模糊神经网络与SARIMA结合的时间序列预测模型

何星星,孙德山
(辽宁师范大学 数学学院,辽宁 大连 116029)

摘 要:模糊神经网络和SARIMA模型分别对非线性和线性时间序列有很好的预测能力,但在实际应用中大多数序列并非稳定、单纯线性或非线性的。为了提高预测精度,提出了一种基于T-S模糊神经网络与SARIMA结合的时间序列预测模型。针对悉尼航班乘客收入数据给出了三种混合模型,并与模糊神经网络、支持向量机、SARIMA和BP神经网络四种单独模型进行比较。实验结果表明,从预测精度和参数选择方面来看,所给模型是有效的。

关键词:模糊神经网络;SARIMA;混合模型预测;时间序列
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2008)08-0061-04

A Time Series Forecasting Model Using a Hybrid Fuzzy Neural Network and SARIMA

HE Xing-xing, SUN De-shan
(College of Mathematics, Liaoning Normal University, Dalian 116029, China)

Abstract: The fuzzy neural network and season autoregressive integrated moving average (SARIMA) model separately have the extraordinary forecasting ability to the non-linear and linear time series. But most time series are not stable, simply linear or non-linear series in practical application. In order to impove precision forecasting, proposes a time series forecasting model using a hybrid fuzzy neural network (FNN) based on Takagi-Sugeno(T-S) and SARIMA. According to the experiment data of Sydney Airport traffic revenue passengers, three kinds of hybrid models are used and compared among four other models, i.e., the FNN model, the Support Vector Machines model, SARIMA and the BP neural network model. The experiment result indicates that the proposed model is effective in the parameter choice and precision forecasting.

Key words: fuzzy neural network; SARIMA; combined forecast; time series

0 引 言

时间序列预测是通过有限个历史观测样本建立模型,并利用模型解释数据的统计规律,以期达到控制和预报目的的一门技术,在工业过程控制、金融经济、信号处理等众多领域都有广泛的应用。对于平稳时间序列的建模和预测,特别是线性模型的研究,有了许多成熟的技术和方法[1]。但在实际问题中,大多数序列并非平稳、线性的,而目前在这类时间序列的分析和处理上没有较为完善的方法,达不到人们所期望的效果。

随着人工智能技术的发展,如神经网络、支持向量机等技术的产生,由于其良好的非线性预测性能和实用性,在旅游[2~5]、金融[6,7]、工业[8]等时间序列预测中都取得了很好的成绩。但现实中,许多序列并不是单纯的线性或非线性模型,且在实际操作中很难判断其为线性或者非线性,再者在预测分类领域,没有哪一个模型适合于任何情况。因此,一些学者提出了建立结合模型的思想。Zhang[9]提出了基于GRG2神经网络与求和自回归滑动平均模型(ARIMA)结合的混合模型预测太阳黑子活动情况和英镑与美元兑换率数据。Tseng等[10]提出了BP网络与ARIMA结合模型预测台湾工业产值和软饮料时间序列。Chen等[11]提出了支持向量机(SVMs)与季节求和自回归滑动平均模型(SARIMA)结合模型预测台湾工业产值,它们的预测精度均高于单一模型的精度。

文中将提出一种基于Takagi-Sugeno(T-S)模糊神经网络(FNN)与SARIMA结合的混合预测模型。针对悉尼航班乘客收入序列给出了三种不同的结合模型,并与FNN模型、模型SVMs和神经网络模型进行比较。实验结果表明,从参数选择和预测精度方面来

收稿日期:2007-11-18
基金项目:辽宁省教育科研计划(2004C068)
作者简介:何星星(1982-),男,湖南常德人,硕士研究生,研究方向为统计学习理论;孙德山,博士,副教授,主要研究方向为统计学习理论、时间序列分析等。

看,所给模型是有效的。

1 SARIMA 模型

对于时间序列 $\{Z_t, t = 1,2,\cdots\}$ 有季节性、趋势性和周期性时,需要建立非平稳季节模型,可表示为 SARIMA$(p,d,q)(P,D,Q)_s$ 模型,其一般形式为:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)a_t$$

