Operations Research and Fuzziology 运筹与模糊学, 2021, 11(1), 105-112
Published Online February 2021 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2021.111013

Hans汉斯

SARIMA与SVR模型在天津市肝炎发病率预测中的比较

张　仙，戴家佳

贵州大学数学与统计学院，贵州　贵阳
Email: 827643042@qq.com

收稿日期：2021年1月11日；录用日期：2021年2月10日；发布日期：2021年2月25日

摘　要

为比较乘积季节差分自回归移动平均模型(SARIMA)与支持向量机回归模型(SVR)对天津市病毒性肝炎发病率的预测效果，本文根据天津市2005年1月至2017年4月病毒性肝炎发病率数据建立SARIMA和SVR预测模型，对2017年5月至12月发病率预测。SVR模型预测的RMSE，MAE和MAPE分别为0.0767，0.0701和4.25%，SVR模型与SARIMA模型中最优模型相比，三个误差评价指标分别下降了0.1089，0.1008和6.04%。预测结果显示SVR模型预测效果优于SARIMA模型，将其用于天津市的病毒性肝炎发病率短期预测，有助于该地区病毒性肝炎的防治工作。

关键词

病毒性肝炎，SARIMA模型，SVR模型，发病率预测

Comparison of SARIMA and SVR Models in Predicting Hepatitis Incidence in Tianjin

Xian Zhang, Jiajia Dai

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
Email: 827643042@qq.com

Received: Jan. 11th, 2021; accepted: Feb. 10th, 2021; published: Feb. 25th, 2021

Abstract

To compare the seasonal autoregressive integrated moving average (SARIMA) model and support vector regression(SVR) model predicted effect on the incidence of viral hepatitis in Tianjin, we

文章引用: 张仙, 戴家佳. SARIMA 与 SVR 模型在天津市肝炎发病率预测中的比较[J]. 运筹与模糊学, 2021, 11(1): 105-112.
DOI: 10.12677/orf.2021.111013

used data collected from January 2005 to April 2017 as training data while the data from May 2017 to December 2017 as testing data. The RMSE, MAE and MAPE predicted by the SVR model are 0.0767, 0.0701 and 4.25%, respectively. Compared with the SARIMA model of optimal model, the three error evaluation indexes of the SVR model decreased by 0.1089, 0.1008 and 6.04% severally. The prediction effect of SVR model is better than that of SARIMA model. Its application to the short-term prediction of the incidence of viral hepatitis in Tianjin is helpful to the prevention and treatment of viral hepatitis in this area.

Keywords

Viral Hepatitis, SARIMA Model, SVR Model, Incidence Prediction

1. 引言

病毒性肝炎(viral hepatitis)简称肝炎，是由肝炎病毒引起的一组以肝脏损害为主要特征的传染性疾病，其具有传染性强、传播途径复杂、发病率高和流行面广等特点[1] [2]。目前，肝炎已作为法定乙类报告传染病，法定报告的肝炎包括甲肝、乙肝、丙肝、戊肝和未分型肝炎[3]。世卫组织报告称，感染肝炎的人口全球大约有 3.25 亿人，肝炎成为全球公共卫生面临的又一重大威胁[4]。我国是肝炎的高流行区，肝炎在天津市乙类传染病报告中发病率一直居于前列，其传播给国家和个人带来了沉重的经济负担，是天津市重点防治的传染病之一。因此探讨天津市肝炎的流行规律、准确有效的预测发病率对天津市肝炎的防治工作有着重要的指导意义。

目前对肝炎发病率预测研究中时间序列方法的 SARIMA 模型最为广泛。汪业胜等[5]用 SARIMA 模型对我国 2009~2018 年肝炎的发病趋势分析和预测，预测值与实际值较一致。陈远方等[6]建立 SARIMA 模型和 BP 神经网络模型对我国乙型肝炎发病预测，结果显示 SARIMA 模型的预测效能和非线性拟合能力略优于 BP 神经网络模型。高云云等[7]和李丽娜等[8]运用季节模型 SARIMA 模型能较好模拟、预测的甲型肝炎发病情况，是一种进行短期预测效果较好的模型，将优化甲肝预防工作。近年来，机器学习算法对非线性时间序列的分析中表现出极大优势，其中较为成熟的支持向量机回归(SVR)模型已在空气质量[9] [10]、农产品价格[11]等时间序列预测方面得到了成功应用，但在肝炎预测方面的应用鲜有报道。鉴于此，本研究用 SARIMA 模型和 SVR 模型对天津市肝炎发病率进行建模和预测，并对建立的模型进行比较。

