**Computer Science and Application** 计算机科学与应用, 2023, 13(12), 2172-2179
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312217

Hans汉斯

# 考虑节假日的ARIMA模型在酸奶销量预测中的应用

**刘丽佳，李晓雪，王海滨，赵林红**

河北地质大学信息工程学院，河北 石家庄

收稿日期：2023年11月10日；录用日期：2023年12月8日；发布日期：2023年12月18日

## 摘　要

**商品的销量往往受多种因素的影响，进行销量预测时也需要将这些因素考虑进去。传统的ARIMA模型进行销量预测时只考虑了时序序列的线性因素，而忽略了一些非线性因素。本文以冷藏酸奶的日销量为研究对象，探究了加入节假日等非线性因素的ARIMA模型预测的精确度。具体地，本文首先提取了2021年冷藏酸奶日销量数据，并运用BIC值对ARIMA模型的参数进行选择，然后对加入节假日等非线性因素的ARIMA模型、普通ARIMA模型和目前流行的ARIMA + SVM组合模型进行了实验评估和对比。实验结果表明，相较于普通的ARIMA模型和目前流行的组合模型，加入非线性因素的ARIMA模型预测结果更加准确。该研究成果对于冷藏酸奶销量预测的实际应用和相关研究具有重要的参考意义。**

## 关键词

**ARIMA模型，节假日，销量预测，BIC值**

# Application of ARIMA Model Considering Holidays in Sales Forecasting

**Lijia Liu, Xiaoxue Li, Haibin Wang, Linhong Zhao**

College of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: Nov. 10th, 2023; accepted: Dec. 8th, 2023; published: Dec. 18th, 2023

## Abstract

**The sales volume of commodities is often affected by a variety of factors, and these factors need to be taken into account when conducting sales volume forecasting. The traditional ARIMA model for sales forecasting only considers the linear factors of time series and ignores some nonlinear fac-**

文章引用: 刘丽佳, 李晓雪, 王海滨, 赵林红. 考虑节假日的 ARIMA 模型在酸奶销量预测中的应用[J]. 计算机科学与应用, 2023, 13(12): 2172-2179. DOI: 10.12677/csa.2023.1312217

**tors. In this paper, we take the daily sales volume of refrigerated yogurt as the research object, and explore the accuracy of the ARIMA model prediction by adding non-linear factors such as holidays. Specifically, this paper firstly extracts the data of daily sales of refrigerated yogurt in 2021 and applies the BIC value to select the parameters of the ARIMA model, and then experimentally evaluates and compares the ARIMA model with nonlinear factors such as holidays, the ordinary ARIMA model, and the combined ARIMA + SVM model. The experimental results show that compared with the ordinary ARIMA model and the current popular combination model, the ARIMA model with nonlinear factors is more accurate in predicting results. The research results are of great reference significance for the practical application of refrigerated yogurt sales forecasting and related research.**

## Keywords

**ARIMA Model, Holiday, Sales Forecast, BIC Value**



## 1. 引言

销量预测一直是企业决策和营销策略制定的重要参考。时间序列模型是一种常用的销量预测方法[1]，而 ARIMA 模型则是其中的一种典型模型[2]。ARIMA 模型具有建模简单、预测精度高等特点，但是在时间序列预测中，ARIMA 模型仅考虑时间序列自身的线性的因素，而未考虑非线性因素的影响，导致在有些时候预测的准确度并不高。在现代消费市场中，消费者的购买行为受到众多因素的影响，比如节假日等因素的影响，将这些非线性因素纳入销量预测模型中，可以更准确地预测销量波动趋势并制定相应的经营策略[3]。因此，为了提高预测的准确度，本文尝试在 ARIMA 模型预测时，将节假日等非线性因素加入到模型中。本文以冷藏酸奶的销量预测为实例，提取冷藏酸奶的销量时序数据中的节假日因素，运用 BIC 值对 ARIMA 模型的参数进行选择，然后分别对加入非线性因素的 ARIMA 模型、传统的 ARIMA 模型[4]、目前较流行的 ARIMA + SVM 组合模型[5]进行训练和预测实验，并将预测结果的准确度进行对比评估。结果表明，在 ARIMA 模型中加入非线性因素进行预测，可以得到更加准确的结果。

## 2. 文献综述

销量预测是企业经营管理中非常重要的一个方面，准确的销量预测可以帮助企业进行合理的生产计划以及进货策略，从而提高企业的经济效益。基于时间序列的销量预测是一种重要的数据分析技术，它是指根据历史数据来估计未来一段时间内的销售量。近年来，很多学者对此进行了研究，主要分为以下三种销量预测的方法。

