文章编号：1004-4353(2021)04-0324-05

SARIMA-SVR 混合模型在电费收入预测中的应用

孙越[1]， 洪义成[1]， 刘鑫[2]， 张志强[2]， 郑雪燕[1]
（1.延边大学 理学院，吉林 延吉 133002；2.国网吉林省电力有限公司 延边供电公司，吉林 延吉 133000）

摘要：针对 SARIMA 模型和 SVR 模型在预测电费收入数据时因存在线性因素和非线性因素所产生的难以精准预测的问题，提出一种将 SARIMA 和 SVR 相结合的 SARIMA-SVR 混合模型. 利用延边供电公司的月电费收入数据对 SARIMA-SVR 混合模型的有效性进行验证显示，SARIMA-SVR 混合模型的平均绝对百分比误差比 SARIMA 模型和 SVR 模型分别低了 13.50%和 73.75%. 该结果表明 SARIMA-SVR 混合模型对电费收入数据具有较好的预测效果.

关键词：电费收入预测；SARIMA-SVR；混合模型；支持向量机；残差分析
中图分类号：N32 **文献标识码**：A

Application of SARIMA-SVR hybrid model in electricity revenue forecasting

SUN Yue[1]， HONG Yicheng[1]， LIU Xin[2]， ZHANG Zhiqiang[2]， ZHENG Xueyan[1]
（1. *College of Science*, *Yanbian University*, *Yanji* 133002, *China*;
2. *State Grid Jinlin Electric Power Co.*, *Ltd.*, *Yanbian Power Supply Company*, *Yanji* 133000, *China*）

Abstract: Aiming at the difficulty of accurate prediction caused by linear and nonlinear factors when SARIMA model and SVR model forecast electricity revenue data, a SARIMA-SVR hybrid model combining SARIMA and SVR was proposed. The validity of the SARIMA-SVR hybrid model was verified by using the monthly electricity revenue data of Yanbian Power Supply Company. The results show that the average absolute percentage error of SARIMA-SVR model is 13.50% lower than that of SARIMA model and 73.75% lower than that of SVR model. The results show that SARIMA-SVR hybrid model has a good prediction effect on electricity income data.

Keywords: electricity revenue forecast; SARIMA-SVR; hybrid model; support vector machine; residual analysis

电费收入是供电企业运营中的一项重要经济指标. 在我国，由于电力商品并不像其他商品采取现场等价交易的方式，而是采用先购买再使用的方式，因此供电企业的电费收入不仅受到用户使用电量的影响，还受到用户缴纳电费全额的影响，即包括了许多随机因素[1-3]. 目前，预测电费收入的方法主要分为两种方法：一是利用时间序列模型（包括AR 模型、ARMA 模型、ARIMA 模型等）进行预测[4-7]，这类模型虽然在操作上方便，但是对数据要求较高；二是利用机器学习方法进行预测，该模型虽然较为复杂，但是在组织和拟合参数方面准确度较高，同时拟合任意非线性趋势的效

收稿日期：2021-09-17
基金项目：延边大学横向项目(20210051)
第一作者：孙越(1997—)，女，在读硕士，研究方向为应用统计.
通信作者：郑雪燕(1989—)，女，硕士，讲师，研究方向为时间系列分析、数据挖掘.

果较好[8-10]. 为进一步提高电费收入的预测效果，本文提出一种将时间序列和机器学习相结合的 SARIMA-SVR 混合模型，并对模型的有效性进行了验证.

1　分析方法

1.1　季节时间序列模型(SARIMA 模型)

时间序列模型[11]是从时间序列中找出变量变化的特征、趋势以及发展规律，以此实现对变量的未来变化进行有效预测的模型. 按照模型中是否包含季节性成分，ARIMA 模型可分为季节模型和非季节模型，其中描述季节性序列的模型又称为季节时间序列模型(seasonal ARIMA model, SARIMA). SARIMA 模型中除了用到一般的差分，还用到了季节性差分 S，即用 S 反映一定的周期(T). 用 t 时刻的值减去 $t-T$ 时刻的值即可得到季节性差分序列.