2. 数据来源与研究方法

2.1. 数据来源

本文收集了 2005 年 1 月~2017 年 12 月天津市肝炎发病率月度数据，共 156 个月，所有数据来源于中国公共卫生科学数据中心。

2.2. 研究方法

2.2.1. SARIMA 模型

ARIMA 模型是时间序列预测的经典模型，它是将非平稳时间序列转化为平稳时间序列，根据因变量、因变量的滞后值、随机误差的现值和滞后值进行回归所建立的模型[12]，它的一般形式为 ARIMA(p,d,q)。

SARIMA$(p,d,q)\times(P,D,Q)_S$模型是在 ARIMA 模型中增加了季节项，称为季节性自回归滑动平均模型。SARIMA 模型首先对季节性因素进行 D 阶差分，其次用差分后周期为 S 的季节性时间序列建立一般的 ARIMA 模型。

SARIMA 的乘积模型表达式为：

$$\nabla^d\nabla_S^D X_t=\frac{\theta(B)\theta_s(B)}{\phi(B)\phi_s(B)}\varepsilon_t \tag{1}$$

其中：

$$\begin{aligned}&\theta(B)=1-\theta_1B-\cdots-\theta_qB^q\\&\phi(B)=1-\phi_1B-\cdots-\phi_pB^p\\&\theta_s(B)=1-\theta_1B^s-\cdots-\theta_QB^{QS}\\&\phi_s(B)=1-\phi_1B^s-\cdots-\phi_PB^{PS}\end{aligned} \tag{2}$$

式中 B 是滞后算子，ε_t 是白噪声序列，$\phi_S(B)$ 和 $\theta_S(B)$ 分别是季节自回归和季节移动平均算子。

2.2.2. SVR 模型

支持向量回归(support vector regression，SVR)是一种基于统计学理论的机器学习算法[13]。SVR 的基本思想是：对于给定的训练样本点，通过 SVR 训练回归一个函数 $f(x)$，使由该函数求出的每一个输入样本的输出值和输入样本对应的目标值不超过误差 ε，同时使回归出的函数平滑[14]。

对于给定训练样本集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，输出值 $y=f(x)=\omega\cdot\varphi(x)+b$。SVR 以 $f(x)$ 为中心构建了宽度为 2ε 的间隔带，落在间隔带以外的样本才计算损失。考虑到误差，引入松弛变量 $\xi_i,\hat{\xi}_i$，和惩罚函数 C，SVR 问题可形式化为：

$$\min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\left(\xi_i+\hat{\xi}_i\right) \tag{3}$$

$$\begin{cases}\text{s.t. } f(x_i)-y_i\le\varepsilon+\xi_i\\y_i-f(x_i)\le\varepsilon+\xi_i\\\xi_i\ge0,\hat{\xi}\ge0,i=1,2,\cdots,n\end{cases} \tag{4}$$

其中 C 为常数，$\xi_i,\hat{\xi}_i$ 代表数据到其对应边界 ε 的距离。引入拉格朗日乘子 λ_i,λ_i^*，将上式转化为

对偶问题，则有：

$$\max\sum_{i=1}^{m}y_i\left(\lambda_i-\lambda_i^*\right)-\varepsilon\sum_{i=1}^{m}\left(\lambda_i+\lambda_i^*\right)-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\lambda_i-\lambda_i^*\right)\left(\lambda_j-\lambda_j^*\right)K(x_i,y_i) \tag{5}$$

$$\begin{cases}\text{s.t.}\sum_{i=1}^{m}\left(\lambda_i-\lambda_i^*\right)=0\\0\le\lambda_i\le C\\0\le\lambda_i^*\le C\end{cases} \tag{6}$$

最后得出表达式：$f(x)=\sum_{i=1}^{m}\left(\lambda_i-\lambda_i^*\right)K(x_i,x_i)+b$。其中 $K(x_i,x_i)$ 为核函数，本文选用的是 RBF 核函数。时间序列数据建立 SVR 模型时，需分别截取固定大小数据段组成滑动窗口，以前面数据为输入，后面数据为输出顺次建立映射关系，即 $X_i=\{x_i,x_{i+1},\cdots,x_{i+k-1}\},i=1,2,\cdots,n-k+1$，$x_{i+k}=f(X_i)$，其中 i 为时序，x_i 为第 i 时刻的数据，输入数据 X_i 随着时间向前推动，训练样本不断地变化，但是样本内元素量 k 不会变动[15]。