### 2.1. 基于统计学方法的销量预测

传统的销量预测方法主要是基于统计学模型的方法，很多学者对此进行了研究，例如，范波等人(2021)根据交换机的历史销售量的时间序列，提出了基于 ARIMA 模型的销量预测模型，并与其他算法进行比较，结果表明该模型具有一定的适用性[6]。李晓彤等人(2022)采用运用季节性预测法对新能源汽车的销量进行预测，得出温斯特加性模型的预测精度最高，可以为企业后续的生产运作提供参考[7]。师丹平等

人(2023)选用全国新能源乘用车月度销量数据，分别建立 SARIMA 模型、Holt-Winters 加法模型和 Holt-Winters 乘法模型，通过对比模型效果评价指标发现，SARIMA 模型的预测效果更好[8]。

传统的统计学方法比较成熟，可以勾画出历史趋势，模型相对简单，预测结果可解释性强；但是对于复杂和非线性的时间序列可能不太适用，不能准确地预测各种复杂因素造成的影响，实际中估计模型往往会受到误差干扰而影响准确性。

### 2.2. 基于机器学习的销量预测

随着数据科学和机器学习的发展，利用更加高级的技术来进行销售预测已经成为了一种新趋势。许多学者基于机器学习的方法进行了预测研究。例如，Liu 等人(2020)基于复杂的销售数据集，提出使用最小描述长度神经网络(MDL-NN)搜索最优模型大小来预测销售值，实验结果表明在所有测试数据集上 MDL-NN 方法比其他常用销售预测方法的误差要小[9]。蒋翠清等人(2021)提出基于消费者关注度构建 Attention-LSTM 模型，进而预测汽车销量，实验表明，引入消费者关注度后的 Attention-LSTM 模型能够有效预测汽车销量的动态变化趋势[10]。徐英卓等人(2023)以影响游戏销量的特征数据为样本，建立基于梯度提升决策树(Gradient Boosting Decision Tree, GBDT)算法的游戏销量预测模型，并将 GBDT 模型预测结果与其他模型进行对比分析。结果表明，游戏销量预测模型具有较高的拟合优度，预测效果更好，且在预测阶段的计算速度快[11]。

机器学习方法对高维数据具有更好的适应性，同时能够自适应地学习新的特征，具有较强的功能，适用性强，预测精度也更高。但是机器学习方法需要大量的数据支持，训练模型需要耗费大量的计算资源，在小样本的情况下，可能不具有很好的预测效果。

### 2.3. 结合多种方法的销量预测

为了提高预测性能，许多研究者提出了基于时间序列的组合预测方法。袁远等人(2021)提出了结合 ARIMA 与 RF 模型的销售预测模型对快销品牌的销量进行预测，针对 ARIMA 模型无法更好地提炼非线性信息的问题，利用随机森林算法对非线性数据特征的学习能力，优化 ARIMA 模型预测残差，构建了实验效果更好、预测精度更高的实验模型[12]。呼雄伟(2022)为了更好的预测液化气钢瓶销量，分别赋予提高精度后的 ARIMA 模型和含有天气因素等多变量 LSTM 模型各自权重，从而构建了 ARIMA-LSTM 组合模型，并将组合模型和单一模型的预测结果进行对比，发现组合模型的误差值均小于其他单一模型，从而验证了组合模型较单一模型的可行性和优越性[13]。

组合销量预测模型将多个单一预测模型的结果进行加权或其他方式的整合，以得到一个综合的预测结果，能够更准确地预测各种销售情况。但是需要在不同方法之间进行适当的平衡，同时需要对不同方法进行清晰的解释，以节省时间和资源，提高预测准确性。

总的来说，销量预测方面现有的研究工作较多，从传统统计学方法到机器学习，再到结合多种方法，每种方法都有其适用的情境，因此需要结合实际需求来选择合适的方法。

## 3. 模型介绍

ARIMA (Autoregressive Integrated Moving averages)模型是一种常用的时间序列模型，它由自回归(AR)、差分(I)和移动平均(MA)三部分组成[14]。AR(autoregressive)部分表示当前时间点的序列值是过去若干时刻的序列值的线性组合，因为不依赖于别的解释变量，只依赖于自己过去的历史值，故称为自回归，如果依赖过去最近的 $p$ 个历史值，称阶数为 $p$，记为 AR($p$)模型，公式为：

