基于误差截尾假设的时序预测可学习性理论与算法

张绍群 张钊钰 姜 远 周志华
(南京大学计算机软件新技术国家重点实验室 南京 210023)

摘 要 在收集和处理时间序列数据的过程中,难免会产生误差,而在很多现实情形中误差是自相关非独立的.已有的预测理论在分析误差自相关的时序数据时,往往需要知道预测算法所输出假设空间的显式表达,而对于一些假设空间不明确的模型,比如神经网络,尚未有系统的求解方法和理论保障来分析其在非平稳且误差自相关时序数据上的预测能力.本文基于误差截尾的假设,提出了时间序列的预测 PAC 可学习理论,并给出了数据依赖情形下的泛化误差界.该界限包含一个时序复杂度度量和一个差异度量,前者描述了序列数据的非平稳性,后者可在适当情形下从数据中估计得到.因此,该误差界并不依赖于假设空间的显式表达,具有较强的普适性.根据上述理论,本文提出了一种基于自回归模型的交替优化算法用于预测非平稳的时间序列数据.我们在真实数据集上进行实验,验证了本文提出算法的有效性.

关键词 机器学习;时间序列分析;自相关误差;预测 PAC 可学习性;差异估计;交替优化
中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2022.02279

Time Series Theory and Algorithm of Predictable Learnability Based on Error Truncation Assumption

ZHANG Shao-Qun ZHANG Zhao-Yu JIANG Yuan ZHOU Zhi-Hua
(*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023)

Abstract In collecting and processing time series data, there are inevitably errors, and in many real-world cases, the errors are autocorrelated and non-independent. Existing methods often rely on the explicit form of the hypothesis space corresponding to the forecasting algorithm. In contrast, there has been no systematic paradigm and guarantee for some models with ambiguous hypothesis space, such as neural networks, to analyze their predictive ability on non-stationary and error autocorrelated time series data. Based on the assumption that errors are autocorrelated and truncated, this paper proposes the predictable PAC learning theory and correspondingly presents the data-dependent learning bound. The bound contains a measure of sequence complexity and a discrepancy; the former indicates the inherent nonstationarity of the concerned time series, and the latter can be estimated from the data under mild assumptions. According to the theoretical results above, we propose an autoregressive model-based alternating optimization algorithm for forecasting non-stationary time series data. The experiments conducted on several real-world data sets confirm the effectiveness of our proposed algorithm.

Keywords machine learning; time series analysis; autocorrelated errors; predictable PAC learnability; discrepancy estimation; alternate optimization

收稿日期:2022-03-10;在线发布日期:2022-07-29.本课题得到国家自然科学基金(62176116,61921006)与科技创新 2030 项目(2020AAA-0109401)资助.**张绍群**,博士,助理教授,主要研究方向为机器学习与数据挖掘.E-mail:zhangsq@lamda.nju.edu.cn.**张钊钰**,硕士,主要研究方向为机器学习与数据挖掘.**姜 远**,博士,教授,主要研究领域为机器学习与数据挖掘.**周志华**,博士,教授,主要研究领域为机器学习与数据挖掘.

1 引 言

近年来,时间序列预测在诸多领域得到了广泛的关注和应用,包括医疗健康[1-3]、天气预报[4]、交通预测[5-6]、量化交易[7-8]等.在预处理和建模时间序列数据的过程中,难免会存在误差,且由于数据变周期[9]、非平稳[10]、影响变量收集不充分[11]、欠拟合等因素的影响,在很多现实情形中误差是自相关非独立的,即当前时刻的误差与先前时刻的误差相关.该领域以往有一些相关工作,比如具有自相关误差的极大似然回归算法[12-13]和具有自相关误差的非参数估计[14]等,往往需要知道预测算法所输出假设空间的显式表达,而对于一些假设空间不明确的模型,比如神经网络,尚未有系统的求解方法和时序理论来分析其在非平稳且误差自相关时序数据上的预测能力.因此,当误差自相关非独立时,如何设计性能良好且有理论保障的时序预测算法是机器学习中非常重要的研究课题.

需明确的是,尽管误差自相关,我们仍需对误差的相关性或者分布有进一步假设;否则若对误差分布一无所知,甚至允许任意变化,那么这样的问题显然不可学.一个自然而然的想法是假设误差的自相关性是截尾的,即当前时刻的误差只与最近有限个时刻的误差相关.本文以 Vanilla RNN 预测股票价格的任务为例说明该假设的普适性(实验细节详见第 5 节).图 1 展示了 400 次迭代训练内误差的偏自相关(Partial Auto-Correlation Function, PACF)图.据观察,预测模型的误差呈现出明显的自相关性,其 PACF 随着训练迭代而变得稳定且表现出 2 阶截尾(2nd order truncation).

