基于自适应小波神经网络的数据挖掘方法研究
——对我国石油产量的预测分析

刘兰娟，谢美萍
（上海财经大学 经济信息管理与工程学院，上海 200433）

摘　要：小波神经网络是近年来在小波分析研究获得突破性进展基础上提出的一种前馈型网络，文章将小波与神经网络相结合，提出了一种基于自适应小波神经网络（SAWNN，self-adaptation wavelet neural network）的数据挖掘方法，并构造了数据挖掘过程的机器学习机制，以提高对问题的处理能力。文章将所构造的自适应小波神经网络用于石油产量的建模预测研究，实证结果表明此预测模型不仅是有效的，而且是可行的。

关键词：石油产量；预测研究；自适应小波神经网络
中图分类号：F224.0　**文献标识码**：A　**文章编号**：1001-9952(2006)03-0114-07

一、引　言

数据挖掘是从大量数据中提取人们感兴趣信息的处理过程，这些信息是隐含的、事先未知的，并且是可信的、新颖的、潜在有用的、能被人们理解的一种“模式”。数据挖掘的目的就是从数据中找出有意义的模式。这种模式可以是一组规则，聚类、决策树或其他方法表示的知识。一般而言，一个典型的数据挖掘过程可以分为四个阶段，即数据预处理、建模、模型评估及模型应用。数据预处理阶段主要包括对数据的理解、属性选择、连续属性离散化、数据中的噪声及丢失值的处理、实例选择等。建模包括学习算法的选择，参数的确定等。模型评估进行模型的训练和测试，对得到的模型进行评价。模型应用是将上述三个阶段的处理应用于特定的研究或预测分析对象。这四个阶段是循环往复的过程直到用户满意为止。

石油是一种不可替代的资源，当今社会没有一种能源能够取代它，石油是“能源中的能源”；是现代工业的“血液”和现代经济的命脉。石油在第二次工业革命时代被广泛使用后，逐渐取代煤成为第一大能源，占到世界能源消耗总

收稿日期：2005-12-30
作者简介：刘兰娟（1960—），女，上海人，上海财经大学经济信息管理与工程管理学院；
谢美萍（1974—），女，江苏如皋人，上海财经大学经济信息管理与工程管理学院。

量的70%左右,而且石油的各种工业成品已经成为人类社会现代生活中必不可少的必需品,如塑料、合成橡胶、动力及能源燃料等。石油和人类的生活密切相关,渗透到人类生活的各个领域,石油与工业、经济环境、国家政策和军事战争等都有着紧密的联系。

我国是一个产油大国也是一个石油需求大国。石油产量的精确预测对指导化工企业生产、合理安排产品结构及计划企业的经营活动等都具有重要的意义。同时,石油产量的精确预测也是石油生产企业合理制订生产任务、避免盲目决策的有力保证。

石油产量属单因子时间序列变量,针对这种类型的变量,目前所采用的主要预测方法有回归分析法、灰色系统预测法、模糊系统预测法等,但这些方法过于形式化和数学化,预测精度难以得到有效提高。为了准确、客观地预测石油产量的变化,本文将小波与已有的神经网络相结合,构造了一种新型的自适应小波神经网络,这种网络有着其显著的优越性,即在建模预测过程中,只需要知道历史数据,通过这些历史数据构造出一个模型,然后进行预测,这样可以避免需要知道状态方程的麻烦,也可以避免陷入局部极小。

因为小波变换在时、频两域都具有表征信号局部特征的能力,突破了传统Fourier分析的局限性,因此,近年来小波变换已经在理论上[1]~[3]得到更广泛的研究和应用,有用于函数逼近[4],故障诊断[5],实时电力的分配和电力转换的检测[6],特征提取[7],非线性系统的识别[8]和其他一些领域[9][10]。本文利用小波分析的特点,将所构造的这种自适应小波神经网络用于石油产量的建模和预测研究,并以此来构造数据挖掘的机器学习机制,以进一步提高处理预测问题的能力。

