计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100077

基于 X12-LSTM 模型的保费收入预测研究

刁 莉[1] 王 宁[2]
1 中央财经大学保险学院 北京 100081
2 北京交通大学计算机与信息技术学院 北京 100044

摘 要 经济新常态下保费收入预测是学术界和业界共同关注的话题。考虑到保费收入时间序列数据具有强烈的季节性特点,文中构建基于长短期记忆(Long Short-Term Memory,LSTM)神经网络的 X12-LSTM 模型以预测保费收入,并与简单 LSTM 模型、SARIMA 模型和 BP 神经网络进行对比。实验结果表明,X12-LSTM 模型对保费收入的预测最准确且稳定度最好。相比简单 LSTM 模型,X12-LSTM 模型在准确度方面提升 8%,在稳定度方面提升 8%,说明 X12-LSTM 模型是对简单 LSTM 模型的有效改进,更适用于具有季节性特征的数据预测。

关键词:X12 季节调整法;长短期记忆神经网络;保费收入预测;季节性;SARIMA
中图法分类号 TP391

Research on Premium Income Forecast Based on X12-LSTM Model

DIAO Li[1] and WANG Ning[2]
1 School of Insurance,Central University of Finance and Economics,Beijing 100081,China
2 School of Computer and Information Technology,Beijing Jiaotong University,Beijing 100044,China

Abstract Under the new normal of economy,the prediction of premium income is a topic of common concern in academia and industry. Considering the strong seasonality of the time series data of premium income,an X12-LSTM model based on long short-term memory neural network is constructed to predict premium income,and compared with simple LSTM model,SARIMA model and BP neural network in this paper. Experimental results show that X12-LSTM model is the most accurate and stable model to predict premium income. Compared with simple LSTM model,the X12-LSTM model achieves an improvement of 8% in accuracy and 8% in stability,which shows that X12-lstm model is an effective improvement on simple LSTM model and is more suitable for data prediction with seasonality.

Keywords X12 seasonal adjustment,Long Short-Term Memoryneural networks,Premium Forecast,Seasonality,SARIMA

1 引言

保险作为四大金融支柱之一,具有分散风险、损失补偿、资金融通和社会管理的职能,有利于保障人民的基本生活水平,维护社会稳定,推动 GDP 稳定增长。自我国保险业复业以来,保费收入规模一直保持高速增长,从 1980 年的 4.6 亿元人民币,增长到 2018 年的 38016.6 亿元人民币,中国也一跃成为仅次于美国和日本的世界第三大保费收入国家。当前我国经济发展进入新常态阶段,保险产品的供给与需求均发生结构性改变,新形势下对保费收入规模的预测成为学术界和业界共同关注的课题。

目前国内外对于保费收入数据的分析多集中于对影响因素的分析,对保费收入预测的研究较少,方法也比较有限,主要包括差分移动平均自回归模型(Autoregressive Integrated Moving Average model,ARIMA)[1-3]、季节性差分移动平均自回归模型(Seasonal Autoregressive Integrated Moving Average model,SARIMA)[4-6]、BP 神经网络[7]、灰色预测模型[8-9]等。Olszowy(2013)选取波兰 2001 年至 2012 年的保费收入数据,采用 SARIMA,对保费收入进行预测,结果显示保费收入存在明显的季节性趋势,SARIMA 模型预测效果准确。范国斌等(2016)选取 1999 年至 2014 年的月度数据,运用预测性回归方法对我国保费收入预测进行了研究,结果表明消费者预期指数和个人存款等微观层面的经济变量对预测保费收入具有显著作用。舒服华(2017)采用 2006 年至 2016 年上海市保费收入年度数据,运用三次平滑指数法对上海市保费收入进行预测,结果表明平滑指数法结合了全期平均和移动平均法的优点,能较为准确地预测保费收入。陈黎明等(2018)选取 2010 年至 2017 年黑龙江省的保费月度数据,采用 SARIMA 模型对保费总收入进行了预测,结果表明保费收入呈持续上升趋势,季节性波动明显,SARIMA 模型预测准确度在短期内表现良好,时间延长后准确度下降。张鑫等(2018)选取 2009 年至 2016 年东北三省的省级数据,运用滚动建模构建灰色最优化模型,通过基因演算法找出合适的背景值,对传统灰色模型进行了改进,预测效果更好。

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:刁莉(diaolibjwz@163.com)