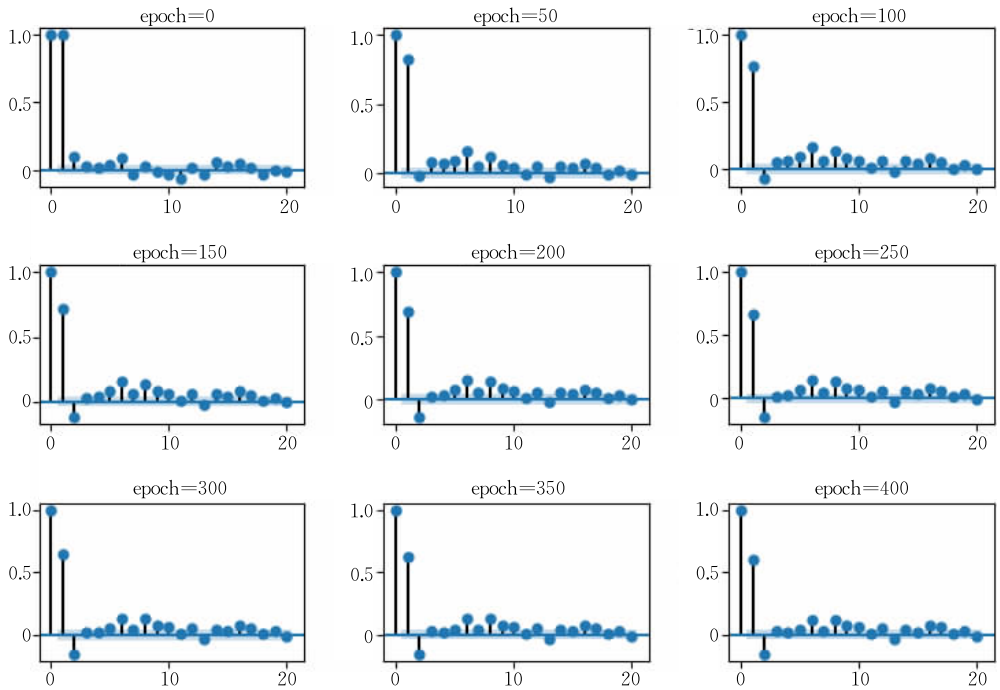

图 1 Vanilla RNN 模型的误差 PACF 图

基于误差截尾的假设,本文提出了时间序列的预测 PAC 可学习理论,并给出了数据依赖情形下的泛化误差界.该界限包含一个序列复杂度度量和一个差异度量;前者描述了序列数据的非平稳性,后者可在适当假设下从数据中估计得到.因此,该误差界并不依赖于预测函数的显式表达,可用于神经网络等模型,具有较强的普适性.根据上述理论,本文提出了一种基于自回归模型的交替优化算法用于预测非平稳的时间序列数据.我们在真实数据集上进行实验,验证了本文提出算法的有效性.

本文第 2 节介绍相关工作;第 3 节和第 4 节分别提出预测 PAC 可学习理论和基于自回归模型的交替优化算法;第 5 节通过在真实数据集上进行实验,验证了算法的有效性;第 6 节总结全文并展望未来工作.

2 相关工作

根据克拉默分解定理(Cramer Decomposition Theorem)[15],当误差随时间独立时,误差序列不可

预测. 然而在很多现实情形中误差是自相关非独立的，这使得基于独立性假设的高斯-马尔可夫定理(Gaussian-Markov Theorem)不再适用. 以最小二乘法为例，当误差非独立时，估计系数的方差增加而标准误差被低估. 因此，模型的预测准确性会降低，并且许多指标在实际上不再具有统计意义. 因此，当误差自相关非独立时，如何提高预测模型的性能则成为时间序列分析中重要的研究问题之一. 组合预测(combination forecasting)是工程上处理该问题的经典方法之一. 该方法的基本原理是建立各种子模型用以模拟不同类型序列的行为，然后组合在一起进行优化和预测，这样组合模型在最终选择时通常不需要估计很多参数就可以做到较好的性能. 比较代表性的工作，如 Taylor 和 Letham[9] 将 Facebook 数据分解为四种模式，即多重强季节性、趋势变化、异常值和假日效应，然后对这四种模式分别建模，最后组合成完整的预测模型. 一些研究[16-17] 表明与单独应用每种预测方法相比，组合预测可以提高预测准确性，并且复合预测集可以产生比任何原始预测都低的均方误差. 混合模型(hybrid model)是另一种广受好评的时间序列预测方法，其在 M-Competitions[18] 和 IEEE-CIS Competition[19] 等主要比赛中取得了巨大成功. 该方法依赖于预处理过程，比如指数平滑、差分(常用于统计)、经验模态分解[20] 和重整化[21] 等，可以消除波动和噪声，并且在模型的最终选择中通常不需要估计很多超参数. 在实际应用中，设计和采用何种预处理技术在很大程度上取决于研究人员的经验. 因而，混合模型仍属于实验密集型的研究.

在线性模型中考虑自相关误差已被广泛研究，比较代表性的工作有具有自相关误差的极大似然回归算法[12-13] 和具有自相关误差的非参数估计[14] 等. 这类方法首先需要利用一些统计方法，如德宾-沃森统计量(Durbin-Watson statistic)[22]、AIC 和 BIC[23] 等，检测一阶自相关误差的存在性并对模型定阶，然后可用类似于极大似然估计的框架求解. 整个算法流程可受传统的统计时间序列分析和预测 PAC 可学习理论支持. 值得注意的是，这类方法必须知道预测算法所输出的预测函数或假设空间的显示表达，而对于一些预测函数表达未知的模型，比如神经网络，尚未有系统的求解方法和时序理论来分析其在非平稳时序数据上的预测能力.

3　预测可学习性理论

本节提出了基于误差截尾假设的时间序列的预测 PAC 可学习理论，包括第 3.1 节中介绍并分析了误差截尾假设、第 3.2 节中数据依赖情形下的泛化误差界以及第 3.3 节中关于样本分布和真实分布之间差异的估计. 在此之前，介绍一些相关概念和基本知识.