二、小波神经网络的结构

1. 小波神经网络。设函数 $\psi(\cdot)$ 满足容许性条件,即有:

$$\int_R \frac{|\hat{\psi}(w)|^2}{|w|}dw < \infty$$

那么可数集合 $\Phi=\{\sqrt{a_k}\Psi(a_k x-b_k): a_k\in R_+, b_k\in R, k\in Z\}$ 满足框架性质,即存在两个常数 $A>0$ 和 $B<\infty$,使得对任意的 $f\in L^2(R)$,都有:

$$A\|f\|^2\leqslant\sum_{\varphi\in\Phi}|<\varphi, f>|^2\leqslant B\|f\|^2$$

上式表明框架 Φ 在 $L^2(R)$ 中是稠密的,即框架 Φ 中元素的所有线性组合的集合

$$f(x)=\sum_{i=1}^{N} w_k\varphi_k(x) \qquad \varphi_k\in\Phi \tag{1}$$

在 $L^2(R)$ 中是稠密的。

于是如下形式的所有有限和的全体

$$f(x)=\sum_{i=1}^{N} w_i \sqrt{a_i}\Psi(a_i x-b_i) \tag{2}$$

在 $L^2(R)$ 中是稠密的。其中 a_i 为任意的伸缩参数，b_i 为任意的平移参数。

比较(1)式和(2)式，显然(2)式中的参数个数比(1)式中的参数个数多，因为在(2)式中 a_i 和 b_i 不一定是可数的。(1)式和(2)式分别称为小波分解和小波网络。在小波分解中，如果基函数固定，则只有系数 w_k 是可调参数，而在小波网络中，w_i、a_i、b_i 均为可调参数，这使得网络学习非线性函数较为灵活，可以满足较高的逼近精度要求。

(2)式与下式是等价的：

$$f(x)=\sum_{i=1}^{N} w_i\Psi(a_i x-b_i)$$

上式即为小波神经网络的结构，它与传统的人工神经网络的区别在于隐层节点激励函数不是 Sigmoid 函数，而是小波函数，这便于对信号进行时频局部化分析，从而对处理诸如突变信号等问题时将显示出较大的优越性。

随着节点数 N 的增加，网络(2)式能充分逼近任意的函数 $f(x)\in L^2(R)$。

由于小波函数 $\Psi(x)$ 是零均值的，为了便于用小波网络逼近非零均值函数，在小波网络(2)式右端加入一个参数 $\bar{f}$，即为：

$$f(x)=\sum_{i=1}^{N} w_i\Psi(a_i x-b_i)+\bar{f}$$

其中 $\bar{f}$ 为 f(x)的均值的估计值。

2. 自适应小波神经网络模型及算法。自适应小波神经网络的拓扑图形如图 1 所示。对于多输入多输出系统 $f:R^n\to R^m$，其输出函数为：

$$\begin{aligned} y_k &= f\Big[\sum_{j=1}^{N} w_j \sum_{i=1}^{N} x_i(t)h((t-b_j)/a_j)\Big] \\ &= f(W,A,B)=f(\cdot) \end{aligned} \tag{3}$$

其中 $y_k(k=1,2,\cdots,m)$ 为系统输出，$x_i(t)$ 为训练样本，$f(\cdot)=1/(1+\exp(\cdot))$，$w_j$ 为网络权值，h 为给定的小波函数，a_j 和 b_j 分别为伸缩和平移参数，$W=\{w_1,w_2,\cdots,w_N\}$，$A=\{a_1,a_2,\cdots,a_N\}$，$B=\{b_1,b_2,\cdots,b_N\}$。

