文章编号:0253-2778(2015)10-0795-09

基于时间序列模型的医院门诊量分析与预测

朱顺痣[1],王大寒[1],何亚男[2],王 琰[1]

(1.厦门理工学院计算机与信息工程学院,福建厦门 3601024; 2.厦门大学王亚南经济研究院,福建厦门 361005)

摘要:医院门诊量分析与预测对医疗资源管理和为高质量医疗护理提供决策有重要作用.当前在门诊量分析与预测方面的研究还没引起足够重视,且研究主要集中在门诊量预测的计算方法,缺少全面深入的数据分析和规律挖掘.为此提出构建ARMAX模型、神经网络模型和ARMAX模型与神经网络的混合模型,用来描述医院门诊量的线性和非线性特征.以时间序列模型全面深入地分析厦门市医院门诊量日度数据的规律,研究发现,医院门诊量有显著的上升趋势、周内日效应以及很强的序列自相关性.通过样本外预测比较发现,采用混合模型进行预测取得的预测结果较好,这是由于混合模型能够同时获取门诊量数据的线性部分和非线性部分,数据信息比较完整.

关键词:门诊量预测;时间序列模型;ARMAX;神经网络;混合模型

中图分类号:TP13 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2015.10.001

引用格式:ZHU Shunzhi, WANG Dahan, HE Yanan et al. Hospital outpatient visit analysis and forecasting using time series models[J]. Journal of University of Science and Technology of China, 2015, 45(10): 795-803.

朱顺痣,王大寒,何亚男,等. 基于时间序列模型的医院门诊量分析与预测[J]. 中国科学技术大学学报,2015,45(10):795-803.

Hospital outpatient visit analysis and forecasting using time series models

ZHU Shunzhi[1], WANG Dahan[1], HE Yanan[2], WNAG Yan[1]

(1. *School of Computer and Informatien Engineering, Xinmen Unversity of Technology, Xiamen* 361024, *China*;
2. *The Wan Yanan Institute for Studies in Economics, Xiamen* 361005, *China*)

Abstract: Analysis and forecasting of hospital outpatient visits are important in making correct and feasible decisions for hospital resources management and high quality patient care provision. However, research in outpatient visit analysis and forecasting has not drawn much attentions so far, and current research mainly focuses on the computational methods for forecasting only, lacking in comprehensive analysis, rules finding, and knowledge discovery for hospital outpatient visits. Thus it was propsed to construct autoregressive moving average models (ARMAX), neural network models, and hybrid models integrating ARMAX and NN for outpatient visit analysis and forecasting. By constructing these models, the rules of the daily outpatient visit of the Xiamen city, China were analyzed comprehensively. It was fund that outpatient visit data show a significantly upward time trend, a significant day-of-week effect, and a significant serial autocorrelation. By comparing the forecasting performance of these time series models, it

收稿日期:2015-08-27;**修回日期**:2015-09-29

基金项目:国家自然科学基金(61373147,61305004).

作者简介:朱顺痣(通讯作者),男,1973年生,博士/教授,研究方向:数据挖掘. E-mail: zhusz66@163.com

was fund that the ARMAX+NN hybrid model achieves better performance, which is mainly due to the fact that the hybrid model can capture both linear and nonlinear parts of the outpatient visit data.
Key words: outpatient visits forecasting; time series models; ARMAX; neural network; hybrid models

0 引言

医院门诊量分析与预测对医疗资源管理和为高质量医疗护理时提供决策有着重要作用. 比如,若能精确分析和预测医院门诊量,则医院可根据预测结果来决定未来一段时间(比如未来一周或未来一个月)该安排多少人力、物力、财力等. 对市卫生管理部门来说,医院门诊量的预测可以作为制定城市医疗政策的重要基础和依据.

目前只有少量的文献在医院门诊量分析与预测方面进行研究. 现有文献主要用人工智能模型(如fuzzy时间序列模型)来预测医院门诊量. 在文献[1]中,Cheng等提出了一种基于weighted-transitional matrix的fuzzy时间序列模型来预测医院门诊量. 在文献[2]中,Hadavandia等提出使用一种基于聚类的fuzzy系统来进行医院门诊量预测. 文献[3]中提出了一种基于fuzzy时间序列模型的计算模型用于门诊量预测,该方法引入一个时间离散过程、一个基于频域的分布过程、一个fuzzy分布过程以及一个fuzzy关系最优化过程来构建计算时间序列模型. 这些方法得到了一些预测结果,但它们主要集中在研究预测的计算方法,而不是针对门诊量预测问题本身的规律进行研究. 此外,这些文献主要处理月度医院门诊量数据,忽略了对日度数据的分析和预测,但实际上日度数据包含不同的动态特征,能表征更多的信息.

国内有少量工作进行了医院门诊量预测方面的研究,研究主要是采用ARMA模型进行预测[4-5],但这些方法也主要是对月度数据的预测,对规律的发现和分析相对较少.

本文通过深入分析厦门市医院门诊量日度数据,并用时间序列模型进行分析和预测,试图发现日度数据的规律和特征,而不仅仅是预测日度门诊量. 为分析医院门诊量数据,本文构建了ARMA模型、神经网络模型(NN模型)以及ARMA模型与NN模型的混合模型. 由于混合模型融合了ARMA模型和NN模型的优点(前者可以获取数据的线性特征,后者可以获取非线性特征),因此能够更好地获取数据的特性.