令 i 表示虚数单位元，其满足 $\mathrm{i}=\sqrt{-1}$. 令 $\mathcal{B}$ 表示后移算子(backshift operator)，使得 $\mathcal{B}x_t=x_{t-1}$. 令 λ 为算子 $\mathcal{B}$ 的特征值. 对于整数 $N>0$，记 $[N]=\{1,2,\cdots,N\}$. 令 $\boldsymbol{z}\geqslant 0$ 表示对于任意 $i\in[T]$，其对应元素满足 $z_i\geq 0$. 用 $\mathcal{M}_{\boldsymbol{z}}(\chi,\mathcal{H},\boldsymbol{\sigma})$ 表示序列 $\boldsymbol{z}=(z_1,\cdots,z_T)$ 在函数类 $\mathcal{H}$ 下的最小序列 χ-覆盖数(minimal sequence χ-covering number)[24-25]，其中二元指示器 $\boldsymbol{\sigma}\in\{-1,+1\}^D$.

令 $X_{-T}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)\in\mathbb{R}^{D\times T}$ 是由 D 个变量在 T 个时间戳上生成的多变量时间序列，其中 $\boldsymbol{x}_t\in\mathbb{R}^{D\times 1}$ 且 $t\in[T]$. 给定观测数据 X_{-t}，我们的目标是预测 h 步超前值 $\boldsymbol{x}_{t+h}$，其中 $h\in\mathbb{N}^+$. 方便起见，将示例集记为 $\{(X_{-t},\boldsymbol{y}_t):\boldsymbol{y}_t=\boldsymbol{x}_{t+h}\}$. 令 $f(\cdot;\Theta)$ 表示预测模型，其中 Θ 表示模型参数，$e_t=\boldsymbol{y}_t-f(X_{-t};\Theta)$ 表示模型 f 在 t 时刻的误差. 给定 $p\in\mathbb{N}^+$，记 $\boldsymbol{e}_t=(e_{t-p},e_{t-p+1},\cdots,e_{t-1})$.

对于任意假设 $\hbar\in\mathcal{H}$ 和序列 $\{\boldsymbol{Z}_t\}$，定义

(1) 平均经验误差(averaged empirical error)

$$\hat{L}_a(\hbar)=\frac{1}{T}\sum_{t=1}^{T-1}\ell(\hbar(\boldsymbol{Z}_{-t}),\boldsymbol{Z}_{t+1}),$$

(2) 路径依赖经验误差(path-dependent empirical error)

$$\hat{L}_p(\hbar)=\frac{1}{T}\sum_{t=1}^{T-1}\gamma_t\ell(\hbar(\boldsymbol{Z}_{-t}),\boldsymbol{Z}_{t+1}),$$

(3) 泛化误差(generalization error)

$$\hat{L}(\hbar)=\mathbb{E}_{Z_{t+1}\in\hat{\boldsymbol{z}}}[\ell(\hbar(\boldsymbol{Z}_{-t}),\boldsymbol{Z}_{t+1})],$$

其中 ℓ 是损失函数，γ_t's 是实数标量，$\hat{\boldsymbol{Z}}$ 表示测试集.

定义 1.　我们称学习算法 $\mathcal{L}$ 对时间序列 X_{-T} 所诱导的随机过程是预测概率近似正确可学习的(Predictive Probably Approximately Correct learning learnable，简称为预测 PAC 可学习的)，若对于 $0<\epsilon,\delta<1$，存在学习算法 $\mathcal{L}$ 和多项式函数 ploy(…)，使得对于目标概念 c 和任何 $T\geq\mathrm{ploy}(1/\epsilon,1/\delta,D,\mathrm{size}(\boldsymbol{c}))$，学习算法 $\mathcal{L}$ 所输出的假设(hypothesis) $\hbar\in\mathcal{H}$ 满足 $\mathbb{P}(\mathbb{E}(\hbar)\leq\epsilon)\geq 1-\delta$. 进一步，如果存在学习算法 $\mathcal{L}$ 对模型 f 的误差序列是预测 PAC 可学习的，则称模型 f 在时间序列 X_{-T} 上是可提升的(improvable).

3.1　误差截尾假设

观察图 1 易知，误差序列的 PACF 呈现出二阶

截尾.受此启发,本文考虑一个更通用的假设形式,即误差序列的自相关系数存在 p 阶截尾,如下

$$e_t=\alpha_1 e_{t-1}+\alpha_2 e_{t-2}+\cdots+\alpha_p e_{t-p}+\epsilon_t \tag{1}$$

其中 $p\in\mathbb{N}^+$,$\alpha_i(i\in[p])$ 是回归系数,$\epsilon_t\in\mathcal{N}(0,\Sigma)$ 且 $\Sigma>0$.注意,式(1)也是一个 p 阶微分方程.因此,易知该方程的解由两部分组成,其一是其齐次形式(homogeneous form,即 $\epsilon_t=0$)的通解 $e_t^{\mathcal{G}}$,其二是其非齐次形式(non-homogeneous form,即 $\epsilon_t\neq0$)的一个特解.对于 $\epsilon_t=0$ 的情况,有

$$e_t-\alpha_1 e_{t-1}-\alpha_2 e_{t-2}-\cdots-\alpha_p e_{t-p}=0 \tag{2}$$

将后移算子 $\mathcal{B}$ 代入式(2),可得到如下特征方程(characteristic equation)

$$\lambda^p-\alpha_1\lambda^{p-1}-\alpha_2\lambda^{p-2}-\cdots-\alpha_{p-1}\lambda-\alpha_p=0 \tag{3}$$

因为式(3)是一个 p 阶的非齐次实值方程.所以该方程有 p 个非零解,依次记为 $\lambda_1,\lambda_2,\cdots,\lambda_p$.凭此,方程(2)的通解 $e_t^{\mathcal{G}}$ 由如下三种基本情况及其变种组成:

① 当 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 为不同的实根,即 $\lambda_1\neq\lambda_2\neq\cdots\neq\lambda_p\in\mathbb{R}$ 时,方程(2)的解为

$$e_t^{\mathcal{G}}=c_1\lambda_1^t+c_2\lambda_2^t+\cdots+c_p\lambda_p^t,$$

其中 $c_1,c_2,\cdots,c_p\in\mathbb{R}$.

