

基于 BPNN 的混合 ARIMA 时间序列数据预测模型

徐海洋¹ 邓文文¹ 李 彤¹
XU Haiyang DENG Wenwen LI Tong

摘 要

传统的时间序列预测模型具有较好的稳定性和可解释性,但也存在一些问题,一方面对于非线性时间序列的适应能力不足,另一方面对于具有季节性变化的时间序列的适应能力不足,需要通过差分操作消除时间序列的趋势和季节性,但这种方法存在一定的局限性。针对以上问题,文章研究并提出了基于 ARIMA/BPNN 的时间序列数据混合预测模型,对数据进行短期预测,使用中国的进口总值当期值数据集来评估所提出的模型。所提出的模型联合用于线性和非线性模型,旨在捕获时间序列数据中的不同关系模式。混合预测模型能够帮助用户更好地理解市场和业务需求,从而做出更准确的决策,减少决策带来的风险和成本,提高资源利用效率。

关键词

深度学习;卷积神经网络;时间序列数据;短期预测;预测模型

doi: 10.3969/j.issn.1672-9528.2025.01.024

0 引言

时间序列预测在众多应用中发挥着重要作用,包括但不限于工业规划、用水量、医疗领域、汇率和消费者价格指数,主要问题是预测精度不够。本文研究提出了混合 ARIMA/BPNN 模型来满足这一需求,对数据进行短期预测。本研究提出的预测模型用于预测中国的进口总值当期值,评估所提出方法的统计标准度量包括均方误差(MAE)、平均绝对误差(MSE)、均方根误差(RMSE),以及平均绝对百分比误差(MAPE)。

1 混合 ARIMA/BPNN 模型介绍

针对传统时间序列数据预测模型对于非线性时间序列的适应能力不足和对于具有季节性变化的时间序列的适应能力不足的问题,提出的混合 ARIMA/BPNN 模型可以提高预测精度,对数据进行短期预测^[1]。

为了最小化误差,BPNN 将通过找到每个权重的误差导数然后从权重值中减去该值来反向传播。BPNN 模型可以识别某些特定时间序列的潜在现象的因果关系,用于各种目的,包括预测、决策和控制^[2]。

1.1 BPNN 架构模型

层数、每层的神经元数以及神经元之间的加权连接决定了神经网络的拓扑结构。拓扑的确定是针对任何给定问题开发模型的非常重要的步骤之一。

BPNN 网络架构一般由 3 层组成,隐藏层、输入层和输出层,这些层通过承载权重的接口完全链接在一起,每个神经元都与上一层的所有神经元连接,并且具有一个可调整的权重。

BPNN 网络的训练过程分为两个阶段:前向传播和反向传播。在前向传播阶段,输入数据通过输入层向前传播至隐藏层和输出层,每个神经元根据输入数据及其对应的权重计算输出值,并将其传递给下一层神经元;反向传播阶段,通过计算误差和调整权重,反向传播从输出层到输入层更新权重,使网络输出更接近期望的输出^[3]。

本文中使用的 BPNN 网络架构是具有反向传播算法的多层网络。在这个网络中,除了输入和输出单元之外,还可以有一个或多个其他单元,称为隐藏层,这些单元是在训练数据中确定的。与单层神经网络相比,多层神经网络有时需要更复杂和更长的训练过程,但它可以解决更复杂的问题。通过确定输入层神经元、隐藏层和输出层神经元的数量来设计人工神经网络体系结构。

BPNN 对隐藏层神经元个数敏感,因此网络中隐藏层节点的数量与预测精度有很大的相关性。众所周知,选择合适的神经元数量可以大大提高 BPNN 的预测能力。在本文中,使用试错法来确定隐藏层的神经元数。在参数相同的情况下,通过对不同隐藏层神经元数量的大量重复实验,确定最佳隐藏层神经元数量。

本文所提出的 BPNN 网络包括 5 个用于输入层的神经元和一个用于输出层的神经元。图 1 说明了所提出的 BPNN 的架构,其中 A_{k-4} 、 A_{k-3} 、 A_{k-2} 、 A_{k-1} 、 A_k 表示网络的输入, A_{k+1}