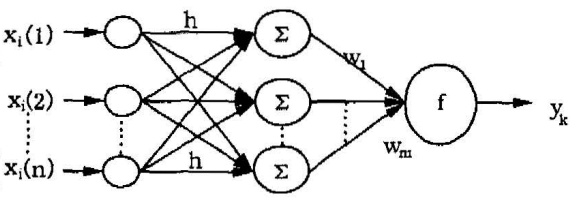

图 1　自适应小波神经网络

由(3)式可知，系统在 k+1 时刻的预测值 $\hat{y}(k+1|k)$ 可写为：

$$\begin{aligned} \hat{y}(k+1|k) &= \hat{y}(k|W,A,B) \\ &= f(y_{k-1},w_{k-1},a_{k-1},b_{k-1},k) \end{aligned} \tag{4}$$

假设由(3)式得到的 m 组数据，则由递推算法可知，给定的指标函数为：

$$J_m(W,A,B)=\frac{1}{m}\sum_{i=1}^{m}J(i,W,A,B)$$

$$=\frac{1}{2m}\sum_{i=1}^{m}\{y_i-f(y_{i-1},W,A,B)\}^2 \tag{5}$$

如果已得参数 W、A、B 的第 k−1 次估值为 W_{k-1}、A_{k-1}、B_{k-1}，则由梯度法，为了求下一次的 W_k、A_k、B_k，应有：

$$W_k=W_{k-1}-R_{k-1}\nabla J_m(W_{k-1},A_{k-1},B_{k-1})|_{W_{k-1}} \tag{6}$$

$$A_k=A_{k-1}-R_{k-1}\nabla J_m(W_{k-1},A_{k-1},B_{k-1})|_{A_{k-1}} \tag{7}$$

$$B_k=B_{k-1}-R_{k-1}\nabla J_m(W_{k-1},A_{k-1},B_{k-1})|_{B_{k-1}} \tag{8}$$

其中：$R_{k-1}=\|\nabla f[k,W_{k-1},A_{k-1},B_{k-1}]\|^{-2}I$，式(6)、式(7)、式(8)中的 $\nabla J_m(\cdot)$ 是指标函数式(5)分别关于 W、A、B 的梯度在 W_{k-1}、A_{k-1}、B_{k-1} 处的值，因此由式(4)有：

$$\nabla J_m(W_{k-1},A_{k-1},B_{k-1})|_{W_{k-1}}=-\frac{1}{m}\sum_{n=1}^{m}\{y_k-f[Y_{k-1},W_{k-1},A_{k-1},B_{k-1},k]\}|_{W=W_{k-1}} \tag{9}$$

$$\nabla J_m(W_{k-1},A_{k-1},B_{k-1})|_{A_{k-1}}=-\frac{1}{m}\sum_{n=1}^{m}\{y_k-f[Y_{k-1},W_{k-1},A_{k-1},B_{k-1},k]\}|_{A=A_{k-1}} \tag{10}$$

$$\nabla J_m(W_{k-1},A_{k-1},B_{k-1})|_{B_{k-1}}=-\frac{1}{m}\sum_{n=1}^{m}\{y_k-f[Y_{k-1},W_{k-1},A_{k-1},B_{k-1},k]\}|_{B=B_{k-1}} \tag{11}$$

将(9)式、(10)式、(11)式和 R_{k-1} 分别相应地代入(6)式、(7)式、(8)式中，就可以求得使指标函数极小化的算法 W_k、A_k、B_k。

三、实例预报

至此，我们将所构造的模型及算法应用于从 1990 年 1 月至 2004 年 6 月间的原油月产量数据(共计 174 个数据，见表 1)来进行建模预报，前 168 个数据作为建模使用，后 6 个数据作为预报比较使用，将预报的数据和原来的数据进行比较，采用均方误差评估。

在进行预报之前，首先将数据先进行归一化处理，即找出所给数据的最大值，然后将所有 30 个数据都除以这个绝对值的最大值，这样就可以使所有数据的模都小于 1。

在实际操作中，使用 Visual C++编程计算，选取的小波函数为 $h(x)=-x\exp(-x^2/2)$，共选用 7 个小波函数，其初始伸缩 a_j 和平移参数 b_j 通过正交化手段得到，初始权值 w_j 均取为 1。根据算法规则(5)式、(6)式、(7)式和(8)式进行编程计算，将 Visual C++计算所得的数据生成数据文件，再用