② 当 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 中存在 d 重实根,即 $\lambda_1=\cdots=\lambda_d\neq\lambda_{d+1}\neq\cdots\neq\lambda_p\in\mathbb{R}$ 时,方程(2)的解为

$$e_t^{\mathcal{G}}=(c_1+c_2t+\cdots+c_dt^{d-1})\lambda_1^t+c_{d+1}\lambda_{d+1}^t+\cdots+c_p\lambda_p^t,$$

其中 $c_1,c_2,\cdots,c_p\in\mathbb{R}$.

③ 当 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 中存在复根时,由于 $\alpha_1,\alpha_2,\cdots,\alpha_p\in\mathbb{R}$,可以断定复数根是以共轭对(conjugate pairs)的形式出现.因此,只需要类比实重根的情形,如

$$\lambda_1=re^{i\omega},\ \lambda_2=re^{-i\omega}(r=|\lambda_1|)$$

为方程(3)的一对复重根,且重数为 m 时,方程(2)的解为

$$e_t^{\mathcal{G}}=\sum_{i=1}^{m}r^t(c_1e^{i\omega t}+c_2e^{-i\omega t}),$$

其中 $c_1,c_2\in\mathbb{R}$.综上,方程(1)的解为

$$\begin{aligned}e_t&=e_t^{\mathcal{G}}+e_t^{\mathcal{S}}\\&=\sum_{k=1}^{d'}r_k^t(a_ke^{i\omega_k}+b_ke^{-i\omega_k})+\sum_{j=n'+1}^{d''}c_j\lambda_j^t+\\&\quad\sum_{i=1}^{d}\Big(\sum_{i'=1}^{n_i}c_{i,i'}t^{i'-1}\Big)\lambda_i^t+e_t^{\mathcal{S}}\end{aligned} \tag{4}$$

式(4)中有 d 组实数重根,每组中有 $n_i(i\in[d])$ 个元素;d' 对复根($r_k^te^{i\omega_k}$,$r_k^te^{-i\omega_k}$);d'' 个非重根的实根.这里 $\{a_k,b_k,c_{i,i'},c_j\}$ 都是实数标量,且有 $\sum_{i=1}^{d}n_i+2d'+d''=p$.

3.2 预测 PAC 可学习性

基于上节中的误差截尾假设,本文给出第一个主要定理,如下.

定理 1. 如果对于 $i\in[p]$ 有 $|\lambda_i|<1$,那么模型 f 在时间序列 X_{-T} 上是可提升的.

根据定义 1 可知,要完成定理 1 的证明,只需证明式(4)中的误差序列 $\{e_t\}_{t\in[T]}$ 是预测 PAC 可学习的即可.本文介绍下列引理来证明该结论.

引理 1. 如果对于 $i\in[p]$ 有 $|\lambda_i|<1$,那么式(4)定义的序列 $\{e_t\}_{t\in[T]}$ 服从一个弱依赖过程(weak-dependence process).

证明. 令 $\mathcal{D}_s^t$ 表示序列 $\{Z_s,Z_{s+1},\cdots,Z_t\}$ 的分布,其中 $0\leq s<t$.定义如下的系数[26]

$$\beta(s)=\sup_t\mathbb{E}_{Z_{-t}}[\|\mathcal{D}_{t+s}^{+\infty}(\cdot|Z_{-t})-\mathcal{D}_{t+s}^{+\infty}(\cdot)\|_\mu],$$

其中 $\mathcal{D}(\cdot|\cdot)$ 表示条件概率分布,μ 表示概率测度或者事件集 $\mathcal{G}$ 上的 σ-代数(σ-algebra),其对于两个概率分布 P 和 Q 满足

$$\|P-Q\|_\mu=\sup_{z\in\mathcal{G}}|P(z)-Q(z)|.$$

根据式(4),有

$$\begin{aligned}e_t^{\mathcal{G}}&=\sum_{k=1}^{d'}r_k^t(a_ke^{i\omega_k}+b_ke^{-i\omega_k})+\sum_{j=n'+1}^{d''}c_j^t\lambda_j+\\&\quad\sum_{i=1}^{d}\Big(\sum_{i'=1}^{n_i}c_{i,i'}t^{i'-1}\Big)\lambda_i^t.\end{aligned}$$

如果对于 $i\in[p]$ 有 $|\lambda_i|<1$,则显然有

$$e_t^{\mathcal{G}}\to0\ \text{ 随着 }\ t\to\infty.$$

根据克拉默分解定理[15],当 t 充分大时,有 $e_t\in\mathcal{N}(0,\sigma^2)$.令 $Z_t=e_t$,则

$$\beta(s)\to0\ \text{ 随着 }\ s\to+\infty.$$

因此,式(4)中的序列 $\{e_t\}_{t\in[T]}$ 是 β-混合的(β-mixing),即服从一个弱依赖过程. 证毕.