对于时间序列预测，研究方法可分为传统预测方法和机器学习方法。传统的时间序列预测方法指确定参数模型、求解参数并进行预测[10]。时间序列参数模型包括移动平均模型（Moving Average，MA）、自回归模型（Auto Regressive，AR）、自回归移动平均模型（Auto Regressive Moving Average，ARMA）等。常见的机器学习方法有支持向量机（Support Vector Machine，SVM）、贝叶斯网络（Bayesian Network，BN）、深度学习中的卷积神经网络（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Networks，RNN）。

金融时间序列数据通常具有不平稳和高噪声的特征[11]，变量与变量间的关系也是动态变化的。传统的时间序列模型要求严格、适用性差并且很难实现精准预测。长短期记忆神经网络能够利用自身的网络结构，“记忆”过去的信息，保存过去的状态，以此刻画当前数据与历史数据之间的关系，在挖掘时间序列数据的长期关系中具有明显优势。

本文针对具有强烈季节性特征的时间序列数据进行研究，提出了 X12-LSTM 模型，利用财产保险保费收入月度数据进行试验，并将预测结果与简单 LSTM 模型、SARIMA 模型和 BP 神经网络进行比较。结果表明，本文提出的 X12-LSTM 模型在季节性特征明显的时间序列预测中，准确度与稳定度均最佳。

2　研究理论与方法

2.1　X12 季节调整法

X12 季节调整法将时间序列分解为相互正交的趋势循环要素、季节要素、随机要素和周工作日要素，减少要素间的相互干扰，以便更准确地进行时间序列分析和预测。本文选取的是月度数据，因此忽略周工作日要素。X12 季节调整法包含 4 种模型：加法模型、乘法模型、伪加法模型和对数加法模型分别如式(1)－式(4)所示。

$$Y_t = TC_t + S_t + I_t \tag{1}$$

$$Y_t = TC_t \times S_t \times I_t \tag{2}$$

$$Y_t = TC_t(S_t + I_t - 1) \tag{3}$$

$$\ln Y_t = \ln TC_t + \ln S_t + \ln I_t \tag{4}$$

若各分离要素相互独立则选择加法模型，具体表现为季节因素产生的波动差值保持不变。若波动的差值随时间增大，说明要素间存在相互作用，则选择乘法模型[12]。

以乘法模型为例，X12 季节调整法的第一步，对季节调整进行初始估计。

$$TC_t^{(1)} = \left(\frac{1}{2}Y_{t-6} + Y_{t-5} + \cdots + Y_t + \cdots + Y_{t+5} + \frac{1}{2}Y_{t+6}\right)/12 \tag{5}$$

计算 SI 的初始估计：

$$(SI)_t^{(1)} = \frac{Y_t}{TC_t^{(1)}} \tag{6}$$

利用 3×3 移动平均计算 S 的初始估计：

$$\hat{S}_t^{(1)} = [(SI)_{t-24}^{(1)} + 2(SI)_{t-12}^{(1)} + 3(SI)_t^{(1)} + 2(SI)_{t+12}^{(1)} + (SI)_{t+24}^{(1)}]/9 \tag{7}$$

消除 S 中的残余趋势：

$$S_t^{(1)} = \hat{S}_t^{(1)} - (\hat{S}_{t-6}^{(1)} + 2\hat{S}_{t-5}^{(1)} + \cdots + 2S_{t+5}^{(1)} + \hat{S}_{t+6}^{(1)})/24 \tag{8}$$

季节调整的初始结果为：

$$(TC \cdot I)_t^{(1)} = \frac{Y_t}{S_t^{(1)}} \tag{9}$$

第二步计算暂定的 TC 和最终的 S。

利用 Henderson 移动平均公式计算暂定的 TC：

$$TC_t^{(2)} = \sum_{j=-H}^{H} h_j^{(2H+1)} (TC \cdot I)_{t+j}^{(1)} \tag{10}$$

其中，$h_j^{(2H+1)}$ 表示 Henderson 移动平均系数。

计算暂定的 SI：

$$(SI)_t^{(2)} = \frac{Y_t}{TC_t^{(2)}} \tag{11}$$

利用 3×5 移动平均计算暂定的 S：

$$\hat{S}_t^{(2)} = [(SI)_{t-36}^{(2)} + 2(SI)_{t-24}^{(2)} + 3(SI)_{t-12}^{(2)} + 3(SI)_t^{(2)} + 3(SI)_{t+12}^{(2)} + 2(SI)_{t+24}^{(2)} + (SI)_{t+36}^{(2)}]/15 \tag{12}$$