2.2.3. 模型评估标准

本文采用均方根误差(RMSE)、平均绝对误差(MAE)和平均绝对百分误差(MAPE)三个综合指标用于比较各模型预测差异。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2} \tag{7}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right| \tag{8}$$

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|\times 100\% \tag{9}$$

其中 y_i 为实际值， $\hat{y}_i$ 为预测的值， n 为预测样本数。

3. 模型的建立

3.1. SARIMA 模型的建立

2005 年 1 月至 2017 年 4 月天津市肝炎发病率的时序图见图 1。首先对序列进行平稳性检验，通过时序图可以看出该肝炎发病率数据有明显长期趋势，其单位根检验 P 值为 0.812，该序列是非平稳时间序列。对肝炎发病率序列做季节分解，序列具有 12 个月为一个周期的季节性。对原始序列作一阶 12 步差分提取原序列的趋势效应和季节效应，对差分后的序列平稳性检验，单位根检验的 P 值为 0.037，该序列是平稳时间序列。对差分后序列做白噪声检验，计算出序列延迟 6 期和 12 期的 Q 统计量的 P 值均小于 0.05，即该序列是非白噪声序列可进行下一步建模分析。差分后序列的自相关和偏自相关图见图 2，初步确定 SARIMA 模型为 $\mathrm{ARIMA}(4,1,3)\times(4,1,3)_{12}$，但通过查看 ACF 和 PACF 图来做出选择不准确[16]，为寻找最优 SARIMA 模型，使用网格搜索法对参数进行选择，AIC 最小的模型作为最优模型，最终确定模型为 $\mathrm{ARIMA}(2,1,4)\times(2,1,1)_{12}$，AIC 为 70.137。采用上述选取的最优预测模型对 2017 年 5 月至 12 月天津市肝炎发病率进行预测，对 2005 年 2 月至 2017 年 4 月肝炎发病率拟合，并与实际发病数据进行对比，结果见图 3。