1. 中国航空工业集团公司西安航空计算技术研究所
陕西西安 710065

表示网络的输出。通过连续的统计实验，该架构中合适的隐藏层数为 5 个元素。

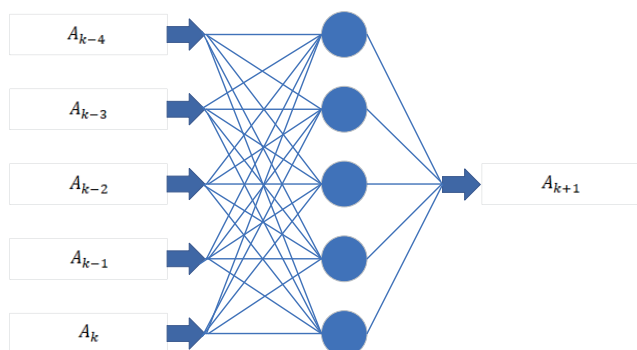


图 1 BPNN 架构

1.2 模型方案

混合模型是两个或多个独立模型的任意组合。混合的目的是提高模型的预测精度。ARIMA 模型处理时间序列的线性特性，而神经网络处理非线性特性。ARIMA/BPNN 混合模型用于寻找一种有效的预测方法。

ARIMA 和 BPNN 模型可以处理线性和非线性域，然而没有一个被认为是适用于所有情况的通用模型。由于很难完全识别真实问题数据的特征，因此具有线性和非线性建模的混合方法是时间序列预测最好的选择^[4]。

在混合模型中，ARIMA 用于通过使用之前的数据来预测目标问题的数据。ARIMA 的预测数据与之前的数据一起用作 BPNN 的输入。ARIMA 不容忍数据中的任何极端值，应在每个新数据点后重新估计。与 ARIMA 相比，BPNN 更能容忍数据中的极值。在混合模型的第一步中，ARIMA 模型用于对时间序列的线性部分建模并创建 ARIMA 预测。在下一步中，ARIMA 的输出被用作人工神经网络的输入，并使用已知的输入和输出训练数据进行训练，以对时间序列的非线性部分建模^[5]。

首先，初始化权重，然后进行反向传播过程算法，该算法由旨在跟踪误差大小的高级计算和反向计算组成，以更新和调整权重。更新权重可以通过两种方式完成，即无动量和有动量。然而，在更新权重时所解释的是在不考虑动量的情况下完成的。因此，在反向传播方法中，必须执行的算法是权重初始化、前馈和反向传播计算以及基于误差限制值或 epoch 限制的数量的停止条件初始化。

混合 ARIMA/BPNN 模型方案包括两个主要阶段。在第一阶段，由于 ARIMA 模型无法捕获非线性数据结构，因此基于时间序列数据作为 ARIMA 模型的输入来分析问题的线性部分。线性模型的残差将包含非线性信息。因此，在第二步中，开发了 BPNN 模型，BPNN 的输入是构建的 ARIMA

模型的产物。ARIMA 模型的这个产品可以包括残差、输出估计或预测。BPNN 模型产生最终的混合模型输出^[6]。

混合 ARIMA/BPNN 模型的开发过程如下：

(1) 预处理数据。预处理数据是通过收集进行数据训练所需的数据来进行的，将原始数据中的错误、缺失或异常值进行清洗和处理，以保证输入 BPNN 的数据质量和准确性。

(2) 数据归一化。由于 BPNN 对输入数据的敏感性较高，因此需要对输入数据进行归一化处理，将数据转化为相同的尺度和范围，避免因数据量级不同而导致预测结果的偏差。

(3) 对 ARIMA 模型进行定阶。定阶是根据已经平稳的实际数据图进行的。如果它不是平稳的，则必须首先使数据平稳。接下来，确定 ARIMA 模型可能的参数组合。从自相关图中确定 MA(q) 的阶数，从偏自相关图中确定 AR(p) 的阶数。

(4) 对 ARIMA 模型进行评估。在这个阶段，对可能的 ARIMA 模型进行了估计和测试，并进行了最佳模型的选择。

(5) 对 ARIMA 模型执行诊断检查。这个阶段是通过确定最佳 ARIMA 模型的方程和预测值来进行的。

(6) 混合模型中使用的函数和参数。激活函数用于限制所需预测输出的值。

(7) 进行数据训练。进行数据训练以通过多次试错来验证 BPNN 网络模型，以便通过确定神经元的数量直到获得低误差值来获得最佳 BPNN 网络。

(8) 测试 BPNN 模型的参数。将训练阶段生成的 BPNN 的权重应用到测试中。

改进的内容：

(1) 使用的激活函数是 logsig-purelin ，均方误差 (MSE) 为 0.1~2.3，而使用的学习率为 0.1，动量常数 (mc) 为 0.95。而为了获得良好的结果，训练迭代进行了 1000 个 epoch。

(2) BPNN 上的训练经过 3 个阶段，它们是高级传播阶段、反向传播阶段以及权重和偏置阶段的变化。训练过程是通过进行重量训练来寻找最佳方程，并且可以试错。将训练数据手动进行，通过逐个插入权重，然后找到最常出现和最稳定或没有显著差异的方程。