其中:$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$,$p$ 为自回归阶数。

$\Phi_P(B^s) = 1 - \Phi_s B^s - \Phi_{2s} B^{2s} - \cdots - \Phi_{Ps} B^{Ps}$,$P$ 为季节自回归阶数。

$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$,$q$ 为移动平均阶数。

$\Theta_Q(B^s) = 1 - \Theta_s B^s - \Theta_{2s} B^{2s} - \cdots - \Theta_{Qs} B^{Qs}$,$Q$ 为季节移动平均阶数。

d,D 分别为普通差分和季节差分的阶数。s 为季节的长度。a_t 为白噪声序列。

SARIMA 模型的建立首先确定 d,D,s,通过普通差分和季节差分得到零均值的平稳时间序列;然后根据 ACF 和 PACF 确定 p,P,q,Q 的值;再次估计式中的各个系数值,并且对得到的模型进行适应性检验;最后利用建立的模型进行预测。

2 基于 T-S 的 FNN 模型

设输入向量 $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^{\mathrm{T}}$,并设 $T(x_i) = \{A_i^1, A_i^2, \cdots, A_i^{m_i}\}$,$i = 1,2,\cdots,n$,式中 $A_i^j (j = 1,2,\cdots,m_i)$ 是 x_i 的第 j 个语言变量值,定义在论域 U_i 上的一个模糊集合,相应的隶属度函数为 $\mu_{A_i^j}(x_i)$。模糊规则后件是输入变量的线性组合,即:R_j:x_1 是 A_1^j,x_2 是 $A_2^j, \cdots, x_n$ 是 A_n^j,则 $y_i = P_{j0} + P_{j1}x_1 + \cdots + P_{jn}x_n$,式中,$j = 1,2,\cdots,m$;$m \leqslant \prod_{i=1}^{n} m_i$;若输入量采用单点模糊集合的模糊化方法,对于给定的输入 $\boldsymbol{x}$,可求每条规则的适应度为:$\alpha_j = \mu_{A_1^j}(x_1) \wedge \mu_{A_2^j}(x_2) \wedge \cdots \wedge \mu_{A_n^j}(x_n)$,或 $\alpha_j = \mu_{A_1^j}(x_1)\mu_{A_2^j}(x_2)\cdots \wedge \mu_{A_n^j}(x_n)$,模糊系统的输出量为每条规则的输出量的加权平均,即

$$y = \frac{\sum_{j=1}^{m} \alpha_j y_j}{\sum_{j=1}^{m} \alpha_j} = \sum_{j=1}^{m} \bar{\alpha}_j y_j, \text{其中 } \bar{\alpha}_j = \frac{\alpha_j}{\sum_{i=1}^{m} \alpha_i}$$

模糊神经网络的系统结构由前件网络和后件网络两部分组成,前件网络用来匹配模糊规则的前件,后件网络用来产生模糊规则的后件。网络学习算法为 BP 网络算法。具体的网络结构层的描述及学习参数的推导参见文献[12]。

3 混合模型

模糊神经网络和 SARIMA 模型分别对非线性和线性时间序列预测有很好的效果,而将两者结合起来建立混合模型,更能体现各自的特性,以达到预测的目的。

设时间序列 $\{Z_t, t = 1,2,\cdots\}$ 由线性部分 L_t 和非线性部分 N_t 构成,即 $Z_t = L_t + N_t$。先建立 SARIMA 模型预测 Z_t 中的线性部分,令 ε_t 为 Z_t 预测后在 t 时刻的剩余部分,即 $\varepsilon_t = Z_t - \hat{L}_t$,其中 $\hat{L}_t$ 为 SARIMA 模型在 t 时刻的预测值;然后利用 FNN 模型对 ε_t 建模,设输入节点数为 n,则 FNN 预测一般模型为 $\varepsilon_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_{t-n}) + e_t$,其中 f 是由 FNN 确定的非线性函数,e_t 为随机误差;记 $\hat{N}_t = \varepsilon_t$,则混合时间序列预测模型为:$\hat{Z}_t = \hat{L}_t + \hat{N}_t$。