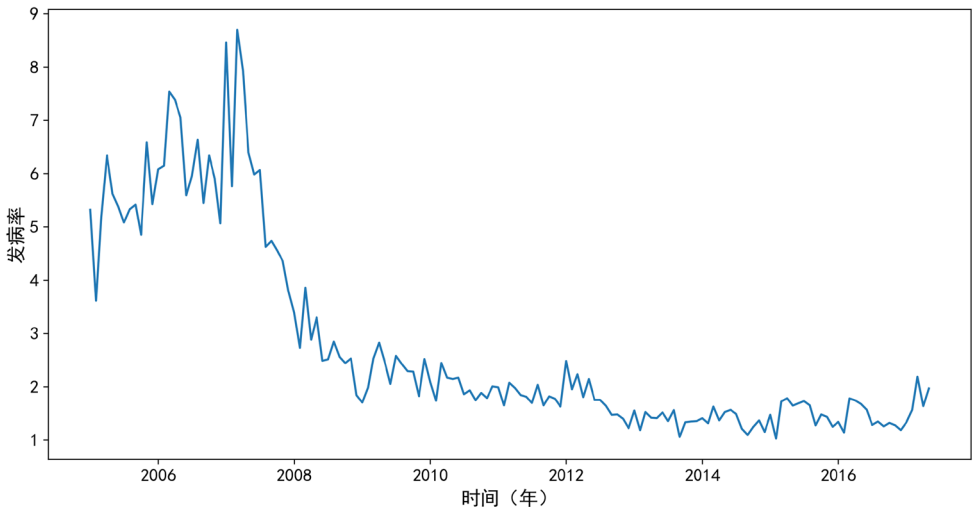

Figure 1. Time series of hepatitis incidence
图 1. 肝炎发病率时序图

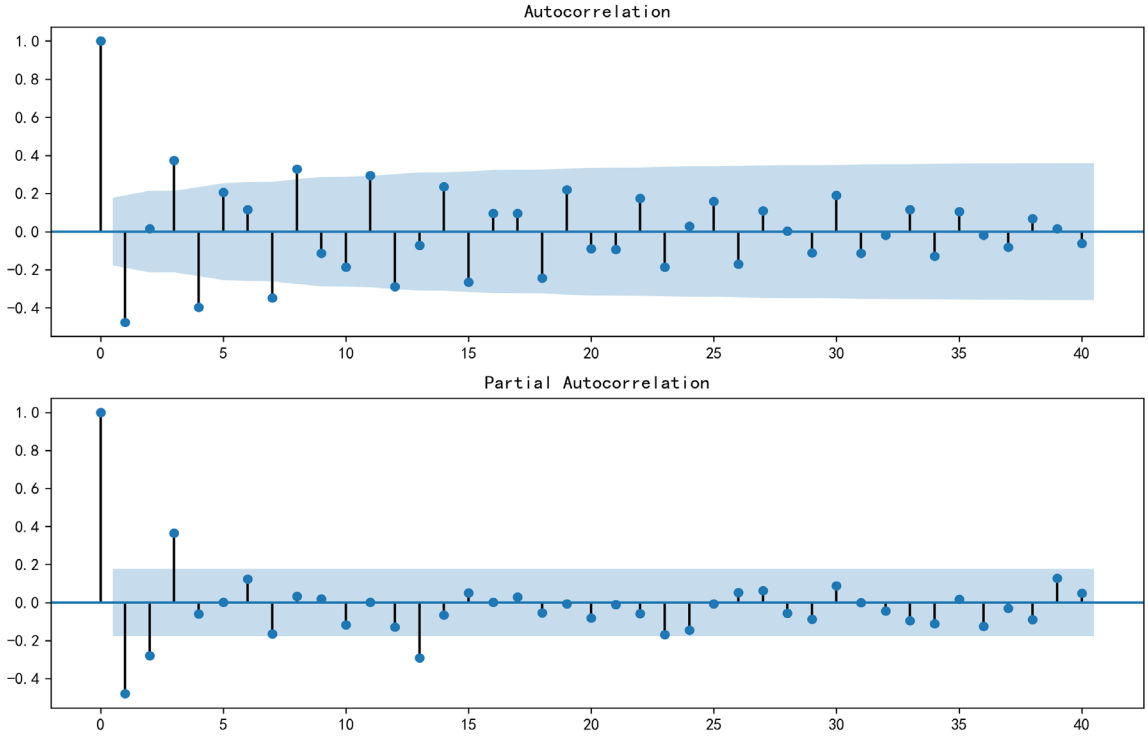

Figure 2. Autocorrelation and partial autocorrelation of the sequence after difference
图 2. 差分后序列的自相关图和偏自相关图

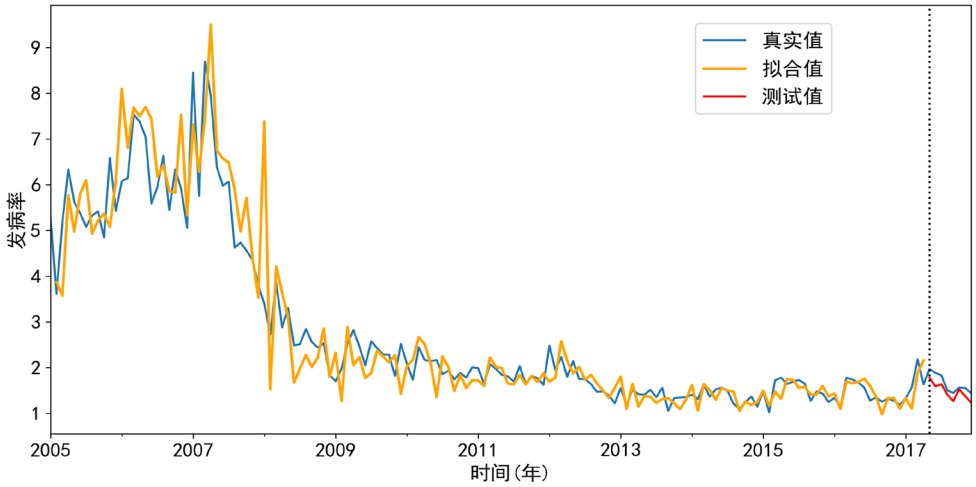

Figure 3. Comparison diagram of SARIMA model prediction
图 3. SARIMA 模型预测对比图

3.2. SVR 模型的建立

对 2005 年 1 月至 2017 年 4 月的肝炎发病率时间序列数据建立 SVR 模型时，输入样本为滚动的时间序列，数据段长度 k 的选取(即滑动窗口选择)会影响模型预测的效果，本文 k 值从 2 依次增加，根据 k 值不同取值分别建立 SVR 模型，建立好模型后对数据拟合，以 RMSE 和 MAPE 为评价指标，若取 k+1 时模型的 RMSE 和 MAPE 比取 k 小，则继续增加 k 值，否则 k 停止增加，k 为最终确定选定的值，见表 1 所示，本文 k 的取值为 8。

