

引用格式:刘亚波,吴秋轩. 基于高斯过程混合模型的时间序列预测算法研究[J]. 微电子学与计算机, 2021, 38(6): 93-98.
[LIU Y B, WU Q X. Research on time series prediction algorithm based on gaussian process mixture model[J]. Microelectron-
ics & Computer, 2021, 38(6): 93-98.]

基于高斯过程混合模型的时间序列预测算法研究

刘亚波, 吴秋轩

(杭州电子科技大学 自动化学院, 浙江 杭州 310000)

摘 要: 针对实时变化且不同时段差异大的时间序列, 提出一种基于高斯过程混合模型的预测算法. 该算法首先对时间序列进行预处理, 并采用密度空间含噪聚类(DBSCAN)去除奇点. 然后针对扩展迪基-福勒(ADF)检验结果将时间序列分为常数项、平稳和非平稳三类, 最后基于高斯过程混合(GPM)模型对各类时间序列进行预测, 并和差分自回归移动平均模型(ARIMA)、支持向量机(SVM)、高斯过程(GP)模型进行性能对比. 以采购商品报价时间序列为例进行的预测结果表明: GP 模型与 GPM 模型均能输出预测置信区间, 给出预测结果的可信程度; GPM 模型的优势是能够更精准刻画时间序列各时段差异, 预测精度更高.

关键词: 时间序列; 预测; 高斯过程; 高斯过程混合模型

中图分类号: TP18

文献标识码: A

文章编号: 1000—7180(2021)06—0093—06

Research on time series prediction algorithm based on gaussian process mixture model

LIU Ya-bo, WU Qiu-xuan

(School of Automation Engineering, Hangzhou Dianzi University, Hangzhou 310000, Zhejiang China)

Abstract: Aiming at the time series that change in real time and differ greatly in different time periods, a prediction algorithm based on Gaussian process mixture model is proposed. Firstly, the time series is preprocessed and the singularities are removed by density-space noise-containing clustering (DBSCAN). Then, according to the results of the extended Diki-Fowler test, the time series are divided into three categories: constant term, stationary term and non-stationary term. Finally, the prediction of various time series is made based on the Gaussian process mixture (GPM) model, and the performance of the different autoregressive moving average model (ARIMA), support vector machine (SVM) and Gaussian process (GP) model is compared. Taking the time series of government procurement commodity quotes as an example, the results show that both the GP model and the GPM model can output prediction confidence intervals and give the credibility of the prediction results. The advantage of the GPM model is that it can more accurately characterize the differences of time series in different periods, and the prediction accuracy is higher.

Key words: time series; prediction; gaussian process; gaussian process mixture model

1 引言

时间序列预测可以采用很多模型或算法. 例如许等^[1]利用改进差分进化算法优化回声状态网

络进而进行时序预测, 在大连月平均气温、降雨量等时序上具有一定预测精度, 但计算相对复杂. 王等^[2]提出一种基于长短时记忆的故障时间序列预测算法, 虽能利用历史阶段故障进行预测, 但网络复杂导致计算时间复杂度高. Arunraj 等^[3]提出季

节性差分自回归移动平均模型(SARIMAX)来预测零售商店中易腐食品日常销售,虽预测性能有所提升,但不具有通用性. Taghizadeh 等^[4]针对天气敏感产品,提出一种基于人工神经网络的产品需求预测方法,虽然预测精准度高,但训练与测试时间长. Jaipuria 等^[5]提出离散小波变换(DWT)-人工神经网络(ANN)组合模型用于不确定环境下供应链需求预测,Flunkert 等^[6]提出一种基于自回归递归网络的模型用于亚马逊商城销量预测. Mukherjee 等^[7]提出 ARMDN 神经网络对零售电商进行预测等.

高斯过程(GP)模型是一种普适的机器学习模型. 相对于神经网络等建模简单,孙等^[8]利用 GP 模型预测短期风速,结果表明预测精度及稳定性均高于 BP 网络模型等. 针对上海市多场景下饱和负荷,彭等^[9]提出一种混沌粒子群与高斯过程组合预测模型,具有较高预测精度及一定预测弹性. Li 等^[10]提出一种改进的高斯过程混合模型,即增强高斯过程混合模型(EGPM)用于预测短期电力负荷,预测精度总体上高于 GP 模型.