通过构建这些模型,分析了厦门市医院门诊量的日度数据. 通过统计分析,本文发现了医院门诊量的周内日效应,即门诊量周期性变化与星期有关系. 通过预测实验证明,混合模型在预测门诊量时具有一定的优越性.

1 相关工作

医院门诊量的分析和预测实际上是一个时间序列分析问题. 时间序列分析在经济和金融、工业工程和计算机科学领域等诸多应用领域已受到广泛关注和研究[6]. 时间序列分析模型大致可以分为三类:传统的统计预测模型、基于人工智能(AI)的模型和混合模型. 统计预测模型中,一般使用线性模型,比如AR模型(autoregressive models)、MA模型(the moving average models)、ARIMA模型(autoregressive integrated moving average models)等. 其中,ARIMA性能较好,使用也最为广泛. ARIMA模型的主要缺陷是它假定数据是线性的,因此当数据存在非线性特性时,ARIMA模型不能较好地模拟和获取数据的非线性特性.

基于AI的时间序列模型如人工神经网络(artificial neural networks)、模糊逻辑(fuzzy logic)、遗传算法(genetic algorithms)等常被用来进行时间序列分析,这主要是由于这些模型能够处理包含非线性特性数据的复杂问题[3,7-10]. 在这类方法中,由于后向传播算法(back propagation algorithm)[11]的提出,神经网络(neural network, NN)一度成为最为流行的时间序列建模和预测方法[10, 12-13]. NN模型的特点是它能够以任意精度无限逼近任一连续函数,也就是说,理论上NN模型能够模拟数据中包含的多种非线性特性[14]. 由于这个优点,NN模型已广泛用来解决许多实际问题中的时间序列预测问题,如电力负荷预测[15]、经济金融预测[13]、太阳黑子序列预测[16]等. 虽然基于AI的模型可以模拟较为复杂的问题,但这类方法就像一个黑箱,解释起来较为困难.

由于统计模型与NN模型各有优缺点,于是提出了综合这两种模型的混合模型,以得到更好的预测结果[10,12,17]. 使用混合模型的原因和优点是,在实

际应用中，很难确定所处理的数据是线性的还是非线性的，因此也很难确定使用哪种时间序列模型更合适. 但通过融合多种不同的模型，则可以同时获取数据的线性和非线性特性，从而得到更为鲁棒的结果.

虽然已经提出了很多时间序列模型，但用时间序列模型对医院门诊量分析与预测问题进行广泛深入的研究还比较缺乏. 因此，本文以厦门市医院门诊量为例，用可解释的统计模型（比如 ARMAX 模型）分析了医院门诊量数据的特点，并使用 ARMAX＋NN 的混合模型进行预测，获得了较好的结果.

2　时间序列模型

为了分析和预测日度的医院门诊量，本文构建了多个时间序列模型来获取线性和非线性特性，并评估了这些模型的性能.

2.1　时间序列分析模型

2.1.1　Simple multiple linear regression (SMLR)

首先考虑的模型是多线性回归模型（simple multiple linear regression model, SMLR). 令 Y_t 表示 t 时刻的医院门诊总量. 考虑到数据可能存在上涨趋势和周内日效应（周内日效应指的是从周一到周日门诊量存在一定的趋势和规律），因此 SMLR 模型包含一个线性趋势变量 T 和六个虚拟变量 D_j ($j=1,\cdots,6$)，这六个虚拟变量用来描述周内日效应（共七天，所以用六个虚拟变量即可）. 简单来说，用 D_1 来估计星期一的周内日效应，用 D_j ($j=2,\cdots,6$) 分别估计星期二到星期五的周内日效应. 在定义变量的值时，采取方式如下：若在星期一，则 $D_1=1$；若在其他几天，则 $D_1=0$. 其他几个变量用同样的方式定义. SMLR 模型可表示为：

$$(\mathrm{SMLR})Y_t=\alpha_0+\gamma_0 T+\sum_{j=1}^{6}\gamma_j D_j+\varepsilon_t \quad (1)$$

式中，α_0 和 γ_j ($j=0, 1, \cdots, 6$) 分别用来估计医院门诊量的时间趋势和周内日效应. 如果 α_0 显著不为零，则说明医院门诊量有显著趋势. 如果 D_j 显著不为零，则说明有显著的周内日效应.

2.1.2　ARMAX(p, q)模型

ARMA 模型是应用非常广泛的时间序列模型，能够很好地描述相关的线性特性，同时具有很强的预测能力. 它包含了自回归项（AR）和移动平均项（MA）两部分，本质是利用序列自身过去的信息来解释和预测序列当前的变动. 一个病人可能连续几天去医院诊断看病，医院门诊量的时间序列数据可能存在自相关，所以用 ARMA 模型来描述医院门诊量是适用的.

