晓红 1967年生. 1990年毕业于北京航空航天大学自动控制系，获学士学位. 1993年获北京航空航天大学自动控制理论及应用专业硕士学位，目前在攻读该专业博士学位. 主要研究兴趣为人工神经元网络基础理论及其在控制、系统辨识及模式识别等方面的应用.

贾英民 1958年生. 1981年毕业于山东大学数学系控制理论专业，分配到河南焦作矿业学院电器工程系任教，1990年和1993年分获北京航空航天大学自动控制理论及应用专业硕士和博士学位，1994年在北京航空航天大学航空与宇航技术博士后流动站从事博士后研究工作期满，现在北京航空航天大学第七研究室工作. 1992年任副教授，1995年任教授. 学术兴趣为多变量系统，鲁棒控制，H_∞理论，智能控制等.