若时序随时间实时变化且不同时段差异较大,GP 模型的均方连续性导致难于拟合. 而高斯过程混合(GPM)模型是将多个高斯过程模型进行加权线性组合,能进一步提升模型在时序上的分段拟合预测性能^[11]. 本文算法研究具体意义体现于:(1)基于 GPM 模型精准刻画各时段时序差异;(2)提高预测精度及泛化能力.

2 时间序列数据预处理

本文以型号为“DPK901P”政采打印机商品报价数据作为一个时间序列实例,其它时间序列数据处理方式可以参照打印机数据进行. 为探索时间序列数据规律,对随时间变化的数据进行可视化. 很多日期并没有相对应的报价,这是由于数据缺失造成,这种现象在 DPK910P 打印机数据中普遍存在. 为保证数据完整准确性,对其进行空缺值处理.

现通过密度空间含噪聚类(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)筛选出时间序列中的奇点.

为更加直观,将每个报价转换为二维坐标点,如报价1 000元转换后的点为(1 000,1 000). 设置价格核心点在半径 $\xi=100$ 内其它价格点数量 $\text{MinPts} \geq 4$,由图 1 可知,DBSCAN 自动将所有价格划分为 6 个类别. 其中,57.23~215 范围内价格为类别 1;所

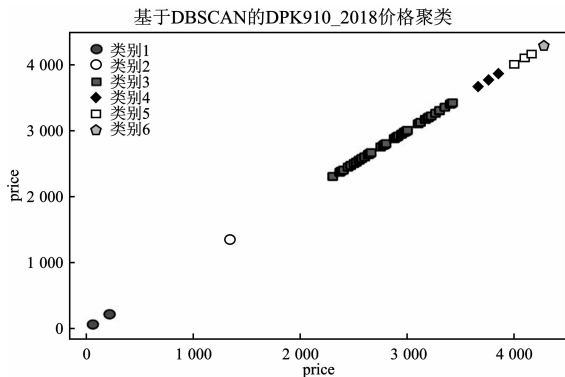


图 1 DBSCAN 价格聚类

有 1349 价格(图中点重叠,只显示出一个)为类别 2;2 300~3 420.27 范围内价格为类别 3;3 662.5~3 850 范围内价格为类别 4;4 000~4 160 范围内价格为类别 5;所有 4 276(理由同上)为类别 6. 将类别 1 归为一类,类别 2~6 归为另外一类,即可将价格奇点筛选出来. 经查询 DPK910P 打印机价格数据,验证价格为 57.23 和 215 元的打印机实质是原始数据中的错误数据.

3 基于机器学习模型的时间序列拟合

3.1 差分自回归移动平均模型拟合

差分自回归移动平均(Autoregressive Integrated Moving Average, ARIMA)模型,应用于此报价时序,即在某一日期报价 x_t 是它若干历史报价 $\{x_{t-1}, x_{t-2}, \dots, x_{t-p}\}$ 与白噪声序列 $\{\xi_t, \xi_{t-1}, \dots, \xi_{t-q}\}$ 的线性组合:

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2} - \dots - \theta_q \xi_{t-q} \quad (1)$$

具体实现步骤如下:1)判定报价时序的平稳性;2)若时序非平稳,进行差分处理使之平稳;3)通过对平稳报价时序自相关系数 ACF 与偏自相关系数 PACF 图分析,选取相应的 p, q 值;4)进行时序拟合,若残差序列的 ACF 与 PACF 均不显著,则进行时序预测.