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \mu,$$

MA (moving average)部分表示当前时间点的序列值是过去若干时刻的随机误差的线性组合，如果序列依赖过去最近的 $q$ 个历史预测误差值，称阶数为 $q$，记为 MA($q$)模型，公式为：

$$X_t = \beta_1\varepsilon_{t-1} + \beta_2\varepsilon_{t-2} + \cdots + \beta_q\varepsilon_{t-q} + \varepsilon_t + \mu,$$

而差分 I (integrated)则指通过对序列进行差分，使得序列变得平稳。因为时间序列分析要求平稳性，不平稳的序列需要通过一定手段转化为平稳序列，一般采用的手段是差分；$t$ 时刻的值减去 $t-1$ 时刻的值，得到新的时间序列称为 1 阶差分序列；1 阶差分序列的 1 阶差分序列称为 2 阶差分序列，以此类推。公式如下所示：

$$\Delta X_t = X_t - X_{t-1} = X_t - LX_t = (1-L)X_t$$

$$\Delta^2 X_t = \Delta X_t - \Delta X_{t-1} = (1-L)X_t - (1-L)X_{t-1} = (1-L)^2 X_t$$

$$\Delta^d X_t = (1-L)^d X_t$$

$d$ 表示差分的阶数。根据 ARIMA 模型中各个分量的参数选择和阶数选择，可以构建不同形式的 ARIMA 模型，例如 ARIMA($p$, $d$, $q$)、季节性 ARIMA 等。

ARIMA 模型的建立过程通常包括模型检验、参数选择和模型评估等步骤。ARIMA 模型假设时间序列具有平稳性，即序列的均值和方差不随时间变化，对时间序列的差分过程来实现序列的平稳化，进而建立模型进行预测。在模型检验过程中，可以通过查看自相关图(ACF 图)和偏自相关图(PACF 图)来判断时间序列是否符合 ARIMA 模型的假设。在参数选择过程中，可以使用 BIC、AIC 等方法进行选择。在模型评估过程中，可以通过计算预测误差、检查模型的拟合程度以及残差的自相关性和白噪声性质等指标来对所得到的 ARIMA 模型进行评估。

ARIMA 模型在实际应用中广泛使用，例如在金融领域的股票市场预测、电力负荷预测、气象灾害预测、商业销售预测等领域，ARIMA 模型都有其独特的应用价值[15]。

## 4. 实证分析

### 4.1. 数据准备

冷藏酸奶销量数据来自某头部零售企业，该企业拥有大量顾客消费的订单信息数据，我们通过使用 SQL 语句对数据表进行筛选，选择商品品类为冷藏酸奶的所有记录，并以购买时间按日进行聚合，累计商品购买数量，最终得出冷藏酸奶每日的销售量。

本文选了取 2021 年 1 月 1 日至 2021 年 12 月 31 日的冷藏酸奶每日的销售量数据，尽管从销售企业中获取的销量数据已经是经过清洗的，但由于一些原因，存在销售数据的缺失。为了解决这个问题，我们使用了线性插值法来补齐缺失数据，具体方法是使用缺失日期前后相邻日期的销量值的平均值来插补缺失值[16]。这种方法比删除不完整的样本或变量要丢失的信息更少。补全缺失后，将数据按照时间顺序排序，并将日期字段转化为时间序列索引。我们绘制出了 2021 年该超市冷藏酸奶每日销量的折线图，如图 1 所示。

同时，为了获取非线性因素，更好地考虑到节假日因素的影响，我们获取了 2021 年每天对应的节假日特征，包括工作日、周末、传统节假日和指定节日(如 618、双十一等)，并将工作日标记为 0，除工作日以外的节假日标记为 1，形成特征变量。最后，我们使用 2021 年 1 月 1 日至 2021 年 11 月 30 日的数据进行模型训练，将 2021 年 12 月 1 日至 2021 年 12 月 31 日的数据作为测试集。这一数据处理和划分方式有助于使我们的模型更准确地捕捉到节假日因素对销售量的影响，从而提高销量预测的准确性。