Table 1. The corresponding error index is selected for different data segment length k
表 1. 数据段长度 k 不同选取对应的误差指标

k	2	3	4	5	6	7	8
RMSE	0.546	0.520	0.482	0.467	0.443	0.428	0.407
MAPE	10.7%	10.2%	10.0%	9.7%	9.3%	9.2%	9.0%

k 选定为 8，2005 年 1 月至 2017 年 4 月的肝炎发病率被分为 140 组，将 140 组数据中前 132 组作为训练样本集，用于建立模型，后 8 组作为试验样本集，用于验证模型。其中，后 8 组中真实的输出值对应的时间序列为 2016 年 9 月~2017 年 4 月的肝炎发病率数据。在验证模型时，每预测一步后，将预测值作为当月的值再进行下一月预测，将后 8 组真实值与预测值对比，此过程中采用网格搜索进行参数寻优。建立好 SVR 模型后对数据进行拟合以及对 2017 年 5 月至 12 月肝炎发病率预测，并与实际发病率数据进行对比，结果见图 4。

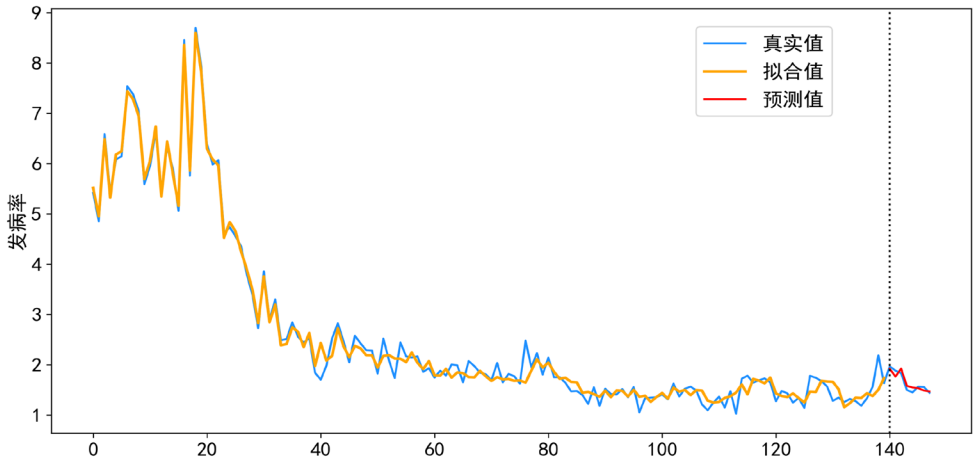

Figure 4. Comparison diagram of SVR model prediction
图 4. SVR 模型预测对比图

3.3. 两模型预测对比

本文测试数据为 2017 年 5 月至 12 月的肝炎发病率，表 2 为两种预测模型预测结果，从表 2 可以看出 SVR 模型预测值较 SARIMA 模型更接近真实值，两种模型预测的 RMSE、MAE 和 MAPE 见表 3，SVR 模型预测的 RMSE，MAE 和 MAPE 分别为 0.0767，0.0701 和 4.25%，SVR 模型与 SARIMA 模型相比，三个误差评价指标分别下降了 0.1089，0.1008 和 6.04%。

Table 2. Comparison of the predicted results of the two models
表 2. 两种模型预测结果对比

时间	实际值	SARIMA 模型预测值	SVR 模型预测值
2017 年 5 月	1.9653	1.7766	1.9181
2017 年 6 月	1.8821	1.5892	1.7625
2017 年 7 月	1.8308	1.6359	1.9197
2017 年 8 月	1.4980	1.4088	1.5782
2017 年 9 月	1.4468	1.2658	1.5456

Continued

2017 年 10 月	1.5620	1.5254	1.5322
2017 年 11 月	1.5556	1.3760	1.4856
2017 年 12 月	1.4404	1.2356	1.4667

Table 3. Two models predict evaluation indexes
表 3. 两种模型预测评价指标

评价指标	SARIMA 模型	SVR 模型
RMSE	0.1856	0.0767
MAE	0.1709	0.0701
MAPE	10.29%	4.25%

4. 结论

肝炎是危害我国人民身体健康的主要传染病，是防控的重点与研究的热点，本文根据天津市肝炎发病率月度数据，建立 SARIMA 和 SVR 模型对 2017 年 5 月~12 月的发病率预测。从预测结果来看，SVR 模型不仅较好地拟合原数据的趋势，其预测精度也比 SARIMA 模型更好。因此相对 SARIMA 模型，SVR 模型更适合用来预测天津市肝炎的发病情况，为天津市的病毒性肝炎防控工作提供有利的帮助。

基金项目

项目名称：贵州省数据驱动建模学习与优化创新团队。

合同编号：黔科合平台人才[2020]5016。

参考文献

[1] 梁晓峰. 我国病毒性肝炎流行特征及对策[J]. 临床肝胆病杂志, 2010, 26(6): 561-564.