对于时间序列 $\{Y_t\}$，SARIMA 模型的一般表达式为：

$$\begin{aligned} &W_t=\nabla^d\nabla_s^D Y_t=\frac{\Theta(B)\Theta_s(B)}{\varphi(B)\varphi_s(B)}\xi_t,\\ &\Theta(B)=1-\theta_1 B-\cdots-\theta_q B^q,\\ &\phi(B)=1-\varphi_1 B-\cdots-\varphi_p B^p,\\ &\Theta_s(B)=1-\theta_1 B^S-\cdots-\theta_Q B^{QS},\\ &\phi_s(B)=1-\varphi_1 B^S-\cdots-\varphi_s B^{PS}. \end{aligned} \tag{1}$$

本文将公式(1)记为 SARIMA$(p,d,q)\times(P,D,Q)_s$ 模型，其中 s 为季节周期，p、d 和 q 为非季节阶数，P、D 和 Q 为季节阶数. SARIMA 模型的建模流程图如图 1 所示.

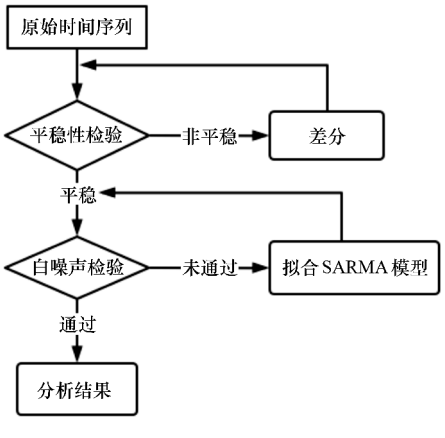

图 1　SARIMA 模型的建模流程

1.2　支持向量回归模型(SVR 模型)

支持向量回归(support vector regression, SVR)模型[12]是在线性函数的两侧建造一个“间隔带”，然后通过最小化“间隔带”的宽度与总损失来优化模型，其中损失函数仅计算间隔带之外的样本. SVR 模型利用非线性函数 $\varphi(x)$ 将给定的原始数据 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ 映射到高维空间，以此形成高维空间的线性函数，其表达式为：

$$f(x)=w^T(x)+b, \tag{2}$$

其中 w 为权重，b 为截距. 假设 SVR 模型允许 $f(x)$ 与 y 之间的最多误差为 ε，且仅当 $f(x)$ 与 y 之间的差的绝对值大于 ε 时才计算损失. 根据结构风险最小化原则可知，求解 $f(x)$ 等效于求解优化问题，即：

$$\begin{aligned} &\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}l_\varepsilon(f(x_i)-y_i),\\ &l_\varepsilon(z)=\begin{cases}0, & \text{if } |z|<\varepsilon;\\ |z|-\varepsilon, & \text{if } |z|\geqslant\varepsilon.\end{cases} \end{aligned} \tag{3}$$

其中 C 为惩罚因子，实质是正则化常数，l_ε 为损失函数. 为了增加容错性，SVR 模型引入了松弛变量 ξ_i 和 $\hat{\xi}_i$，由此式(3)可写为：

$$\min_{w,b,\xi_i,\hat{\xi}_i}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i). \tag{4}$$

将回归问题转换为求解目标函数的最小化问题时，SVR 模型引入了拉格朗日乘法算子，由此回归问题转换为较为易解的拉格朗日函数：

$$\begin{aligned} L(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},\mu,\hat{\mu})=&\frac{1}{2}\|w\|^2+\\ &C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)-\sum_{i=1}^{m}\mu_i\xi_i-\sum_{i=1}^{m}\hat{\mu}_i\hat{\xi}_i+\\ &\sum_{i=1}^{m}\alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)+\\ &\sum_{i=1}^{m}\hat{\alpha}_i(y_i-f(x_i)-\varepsilon-\hat{\xi}_i). \end{aligned} \tag{5}$$

利用对偶原理可得式(5)的对偶问题为：

$$\begin{aligned} &\max_{\alpha,\hat{\alpha}}\sum_{i=1}^{m}y_i(\hat{\alpha}_i-\alpha_i)-\varepsilon(\hat{\alpha}_i+\alpha_i)-\\ &\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)x_i^Tx_j,\\ &\text{s.t.}\ \sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)=0,\\ &0\leqslant\alpha_i\leqslant\alpha_i+\mu_i,\quad 0\leqslant\hat{\alpha}_i\leqslant\hat{\alpha}_i+\hat{\mu}_i, \end{aligned} \tag{6}$$