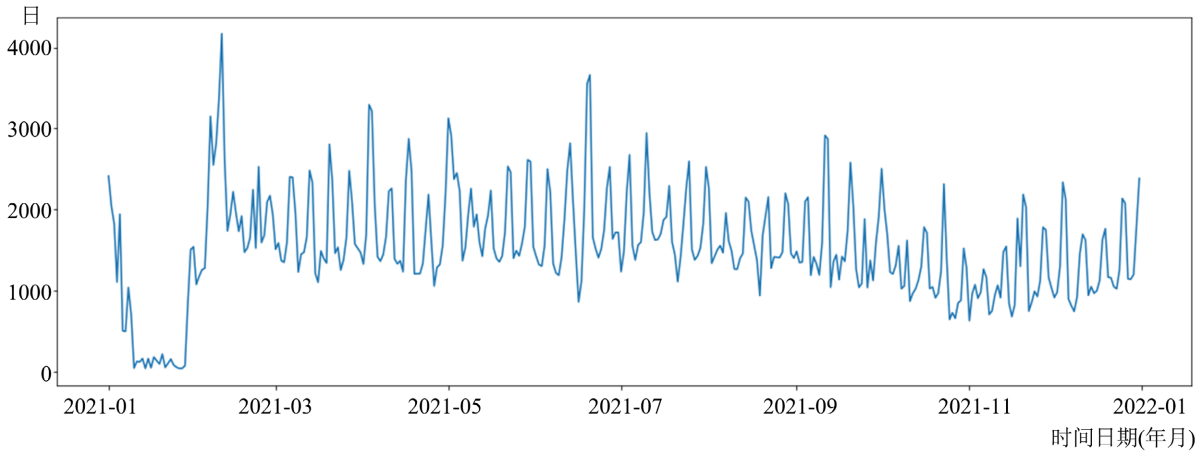


**Figure 1.** Yogurt sales statistics chart
**图 1.** 酸奶销量统计图

## 4.2. 平稳性检验

在使用 ARIMA 模型之前，要保证时间序列数据是稳定的，如果不稳定，则需要对时序数据进行差分，使其平稳。所以数据平稳性检验是一个非常重要的步骤，可以为 ARIMA 模型的参数估计打下基础。这里，我们使用假设检验方法中最常用的 ADF 检验(Augmented Dickey-Fuller Testing)，来检验酸奶日销量时间序列数据是否平稳。首先假设序列是非平稳的，然后通过检验序列是否存在单位根来判断序列是否是平稳的，对于平稳序列，就需要在给定的置信水平上显著，拒绝原假设。对冷藏酸奶日销量的时间序列进行 ADF 检验，结果如表 1 所示，P 值为 0.000884 < 0.05，时间序列是平稳的，无需进行平稳化处理。

**Table 1.** Results of ADF test
**表 1.** ADF 检验结果

| 项目 | t 统计量 | P 值 |
|---|---|---|
| ADF 检验统计量 | −4.123890 | 0.000884 |
| 1%显著性水平 | −3.449227 | |
| 5%显著性水平 | −2.869857 | |
| 10%显著性水平 | −2.571201 | |

## 4.3. ARIMA 模型参数选择

BIC (Bayesian Information Criterion)是一种选择模型的常用准则，它可以平衡模型的拟合程度和复杂度，帮助选择最优模型。我们根据贝叶斯信息准则(BIC)，选择最优的 ARIMA 模型参数。给定一组 ARIMA 参数，用这组参数拟合时序数据，并计算 BIC 值，选择 BIC 值最小的那组参数作为 ARIMA 模型的最优参数，经过实验，最小的 BIC 值为 4457.34 ，最好模型参数为 $p = 3$，$q = 4$。

## 4.4. 模型预测结果及分析

在选择好最优的模型参数之后，我们对模型进行测试和评估。根据分离出来的冷藏酸奶的日销量测试集，我们使用普通的 ARIMA 模型、加入非线性因素的 ARIMA 模型以及 ARIMA + SVM 组合模型进行分别进行预测实验。三种模型对冷藏酸奶日销量的预测值和实际值的对比结果如图 2，图 3，图 4 所示。