3.2 支持向量机拟合

在此应用场景下,假定报价时序组成训练样本集为 $\{x_k, y_k\}_{k=1}^d, x_k, y_k$ 为历史报价与目标报价. 支持向量机(Support Vector Machine, SVM)实现步骤如下:1)将 x_k 映射到高维特征空间,得到输入变量 $f(x)$ 与 x_k 之间关系为 $f(x_k) = w^T \varphi(x) + b$,其中 w, b 分别为权重,偏置项;2)迭代训练过程中,使 $f(x_k)$ 与 y_k 之间误差最小即使如式(2)所示的损失

函数最小:

$$J = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \beta \sum_{i=1}^d L(f(x_i), y_i) \quad (2)$$

式中, β 为惩罚系数; ξ 不敏感函数 L 定义为 $L(f(x_i), y_i) = \max(0, |f(x_i) - y_i| - \xi)$; 3) 迭代训练完成之后, 便可对新输入的报价时序进行预测。

3.3 高斯过程模型拟合

针对本文报价时序进行回归预测, GP 模型实现步骤如下:

1) 将报价时序中每一日期对应的报价制作成 N 个训练样本 $\{x_k, t_k\}_{k=1}^N$, 其中 x_k, t_k 分别为历史报价和目标报价。

2) 利用 GP 模型求得回归函数 $f(x)$, 进而使用 $f(x)$ 预测非训练样本中的历史报价 x_{N+1} 的函数值 $f(x_{N+1})$ 。假设 $f(x)$ 是一个高斯过程, 将目标报价 $t_N = (t_1, t_2, \dots, t_N)^T$ 视为另一个高斯过程 $t(x)$ 在历史报价 $X_N = (x_1, x_2, \dots, x_N)^T$ 上的实现, 并且假设 $f(x), t(x)$ 满足 $t(x) = f(x) + e(x)$, $e(x)$ 是方差为 σ_e^2 的高斯白噪声。现将 $f(x)$ 等价于 $f(x) = \sum_{h=1}^H \omega_h \Phi_h(x)$, H 为基函数个数, 随机权重向量 $\mathbf{w} = (\omega_1, \omega_2, \dots, \omega_H)^T$ 满足 $\mathbf{w} \sim N(0, \sigma_w^2 \mathbf{I}_H)$ 。由 $f(x), t(x)$ 间关系得 $t_N \sim N(0, \mathbf{C}_N)$, 矩阵 $\mathbf{C}_N$ 由式(3)协方差函数决定, 其中若 $n = m, \delta_{nn} = 1$, 否则 $\delta_{nn} = 0$ 。

$$C(x_n, x_m) = \sigma_w^2 \sum_{h=1}^H \Phi_h(x_n) \Phi_h(x_m) + \sigma_e^2 \delta_{nn} \quad (3)$$

3) 预测给定历史报价 x_{N+1} 对应的目标报价 t_{N+1} 时, 有 $P(t_{N+1} | t_N) = P(t_{N+1}, t_N) / P(t_N)$, 根据 GP 模型, 式中分子表示式如式(4)所示:

$$P(t_{N+1}, t_N) \propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} t_N^T & t_{N+1} \end{bmatrix} \mathbf{C}_{N+1}^{-1} \begin{bmatrix} t_N \\ t_{N+1} \end{bmatrix} \right\} \quad (4)$$

现将矩阵 $\mathbf{C}_{N+1}$ 分解为子块 k, k^T, κ 等, 进而可得目标报价 t_{N+1} 的预测分布为 $P(t_{N+1} | t_N) \sim N(\hat{t}_{N+1}, \sigma_{\hat{t}_{N+1}}^2)$, 其中 $\hat{t}_{N+1} = k^T \mathbf{C}_N^{-1} t_N, \sigma_{\hat{t}_{N+1}}^2 = \kappa - k^T \mathbf{C}_N^{-1} k$ 。即 $f(x_{N+1}) = \hat{t}_{N+1}$ 为历史报价 x_{N+1} 对应的目标报价。

3.4 高斯过程混合模型拟合

高斯过程混合模型 (Gaussian Process Mixture, GPM) 模型是多个 GP 模型的加权线性混合, 若每一个 GP 模型看作一个专家, 则 GPM 模型也可视为一种特殊混合专家模型。对于 3.3 节的训练样本 $\{x_k, t_k\}_{k=1}^N$, GPM 模型实现步骤如下:

1) 首先, 通常用隐变量 z_k 表示第 k 个历史报价

所属组别, 若 $z_k = c$ 表示第 k 个历史报价组别为 c , 即 x_k 服从第 c 个高斯过程, $c = 1, 2, \dots, C, C$ 为混合的 GP 模型总数。若 c 组的历史报价与目标报价集合分别为 $X_c = \{x_k; z_k = c\}, T_c = \{t_k; z_k = c\}$, 由 GP 模型假设, 则有:

$$T_c | X_c \sim N(u_c(X_c), K_c(X_c, X_c) + \sigma_c^2 \mathbf{I}_N) \quad (5)$$

式中, $u_c(\cdot), K_c(\cdot, \cdot)$ 分别是 c 组的均值函数与协方差函数; σ_c^2 为 c 组的噪声方差; $\mathbf{I}_N$ 是 N 阶单位阵。

2) 基于聚类算法将训练样本分为 C 组, 假定每组样本服从其相应的高斯过程且相互独立, 即:

$$(x_k | z_k = c) \sim N(u_c, S_c) \quad (6)$$

当隐变量与输入的历史报价确定之后, 协方差函数 $K_c(\cdot, \cdot)$ 采用如式(7)所示的高斯核函数:

$$K_c(x_i, x_j) = l_c^2 \exp(-||x_i - x_j||^2 / 2\delta_c^2) \quad (7)$$

由式(6)可得 GPM 模型的输出, 各个高斯过程的差异由 $u_c, S_c, l_c^2, \delta_c^2, \sigma_c^2$ 随 c 变化来体现。对于测试样本 x_{N+1} 而言, 将其放入相对应的第 c 个高斯过程, 借助高斯过程模型输出 $f(x_{N+1})$, 可得目标报价 t_{N+1} 的预测分布。

3.5 拟合结果分析

为探索各类供货商报价变化规律且便于预测, 现基于扩展迪基 - 福勒 (Augment Dickey-Fuller, ADF) 检验结果将报价时序分为常数项时序、非平稳时序以及平稳时序。然后分别对常数项时序、平稳时序与非平稳时序拟合。拟合的报价与真实报价之间的误差通过如式(8)的均方根误差 (RMSE) 来衡量, 其中 $x_i, h(x_i), y_i$ 分别为当日的报价, 拟合的报价以及真实报价。

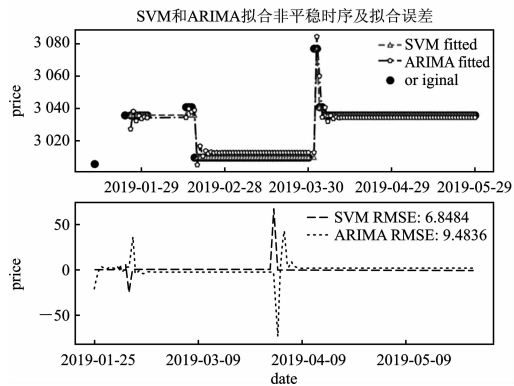
$$RMSE(X, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2} \quad (8)$$

在预测中设置以一星期历史报价来预测第八天的报价。SVM、ARIMA、GPM 模型拟合同一报价非平稳时序结果如图 2 所示, SVM 与 ARIMA 的拟合结果合并展示。ARIMA 先要进行对数变换, 然后使用一阶差分处理, 使报价时序平稳后再进行拟合, 并且还需对拟合结果进行差分及对数还原等处理, 而其它模型只需要调整参数。

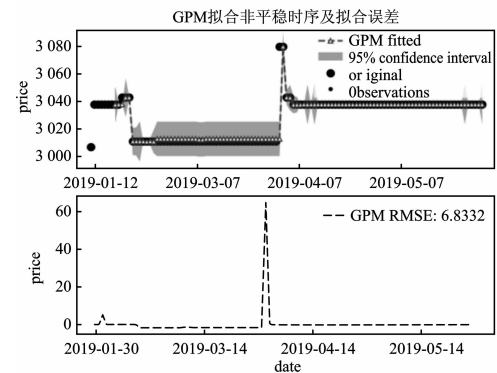
表 1 基于四种模型的两类报价时序拟合 RMSE 对比

拟合模型	平稳时序拟合 RMSE	非平稳时序拟合 RMSE
ARIMA	153.449 1	9.483 6
SVM	109.946 9	6.848 4
GP 模型	97.646 1	6.435 6
GPM 模型	97.460 0	6.433 2

两个 GP 模型加权线性组合.