一个 ARMA(p, q)过程可以表示为：

$$Y_t=\mu+\gamma_1 Y_{t-1}+\gamma_2 Y_{t-2}+,\cdots,+\gamma_p Y_{t-p}+\varepsilon_t-\theta_i\varepsilon_{t-1}-,\cdots,-\theta_q\varepsilon_{t-q} \quad (2)$$

式中，Y_t 和 ε_t 分别表示医院门诊量和随机误差，$\gamma_i(i=1,2,\cdots,p)$ 和 $\theta_i(i=1,2,\cdots,q)$ 是模型参数. $\{\varepsilon_t\}$一般假定为独立同分布，且均值为 0，方差为 σ^2. 从式(2)可以看出，ARMA(p,q)模型包含 p 个自回归项和 q 个移动平均项. 滞后阶数 p 和 q 根据 AIC 和 BIC 信息准则确定.

文中，我们在 ARMA 模型中还考虑了线性趋势和周内日效应，本文将这样的 ARMA 模型记为 ARMAX 模型. 医院门诊量的 ARMAX 模型为：

$$(\mathrm{ARMAX})Y_t=\alpha_0+\gamma_0 T+\sum_{j=1}^{6}\gamma_j D_j+\gamma_j\sum_{j=1}^{p}\alpha_j Y_{t-j}+\sum_{j=1}^{q}\beta_j\varepsilon_{t-j}+\varepsilon_t \quad (3)$$

式中，Y_{t-j} 为 j 阶滞后项，即 j 天之前的医院门诊量，ε_t 为随机扰动项.

2.2　神经网络

上述的 SMLR 模型和 ARMAX(p, q)模型可以很好地获取数据的线性特性. 而神经网络则可以用来获取时间序列的非线性特性. 在各种神经网络模型中，单隐层前馈神经网络常被用来进行时间序列建模和预测[18]. 单隐层神经网络包含三层处理节点，其中包含输入层、输出层和一层隐层节点. 输入层的节点个数等于输入的变量个数（或叫特征维数），而输出层则只有 1 个节点（即输出 1 个预测值）. 隐层节点的节点数则决定了模型的复杂度.

输出值 Y_t 与输入(Y_{t-1}, Y_{t-2}, …, Y_{t-p})的关系用 NN 表示为：

$$Y_t=\alpha_0+\sum_{j=1}^{q}\alpha_j h\left(\beta_{0j}+\sum_{i=1}^{p}\beta_{ij}Y_{t-i}\right)+\varepsilon_t \quad (4)$$

式中，$\alpha_j(j=0,1,\cdots,q)$ 和 β_{ij} ($i=0,1,\cdots,p$；$j=1,2,\cdots,q$)是模型参数，p 表示输入层节点个数，q 表示隐层节点个数. 在时间序列分析里，p 也表示滞后量，即用当前数据的前 p 个数据来表示当前数据. h 表示隐层节点的应激函数，比较常用的就是逻辑回归函数：

$$h(x)=\frac{1}{1+\exp(-x)} \quad (5)$$

这里，$h(x)$ 是个非线性函数.

在时间序列预测问题里，预测模型实际上是从过去的观测值(Y_{t-1}, Y_{t-2}, …, Y_{t-p})到当前值 Y_t 的非线性映射：

$$Y_t = f(Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p}, \omega) + \varepsilon_t \quad (6)$$

式中，ω 表示参数集合，函数 f 则由神经网络的结构和参数决定，模型的参数采用经典的 BP 算法学习得到.

2.3 混合模型

在实际应用中，得到的时间序列数据很少是纯线性或者非线性的，因此通过融合不同的模型，可以有效获取数据的不同特性. 混合模型一般由两部分组成：一个线性自回归结构和一个非线性部分，即

$$Y_t = L_t + N_t \quad (7)$$

式中，L_t 表示线性部分，N_t 表示非线性部分. 本文提出采用混合模型来预测医院门诊的日度访问量，其中的线性部分用 ARMAX 模型来表示，非线性部分用单隐层前馈神经网络来表示.

用混合模型预测日度的医院门诊量时的主要过程如下：

(Ⅰ)用 ARMAX 模型进行样本外预测(样本外预测指的是在测试集上进行预测)来提取线性部分. 在预测的同时，计算训练样本的样本内残差值(样本内指的是在训练集上计算残差). 这一步实际上是提取了线性部分，并得到了残差值(实际上是可能的非线性部分).

$$e_t = Y_t - \hat{L}_t \quad (8)$$

(Ⅱ)在上一步中，通过计算残差值，实际上是得到了非线性部分的训练样本，用残差值来训练 NN 模型即可得到非线性模型. 在训练得到 NN 模型后，再进行样本外预测即可提取得到非线性部分. 非线性部分模型表示为：

$$e_t = f(e_{t-1}, e_{t-2}, \cdots, e_{t-n}) + \varepsilon_t \quad (9)$$

这里，f 由神经网络结构和参数决定，ε_t 表示随机误差项.

(Ⅲ)计算得到线性部分和非线性部分后，则最后的预测值为：

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t \quad (10)$$

3 实验结果

本节将用上述时间序列模型对厦门医院门诊量日度数据进行分析和预测. 由于本文针对的是医院门诊量预测问题，且使用了较新的数据，在本文所使用数据上还没有其他文献报道过预测结果. 因此本文暂不考虑与其他文献的方法进行比较，与其他方法进行比较可作为下一步的研究.