因此,建立混合模型对时间序列进行预测可分两步:第一步,利用 SARIMA 模型针对时间序列中的线性部分 L_t 进行建模;第二步,对剩余的非线性部分 N_t 建立 FNN 模型。

4 实 验

4.1 数据及预处理

采用悉尼航班乘客收入作为实验数据。序列为 1999 年 1 月到 2007 年 1 月悉尼航班乘客收入,如图 1 所示。

其中由于悉尼奥运会的影响,从 2000 年 5 月到 10 月有显著的递增趋势,总共 79 个样本,取前 69 个数据为训练样本,剩余的 10 个样本为测试样本。为了提高预测效果,将原数据归一化:

$$Y_t = \frac{Z_t - Z_{\min}}{Z_{\max} - Z_{\min}} \times 0.7 + 0.15 \quad t = 1,2,\cdots,69$$

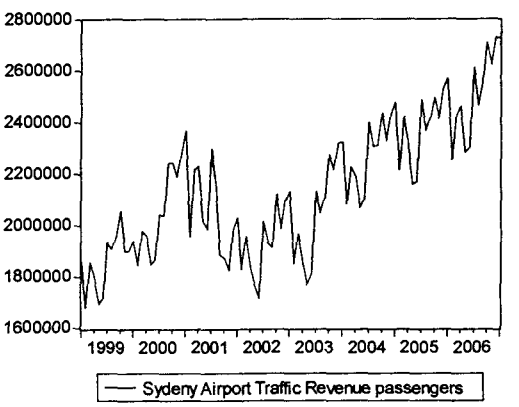

图 1 悉尼航班乘客收入

4.2 建立模型

根据实验数据 $\{Z_t, t = 1,2,\cdots\}$ 的特征,给出以下三种混合模型:

(1)FNN - SARIMA1 模型:

$Z_t = f(Z_{t-1}, Z_{t-12}, \hat{\varepsilon}_t)$

(2)FNN - SARIMA2 模型:

$Z_t = f(\hat{L}_t, \hat{\varepsilon}_t)$

(3)FNN - SARIMA3 模型:

$Z_t = f(\hat{L}_t)$

其中 $\hat{L}_t, \hat{\varepsilon}_t$,分别为 SARIMA 模型在 t 时刻的预测值和误差。

4.3 参数选择

对于 SARIMA 模型,首先为了消除周期性和趋势性,对 Y_t 进行一阶普通差分和一阶季节差分(季节为 12 个月)得到序列 X_t;然后使用 Eviews5.1 软件包对序列建模,根据 AIC 准则,建立的模型为:SARIMA(1,1,1)$(1,1,1)_{12}$,回归方程为:

$$(1 + 0.3578B^{12})(1 + 0.1618B)X_t = -0.0004 + (1 - 0.0843B)(1 - 0.9384B^{12})a_t$$

SARIMA 模型建模估计如表 1 所示。

表 1　SARIMA 模型建模估计

Variable	Coefficient	Std. Error	t - Statistic	Prob
C	-0.000435	0.007573	-0.057497	0.9544
AR(1)	-0.161777	0.282448	-0.572767	0.5691
SAR(12)	-0.357800	0.068687	-5.209169	0.0000
MA(1)	0.084295	0.263616	0.319764	0.7503
SMA(12)	0.938382	0.024858	37.74928	0.0000
R - squared	0.878249	Mean dependent var		-0.001361
Adjusted R - squared	0.855267	S.D. dependent var		0.078397
S.E. of regression	0.046030	Akaike info criterion		-3.240630
Sum squared resid	0.118651	Schwarz criterion		-3.067607
Log likelihood	103.8392	F - statistic		29.51198