(a)SVM 与 ARIMA 拟合非平稳时序对比



(b)GPM 拟合非平稳时序

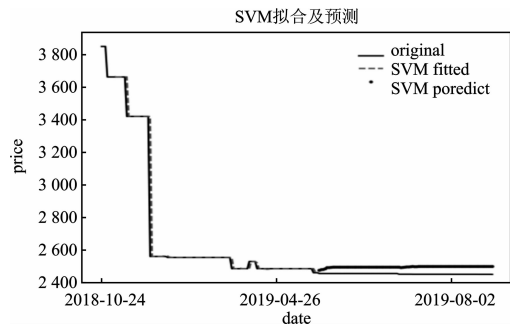
图 2 不同模型拟合非平稳时序对比

从上图可知,与 ARIMA 和 SVM 拟合结果相比,GPM 模型对非平稳时序拟合曲线输出了拟合的置信区间,也即定量给出了拟合的可信程度.从表 1 看出,GPM 模型对此两类报价时序的拟合误差总体小于其它模型,但只略小于 GP 模型.

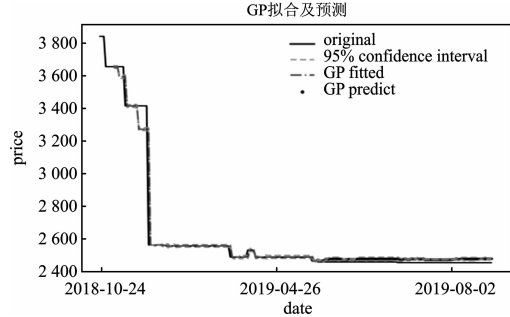
综上,ARIMA 建模复杂,相较于 ARIMA 和 SVM,GP 与 GPM 建模便捷,参数少且易调整,无需平稳性检验,对平稳及非平稳报价时序拟合效果更优.由于选取报价时序实时变化小且不同时间段差异甚微,导致 GPM 模型与 GP 模型拟合 RMSE 相差甚小.对下一节中选取的报价时序进行预测,GPM 模型的优势更能够体现.

4 基于高斯过程混合模型的时间序列预测

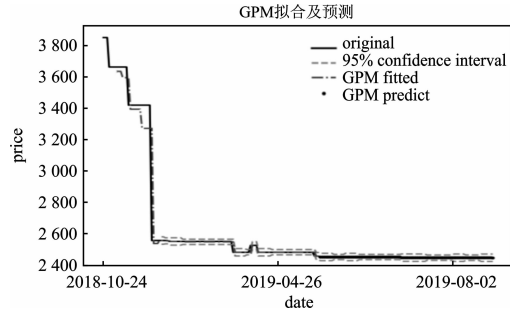
GPM 模型在各类报价时序上均取得了不错的拟合效果.进一步验证其性能,现对“文一朝阳”供货商在广东省政采商城中的报价时序进行预测.此序列时间维度长,报价实时变化且不同时段差异大.图 3 分别展示了 SVM、GP 模型以及 GPM 模型对此序列的拟合及未来 90 天报价预测结果,其中嵌入维度设置为 $d=7$. GPM 模型参数 $C=2$,即 GPM 模型由



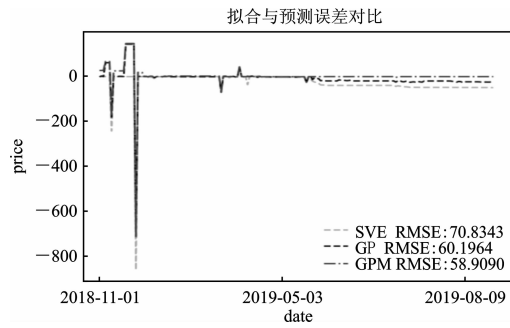
(a)SVM 拟合及预测



(b)GP 模型拟合及预测



(c)GPM 模型拟合及预测



(d)拟合及预测误差对比

图 3 三种模型的拟合与预测结果

由图 3 可知,SVM 拟合及预测并无置信区间,而 GP、GPM 模型输出 95%置信区间. SVM 预测未来报价均高于实际报价,GP 模型预测的报价略高于真实报价,而且在一定范围内波动. GPM 模型预测的报价几乎与此供货商真实报价吻合.由图 3(d)