3.1 数据描述和分析

实验所采用数据为厦门市 2012～2013 两年的医院门诊量日度数据，共 730 条数据，其中包括整个厦门市和厦门市六个区的门诊量数据. 图 1 所示为厦门市行政地图，包含六个区，分别记为 SM、HL、JM、HC、XAN 和 TAN. 其中 SM 和 HL 为市中心区域(岛内)，另外四个为郊区(岛外).

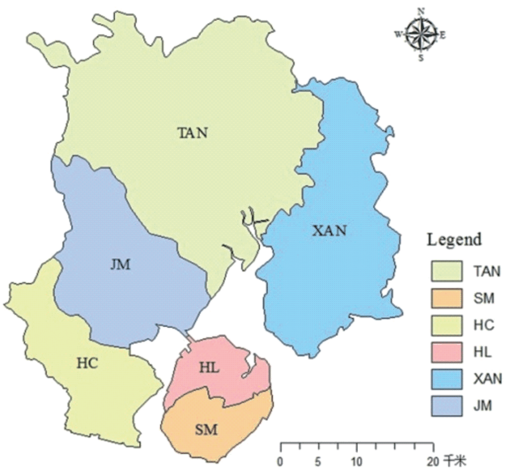

图 1 厦门市行政区域简要地图

Fig. 1 Brief map of the administrative area of Xiamen city

图 2 所示为厦门市医院门诊量日度数据，图 3 所示为厦门市六个区的门诊量日度数据，其中不同的区用不同的灰度和线型表示. 在图 2 和图 3 中，横轴为序列，纵轴为日度门诊量.

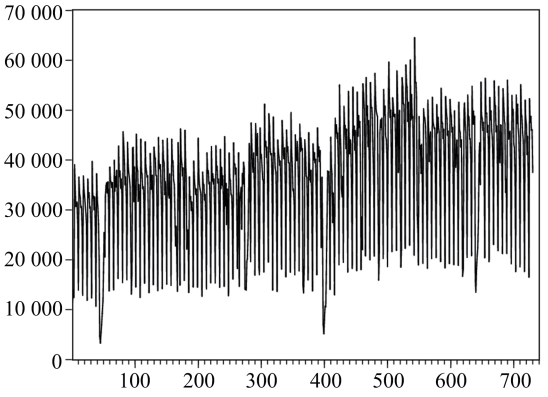

图 2 厦门市 2012 和 2013 两年医院门诊量日度数据

Fig. 2 Data of the daily outpatient visits of the Xiamen city from years 2012 to 2013

上述时间序列数据其统计特性如表 1 所示. 对于任何一个门诊量数据，最小值都远远小于最大值.

这与图 2 所显示的一致，说明每天的门诊人数变化非常大. 表 2 给出了 ADF[19] 和 Phillips-Perron[20] 单位根检验结果，结果表明各门诊变量都是平稳的.

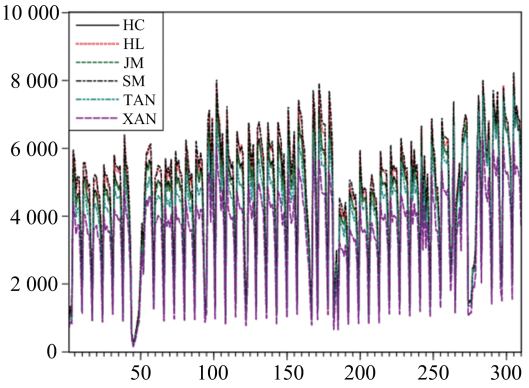

图 3　厦门市六个行政区域 2012 年和 2013 年医院门诊量日度数据

Fig. 3　Data of the daily outpatient visits in six administrative areas of Xiamen city from years 2012 to 2013

表 1　厦门市医院门诊量日度数据基本统计特性(TOPV 表示厦门市整体日度数据)

Tab. 1　Basic statistical characteristics of data of the daily outpatient visits amount of Xiamen hospital

	Mean	Std.	Min	Max	Skewness	Kurtosis
TOPV	36 355	11 799	3 283	64 554	−0.56	2.60
TAN	4 547	2 111	218	1 1807	0.30	3.52
XAN	3 713	1 583	157	7 501	−0.31	2.51
JM	4 654	1 936	212	8 974	−0.38	2.47
HC	4 680	1 899	235	8 517	−0.50	2.45
HL	4 943	2 000	232	8 906	−0.53	2.44
SM	4 993	2 003	250	8 824	−0.59	2.44

表 2　厦门市医院门诊量单位根检验

Tab. 2　Root test of outpatient unit of Xiamen hospital

Variable	ADF	PP
TOPV	−5.45***	−23.27***
TAN	−5.28***	−13.37***
XAN	−5.39***	−13.34***
JM	−5.26***	−13.38***
HC	−5.27***	−13.33***
HL	−5.31***	−13.28***
SM	−5.39*	−13.40***

注：星号***，** 和* 分别表示在 1%，5%，10% 的水平上显著.