引理 2. 如果对于 $i\in[p]$ 有 $|\lambda_i|<1$,那么对于 $\delta>0$,$\delta'=\delta-q\beta(k)>0$,$q=\lceil N/k\rceil$ 和任意假设 $h\in\mathcal{H}$,以至少 $1-\delta>0$ 的概率成立

$$\begin{aligned}\mathbb{L}(h)\leq\hat{\mathbb{L}}_a(h)+C_\ell\sqrt{\frac{\log(2/\delta')}{8q}}+\\\frac{2}{k}\sum_{i=1}^{2k}\mathfrak{R}_q(\mathcal{E}_i)+\frac{2}{T}\sum_{t=1}^{T}d(t,T+1)\end{aligned} \tag{5}$$

其中,$\mathbb{L}(h)$ 和 $\hat{\mathbb{L}}_a(h)$ 分别表示泛化误差和平均经验误差,C_ℓ 表示损失函数 ℓ 的上界(即对于任意 $h\in\mathcal{H}$ 有 $|\ell(h,\cdot)|\leq C_\ell$).泛化误差界(5)的第二项是每个子集 $\mathcal{E}_i$ 的 Rademacher 复杂度的总和

$$\mathfrak{R}_q(\mathcal{E}_i)=\frac{1}{q}\mathbb{E}\Big[\sup_{h\in H}\sum_{j:\boldsymbol{x}_j\in\mathcal{E}_i}\sigma_j\ell(h(\boldsymbol{e}_j),e_j)\Big],$$

其中 σ_j's 是 Rademacher 随机变量,$\{\mathcal{E}_i\}$ 是对训练集

的一种划分 $\mathcal{E}_i=\{e_{i(q-1)+1},e_{i(q-1)+2},\cdots,e_{iq}\}$. 泛化误差界(5)的第三项测量了不同时间戳 s 和 t 上两个概率分布之间的差异

$$d(s,t)=\sup_{\hbar\in H}\left|\mathbb{E}_{e_s}[\ell(\hbar(\boldsymbol{e}_s),e_s)]-\mathbb{E}_{e_t}[\ell(\hbar(\boldsymbol{e}_t),e_t)]\right|.$$

值得注意的是,引理 2 中的 $d(s,t)$ 描述了 s 时刻预测产生的误差分布与 t 时刻预测产生的误差分布之间的差异,进而式(5)中泛化误差界的第三项 $2/T\sum_{t=1}^{T}d(t,T+1)$ 表示了各个观测时刻预测产生的误差分布与测试时刻 $(T+1)$ 预测产生的误差分布之间的差异和. 该差异和描述了学习器利用已观测数据对测试数据进行预测时的困难程度,这也是学习者在处理非平稳时序数据时可能会面临的挑战. 该引理是引理 1,文献[27]中定理 2 和文献[28]中定理 1 的直接推导. 其证明略.

引理 3.　如果对于 $i\in[p]$ 有 $|\lambda_i|<1$,那么对于 $\delta>0,C>0,\chi>0$ 和任意假设 $\hbar\in\mathcal{H}$,以至少 $1-\delta>0$ 的概率成立

$$\mathbb{L}(\hbar)\leq\hat{\mathbb{L}}_p(\hbar)+\mathfrak{R}_p+C_\ell\|\boldsymbol{\gamma}\|_2\sqrt{2\log\left(\frac{\mathbb{E}_{\boldsymbol{\sigma}}[\mathcal{M}_e(\chi,\mathcal{H},\boldsymbol{\sigma})]}{\delta}\right)},$$

其中,$\mathbb{L}(\hbar)$ 和 $\hat{\mathbb{L}}_p(\hbar)$ 分别是泛化误差和路径依赖经验误差,C_ℓ 表示损失函数 ℓ 的上界(即对于任意 $\hbar\in\mathcal{H}$ 有 $|\ell(\hbar,\cdot)|\leq C_\ell$),$\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_T)$,$\boldsymbol{\sigma}$ 是 Rademacher 随机变量,且

$$\mathfrak{R}_p=\sup_{\hbar\in\mathcal{H}}\left[\mathbb{E}_{e_{T+1}\in\hat{E}}[\ell(\hbar(\boldsymbol{e}_{-(T+1)}),e_{T+1})]-\frac{1}{T}\left(\sum_{t=2}^{T}\gamma_t\mathbb{E}_{e_t\in E}[\ell(\hbar(\boldsymbol{e}_{-t}),e_t)]+C\right)\right].$$

此处 E 和 $\hat{E}$ 分别表示训练集和测试集.

引理 2 和引理 3 分别利用不同的经验误差提供了预测误差序列 $\{e_t\}_{t\in[T]}$ 时的泛化误差界. 该结论指明了,由序列 $\{e_t\}_{t\in[T]}$ 诱导的随机过程是可预测 PAC 可学习的,因为给定 $|\lambda_i|<1(i\in[p])$,则 $\{e_t\}_{t\in[T]}$ 服从一个混合随机过程且具备 p 阶 PACF. 当给定如下引理时,引理 3 可由文献[29]中的定理 1 直接推得.