[2] Zhu, B., Liu, J.L., Fu, Y., Zhang, B. and Mao, Y. (2018) Spatio-Temporal Epidemiology of Viral Hepatitis in China (2003-2015): Implications for Prevention and Control Policies. *International Journal of Environmental Research and Public Health*, **15**, 661. https://doi.org/10.3390/ijerph15040661

[3] 田园, 王伟健, 张旭. 2010~2019 年锦州市病毒性肝炎流行病学特征分析[J/OL]. 微生物学免疫学进展, 2020(6): 1-6.

[4] World Health Organization (2017) Global Hepatitis Report 2017. World Health Organization, Geneva

[5] 汪业胜, 王胜难, 潘金花. 王伟炳. 我国 2009~2018 年病毒性肝炎的发病趋势分析和预测研究[J]. 中华流行病学杂志, 2020, 41(9): 1460-1464.

[6] 陈远方, 张熳, 王小莉, 戎毅, 彭海燕, 管芳. ARIMA 模型和 BP 神经网络模型在我国乙型肝炎发病预测中的应用[J]. 江苏预防医学, 2015, 26(3): 23-26. http://dx.chinadoi.cn/10.13668/j.issn.1006-9070.2015.03.008

[7] 高云云, 李军, 杨海燕, 陈帅印. ARIMA 模型在河南省甲型病毒性肝炎发病数预测中的应用[J]. 现代预防医学, 2017, 44(7): 1294-1298.

[8] 李丽娜, 李宁, 胡樱, 刘娜, 宇传华, 王雷, 等. 2004~2019 年湖北省甲型肝炎流行特征分析及预测[J]. 中华疾病控制杂志, 2020, 24(10): 1165-1169. http://dx.chinadoi.cn/10.16462/j.cnki.zhjbkz.2020.10.011

[9] 张楠, 王鹏, 白艳萍, 王永杰. 基于 MGWO-SVR 的空气质量预测[J]. 数学的实践与认识, 2018, 48(8): 159-165.

[10] 李萍, 倪志伟, 朱旭辉, 宋娟. 基于改进萤火虫优化算法的 SVR 空气污染物浓度预测模型[J]. 系统科学与数学, 2020, 40(6): 1020-1036.

[11] 张宝文, 王川, 杨春英, 王来刚. 基于时间序列 SVR 模型的玉米价格预测研究[J]. 中国农学通报, 2020, 36(31): 115-120.

[12] Yan, D.F., Zhou, J.W., Zhao, Y. and Wu, B. (2017) Short-Term Subway Passenger Flow Prediction Based on ARIMA.

International Conference on Geo-Spatial Knowledge and Intelligence, Chiang Mai, 8-10 December 2017, 464-479. https://doi.org/10.1007/978-981-13-0893-2_49

[13] Vapnik, V. (2000) The Nature of Statistical Learning Theory. 2nd Edition, Springer, New York, 69-91. https://doi.org/10.1007/978-1-4757-3264-1

[14] 刘有冠, 陈冬玲. 基于支持向量机回归算法的钢铁企业备件预测研究[J]. 轻工科技, 2017, 33(6): 100-101+103.

[15] 陈涛. 基于滚动时间序列 SVR 的地铁咽喉区小净距隧道围岩位移预测[J]. 铁道科学与工程学报, 2020, 17(9): 2338-2345. http://dx.chinadoi.cn/10.19713/j.cnki.43-1423/u.T20191066

[16] Yu, S.J., Dong, H.W., Chen, Y.G., He, Z. and Shi, X.J. (2019) Clothing Sales Forecast Based on ARIMA-BP Neural Network Combination Model. 2019 *IEEE International Conference on Power*, *Intelligent Computing and Systems*, Shenyang, 12-14 July 2019, 367-372. https://doi.org/10.1109/ICPICS47731.2019.8942427