其中 α_i 和 $\hat{\alpha}_i$ 为拉格朗日乘数.求解式(6)可得如下非线性映射SVR模型:

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x+b. \tag{7}$$

式(7)中的 x 可以利用核函数将其表示为 $\varphi(x_i)$,从而SVR模型的最终表达式为:

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\varphi(x_i)^T\varphi(x)+b. \tag{8}$$

理论上来说,求解 b 值可通过选取任意一个满足 $0<\alpha_i<C$ 的样本后再利用式(9)进行计算即可.但在实际中求解 b 值常采用的方法是选取多个或者所有满足条件 $0<\alpha_i<C$ 的样本后将每个样本代入式(9)中求解 b_i(第 i 个样本求得的解表示为 b_i),然后再将所有的 b_i 取平均值后作为 b 值,即:

$$b=y_i+\varepsilon-\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x. \tag{9}$$

1.3 SARIMA-SVR混合模型

基于SARIMA和SVR模型的优缺点,本文提出了SARIMA-SVR混合模型.该模型首先使用SARIMA模型预测月电费收入,以此获得月电费收入的SARIMA模型预测序列 $\{\hat{Y}_t\}$ 及残差序列 $\{R_t\}$,其中预测序列包含数据中的线性规律,残差序列包含数据中的非线性规律.然后用SVR模型预测SARIMA模型的残差序列,以使非线性规律包含在SVR模型的预测结果中,并得到残差序列的预测值 $\{\hat{R}_t\}$.最后相加SARIMA模型的预测结果和SVR模型的预测结果,由此即可得到混合预测模型的预测值 $\{\hat{F}_t\}$,即:

$$\hat{F}_t=\hat{Y}_t+\hat{R}_t. \tag{10}$$

SARIMA-SVR混合模型的流程如图2所示.

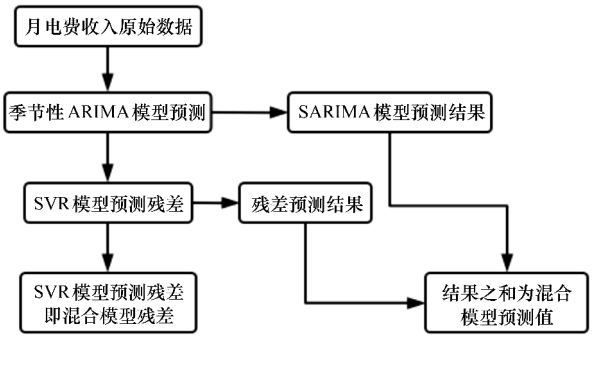

图2 SARIMA-SVR混合模型的建模流程

2 实证分析

2.1 数据处理

2.1.1 数据集

本文采用的数据资料是国网延边供电公司2010年7月至2021年7月的月电费回收数据,该数据的时序图如图3所示.

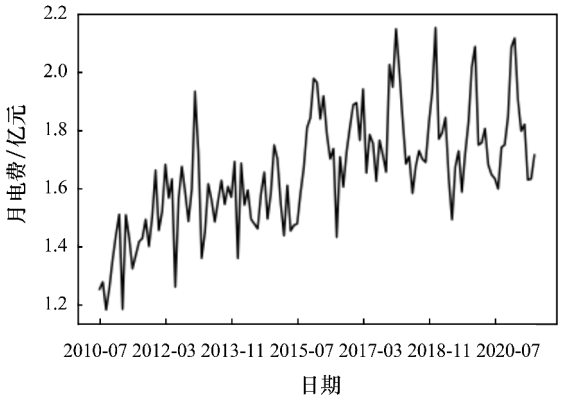

图3 月电费收入的时序图

实验时,本文将全部数据按时间段划分为训练集和测试集.其中2010年7月至2020年12月的月电费收入为训练集,2021年1月至2021年7月的月电费收入为测试集.考虑到电费收入数据在不同季节和特殊日期的波动情况,本文采用3种影响特征(见表1)预测SARIMA-SVR混合模型的有效性.