可知,GPM 拟合及预测的误差均低于 SVM 与 GP 模型. 另外从图中看出,GPM 模型的 RMSE 为 58.909 0,GP 模型的 RMSE 为 60.196 4,SVM 的 RMSE 为 70.834 3.

另外由表 2 可知,当嵌入维度 $d=7,C=2$ 时,SVM 预测 RMSE 为 43.411 0,GP 模型预测 RMSE 为 20.317 9,而 GPM 模型预测 RMSE 只有 0.432 4,此时 GPM 模型的预测精度要大幅高于其它模型.

为定量分析嵌入维 d 对模型预测性能影响,现取嵌入维的 8 组典型值,分别训练三种模型然后进行预测,预测均方误差(RMSE)见表 2. 在训练中 GPM 模型的其它内含参数均取最优值,仅改变 GPM 模型内含的混合成分个数 C .

表 2 不同嵌入维度下预测 RMSE 对比

d	SVM RMSE	GP 模型 RMSE	GPM 模型 ($C=2$)RMSE	GPM 模型 ($C=3$)RMSE	GPM 模型 ($C=4$)RMSE
1	16.506 7	10.218 7	0.560 3	0.560 2	0.560 3
2	17.292 7	11.701 4	0.509 4	3.389 5	3.389 5
3	24.453 1	17.051 7	0.436 3	8.952 0	7.899 1
4	32.127 7	16.124 6	4.072 2	6.156 5	5.665 2
5	34.272 0	14.712 6	6.955 4	7.429 8	7.672 6
6	37.868 0	18.159 7	0.434 3	7.672 6	5.304 5
7	43.411 0	20.317 9	0.432 4	0.432 4	0.432 3
8	48.818 2	21.856 5	0.431 9	0.431 9	0.431 8

不同维度的嵌入值,意味着训练样本来自于报价序列的不同时段. 从表 2 可以看出:对 SVM 而言,预测 RMSE 随着嵌入维度增大而增加. GP 模型预测 RMSE 受嵌入值维度影响较小,基本稳定在 20 左右. 具有不同混合成分的 GPM 模型预测误差相对于其它模型均能达到较小值,这体现了该模型能更精准反映报价时序各时段的差异. 通过对比不同混合成分 GPM 模型的预测均方根误差,可知 GPM 模型的最优混合成分个数为 2.

GPM 模型的预测结果能够准确反映此供货商未来随时间变化的报价规律,其实际意义在于:将报价突变因素考虑在内,若报价高于或低于 1.5 倍的预测报价,则表明“文一朝阳”存在故意过高或者过低报价,对其它供货商而言,此行为属于不公平商业竞争. 现广东省政府采购商城事先预测了“文一朝阳”报价变化趋势,便能阻止发生此种不正当商业竞争. 对采购人而言,购买商品的报价是合理的,其利益也得到了最大保证.

5 结束语

本文实现基于 GPM 模型的时间序列预测算法,以 69 415 条 DPK910P 打印机报价数据为例,对其报价进行预测. 首先进行了报价数据预处理:包括空缺报价处理以及基于 DBSCAN 聚类价格奇点去除. 各供货商在各商城中报价组成的报价时序数量太多,对报价时序基于 ADF 检验结果,将其分为常数项、平稳、非平稳报价时序,以便分类别进行拟合预测及评价. 针对各类报价,先分别以 SVM、ARI-MA、GP 和 GPM 模型进行拟合. 鉴于 GPM 模型强大的数据表达和预测能力,将之用于报价时序预测,结果表明 GPM 模型是一种时间序列预测有效模型,具有较好的实际应用价值. 未来将重点研究时序中突变情形下的算法预测性能.