3.2　ARMAX 模型构建及分析

厦门市整体数据样本估计 SMLR 模型和 ARMAX 模型的结果见表 3. 从表 3 第 2 列 SMLR 的估计结果可以看出，时间趋势项的估计参数 γ_0 为正，且在 1%的水平上显著不为零. 这表明医院门诊量有显著上升趋势，平均每天增加 22 人. 估计参数 D_j 也在 1%的水平上显著不为零，说明周内日效应明显. 医院门诊量在周一最多(大约为 36 412 人)，其后几天逐渐减少，在周日最低(约 9 245 人). 一周之内不同工作日的门诊人数相差较大. 例如，周二的门诊人数比周一少 2 814 人(日平均门诊量的 8%). 周日的门诊人数比周六低 6 003(占日平均门诊量的 17%). 该线性模型调整 R^2 为 0.67，拟合程度很高，说明时间趋势和周内日效应能够解释医院门诊人数 67%的变动. 不过 Ljung-Box Q 检验统计量在 1%的水平上显著，说明模型残差有明显的序列自相关性，SMLR 模型还不能完全描述门诊量的线性特征.

表 3　SMLR 模型和 ARMAX 模型在整体样本上的估计结果

Tab. 3　Estimation results of the SMLR model and the ARMAX model in the overall sample

	SMLR model	ARMAX model
α_0	9245***	9609***
γ_0	22***	21***
γ_1	27167***	27144***
γ_2	24353***	24313***
γ_3	23708***	23657***
γ_4	21812***	21753***
γ_5	21926***	22036***
γ_6	15248***	15301***
α_1		0.72***
β_1		−0.26***
R_2	0.67	0.77
$Q(30)$	528***	26

表 3 的第三列给出了 ARMAX(1,1)模型的估计结果. Ljung-Box Q 检验统计量(表 3 最下面一行 $Q(30)$)不显著，说明该模型残差不存在残差自相关性，完全能够描述门诊量的线性特征. 样本内拟合指标调整 R^2 为 0.77，说明该模型可以解释门诊量变动的 77%，比 SMLR 模型提高了 10 个百分点. 这说明门诊量的滞后信息有助于对当前的门诊人数的估

计与预测. 滞后一阶的系数估计值为 0.72，说明门　　诊量的日度时间序列有比较高的持久性.

表 4　用厦门市六个区的日度门诊量数据估计参数结果

Tab. 4　Estimation of the parameters with six districts of Xiamen city

	Sm	H1	Jm	Hc	Tan	Xan
α_0	1244***	1212***	966***	1070.44***	778***	763***
γ_0	2.72*	2.87*	3.67***	3.26**	4.20***	2.76***
γ_1	4217***	4142***	3877***	3905***	3656***	3143***
γ_2	4063***	3992***	3725***	3754***	3515***	2981***
γ_3	3798***	3718***	3475***	3483***	3245***	2780***
γ_4	3689***	3623***	3376***	3395***	3158***	2719***
γ_5	3706***	3627***	3360***	3395***	3183***	2727***
γ_6	2390***	2324***	2192***	2036***	1273***	
α_1	0.62	0.63***	0.63***	0.63***	0.64***	0.61***
β_1	−0.19	−0.19	−0.20	−0.19	−0.19	−0.18
R_2	0.67	0.67	0.68	0.67	0.68	0.67
$Q(30)$	24.67	23.93	25.06	23.63	25.43	26.64

此外，我们还构建了厦门六个区的日度门诊量模型. 表 4 列出了在全部样本上的估计结果. ARMAX(1,0)模型足以描述各个区的日度门诊数据. 从表 4 可以看出，时间趋势和周内日效应可以解释大概 55%的变动. 与厦门市整体门诊量类似，区医院门诊量也存在比较强的自相关性.

3.3　厦门市整体日度门诊量预测

本节将用 ARMAX 模型、NN 模型以及 ARMAX+NN 混合模型进行厦门市医院门诊量预测，其中用前 500 条数据学习模型，用后 230 条数据进行样本外测试. 这里采用的 NN 模型为包含 1 个隐层的多层感知机(multi-layer perception，MLP)，主要是用 MLP 进行学习和预测比较方便. 实验中，使用均方根误差 (RMSE)来评价模型的预测性能.

SMLR 模型和 ARMAX 模型的预测结果如图 4(a)和图 4(b)所示. 计算得出的 RMSE 分别为 SMLR：6544 和 ARMAX：5845. 为了进行参考和比较，本文也测试了用随机游走方法进行门诊量预测的性能，用其结果作为基准，其性能为 RMSE 值 15 081. 从这些结果可以看出 ARMAX 模型取得了较好的性能. 对比图 4(a)和 4(b)的预测结果也能看出 ARMAX 的预测结果更优.

在用 NN 模型进行预测时，首先要用训练数据训练 NN 模型. 这一步中，对某一天的数据，若设滞后量记为 DL(即用该数据的前 DL 个数据来预测该数据)，则该数据的前 DL 个数据作为特征，该数据作为待预测的值(或真实值). 这样就可以得到(500-DL)个训练样本. 训练 NN 模型时，采用经典的后向传播算法(BP 算法)来学习参数.

除了 DL 会影响 NN 模型的预测性能外，神经网络的隐层节点数(本文记为 HN)也会影响其性能，因此本文测试了不同的 DL 值和不同的 HN 值对预测性能的影响. 首先，根据经验将 DL 固定为 15，然后用不同的 HN 进行测试，HN 的值从 10～20. 表 5 列出了用不同的隐层节点数(HN)进行预测时的性能. 从表 5 可以看出，当 HN 不同时，预测性能有较大差异，且在大部分情况下比 ARMAX 的性能差. 当 HN 等于 17 时，其性能优于 ARMAX 的性能差. 这些结果表明 NN 模型能取得较好的性能，但较难选择合适的 HN 值.