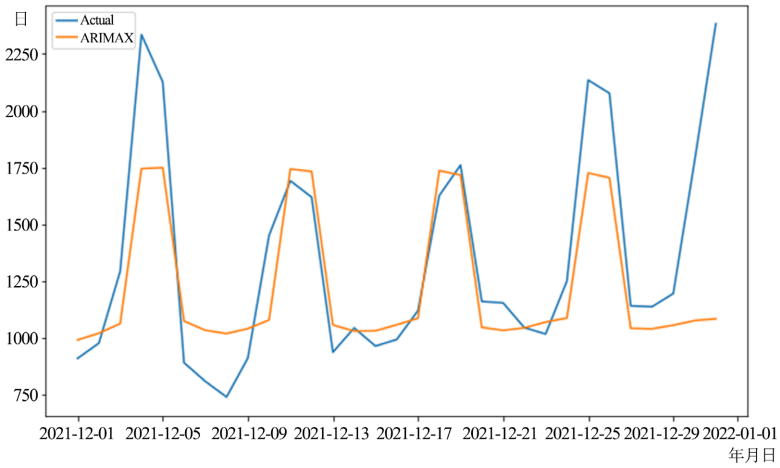


**Figure 2.** ARIMA model prediction results with the inclusion of nonlinear factors
**图 2.** 加入非线性因素的 ARIMA 模型预测结果

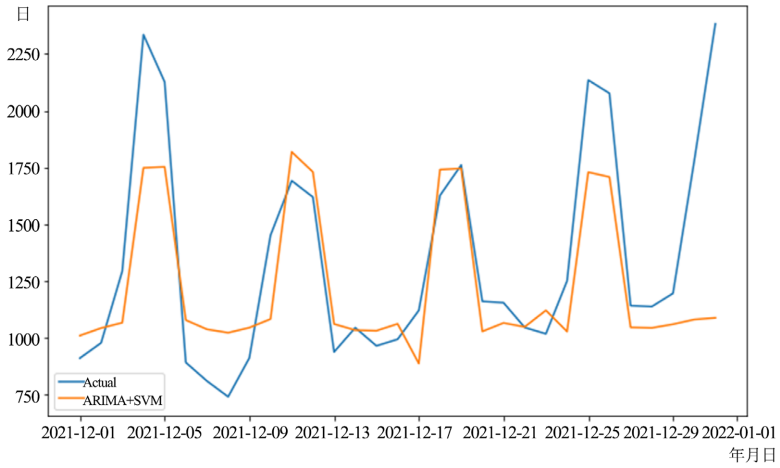


**Figure 3.** Combined ARIMA + SVM model prediction results
**图 3.** ARIMA + SVM 组合模型预测结果

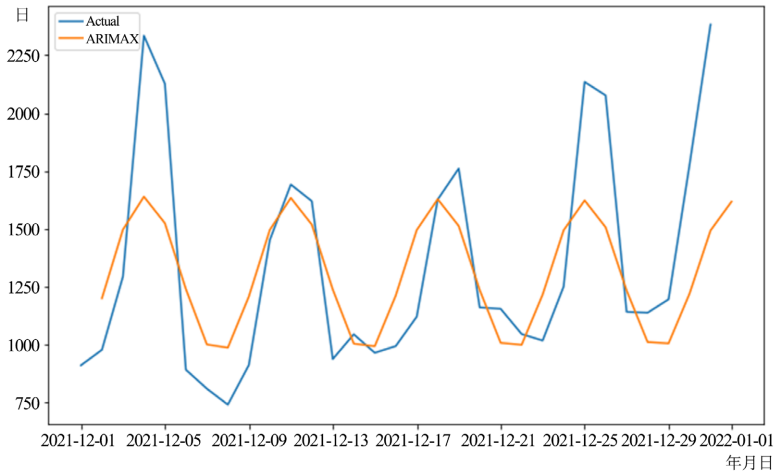


**Figure 4.** Ordinary ARIMA model prediction results
**图 4.** 普通 ARIMA 模型预测结果

同时，我们使用均方根误差(RMSE)、平均绝对误差(MAE)和平均绝对百分比误差(MAPE)这三个指标来判断三种模型预测的准确性，各个指标的值越小表明预测精度越高，预测结果也就越好。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2} \tag{1}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right| \tag{2}$$

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{3}$$

$n$ 代表样本个数，$\hat{y}_i$ 为真实值，$y_i$ 为预测值。

经过计算，三种模型的三个指标结果如表 2 所示，我们发现，加入非线性因素，即节假日因素的 ARIMA 模型预测结果的 RMSE、MAE、MAPE 分别为 335.911、215.864、14.450，相较于普通的 ARIMA 模型和 ARIMA + SVM 组合模型，加入非线性因素的 ARIMA 模型预测误差明显降低，准确度更高。

**Table 2.** Results of model error evaluation
**表 2.** 模型误差评估结果

| 模型 | RMSE | MAE | MAPE (%) |
|---|---|---|---|
| 加入非线性因素的 ARIMA 模型 | **335.911** | **215.864** | **14.450** |
| ARIMA + SVM 组合模型 | 338.447 | 227.404 | 15.548 |
| 普通 ARIMA 模型 | 433.974 | 357.147 | 26.716 |