计算最终的 S：

$$S_t^{(2)} = \hat{S}_t^{(2)} - (\hat{S}_{t-6}^{(2)}) + 2\hat{S}_{t-5}^{(2)} + \cdots + 2\hat{S}_{t+5}^{(2)} + \hat{S}_{t+6}^{(2)})/24 \tag{13}$$

季节调整的第二次估计结果为：

$$(TC \cdot I)_t^{(2)} = \frac{Y_t}{S_t^{(2)}} \tag{14}$$

第三步计算最终的 TC 和最终的 I。

利用 Henderson 移动平均公式计算最终的 TC：

$$TC_t^{(3)} = \sum_{j=-H}^{H} h_j^{(2H+1)} (TC \cdot I)_{t+j}^{(2)} \tag{15}$$

计算最终的 I：

$$I_t^{(3)} = \frac{(TC \cdot I)_t^{(2)}}{TC_t^{(3)}} \tag{16}$$

2.2　LSTM 模型

LSTM 神经网络由 Hochreiter 等在 1997 年提出[13]，是一种改进的循环神经网络。其起源于 RNN 又区别于 RNN，在标准的 RNN 模型中，隐藏层只有一个状态，记为 h，它用于记忆短期的状态。LSTM 神经网络对隐藏层的节点进行改进，在原有状态 h 的基础上增加一个新的状态，记为 c，用于记忆长期的状态，很好地解决了 RNN 的长期依赖问题。并且 Graves 将其成功应用到多个领域，如无约束手写识别[14]、语音识别和手写生成[15]。

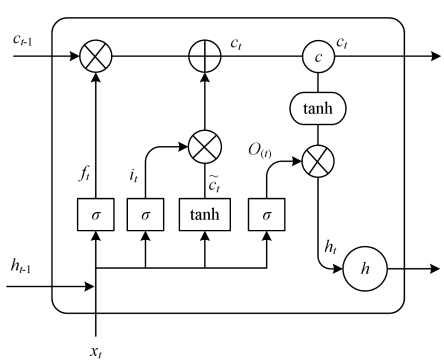

图 1　LSTM 神经网络结构

Fig. 1　LSTM neural network structure

LSTM 神经网络通过使用输入门、遗忘门和输出门这 3 个“门”结构，有选择性地控制信息的输入量，从而实现遗忘或记忆信息的功能，影响神经网络每个时刻的状态，其结构如

图 1 所示。其中的“门”结构使用激活函数控制可以通过此结构的信息量，当激活函数值为 0 时(门关闭)，输入量为 0；当其值为 1 时(门打开)，输入全部信息，类似一扇“门”的作用。

“遗忘门”的作用是让神经网络忘记过去的无用信息，根据当前的输入 x_t 和上一时刻的输出 h_{t-1} 输出一个 0－1 之间的数值 f_t，然后将其赋值给当前状态 c_t，f_t 的计算如下：

$$f_t=\sigma(W_f*[h_{t-1},x_t]+b_f) \tag{17}$$

“输入门”的作用是用于补充最新的信息，填补之前被遗忘的部分，同样根据 x_t 和 h_{t-1} 决定输入值 i_t，然后将其赋值给当前状态 c_t，i_t 的计算如下：

$$i_t=\sigma(W_i*[h_{t-1},x_t]+b_i) \tag{18}$$

当前时刻状态 c_t 的更新过程如下：

$$c_t=f_t*c_{t-1}+i_t*(c_t) \tag{19}$$

$$\tilde{c_t}=\tanh(W_c*[h_{t-1},x_t]+b_c) \tag{20}$$

“输出门”的作用是根据 x_t，h_{t-1} 和当前时刻的 c_t 共同决定输出此时刻状态 h_t，计算公式如下：

$$h_t=o_t*\tanh(c_t) \tag{21}$$

$$o_t=\sigma(W_o*[h_{t-1},x_t]+b_o) \tag{22}$$

因此在每一时刻，LSTM 神经网络的输入有 3 个：当前时刻的输入 x_t、上一时刻的输出 h_{t-1} 和上一时刻的状态 c_{t-1}。相比于标准的 RNN，LSTM 神经网络在每一时刻多一个输入值，即上一时刻用于记录长期信息的状态 c 的值，因此能得到更好的模型性能。