FNN 模型中参数的选择仅有隶属度函数个数、隶属度函数类型及最大迭代次数,简单易操作。由于实验中与 BP 神经网络和 SVMs 模型进行比较,下面给出它们的参数,BP 神经网络模型:学习率为 0.2,网络层为 3,隐层节点数为 3,动量因子 0.85;SVMs 模型:$\sigma^2 = 39.1632, c = 6.98, \varepsilon = 0.00198$。其中 σ^2 为高斯核函数的宽度,c 为惩罚因子,ε 为误差值。

4.4 评价标准

SARIMA 模型中笔者采取 1 - 步预测,FNN、SVMs、BP 神经网络模型中采取多步预测。误差函数分别为:平均绝对误差($MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$)、均方误差($MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$)、平均绝对误差百分比($MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{| y_i - \hat{y}_i |}{| y_i |}$)、正则均方误差($NMSE = \frac{1}{\delta^2 * n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \delta^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i$)、$R$ 统计量($R = \frac{\sum_{i=1}^{n}(y_i \times \hat{y}_i)}{\sqrt{\sum_{i=1}^{n} y_i} \times \sqrt{\sum_{i=1}^{n} \hat{y}_i}}$),其中,$y_i$ 和 $\hat{y}_i$ 分别为实际值和预测值。

4.5 实验结果及分析

实验结果及预测效果分别如表 2 和图 2 所示。

表 2　实验误差比较

	MAE	MSE	MAPE	NMSE	R
SARIMA	3.3613e+005	1.1622e+011	13.3179	4.2255	0.9977
SVMs	5.5117e+004	4.4678e+009	2.2155	0.1624	0.9997
BP	5.3326e+004	4.5736e+009	2.0578	0.1663	0.9997
FNN	5.7923e+004	4.6210e+009	2.2981	0.1680	0.9997
FNN-SARIMA1	4.7308e+004	2.9638e+009	1.8411	0.1078	0.9998
FNN-SARIMA2	4.7197e+004	2.9608e+009	1.8370	0.1076	0.9998
FNN-SARIMA3	5.4744e+004	5.2754e+009	2.0989	0.1918	0.9997

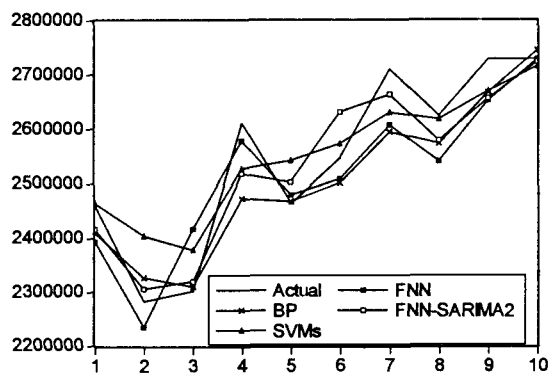

图 2　四种模型预测效果图

从实例结果可以看出,一方面,FNN - SARIMA2 模型从误差评价函数 MAE、MSE、MAPE、NMSE 来看预测效果都优于单独模型和其它混合模型;并且,在参数选择上,FNN 结合模型也有明显优势。BP 网络模型在选取初始值时,易陷入局部极小,且学习率、动量因子、网络结构及学习函数的选择上没有统一的方法;SVMs 模型对参数 σ^2、c、ε 的选择也尚无最优方法[5],如提出的结合遗传算法确定参数,其计算量也是相当大的,并且预测结果对参数的选择是十分灵敏的。而 FNN 结合模型参数选择主要是隶属度个数和类型,简单易操作,且算法稳定,所以结合两者来看,该模型是有效的。

另一方面,从 FNN - SARIMA3 模型预测效果来看,其性能还不如单独使用 FNN,因此在实际应用中,

针对具体时间序列特征，如何建立合理的混合模型应是日后研究的重点。

5 结束语

模糊神经网络将模糊系统的逻辑推理能力和神经网络的自适应能力相结合，在预测、智能控制的方向有广泛的应用。时间序列预测自从 Box 和 Jenkins 提出建模方法后，一直是金融、工业预测中最常用的工具。文中将二者结合，提出一种基于 T－S 模糊神经网络与 SARIMA 结合的混合预测模型。从实验结果可以看出，该模型预测性能优于其它几个模型。另外，在参数确定上比 BP、SVMs 模型也易于操作。

参考文献：

[1] 何书元. 应用时间序列分析[M]. 北京：北京大学出版社，2003.