Excel 将生成的数据文件绘制成预测图形。具体算法步骤如下：

表 1　1990 年 1 月至 2004 年 6 月原油月产量数据①

时间	月原油产量（万吨）	时间	月原油产量（万吨）	时间	月原油产量（万吨）	时间	月原油产量（万吨）	时间	月原油产量（万吨）
1990 年 1 月	1 152.80	1993 年 1 月	1221.20	1996 年 1 月	1 271.51	1999 年 1 月	1 363.47	2002 年 1 月	1 428.96
2 月	1 060.50	2 月	1 099.20	2 月	1 270.68	2 月	1232.85	2 月	1 277.03
3 月	1 174.50	3 月	1 222.00	3 月	1 339.56	3 月	1 355.28	3 月	1 422.30
4 月	1 125.40	4 月	1 187.20	4 月	1 267.05	4 月	1 319.89	4 月	1 369.64
5 月	1 168.00	5 月	1 239.40	5 月	1 352.59	5 月	1 385.21	5 月	1 430.23
6 月	1 139.00	6 月	1 212.50	6 月	1 330.06	6 月	1 322.53	6 月	1 480.66
7 月	1 156.00	7 月	1 240.90	7 月	1 359.25	7 月	1 375.46	7 月	1 442.60
8 月	1 169.60	8 月	1 209.60	8 月	1 330.64	8 月	1 361.70	8 月	1 481.18
9 月	1 142.30	9 月	1 186.00	9 月	1 297.01	9 月	1 287.16	9 月	1 409.93
10 月	1 180.70	10 月	1 260.10	10 月	1 371.17	10 月	1 339.96	10 月	1 463.69
11 月	1 152.80	11 月	1 207.60	11 月	1 333.64	11 月	1 334.90	11 月	1 391.16
12 月	1 173.90	12 月	1 227.20	12 月	1 344.36	12 月	1 359.36	12 月	1 431.35
1991 年 1 月	1 172.80	1994 年 1 月	1 326.14	1997 年 1 月	1 374.46	2000 年 1 月	1 346.52	2003 年 1 月	1 433.75
2 月	1 071.80	2 月	1 155.41	2 月	1 235.71	2 月	1 299.70	2 月	1 311.60
3 月	1 180.90	3 月	1 237.49	3 月	1 385.96	3 月	1 407.54	3 月	1 455.84
4 月	1 144.50	4 月	1 189.29	4 月	1 348.25	4 月	1 338.56	4 月	1 415.94
5 月	1 187.40	5 月	1 258.26	5 月	1 389.20	5 月	1 371.09	5 月	1 457.34
6 月	1 146.80	6 月	1 223.11	6 月	1 343.98	6 月	1 344.35	6 月	1 419.40
7 月	1 175.20	7 月	1 255.58	7 月	1 379.44	7 月	1 367.55	7 月	1 436.81
8 月	1 180.50	8 月	1 239.32	8 月	1 357.43	8 月	1 355.21	8 月	1 439.17
9 月	1 153.40	9 月	1 179.21	9 月	1 337.33	9 月	1 336.32	9 月	1 385.14
10 月	1 190.30	10 月	1 259.20	10 月	1 375.06	10 月	1 363.37	10 月	1 438.71
11 月	1 148.00	11 月	1 250.62	11 月	1 325.28	11 月	1 317.69	11 月	1 391.51
12 月	1 188.00	12 月	1 284.11	12 月	1 352.88	12 月	1 374.93	12 月	1 459.11
1992 年 1 月	1 199.40	1995 年 1 月	1 236.29	1998 年 1 月	1 370.84	2001 年 1 月	1 351.77	2004 年 1 月	1 452.44
2 月	1 133.00	2 月	1 163.32	2 月	1 185.31	2 月	1 276.48	2 月	1 379.96
3 月	1 200.30	3 月	1 281.61	3 月	1 342.02	3 月	1 433.60	3 月	1 440.51
4 月	1 169.50	4 月	1 218.40	4 月	1 287.52	4 月	1 356.86	4 月	1 411.77
5 月	1 201.00	5 月	1 240.82	5 月	1 356.64	5 月	1 405.69	5 月	1 473.73
6 月	1 159.50	6 月	1 241.08	6 月	1 341.15	6 月	1 361.02	6 月	1 440.87
7 月	1 196.90	7 月	1 280.56	7 月	1 366.27	7 月	1 385.19		
8 月	1 191.60	8 月	1 277.22	8 月	1 349.44	8 月	1 402.53		
9 月	1 172.30	9 月	1 258.57	9 月	1 317.82	9 月	1 351.04		
10 月	1 217.70	10 月	1 309.11	10 月	1 361.57	10 月	1 406.98		
11 月	1 166.10	11 月	1 258.27	11 月	1 350.37	11 月	1 362.72		
12 月	1 180.30	12 月	1 210.15	12 月	1 386.30	12 月	1 389.48		