表1 影响电费收入预测的因素

影响因素	定义
每月实际电费收入	银行到款的每月电费收入
是否为节假日	节假日特征值为1,否则为0
是否为采暖期	采暖期间特征值为1,否则为0
月份特征	1月至12月分别用1-12表示

2.1.2 数据平稳性检验

由于需要判断原始时间序列数据是否平稳,因此需要对序列的平稳性进行检验.平稳性检验的方法有两种:一种是通过时序图的形状和走势来判断平稳性;另一种是通过构造检验统计量来判断平稳性.由于第2种方法中的单位根检验(ADF)能够准确地判断序列平稳性,因此本文采用单位根检验方法来判断原始序列和差分之后的序列是否平稳.

2.1.3　最优参数

赤池信息准则(Akaike information criterion，AIC)和贝叶斯信息准则(Bayesian information criterion，BIC)是衡量统计模型拟合是否优良的常用标准，其表达式为：

$$AIC = 2k - 2\ln L, \tag{11}$$

$$BIC = k\ln n - 2\ln L. \tag{12}$$

其中，k 为模型参数个数，n 为样本数量，L 为似然函数. 本文以 AIC 准则和 BIC 准则为依据，使用 AUTO-ARIMA 函数(python 3.8 版)选取最优参数，得到的具体参数设置如表 2 所示.

表 2　SARIMA-SVR 混合模型的参数设置

参数	值域
(p,d,q)	(0,1,1)
(P,D,Q,s)	(0,1,1,12)
kernel	rbf
C	80
gamma	0.1

2.2　结果分析

2.2.1　误差分析

为了对比分析模型的预测结果，在进行精准度评估时，本文选取平均绝对百分比误差(MAPE)作为模型的评价标准. 平均绝对百分比误差的计算公式为：

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - y_i^*|}{y_i} \times 100\%, \tag{13}$$

式中 y_i 为实际值，y_i^* 为预测值. MAPE 值越小表示预测值越准，即表示模型的预测效果越好.

2.2.2　平稳性检验

对原始数据进行单位根检验显示，其 P 值(0.137 728)远大于 0.05，说明原始数据是不平稳序列，需要进行差分处理. 为此本文利用表 2 中的参数对原始数据进行差分处理，然后再对差分后的数据进行计算得到了时间序列数据的自相关系数(ACF)和偏自相关系数(PACF)，如图 4 所示. 由图 4 可以看出，差分后的数据已趋于平稳. 另外，根据平稳性检验原理对差分后的数据进行单位根检验得其 P 值远小于 0.05，这进一步说明差分后的数据是平稳的.

2.2.3　残差分析

对预测数据的残差进行白噪声分析后得其 P 值远小于标准值 0.05，由此表明得到的 SARIMA 模型的残差不是一组白噪声序列. 这说明残差中还有有用的信息，需进一步提取有效信息. 提取有效信息的方法是：首先对残差序列进行支持向量回归分析，以此得到残差的预测值；然后将残差的预测值和 SARIMA 的预测值相加，以此得到更为接近实际值的预测值.

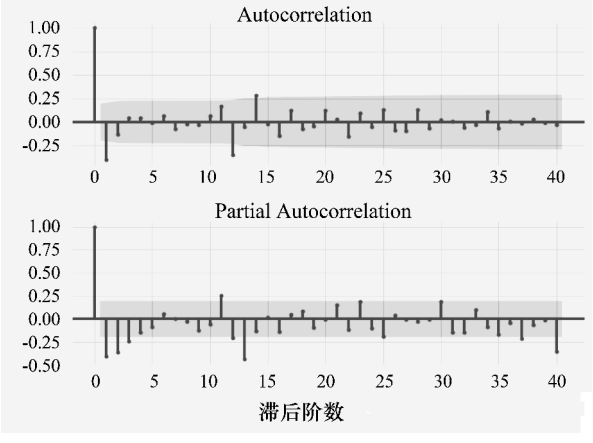

图 4　序列的自相关系数和偏自相关系数

2.2.4　对比分析

为了验证 SARIMA-SVR 混合模型的预测准确性，在相同的测试集下将 SARIMA-SVR 混合模型与 SARIMA 模型、SVR 模型进行了对比实验. 两种模型的参数值如表 3 所示，该参数值可以使 SARIMA 模型和 SVR 模型的整体效果达到最佳.