1.3 模型缺点

混合 ARIMA/BPNN 模型也存在一些问题，一是容易陷入局部最优解：BPNN 网络的训练过程中容易陷入局部最优解，而非全局最优解，导致模型预测效果不佳。二是需要大量数据：BPNN 网络需要大量的数据来进行训练，否则容易出现过拟合或欠拟合现象。三是对初始权重敏感：BPNN 网络的训练结果会受到初始权重的影响，因此需要一定的经验来确定初始权重值，否则可能导致训练结果不稳定^[7]。

2 预测实验

本节介绍用于进行实验的数据集描述。进行了几次实验和比较分析来评估所提出的预测方法的性能。所得结果及相关讨论如下。

2.1 数据集

数据集是从2003年2月至2020年12月中国的进口总值当期值数据集,如图2所示,时间序列具有不同的统计特征。

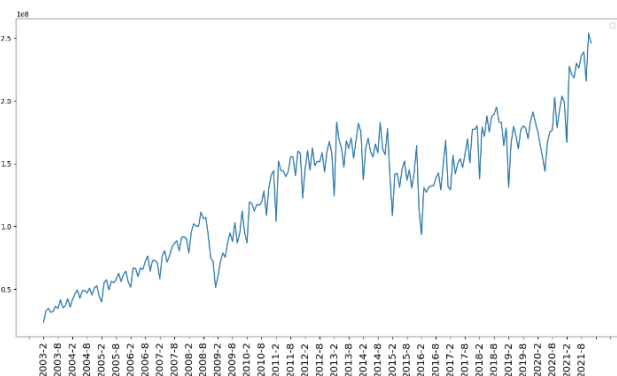


图2 中国进口总值当期值数据集

评估不同模型的预测性能考虑到每个数据集分为训练和测试两个样本。输入和输出数据集是真实值,基本权重是随机选择的。

2.2 数据预处理

本次实验以中国的进口总值作为数据集,然而,在对数据进行统计的时候,可能由于人为因素等问题,导致数据出现缺失值等问题,如果直接使用这些原始数据进行预测,会对模型预测造成影响,导致预测的结果不准确,无法达到实验的目的。所以为了解决上述的问题,就要对原始的数据进行预处理操作。以下部分是对数据进行缺失值处理和归一化处理^[8]。

(1) 缺失值处理

数据中的缺失值是指在收集数据时,某些变量的取值并没有被记录下来,或者记录的数据有误,导致一部分数据缺失。缺失值处理是数据预处理的重要步骤之一,以保证数据的完整性和准确性^[9]。

本文使用了平均值填充的方法来对缺失值进行填充,以缺失值为原点,计算出前后两个月的平均值作为缺失值进行填充操作。

(2) 数据归一化

使用的方法是最小-最大归一化,它的优点是简单易懂、实现方便,且保留原始数据的分布信息。但是它对异常值比较敏感,容易受到数据集中的极端值的影响,因此需要先对数据集进行去除异常值的处理。此外,由于该方法将所有数据映射到[0, 1]之间,如果数据集中的值本来就分布在较小

的范围内,归一化后的值区间可能会过于狭窄,从而可能会对数据分析和建模产生负面影响^[10]。

2.3 使用混合模型预测进口总值数据集

使用混合ARIMA/BPNN模型预测中国进口总值当期值数据集获得的结果如图3所示。



图3 使用混合模型预测中国进口总值数据集结果

其中,X轴表示时间(以月为单位);Y轴表示进口总值当期值(以千美元为单位)。

2.4 预测结果对比

选择均方误差(MAE)、平均绝对误差(MSE)、均方根误差(RMSE),以及平均绝对百分比误差(MAPE)作为预测准确性的度量。中国进口总值当期值数据集的预测结果如表1所示。

混合模型在线性和非线性建模中获得了比ARIMA和BPNN更大的优势,当使用混合模型对中国进口总值时间序列数据进行预测时,模型的准确度比单独应用ARIMA模型提高了47.8%,比单独应用BPNN模型提高了26.8%。与混合模型相比,时间序列模型发生的变化越多,单独使用预测模型的效率就越低。

3 结语

本文对时间序列预测模型进行了研究,并提出了混合ARIMA/BPNN模型来提高预测精度,对数据进行短期预测。首先对现有问题进行了分析,针对传统时间序列数据预测模型对于非线性时间序列的适应能力不足,和对于具有季节性变化的时间序列的适应能力不足的问题提出了混合预测模型,最后使用中国进口总值数据集来评估所提出的模型。对于每个模型,根据包括MAE、MSE、RMSE和MAPE在内的统计标准度量给出和分析结果。结果表明,混合预测模型降低了随机性、影响时间序列的变化和数据非线性。真实数据集的结果证实了ARIMA/BPNN模型通过提高预测准确性和降低整体误差,优于本研究中引入单一模型的预测准确度。