第一步：初始化 w_j, a_j, b_j；

第二步：按照(6)式、(7)式、(8)式计算第一步中的参数；

第三步：将第一步所求参数代入(4)式中，求得 $\hat{y}(k+1|k)$；

第四步：计算(5)式；

第五步：重复第二步到第四步，直到满足精度要求为止。

预测结果如图 2 所示。

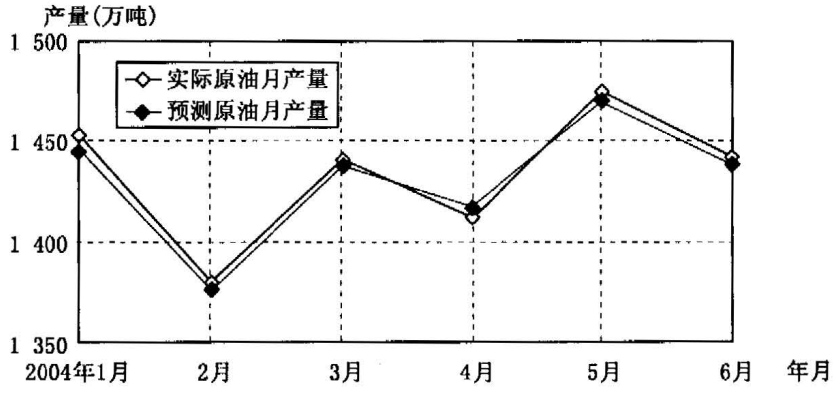

图 2　2004 年 1 月至 2004 年 6 月实际原油产量与预测原油产量比较图

实际原油产量与预测原油产量绝对相对误差如表 2 所示。

表2 2004年1月至2004年6月实际原油产量与预测原油产量的绝对相对误差

时间	实际原油产量(万吨)	预测原油产量(万吨)	相对绝对误差(%)
2004年1月	1 452.44	1 445.39	0.49
2月	1 379.96	1 376.51	0.25
3月	1 440.51	1 437.19	0.23
4月	1 411.77	1 417.27	0.39
5月	1 473.73	1 469.36	0.30
6月	1 440.87	1 437.25	0.25

经过计算,最后所得的均方误差为0.015,结果说明本预测算法不仅是有效的,而且是可行的。

四、结论及展望

数据挖掘作为一个独立的、有别于其他预测工具的研究方法,有其鲜明的特点,尤其是当处理的数据具有庞大、噪声、不确定和稀疏等特性时,更能显现出它的优越性。因此,在对所挖掘规则的评价中,算法的计算复杂度也是一个不可忽视的指标。小波分析是继Fourier分析方法后又一个有力的数学分析工具,在理论研究和数值计算中都有着广泛的应用。本文分析和总结了小波分析的应用成果,所给出的自适应小波神经网络的算法预测问题是基于数据挖掘和知识发现的初步应用,将该算法应用于石油产量的预测分析中,其结果说明这种算法比较有效、可行。