引理 4.　若 $|\lambda_i|<1(i\in[p])$,则对于任意 $\delta>0$ 和 $C>0$,以至少 $1-\delta>0$ 的概率成立

$$\mathfrak{R}_p\leq\sup_{\hbar\in\mathcal{H}}\left[\sum_{t=2}^{T}(\gamma_t^u-\gamma_t)\mathbb{E}_t[\ell(\hbar(\boldsymbol{e}_{-t+1}),e_{t+1})]\right]+2C+C_\ell\left[\frac{p+1}{2}\beta(1)+\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^u\|_2\sqrt{2\log\left(\frac{2\mathbb{E}_{\boldsymbol{\sigma}}[\mathcal{M}_e(\chi,\mathcal{H},\boldsymbol{\sigma})]}{\delta}\right)}\right],$$

其中向量 $\boldsymbol{\gamma}^u=(\gamma_1^u,\cdots,\gamma_T^u)$ 服从某个均匀分布,它对应于平均经验误差.

证明.　根据三角不等式,可得

$$\mathfrak{R}_p\leq\mathfrak{R}_1+\mathfrak{R}_2 \tag{6}$$

其中

$$\mathfrak{R}_1=\sup_{\hbar\in\mathcal{H}}\left[\mathbb{E}_{T+1}[\ell(\hbar(\boldsymbol{e}_{-(T+1)}),e_{T+1})]-\frac{1}{p}\left(\sum_{t=T-p+1}^{T}\mathbb{E}_t[\ell(\hbar(\boldsymbol{e}_{-t}),e_t)]\right)\right],$$

和

$$\mathfrak{R}_2=\sup_{\hbar\in\mathcal{H}}\left[\frac{1}{p}\left(\sum_{t=T-p+1}^{T}\mathbb{E}_t[\ell(\hbar(\boldsymbol{e}_{-t}),e_t)]\right)-\frac{1}{T}\left(\sum_{t=2}^{T}\gamma_t\mathbb{E}_t[\ell(\hbar(\boldsymbol{e}_{-t}),e_t)]+C\right)\right].$$

对于第一项 $\mathfrak{R}_1$,有

$$\mathfrak{R}_1\leq\frac{1}{p}\sum_{t=T-p+1}^{T}\sup_{\hbar\in\mathcal{H}}\Delta_{1,T}(t)\leq\frac{1}{p}\sum_{t=T-p+1}^{T}\beta_{1,T}(t)\leq\frac{1}{p}\frac{p(p+1)}{2}C_\ell\beta(1)=C_\ell(p+1)\beta(1)/2 \tag{7}$$

其中,

$$\Delta_{1,T}(t)=\left[\mathbb{E}_{T+1}[\ell(\hbar(\boldsymbol{e}_{-(T+1)}),e_{T+1})]-\mathbb{E}_t[\ell(\hbar(\boldsymbol{e}_{-t}),e_t)]\right]$$

和

$$\beta_{1,T}(t)=\sup_{\hbar\in\mathcal{H}}\|\mathcal{D}_{T+1}^{+\infty}(\cdot|\boldsymbol{e}_{-(T+1)})-\mathcal{D}_{t+1}^{+\infty}(\cdot|\boldsymbol{e}_{-t})\|_\mu.$$

式(7)的第三个不等式成立于 $|\lambda_i|<1(i\in[p])$ 和任意 $s=T-p+1,\cdots,T$. 这说明 $\{e_t\}_{t\in[T]}$ 对应的随机过程具有较小程度的非平稳性. 换句话说,后 p 个观测值的分布接近于 $T+1$ 时刻处的真实分布. 从另一方面来看,$\mathfrak{R}_2$ 显然可以从数据中计算得到. 根据文献[29]中的定理 3,对于任意 $\delta>0$ 和 $C>0$,以至少 $1-\delta>0$ 的概率成立

$$\sup_{\hbar\in\mathcal{H}}\left[\sum_{t=2}^{T}(\gamma_t^u-\gamma_t)\mathbb{E}_t[\ell(\hbar(\boldsymbol{e}_{-t+1}),e_{t+1})]\right]\leq\sup_{\hbar\in\mathcal{H}}\left[\sum_{t=2}^{T}(\gamma_t^u-\gamma_t)\ell(\hbar(\boldsymbol{e}_{-t+1}),e_{t+1})\right]+2C+C_\ell\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^u\|_2\sqrt{2\log\left(\frac{2\mathbb{E}_{\boldsymbol{\sigma}}[\mathcal{M}_e(\chi,\mathcal{H},\boldsymbol{\sigma})]}{\delta}\right)} \tag{8}$$

将式(7)和式(8)代入不等式(6),即可得证.　证毕.

3.3　差异估计

上节给出了数据依赖情形下的泛化误差界,该界限包含一个序列复杂度度量和一个差异度量. 本小节将针对具体问题并在适当假设下对该差异度量进行估计. 我们采用简单的自回归模型 AR(p)作为误差序列 $\{e_t\}_{t\in[T]}$ 的预测模型,则有如下结论.

定理 2. 给定均方误差(square loss)ℓ和参数为$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_p)$的 AR($p$)预测器$\hbar$,如果对于任意$i\in[p]$,有$|\alpha_i|<1$和$\|\boldsymbol{\alpha}\|_2\leq M$,那么对于任意$\delta>0$,$0<\|\boldsymbol{\gamma}^u-\boldsymbol{\gamma}\|_1\leq 1$和假设$\hbar$,以至少$1-\delta$的概率成立

$$\mathbb{E}[\hbar(\boldsymbol{e}_T;\boldsymbol{\alpha})-e_{T+1}\,|\,\mathcal{D}_1^T]\leq \sum_{t=1}^{T}(\hbar(\boldsymbol{e}_t;\boldsymbol{\alpha})-e_{t+1})^2+\mathfrak{R}_p+C_\delta \tag{9}$$