表 1 中国进口总值数据集的预测比较

时间	实际值	BPNN 架构模型	ARIMA 模型	混合 ARIMA/BPNN 模型
2021-1	198 640 212	205 645 116	203 324 249	201 486 346
2021-2	166 975 331	184 556 131	181 641 328	178 912 654
2021-3	227 336 197	249 874 651	241 876 123	238 714 872
2021-4	221 063 587	242 138 561	239 745 231	235 792 163
2021-5	218 382 769	238 974 651	235 764 125	231 748 961
2021-6	229 890 607	258 756 234	252 789 641	247 939 677
2021-7	226 073 489	264 563 123	258 712 537	247 912 533
2021-8	235 984 248	274 531 233	267 861 234	257 912 379
2021-9	238 979 506	275 461 234	268 716 185	259 876 451
2021-10	215 681 597	257 896 134	251 789 615	245 498 314
2021-11	253 813 881	277 861 253	272 753 187	268 465 132
2021-12	246 035 373	267 861 231	263 748 614	258 165 341
2022-1	241 914 340	257 896 412	253 845 324	249 684 653
2022-2	186 832 863	237 896 123	226 781 346	219 856 431
2022-3	228 704 260	257 845 321	247 812 364	234 984 564
2022-4	222 500 588	267 811 235	257 961 462	248 947 487
2022-5	229 490 725	257 896 121	247 891 263	242 987 464
2022-6	233 322 854	267 123 456	257 891 236	252 987 898
2022-7	231 696 993	271 235 463	259 678 923	248 879 478
2022-8	235 527 165	267 123 612	254 213 678	245 489 183
2022-9	238 011 751	257 861 261	251 781 267	248 413 659
2022-10	213 217 901	247 861 237	237 892 179	232 749 842
2022-11	226 252 602	257 489 126	247 892 137	234 894 655
2022-12	228 066 200	251 786 123	242 897 896	238 946 514
MAE		22 119 421	29 314 583	15 702 150
MSE		561 989 070 740 955	96 296 337 5046 195	299 277 628 046 160
RMSE		23 706 308	31 031 651	17 299 642
MAPE		0.09	0.13	0.07

参考文献:

[1]SOARES E , COSTA J P , COSTA B ,et al.Ensemble of evolving data clouds and fuzzy models for weather time series prediction[J].Applied soft computing, 2018, 64(3):445-453.

[2]YIN L F , XIE J X . Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems[J]. Applied energy, 2021, 283(2):116328.

[3]SHAHI S , FENTON F H , CHERRY E M .Prediction of chaotic time series using recurrent neural networks and reservoir computing techniques: A comparative study[J].Machine learning with applications, 2022, 8(6):100300.

[4]LONG L F , LIU Q, PENG H ,et al.A time series forecasting approach based on nonlinear spiking neural systems[J].International Journal of Neural Systems, 2022,32(8):2250020.

[5]LIU F , CHEN L Z , ZHENG Y F ,et al.A prediction method with data leakage suppression for time series[J].Electronics, 2022,11(2):3701.

[6]罗志丹 . 数据驱动金融时间序列混合预测模型研究与实证分析 [D]. 北京 : 对外经济贸易大学 ,2020.

[7]杜懿 , 麻荣永 . 不同改进的 ARIMA 模型在水文时间序列预测中的应用 [J]. 水力发电 , 2018, 44(4):12-14+28.

[8]韩苑 .ARIMA 乘积季节模型及 Prophet 预测模型在交通运输客运量预测中的应用和比较 [J]. 环球市场 , 2019(6): 315-317

[9]李政毓 . 基于 EEMD-ARIMA-LSTM 组合模型对原油期货价格预测 [D]. 济南 : 山东大学 ,2021.

[10]马浩 . 基于粒子群优化算法的 ARMA-SVR 短期用电负荷预测与应用 [D]. 佛山 : 佛山科学技术学院 ,2021.

【作者简介】

徐海洋（1998—），男，陕西西安人，硕士，助理工程师，研究方向：深度学习、神经网络。

邓文文（1998—），男，四川资阳人，硕士，助理工程师，研究方向：深度学习、神经网络。

李彤（1996—），女，陕西咸阳人，硕士，工程师，研究方向：深度学习、神经网络。

（收稿日期：2024-09-24）