[2] Law R, Au N. A neural network model to forecast Japanese demand for travel to Hong Kong[J]. Tourism Management, 1999,20:89－97.

[3] Law R. Back－propagation learning in improving the accuracy of neural network－based tourism demand forecasting[J]. Tourism Management,2000,21:331－340.

[4] Lim C, McAleer M. Time Series forecasts of international travel demand for Australia[J]. Tourism Management,2002, 23:389－396.

[5] Chen Kuan－Yu, Wang Cheng－Hua. Support vector regression with genetic algorithms in forecasting tourism demand [J]. Tourism Management,2007,28:215－226.

[6] Zhang G,Hu M Y. Neural Network Forecasting of the British Pound/Us Dollar Exchange Rate[J]. Omega, 1998, 26(4): 495－506.

[7] Tay F E H, Cao Lijuan. Application of support vector machines in financial time series forecasting[J]. Omega, 2001, 29:309－317.

[8] Pai Ping－Feng, Lin Chih－Sheng. Using support vector machines to forecast the production values of the machinery industry in Taiwan[J]. Int J Adv Manuf Technol,2005,27:205－210.

[9] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 2003, 50: 159－175.

[10] Tseng F M, Yu H C, Tzeng G H. Combining neural network model with seasonal time series ARIMA model[J]. Technological Forecasting and Social Change, 2002,69:71－87.

[11] Chen Kuan－Yu, Wang Cheng－Hua. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan[J]. Expert Systems with Applications,2007,32:254－264.

[12] 李国勇. 智能控制及其 MATLAB 实现[M]. 北京：电子工业出版社,2005:267－274.

(上接第 60 页)

Web 研究中的一个难点。提出了一种基于搜索引擎的 Deep Web 数据源发现方法。该方法在实验中取得了较好的效果。由于传统搜索引擎之间在搜索能力上存在着差异，今后，将以其他搜索引擎为实验平台，归纳总结出更合理的查询关键词。

参考文献：

[1] Ghanem T M, Aref W G. Databases Deepen the Web[J]. IEEE Computer,2004,73(1):116－117.

[2] Bergman M K. Deep Web White Paper[EB/OL]. 2004. http://brighplanet.com/technology/deepweb.asp.

[3] Chang K C C, He B, Li C, et al. Structured Databases on the Web: Observations and Implications[J]. SIGMOD Record, 2004,33(3):61－70.

[4] Chang K C C, He B, Zhang Z. Toward Large－Scale Integration: Building a MetaQuerier over Databases on the Web [C]// Proceedings of the Second Conference on Innovative Data Systems Research (CIDR 2005). Asilomar, California: [s. n.],2005:44－55.

[5] Barbosa L, Freire J. Searching for Hidden－Web Databases [C]// The Eighth International Workshop on the Web and Database (WebDB 2005). Baltimore, MD: [s. n.], 2005: 1－6.

[6] Barbosa L, Freire J. An Adaptive Crawler for Locating Hidden－Web Entry Points[C]//In Proceedings of the 16th International World Wide Web Conference(WWW 2007). Banff: [s. n.],2007:441－450.

[7] Lage J P, da Silva A S, Golgher P B, et al. Automatic generation of agents for collecting hidden Web pages for data extraction[J]. Data & Knowledge Engineering, 2004, 49: 177－196.

[8] 刘 伟，孟小峰，孟卫一. Deep Web 数据集成问题研究[R]. [出版地不详]：WAMDM 实验室，2006:18－34.

[9] 高 岭，赵朋朋，崔志明. Deep Web 查询接口的自动判定[J]. 计算机技术与发展，2007,17(5):148－151.

[10] Baeza－Yates R, Hurtado C, Mendoza M. Query recommendation using query logs in search engines[C]//Current Trends in Database Technology. Berlin, Germany: Springer－Verlag, 2004:588－596.