其中

$$C_\delta=\mathcal{O}\left(\log^3 T\sqrt{\log_2\frac{2}{\|\boldsymbol{\gamma}^u-\boldsymbol{\gamma}\|_1}}\left(\frac{M}{\sqrt{T}}+\|\boldsymbol{\gamma}^u-\boldsymbol{\gamma}\|_1\right)\right).$$

定理 2 对定理 1 的结论提供了一个更具体的分析.为了证明该定理,只需界定住下列顺序 Rademacher 复杂度(sequential Rademacher complexity)即可

$$\mathfrak{R}_p^{seq}=\sup_{\boldsymbol{\gamma}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\hbar\in\mathcal{H}}\sum_{t=1}^{T}\sigma_t\gamma_t\ell(\hbar(\boldsymbol{e}_t;\boldsymbol{\alpha}),e_t)\right],$$

其中$\hbar(\boldsymbol{e}_t;\boldsymbol{\alpha})=\sum_{i=1}^{p}\alpha_i e_{t-i}$.其证明略.

沿用 Kuznetsov 和 Mohri[30]的思路,定理 2 的结论可用于指导如下优化问题的求解:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}}\sum_{t=1}^{T}\gamma_t\ell(t)+\mu_e\sum_{t=1}^{T}\gamma_t\Delta_t+\mu_\alpha\|\boldsymbol{\alpha}\|^2 \tag{10}$$

其中$\mu_e,\mu_\alpha\in\mathbb{R}$,$\|\boldsymbol{\alpha}\|^2$是参数$\boldsymbol{\alpha}$的正则化项,损失函数采用均方误差$\ell(t)=(\hbar(\boldsymbol{e}_t;\boldsymbol{\alpha})-e_t)^2$,以及$\sum_{t=1}^{T}\gamma_t\Delta_t$界定住了经验差异(empirical discrepancy)

$$\Delta_t=\sup_{\boldsymbol{\alpha}}\left|\frac{1}{t}\sum_{s=1}^{t}\gamma_s^u\ell(s)-(\hbar(\boldsymbol{e}_t;\boldsymbol{\alpha})-e_s)^2\right|.$$

对于任意$t\in[T]$,Δ_t可以使用 DC-规划法(DC-Programming)[31]预先求得.假设限制$\boldsymbol{\gamma}\geqslant 0$以确保该优化问题的凸性,则问题(10)可由一个两阶段子优化过程求得,如下

$$\min_{\boldsymbol{\alpha}}\sum_{t=1}^{T}\gamma_t^*(\hbar(\boldsymbol{e}_t;\boldsymbol{\alpha})-e_t)^2+\mu_\alpha\|\boldsymbol{\alpha}\|^2,$$

其中,$\boldsymbol{\gamma}^*=(\gamma_1^*,\gamma_2^*,\cdots,\gamma_T^*)$且

$$\boldsymbol{\gamma}^*=\arg\min_{\boldsymbol{\gamma}\geqslant 0}\left\{\sup_{\boldsymbol{\alpha}}\left(\sum_{t=1}^{T}(\gamma_t-\gamma_t^u)\ell(t)\right)\right\}.$$

4 基于自回归模型的交替优化算法

根据上述分析,本节正式提出基于自回归模型的交替优化算法.方便起见,这里仅考虑一阶 PACF 的简单情况,即$e_t=\alpha e_{t-1}+\epsilon_t$.那么,预测函数可表示为

$$\boldsymbol{y}_t=f(X_{-t};\Theta)+\alpha(\boldsymbol{y}_{t-1}-f(X_{-(t-1)};\Theta))+\epsilon_t \tag{11}$$

将式(11)代入优化问题(10),可得到一个最小化带正则化项的平方损失问题

$$\min_{\alpha,\Theta}\sum_{t}[\boldsymbol{y}_t-f(X_{-t})-\alpha(\boldsymbol{y}_{t-1}-f(X_{-(t-1)}))]^2+\mu_\alpha\mathcal{R}(\alpha)+\mu_\Theta\mathcal{R}(\Theta) \tag{12}$$

这里我们将预测模型$f(X_{-t};\Theta)$简写为$f(X_{-t})$,其中$\mathcal{R}(\alpha)$和$\mathcal{R}(\Theta)$分别是参数α和Θ的正则化器,μ_α和μ_Θ是相对应的正则化系数.注意,如果预测模型f是神经网络,则$\mathcal{R}(\Theta)$通常表示复杂正则化项[32-33],而$\mathcal{R}(\alpha)$防止α过大.本文推荐使用一些反双曲函数(inverse hyperbolic function)作为$\mathcal{R}(\alpha)$,例如$\mathcal{R}(\alpha)=\operatorname{artanh}(\eta\alpha)$.其中$\eta$是一个比例系数.沿用 Yu 等人[34]中的思路,可以采用交替优化的方法求解优化问题(12).这样,每个交替步骤都简化为一些众所周知的方法.具体过程如下:

(1)初始化.从一个预定义的分布,比如均匀分布$\mathcal{U}[-1,1]$,中采样$\hat{\alpha}$.

(2)更新Θ.通过求解下式来更新$\hat{\Theta}$

$$\arg\min_{\Theta}\sum_{t}\mathcal{L}_\Theta(t)+\mu_\Theta\mathcal{R}(\Theta),$$

其中$\boldsymbol{Y}$表示监督信号,且

$$\mathcal{L}_\Theta(t)=[\boldsymbol{y}_t-f(X_{-t})-\hat{\alpha}(\boldsymbol{y}_{t-1}-f(X_{-(t-1)}))]^2.$$

本文采用了 Adam 算法[35]来加速求解过程.