接下来固定 HN 为 17，用不同的 DL 值进行预测. 表 6 列出了用不同的 DL 值进行预测的结果. 从表 6 可以看出，NN 模型的预测性能对滞后值也是较为敏感的，选择合适的 DL 值也是一个较为困难的问题. 在大部分情况下 NN 模型的性能低于 ARMAX 模型，当 DL 等于 15 时，NN 模型高于 ARMAX 模型的性能.

在测试 DL 值和 HN 值对预测性能的影响时，本文也同时用 ARMAX+NN 的混合模型进行预测. 实验结果分别列于表 5 和表 6 中. 从表 5 和表 6

的结果可以看出，本文提出的 ARMAX＋NN 的混合模型的性能相当稳定，对 DL 值和 HN 值的敏感性明显降低. 同时，混合模型的性能相比于只用 ARMAX 模型或 NN 模型时，都有较大程度的提高，表明了混合模型的有效性和优越性. 这主要是由于混合模型能同时获取数据的线性部分和非线性部分，因此能获得更好的性能. 由于 ARMAX 模型作为混合模型稳定的一部分，因此其性能也更稳定，对参数不敏感.

图 4(c)画出了 NN 模型的性能(HN＝17，DL＝17时的性能，RMSE＝5 722)；图 4(d)画出了 ARMAX＋NN 混合模型的性能(HN＝17，DL＝11 时的性能，RMSE＝5 629). 对比图 4(a)至图 4(d)，可以看出混合模型的预测效果更理想(预测曲线和真实曲线贴合更好).

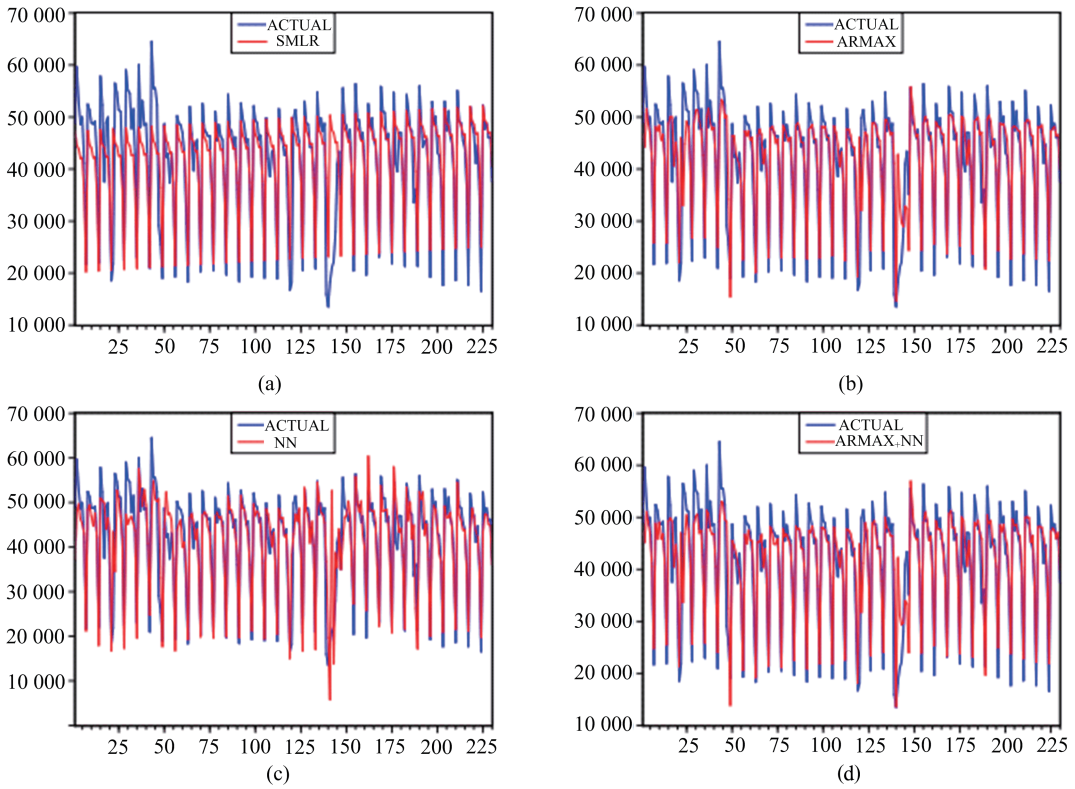

图 4　用时间序列模型预测厦门市医院日度门诊量的结果

Fig. 4　Prediction results of the daily outpatient visits amount Xiamen hospital with time series model

表 5　固定 DL 为 15，用不同的隐层节点数(HN 值)进行预测的性能(用 RMSE 评价)

Tab. 5　The performance of the prediction with different hidden layer node number (HN) by fixed DL＝15

hidden	10	11	12	13	14	15	16	17	18	19	20
NN	6 098	5 895	5 950	6 351	6 014	5 818	6 002	5 722	6 029	5 931	6 048
ARMAX＋NN	5 700	5 637	5 726	5 757	5 659	5 654	5 662	5 705	5 638	5 773	5 778