3 X12-LSTM 模型的构建与预测

3.1 数据概况及预处理

本文选取了原中国保监会网站 2004 年 7 月至 2018 年 11 月财产保险保费收入。先对原始累计数据进行差分处理，获得月度保费收入数据。时间序列变化如图 2 所示。保费收入总体呈波动上升趋势，具有明显的季节性。以年为单位观察，总保费收入自 2004 年以来一直保持高速增长，增长率均高于 10%，2006 年至 2010 年呈现爆发式增长，平均增速超过 25%，最高达到 34%。以月为单位观察，每年的 1 月、3 月、6 月、12 月保费收入较高，2 月最低。

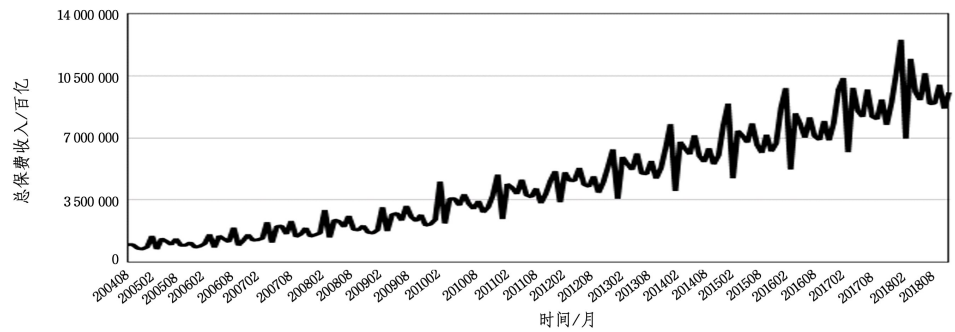

图 2 月度保费收入数据序列

Fig. 2 Monthly premium income data series

3.2 X12-LSTM 模型构建

X12-LSTM 模型预测流程如图 3 所示。

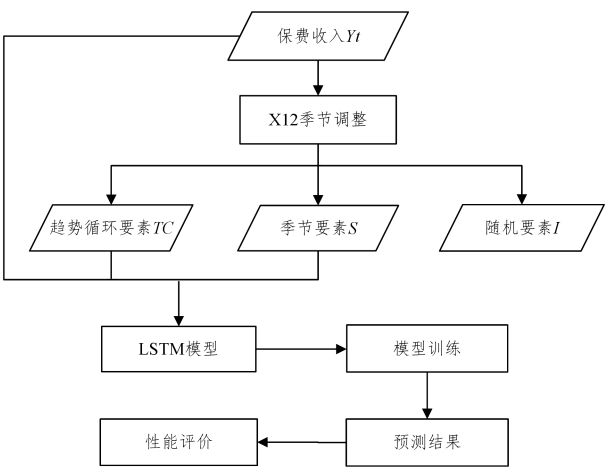

图 3 X12-LSTM 模型预测流程

Fig. 3 X12-lLSTM model prediction process

本文首先利用 X12 季节调整法将保费月收入时间序列 Y_t 分解，获得趋势循环要素 TC、季节要素 S 和随机要素 I。再将 Y_t，TC，S 作为特征输入 LSTM 模型进行训练和测试，排除了随机因素对预测的影响。将 2004 年至 2018 年的数据作为训练集，将 2018 年 1 月至 11 月的数据作为测试集，最终确定 X12-LSTM 模型的参数并输出预测结果。

X12-LSTM 模型需要确定的参数包括网络层数、每层的神经元个数以及训练次数。效果评价指标选择平均绝对百分比误差(Mean Absolute Percentage Error，MAPE)衡量预测的准确度，选择绝对百分比误差标准差(Standard Deviation of Absolute Percentage Error，SDAPE)衡量预测的稳定度。

$$MAPE=\frac{1}{N}*\sum_{i=1}^{N}\left|\frac{T_i-A_i}{A_i}\right| \tag{23}$$

$$SDAPE=\sqrt{\frac{1}{N}*\sum_{i=1}^{N}\left(\left|\frac{T_i-A_i}{A_i}\right|-MAPE\right)^2} \tag{24}$$

其中，T_i 代表保费预测值，A_i 代表保费真实值，N 代表月份。$MAPE$ 越小，说明预测值与真实值的差异越小，预测效果越好。SDAPE 越小，说明预测的稳定度越好。