表 3　SARIMA、SVR 模型的参数设置

模型名称	参数	值域
SARIMA	(p,d,q)	(0,1,1)
SARIMA	(P,D,Q,s)	(0,1,1,12)
SVR	kernel	rbf
SVR	C	80
SVR	Gamma	0.1

为了更加直观地观察预测结果，将各模型的实际值与预测值进行了可视化处理，如图 5 所示. 由图 5 可以看出，SARIMA-SVR 混合模型的预测精准度与实际值最为接近，由此表明混合模型的拟合效果较好.

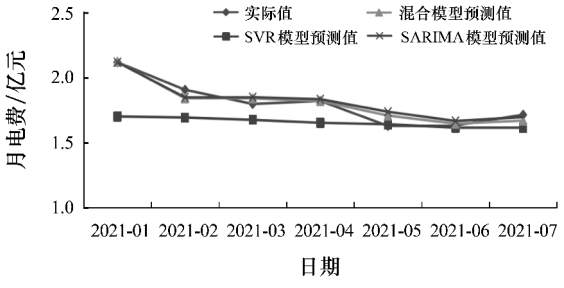

图5 各模型的预测结果

3种模型的MAPE值如表4所示.由表4可以看出,SARIMA-SVR混合模型的数据预测效果显著优于其他两种模型,其中SARIMA-SVR混合模型的MAPE值比SARIMA模型降低了13.50%,比SVR模型降低了73.75%.其原因是SARIMA-SVR混合模型将电费时间序列中包含的主要趋势融入到了模型中进行了残差分析,由此使得SARIMA模型的预测结果得到进一步修正,从而达到了更好的预测效果.

表4 3种模型的预测效果

模型名称	MAPE/%
SARIMA	2.37
SVR	7.81
SARIMA-SVR	2.05

3 结论

利用本文构建的SARIMA-SVR混合模型对电费收入进行预测表明,SARIMA-SVR混合模型的预测精度显著优于单一的SARIMA模型和SVR模型,因此该模型可为今后电费收入预测方面的研究提供参考.由于本文在研究中使用的月电费收入数据相对较少,在寻找训练模型的变量特征方面仍存在不足;因此,在今后的研究中,我们将进一步挖掘数据,如量化居民缴费的心理因素、流动人口、消费者指数等,以此得到更多、更合适的变量特征来训练模型,从而进一步提高模型的准确度.

参考文献:

[1] 曹家伟.供电企业电费回收管理方法研究[D].青岛:青岛大学,2016.

[2] 刘东东,胡少柔,陈荣腾.基于大数据技术的电费回收研究[J].江苏科技信息,2016(27):55-56.

[3] MEMARZADEH G, KEYNIA F. Short-term electricity load and price forecasting by a new optimal LSTM-NN based prediction algorithm[J]. Electric Power Systems Research, 2021,192:106995.

[4] 胡泽月.基于ARIMA模型的企业自由现金流预测研究[D].石家庄:河北经贸大学,2021.

[5] 范恒瑞,任黎秀.ARIMA与指数平滑法在江苏省GDP预测中的应用[J].江西农业学报,2011,23(2):187-189.

[6] 刘松,张帅.运用ARIMA模型对股价预测的实证研究[J].经济研究导刊,2021(25):76-78.

[7] AL-MUSAYLH M S, DEO R C, ADAMOWSKI J F, et al. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia[J]. Advanced Engineering Informatics, 2018,35:1-16.

[8] 张远汀,龚伟伟,叶钰,等.应用机器学习技术预测强雨雪天气过程中的积雪[J].科学技术与工程,2019,19(15):58-69.

[9] LI W, BECKER D M. Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling[J]. Energy, 2021,237:121543.

[10] 王振,高茂庭.基于卷积神经网络的图像识别算法设计与实现[J].现代计算机(专业版),2015(20):61-66.

[11] 王燕.应用时间序列分析[M].北京:中国人民大学出版社,2015.

[12] 周志华.机器学习[M].北京:清华大学出版社,2016.