因此，无论从实验结果还是理论基础上来看，在 ARIMA 模型的基础上加上非线性因素进行预测，可以提高冷藏酸奶销量预测的效率与准确性，从而为零售商提供了更为精细和高效的销售计划安排以及有效的库存控制，具有一定的理论与现实应用价值[17]。

## 5. 结语

本文研究了加入非线性因素的 ARIMA 模型在冷藏酸奶日销量预测方面的应用。通过实证分析和对比，我们证明了加入节假日因素等非线性因素后的 ARIMA 模型相较于传统 ARIMA 模型和目前流行的 ARIMA 组合模型，在一定程度上可以提高销量预测的准确性和精确度，为销售商提供了一种基于非线性因素的销量预测方法，以更好地应对消费者需求的变化和销售季节性的变化。

## 参考文献

[1] 杨海民, 潘志松, 白玮. 时间序列预测方法综述[J]. 计算机科学, 2019, 46(1): 21-28.

[2] 薛冬梅. ARIMA 模型及其在时间序列分析中的应用[J]. 吉林化工学院学报, 2010, 27(3): 80-83. https://doi.org/10.16039/j.cnki.cn22-1249.2010.03.004

[3] 彭俊, 张肖建, 徐超, 等. 基于多特征集成决策树算法的门诊需求预测[J]. 北京生物医学工程, 2021, 40(1): 68-73.

[4] 陈科秀, 刘娟. 基于 ARIMA 模型的欧拉黑猫新能源汽车销量预测[J]. 现代工业经济和信息化, 2022, 12(3): 169-171. https://doi.org/10.16525/j.cnki.14-1362/n.2022.03.064

[5] 陈昆, 曲大义, 贾彦峰, 等. 基于 ARIMA-SVM 模型的短时交通流量预测研究[J]. 青岛理工大学学报, 2022, 43(5): 104-109.

[6] 范波, 宋文彬. 基于 ARIMA 模型的产品销售量预测研究[J]. 工业控制计算机, 2021, 34(5): 128-129+125.

[7] 李晓彤, 王淑萍. 基于季节性预测法的五菱宏光 MINI EV 销量预测分析[J]. 投资与合作, 2022(3): 179-181.

[8] 师丹平, 周丽娟, 王其缘, 等. 基于时间序列模型的新能源乘用车销量预测研究[J]. 统计与咨询, 2023(1): 30-34. https://doi.org/10.19456/j.cnki.tjyzx.2023.01.008

[9] Liu, P., Wei, M. and Bin, H. (2020) Sales Forecasting in Rapid Market Changes Using a Minimum Description Length Neural Network. *Neural Computing and Applications*, **33**, 937-948. https://doi.org/10.1007/s00521-020-05294-8

[10] 蒋翠清, 王香香, 王钊. 基于消费者关注度的汽车销量预测方法研究[J]. 数据分析与知识发现, 2021, 5(1): 128-139.

[11] 徐英卓, 郭博, 王六鹏. 基于 GBDT 算法的游戏销量预测模型研究[J]. 智能计算机与应用, 2023, 13(1): 182-185.

[12] 袁远, 郭天添. ARIMA-RF 组合模型的销售预测研究[J]. 软件导刊, 2021, 20(9): 33-38.

[13] 呼雄伟. 基于 ARIMA-LSTM 组合模型在钢瓶销量预测中的研究与应用[D]: [硕士学位论文]. 上海: 上海第二工业大学, 2022. https://doi.org/10.27916/d.cnki.ghdeg.2022.000045

[14] 葛娜, 孙连英, 赵平, 等. 基于 ARIMA 时间序列模型的销售量预测分析[J]. 北京联合大学学报(自然科学版), 2018, 32(4): 27-33. https://doi.org/10.16255/j.cnki.ldxbz.2018.04.006

[15] Catal, C., *et al*. (2019) Benchmarking of Regression Algorithms and Time Series Analysis Techniques for Sales Forecasting. *Balkan Journal of Electrical and Computer Engineering*, **7**, 20-26. https://doi.org/10.17694/bajece.494920

[16] 涂俐兰, 黄丹. 插值法在数据修正中的应用[J]. 数学理论与应用, 2012, 32(3): 110-116.

[17] 张婷婷. 基于 ARMA 模型的时间序列挖掘[D]: [硕士学位论文]. 合肥: 合肥工业大学, 2013.