4 试验结果与模型比较

4.1 X12-LSTM 模型结果

4.1.1 X12 分解结果

运用 RStudio 软件将保费收入时间序列分解，结果如图 4 所示。分解结果表明，保费收入长期呈上升趋势，上升速度逐渐增加。季节波动明显且具有规律，波动周期以年为单位，每年出现多个波峰和波谷，不同月份保费收入差距较大。随机要素序列表现稳定，说明保费收入受政策、自然灾害等不可抗力因素影响较小。

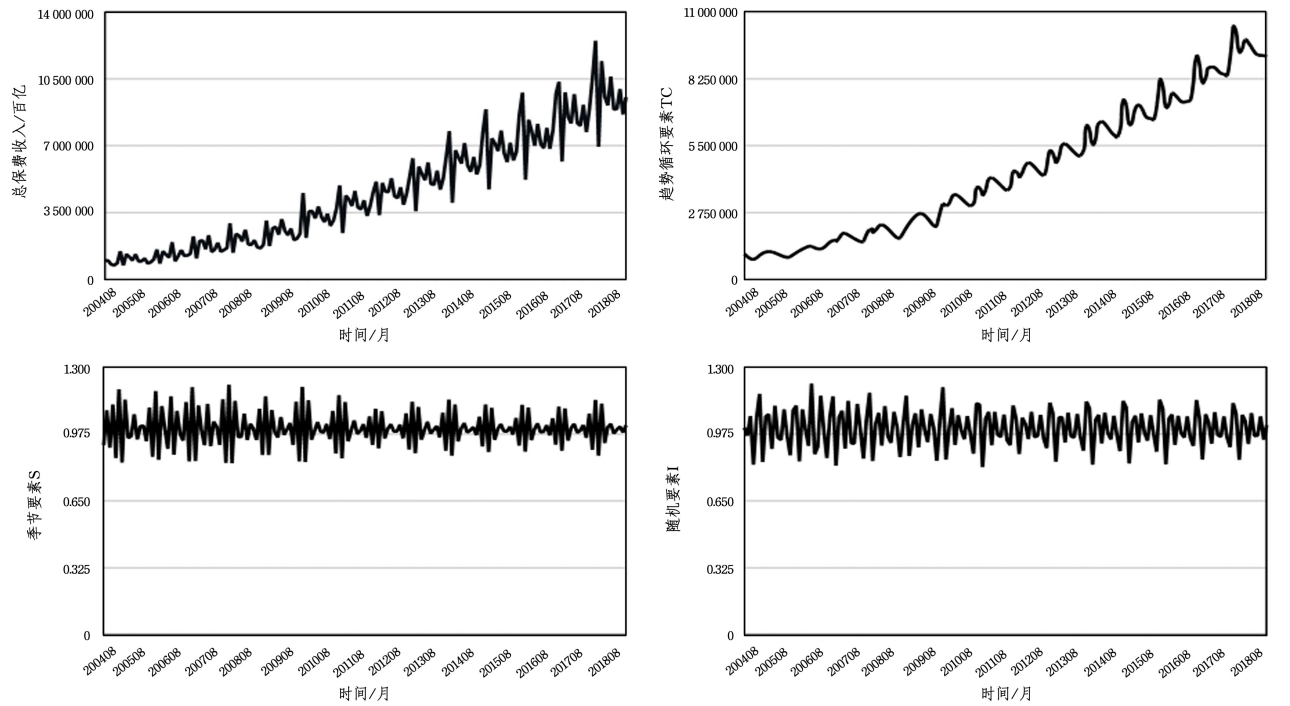

图 4　保费收入分解结果

Fig. 4　Breakdown of premium income

4.1.2　参数选择

第一步，找出最佳的层数与神经元个数。本文固定训练次数为1 000，设定网络层数选取范围为 1－4，神经元个数选取范围为 10，30，50，70，100，共进行 20 组实验，结果如表 1 所列。可以看出，当网络层数为 3，每层神经元个数为 50 时，预测的准确度和稳定度最好。

表 1　不同参数实验结果

Table 1　Experimental results of different parameters

（单位：%）

神经元数		层数			
		1	2	3	4
10	MAPE	5.4761	3.4353	3.8006	3.3354
	SDAPE	5.2677	2.9599	3.9631	3.6528
30	MAPE	3.1711	2.6788	3.6864	3.1223
	SDAPE	3.3394	2.2011	2.6566	2.4991
50	MAPE	4.4216	2.6178	**2.4245**	2.5033
	SDAPE	4.3292	1.7474	**1.3956**	1.8325
70	MAPE	3.9085	3.3869	3.2642	2.9545
	SDAPE	2.7507	3.2836	1.7413	1.7629
100	MAPE	3.2219	3.3395	3.1986	3.2428
	SDAPE	3.1106	2.7114	1.8675	3.7236