表 6　固定 HN 为 17，用不同的滞后(DL)进行预测的性能(用 RMSE 评价)

Tab. 6　The performance of the prediction with different hysteresis (DL) by fixed HN＝17

dely	10	11	12	13	14	15	16	17	18	19	20
NN	6 633	6 777	6 921	6 724	5 857	5 722	5 725	6 037	6 150	6 437	6 257
ARMAX＋NN	5 748	5 629	5 715	5 664	5 631	5 705	5 652	5 690	5 664	5 775	5 723

3.4 厦门市六个区域的日度门诊量预测

本节将分别预测厦门市六个区域的日度门诊量. 仍然用 ARMAX 模型、NN 模型和 ARMAX+NN 混合模型进行预测. 在实验中,在用 NN 模型预测时,采用了不同的 DL 值和 HN 值. 最后设定 HL=15和 HN=16 以得到较好的实验结果.

表 7 列出了厦门市六个区的日度门诊量预测结果. 从表 7 可以看出,NN 模型的性能比 ARMAX 模型和混合模型的性能差很多,而混合模型与 ARMAX 的模型的性能接近. 这些结果表明,在预测厦门市六个区的医院门诊量时,NN 模型不能较好地拟合门诊量数据,而在混合模型中,用 NN 模型模拟数据的非线性特征的效果甚微. 原因可能是六个区的日度门诊量数据本身是线性的,能用 ARMAX 模型较好地模拟. 图 5 显示了用混合模型预测这六个区医院门诊量的结果. 由上述可以看出,本文构建的模型的预测结果还是比较理想的.

表 7 厦门市六个区的日度门诊量预测结果

Tab. 7 Prediction results of the daily outpatient amount of six districts in Xiamen city

	TAN	XAN	JM	HC	HL	SM
NN	1 401	1 202	1 585	1 508	1 561	1 617
ARMAX	1 028	872	1 074	1 057	1 119	1 140
ARMAX+NN	1 028	873	1 073	1 056	1 116	1 137

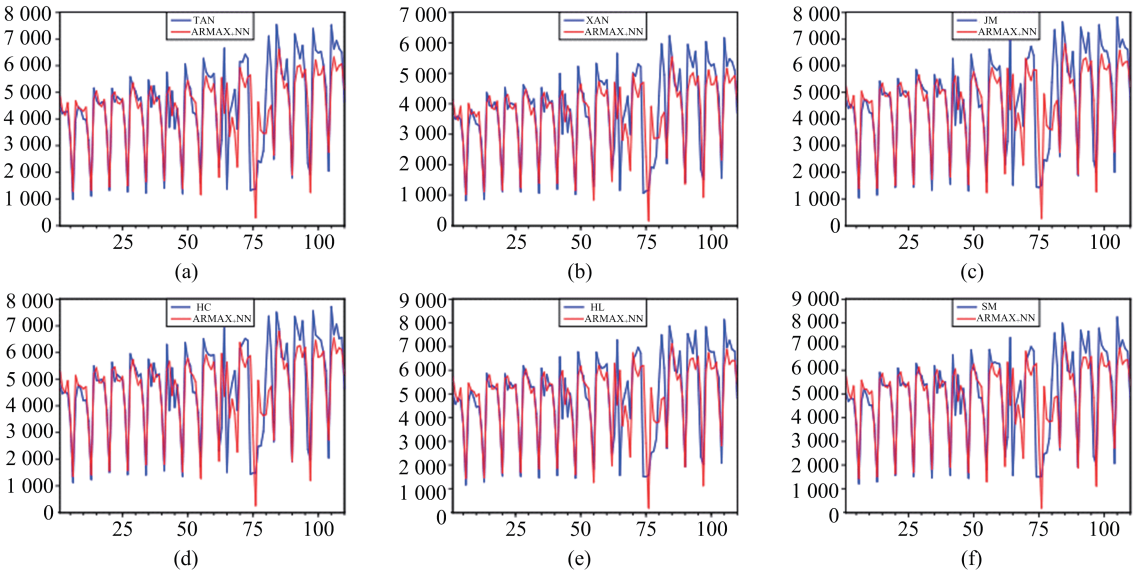

图 5 用 ARMAX+NN 混合模型预测厦门市六个区的医院日度门诊量的结果

Fig. 5 Prediction results of the daily outpatient visits amount with six districts of Xiamen hospital by ARMAX+NN model

3.5 算法复杂度分析

本文使用了神经网络与 ARMAX 模型的混合模型来进行医院门诊量预测. 由于 ARMAX 模型为线性模型,因此其计算复杂度也较低. 神经网络虽为非线性,但本文采用的是只包含 1 个隐层的 MLP 模型(见 2.2 节的神经网络的公式),且使用 BP 算法进行学习,因此能快速地进行学习. 至于预测的计算效率,由于输入变量较少(DL=11 时,输入为 11 个变量,或者 11 维的特征),因此其预测速度也是较快的. 在数据量较大且输入变量巨大(比如几百维甚至上千维)时,若神经网络结构比较复杂(隐层数多且隐层结点个数较多)时,学习和预测速度会受到影响. 这种情况下(比如采用深度神经网络进行大尺度高维数据的分析与预测),可以考虑采用并行学习的方法来提速,这也是当前的研究热点,这里不予赘述.