随着小波分析在理论方面的研究越来越深入,如在小波基的选取方面、小波基对收敛性的影响和小波解的逼近误差等方面的研究都可能成为未来的主要研究方向。与此同时,基于小波分析的数据挖掘方法的应用也将越来越广泛,今后在更多的行业中都将会看到基于小波分析的数据挖掘方法的应用。尽管目前这种如何应用于经济领域、特别是经济预测领域的方法探索,是一个比较新的研究课题,本文只是做了一些尝试,但是,基于小波分析的数据挖掘方法在经济预测方面的优越性和适用性将会充分体现出来,今后这方面的应用研究也会越来越多。

注释:

①数据来源中国统计局信息中心。

参考文献:

[1]B Delyon A, Juditsky A. Benveniste, accuracy analysis for wavelet approximations, IEEE trans[J]. On Neural Networks, 1995, 6(2): 332~358.

[2]Yongyong He, Fulei Chu, Binglin Zhong. A hierarchical evolutionary algorithm for constructing and training wavelet networks[J]. Neural Comput & Applic, 2002, 10: 336~357.

[3]Chris C Holmes, Bani K. Mallick, bayesian wavelet networks for nonparametric regres-

sion[J]. IEEE Trans. On Neural Networks, 2000, 11(1): 27～35.

[4]Jun Zhang, Gilbert G Walter, Yubo Miao, Wan Ngai Wayne Lee. Wavelet neural networks for function learning[J]. IEEE Trans. Signal Processing, 1995, 43(6): 1485～1496.

[5]Y C Huang. Fault identification of power transformers using genetic-based wavelet networks[J]. IEE Proc-Sci. Meas. Technol, 2003, 150(1): 25～30.

[6]Wei-ming Wang, Chao-ming Huang. An evolutionary based wavelet network for real-time power dispatch[J]. Electric Power Components and Systems, 2002, (30): 1151～1166.

[7]Takashi Samatsu, Eiji Uchino, Takeshi Yamakawa. Feature extraction of a vectorcardiogram by employing a wavelet network guaranteeing a global minimum[J]. Journal of Intelligent and Fuzzy Systems, 2000, (8): 221～227.

[8]G P Liu S A, Billings V, Kadirkamanathan. Nonlinear system identification using wavelet networks[J]. International Journal of Systems Science, 2000, 3(12): 1531～1542.

[9]Leonardo M Reyneri. Unification of neural and wavelet networks and fuzzy systems, IEEE trans[J]. On Neural Networks, 1999, 10(4): 801～814.

[10]Stephen A Billings, Hua-Liang Wei. A new class of wavelet networks for nonlinear system identification, IEEE Trans[J]. On Neural Networks, 2005, 16(4): 862～874.

Research of Data Mining Method Based on Self-adaptation Wavelet Neural Network-Prediction Analysis of Petroleum Yield

LIU Lan-juan, XIE Mei-ping

(*School of Information Management & Engineering, Shanghai University of Finance & Economics, Shanghai* 200433, *China*)

Abstract: Wavelet neural network, which is based on wavelet analysis, is sort of feed forward network developed in recent years. In this paper, combining the theories of wavelet and neural network together, a new method of the self-adaptation wavelet neural network for data mining is proposed and a machine study mechanism is then constructed in order to improve the capability of the former in tackling problems. Later on, the self-adaptation wavelet neural network is used to model and predict the petroleum yield, and the following results successfully prove that such an application is effective and feasible.

Key words: petroleum yield; prediction study; self-adaptation wavelet neural network (责任编辑 许 柏)