固定网络层数为 3，神经元个数为 50，选取最佳的训练次数，实验结果如图 5 所示。可以看出训练 1 000 次效果最好，准确度与稳定度均最佳。

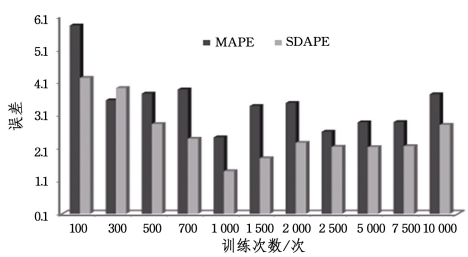

图 5　训练次数实验结果

Fig. 5　Experimental results of different training times

4.1.3　X12-LSTM 模型结果

根据 4.1.2 确定的参数，运用 X12-LSTM 模型预测 2018 年 1 月至 11 月的保费收入。将预测结果与真实值进行对比，预测的 *MAPE* 为 2.42%，*SDAPE* 为 1.4%。结合图 6 可以发现预测值与真实值基本重合，预测误差小且模型稳定。

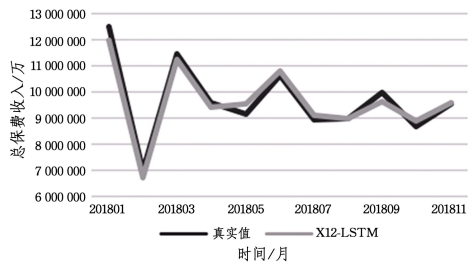

图 6　X12-LSTM 预测效果

Fig. 6　Predicted results of X12-LSTM

4.2　模型比较

保费收入数据常用的预测方法有以 SARIMA 模型为代表的参数模型方法和以 BP 神经网络为代表的机器学习方法。分别运用 X12-LSTM 模型、简单 LSTM 模型、SARIMA 模型和 BP 神经网络预测 2018 年 1 月指 11 月的保费收入，结果如表 2 和图 7 所示。4 种模型中，X12-LSTM 模型的效果最好，SARIMA 次之，BP 神经网络的效果最差。

表 2　模型效果对比

Table 2　Model comparison

（单位：%）

	X12-LSTM	LSTM	SARIMA	BP 神经网络
MAPE	2.4245	10.7868	3.5748	14.8449
SDAPE	1.3956	9.6941	2.4944	19.0310

对于季节特征明显的时间序列数据，能否对其进行准确识别并模拟其季节性规律决定着模型的预测效果。BP 神经网络作为传统的机器学习方法，在进行时间序列数据预测时

容易陷入局部极小值问题，达不到全局最小，因此不能准确学习数据的变化趋势，预测效果较差。

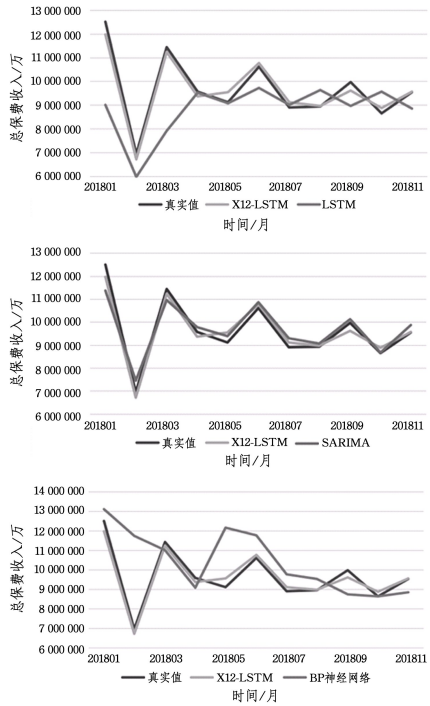

图 7 模型结果对比

Fig. 7 Model results comparison

LSTM 模型是比 BP 神经网络更复杂的深度学习方法，其自身的网络结构决定了对历史数据的记忆能力，并可以有选择地记忆或遗忘特定信息，一定程度上解决了梯度消失问题，提高了预测精度。但 LSTM 模型的预测效果取决于特征选取和参数设定，若不能选出具有代表性的特征作为输入，模型的预测能力将会下降。对于单一的保费收入数据，并不能选出适当的特征用于构建 LSTM 模型，因此需要进行改进。