4 结论

针对医院门诊量分析与预测问题,本文构建了多个时间序列模型,并对厦门市医院门诊量日度数据进行了全面深入的分析和预测. 构建的时间序列包括 ARMAX 模型、NN 模型和 ARMAX+NN 混合模型. 在构建 ARMAX 模型时,考虑了线性趋势变量和六个虚拟变量. 通过分析发现,厦门市医院门诊量日度数据有显著的上升趋势、周内日效应以及很强的序列自相关性. 通过样本外预测实验发现,混合模型由于能同时获取数据的线性特性和非线性特性,因此取得了更为优良的预测结果.

在未来的研究中,我们将考虑在构建时间序列模型时加入更多的变量,比如区域之间的相关性、某一疾病的门诊量分析以及教育水平、家庭住址、病人年龄等因素对门诊量的影响等.此外,建立更加鲁棒的时间序列模型也是将来的研究重点,并与其他时间序列预测模型进行深入的对比和分析.

参考文献(References)

[1] Cheng C H, Wang J W, Li C H. Forecasting the number of outpatient visits using a new fuzzy time series based on weighted-transitional matrix [J]. Expert Systems with Applications, 2008, 34(4): 2568-2575.

[2] Hadavandia E, Shavandi H, Ghanbaric A, et al. Developing a hybrid artificial intelligence model for outpatient visits forecasting in hospitals[J]. Applied Soft Computing, 2012, 12(2): 700-711.

[3] Garg B, Beg M S, Ansari A. A new computational fuzzy time series model to forecast number of outpatient visits [C]// 2012 Annual Meeting of the North American. Berkeley, USA: Fuzzy Information Processing Society, 2012: 1-6.

[4] 李婧, 陈瑛瑛, 霍永胜, 等. 新疆某三级综合医院门诊量预测模型构建及应用[J]. 中国医院统计, 2015, 22(3): 183-185, 189.
Li J, Chen Y Y, Huo Y S, et al. Construction and application of an ARIMA model for predicting the number of outpatient visits in a Xinjiang tertiary general hospital[J]. Chinese Journal of Hospital Statistics, 2015, 22(3): 183-185, 189.

[5] 周忠彬, 吕红梅, 邹郢. ARIMA 干预模型在医院门诊量预测中的应用[J]. 中国医院统计, 2008, 15(2): 110-112.
Zhou Z B, Lu H M, Zou Y. Time series analysis by ARIMA interfering model to forecast amount of outpatient[J]. Chinese Journal of Hospital Statistics, 2008, 15(2): 110-112.

[6] Box G E P, Jenkins G M. Reinsel G C. Time Series Analysis: Forecasting and Control [M]. 3ed Englewood Cliffs, USA: Prentice-Hall, 1994.

[7] Yu L, Lai K K, Wang S Y. Multistage RBF neural network ensemble learning for exchange rates forecasting[J]. Neurocomputing, 2008, 71(16-18): 3295-3302.

[8] Niu D X, Liu D, Wu D S. A soft computing system for day-ahead electricity price forecasting. Applied Soft Computing, 2010, 10(3): 868-875.

[9] Chang J R, Wei L Y, Cheng C H. A hybrid ANFIS model based on AR and volatility for TAIEX forecasting[J]. Applied Soft Computing, 2011, 11(1): 1388-1395.

[10] Xiao Y, Xiao J, Liu J, et al. A multiscale modeling approach incorporating ARIMA and ANNs for financial market volatility forecasting[J]. Journal of Systems Science and Complexity, 2014, 27(1): 225-236.

[11] Rumelhart D, Hinton G, Williams R. Learning representations by back-propagating errors [J]. Nature, 1986, 323: 533-536.

[12] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model [J]. Neurocomputing, 2003, 50(1): 159-175.

[13] Huang W, Lai K, Nakamori Y, et al. Neural networks in finance and economics forecasting[J]. International Journal of Information Technology & Decision Making, 2007, 6(1): 113-140.

[14] Hornik K, Stinnchcombe M, White H. Multilayer feed forward networks are universal approximators [J]. Neural Networks, 1989, 2(5): 359-366.

[15] Hippert H S, Pedreira C E, Souza R C. Neural networks for short-term load forecasting: A review and evaluation[J]. IEEE Transactions on Power Systems, 2001, 16(1): 44-55.

[16] Xie J X, Cheng C T, Chau K W, et al. A hybrid adaptive time-delay neural network model for multi-step-ahead prediction of sunspot activity [J]. International Journal of Environment and Pollution, 2006, 28(3/4): 364-381.

[17] Yu L, Wang S Y, Lai K K. A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates[J]. Computers & Operations Research, 2005, 32(10): 2523-2541.

[18] Zhang G Q, Patuwo B E, Hu M Y. Forecasting with artificial neural networks: The state of the art[J]. International Journal of Forecasting, 1998, 14(1): 35-62.

[19] Dickey D A, Fuller W A. Distribution of the estimators for autoregressive time series with a unit root [J]. Journal of the American Statistical Association, 1979, 74(366): 427-431.

[20] PhillipsP C B, Perron P. Testing for a unit rootin time series regression [J]. Biomètrika, 1986, 75(2): 335-346.