参考文献:

[1] 许美玲,王依雯. 基于改进差分进化和回声状态网络的时间序列预测研究[J/OL]. 自动化学报, 1-10 [2021-03-16]. <https://DOI.org/j.aas.c180549>. DOI: 10.16383/j.aas.c180549.
XU M L, WANG Y W. Time series prediction based on improved differential evolution and echo state network[J]. Acta Automatica Sinica, 1-10[2021-03-16]. <https://DOI.org/j.aas.c180549>. DOI: 10.16383/j.aas.c180549.

[2] 王鑫,吴际,刘超,等. 基于 LSTM 循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报, 2018, 44(4): 772-784. DOI: 10.13700/j.bh.1001-5965.2017.0285.
WANG X, WU J, LIU C, et al. Exploring LSTM based recurrent neural network for failure time series prediction[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(4): 772-784. DOI: 10.13700/j.bh.1001-5965.2017.0285.

[3] ARUNRAJ N S, AHRENS D, FERNANDES M. Application of SARIMAX model to forecast daily sales in food retail industry[J]. International Journal of Operations Research and Information Systems, 2016, 7(2): 1-21. DOI: 10.4018/IJORIS.2016040101.

[4] TAGHIZADEH E. Utilizing artificial neural networks to predict demand for weather-sensitive products at retail stores[EB/OL]. [2020-04-03]. <https://arxiv.org/abs/1711.08325>.

[5] JAIPURIA S, MAHAPATRA S S. An improved demand forecasting method to reduce bullwhip effect in supply chains[J]. Expert Systems with Applications,

- 2014, 41(5): 2395-2408. DOI: 10.1016/j.eswa.2013.09.038.
- [6] SALINAS D, FLUNKERT V, GASTHAUS J, et al. DeepAR: probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 2020, 36(3): 1181-1191. DOI: 10.1016/j.ijforecast.2019.07.001.
- [7] MUKHERJEE S, SHANKAR D, GHOSH A, et al. ARMDN: associative and recurrent mixture density networks for etail demand forecasting [EB/OL]. [2020-04-03]. <https://arxiv.org/abs/1803.03800>
- [8] 孙斌, 姚海涛, 刘婷. 基于高斯过程回归的短期风速预测[J]. *中国电机工程学报*, 2012, 32(29): 104-109. DOI: 10.13334/j.0258-8013.pcsee.2012.29.016.
- SUN B, YAO H T, LIU T. Short-term wind speed forecasting based on Gaussian process regression model [J]. *Proceedings of the CSEE*, 2012, 32(29): 104-109. DOI: 10.13334/j.0258-8013.pcsee.2012.29.016.
- [9] 彭虹桥, 顾洁, 胡玉, 等. 基于混沌粒子群-高斯过程回归的饱和负荷概率预测模型[J]. *电力系统自动化*, 2017, 41(21): 25-32. DOI: 10.7500/AEPS20170119007.
- PENG H Q, GU J, HU Y, et al. Forecasting model for saturated load based on chaotic particle swarm optimization-Gaussian process regression [J]. *Automation of Electric Power Systems*, 2017, 41(21): 25-32. DOI: 10.7500/AEPS20170119007.
- [10] LI L L, SUN J, WANG C H, et al. Enhanced Gaussian process mixture model for short-term electric load forecasting [J]. *Information Sciences*, 2019, 477: 386-398.
- [11] 周亚同, 陈子一, 马尽文. 从高斯过程到高斯过程混合模型: 研究与展望 [J]. *信号处理*, 2016, 32(8): 960-972. DOI: 10.16798/j.issn.1003-0530.2016.08.11.
- ZHOU Y T, CHEN Z Y, MA J W. From Gaussian processes to the mixture of Gaussian processes: a survey [J]. *Journal of Signal Processing*, 2016, 32(8): 960-972. DOI: 10.16798/j.issn.1003-0530.2016.08.11.

作者简介:

刘亚波 男, (1994-), 硕士研究生. 研究方向为机器学习.

吴秋轩(通讯作者) 男, (1978-), 博士, 副教授. 研究方向为仿生机器人建模与控制 and 机器学习.

E-mail: wuqx@hdu.edu.cn