SARIMA 作为传统的时间序列预测方法，准确度与稳定度均表现良好，但也存在明显的缺点。1)SARIMA 对数据要求严格，数据需要通过平稳性检验，若不满足平稳性检验，应先进行差分处理，转化为平稳序列后再进行后续操作。2)模型参数较多，需根据序列的自相关系数、偏自相关系数、差分步数等统计量确定模型参数，并进行参数的显著性检验。最后还需检查残差序列是否为白噪声序列。

X12-LSTM 模型结合了 X12 季节调整法和 LSTM 模型的优点，通过分解时间序列，既排除了随机因素的干扰，又为 LSTM 模型提供了可靠的特征序列，利用 LSTM 优秀的学习能力，充分学习数据中的总体趋势和季节特性，实现精准预测。相比简单 LSTM 模型，X12-LSTM 模型在准确度方面提升 8.36%，在稳定度方面提升 8.29%，说明 X12-LSTM 模型是对简单 LSTM 模型的有效改进，更适用于具有季节性特征的数据预测。

结束语 分析保费收入的特征和规律，准确预测未来保费收入，对保险业和金融业都具有重要的意义。本文构建了季节性 X12-LSTM 模型，从保费收入序列中分解出趋势循环要素和季节要素输入 LSTM 模型进行训练和预测，并与简单 LSTM 模型、SARIMA 模型、BP 神经网络进行对比实验。实验结果表明，X12-LSTM 模型性能最好，可推广于具有季节性特征的时间序列数据预测。

本文只考虑了一种时间序列分解方法与 LSTM 模型的结合，未来可以引入新的时间序列处理方法，尝试选取效果更好的特征序列，以提高模型预测的准确度。

参 考 文 献

[1] BURCAAM, BATRINCAG. Application of Autoregressive Models for Forecasting Marine Insurance Market[C]// Ovidius University Annals, Economic Sciences Series. 2013.

[2] FUY H. An Application of ARIMA Model to the Prediction of the Property Insurance Premiums in China[J]. Statistical Education, 2010(10): 54-57.

[3] SUN J Y, TIAN L N, LI B Q, et al. Time Series Forecasting of Chinese Insurance Corporation Premium Income[J]. Journal of Gansu Sciences, 2011, 23(4): 143-147.

[4] HE X, ZHANG H M. The Trend Analysis of Life Insurance Premium Income Time Series Based on SARIMA Model in Underdeveloped Areas—A Case Study of Guizhou Province[C]// International Seminar on Education Innovation and Economic Management. 2017.

[5] LIU Y H. Breakpoint Analysis of Life Insurance Premium Time Series Based on SARIMA Model[J]. Collected Essays on Finance and Economics, 2013(2): 73-81.

[6] WIKTOR O. Time Series Forecasting of the Development of the Insurance Industry in Poland[C]// Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultéät. 2013.

[7] TIAN X Y. The Influential Factors Analysis and the Time Series Analysis of the Chinese Insurance Indusry Development [D]. Beijing: Beijing Univerity of Technology, 2017.

[8] ZHANG J L. Premium Proportion Prediction of China Insurance Based on GM(1,1)[J]. Technoeconomics & Management Research, 2010(1): 121-124.

[9] HAN W Q. The application of grey prediction model in the insurance premium prediction of Shandong Province [J]. Journal of Insurance Vocational College, 2017(1): 37-39.

[10] YANG H M, PAN Z S, BAI W. Review of Time Series Prediction Methods[J]. Computer Science, 2019, 46(1): 28-35.

[11] WANG J T. Financial Time Series Prediction Based on LSTM Hybrid Model [D]. Zhengzhou: Zhengzhou University, 2019.

[12] SONG Y H, HU L J. Research on the prediction method of local economic GDP [J]. Economic Research Guide, 2017(32): 4-9.

[13] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[14] GRAVES A, SCHMIDHUBER J. Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks [C]// Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2008: 8-11.

[15] GRAVES A, JAITLYN, MOHAMED A R. Hybrid speech recognition with Deep Bidirectional LSTM [C] // 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2013.

DIAO Li, born in 1992, Ph. D. Her main research interests include risk-management andinsurance.