(3)更新α.给定$\hat{\Theta}$,通过求解式(13)来求得新参数$\hat{\alpha}$

$$\arg\min_{\alpha}\sum_{t}\mathcal{L}_\alpha(t)+\mu_\alpha\mathcal{R}(\alpha) \tag{13}$$

其中$\boldsymbol{Y}$表示监督信号$\hat{f}(X_{-t})=f(X_{-t};\hat{\Theta})$,且

$$\mathcal{L}_\alpha(t)=[\boldsymbol{y}_t-\hat{f}(X_{-t})-\alpha(\boldsymbol{y}_{t-1}-\hat{f}(X_{-(t-1)}))]^2.$$

特殊地,根据高斯-马尔科夫定理,当$\mu_\alpha=0$时,优化问题(13)存在如下闭式解

$$\hat{\alpha}=\frac{\sum_{t=2}^{T}[\boldsymbol{y}_t-\hat{f}(X_{-t})][\boldsymbol{y}_{t-1}-\hat{f}(X_{-(t-1)})]}{\sum_{t=2}^{T}[\boldsymbol{y}_{t-1}-\hat{f}(X_{-(t-1)})]^2}.$$

5 实 验

本节在三个数据集上测试了基于自回归模型的优化算法,其目的是验证该方法的有效性以及与几种实用技术的兼容性.

数据集.(1)雅虎财经(Yahoo! Finance)[36]的股票价格,其包含 2007 年至 2016 年 10 个板块中 50 只股票的每日开盘价,本文选择每个板块市值前

5 名的公司.该数据具有高噪音非平稳的特点；(2)柏林天气数据①，其记录德国柏林从 1995 年到 2004 年的每日和每月的天气数据.天气数据含周期性，具有高噪音近似平稳的特点；(3)盐城上牌量数据②，其包括近 1000 天 5 个汽车品牌的每日汽车上牌量.本文只考虑 5 个汽车品牌的汽车上牌量总数，不关心其具体日期信息，从而该数据集构成一个真实的单变量时间序列.该数据是低噪音且非平稳的.

对比方案. 采用循环神经网络(Recurrent Neural Network，RNN)、时间卷积网络(Temporal Convolution Network，TCN)[37]、门循环单元(Gate Recurrent Unit，GRU)[38]、LSTM[39]、DSANet[40] 和 FTNet[41-42] 作为基础模型.所有的基础模型使用 Adam 优化器配置 0.001 的学习率进行优化，且皆在 300 次迭代内收敛.上标 d 和 es 分别表示时间序列预测中两种实用的预处理技术[43]，一阶差分(first-order differencing)和指数光滑(exponential smoothing)，而本文所提出的方法采用下标 α 标记.本文还添加一步延迟序列(one-step-delayed sequence)作为基准线，即 $\hat{y}_t=y_{t-1}(t\in[T])$，记为 Naive.详细的网络配置参数被列举在表 1 中，其中 DSANet 采用自定义配置[44].

表 1　基础模型的配置

模型	雅虎财经	盐城汽车上牌量	柏林天气	学习率	训练轮数
RNN	5×64×1	1×64×1	1×64×1	0.001	300
GRU	5×64×1	1×64×1	1×64×1	0.001	300
LSTM	5×64×1	1×64×1	1×64×1	0.001	300
FTNet	5×64×1	1×64×1	1×64×1	0.001	300
DSANet	—	—	—	0.001	300
TCN	7×10×10×10×1	5×10×10×10×1	7×10×10×10×1	0.001	300

定制化. 整个数据集按照 60%：20%：20%的样本比划为三部分：训练集、验证集和测试集.在验证集上测试各种配置，并挑选出具有最佳验证性能的配置对测试集进行预测.本文采用 4 个评估指标评估预测性能：令 $\hat{y}_t$ 和 y_t 分别表示预测值和监督信号，定义

(1)均方误差(Mean Squared Error，MSE)，

$$MSE=\Big(\sum_{t=T+1}^{T+S}(\hat{y}_t-y_t)^2\Big)/S.$$

(2)平均绝对误差(Mean Absolute Error，MAE)，

$$MAE=\Big(\sum_{t=T+1}^{T+S}|\hat{y}_t-y_t|\Big)/S.$$

(3)混淆精度(Confusion Accuracy，CA)，

$$CA=\frac{TP+TN}{TP+TN+FP+FN},$$

其中 TP，FP，TN 和 FN 分别表示超前一步增量的真正例(True Positives)、假正例(False Positives)、真反例(True Negatives)和假反例(False Negatives).易知，CA 可以更加细致地评价模型对时序趋势走向(+/−)的预测精度：CA 值越高表示模型对时序趋势走向预测越准确；反之，预测精度越低.

假设“预测误差是否独立(Independent Prediction Errors，IPE)”被用于估计预测误差 $\hat{y}_t-y_t$ 是否随时间独立.反馈“√”表示预测误差独立；否则，“×”表示预测误差与时间相关.该指标可以用来检测预测误差是否独立.根据克拉默分解定理，若误差独立，则说明预测模型已达到最好预测性能.

最后，本文使用组合指标符$[MSE\pm std, MAE\pm std, CA, IPE?]$来评估预测性能.

实验结果. 表 2 列出了各对比方案和本文所提出的方法在雅虎财经股票价格、盐城汽车上牌量和柏林天气预测任务上的性能.在每个数据集上，均进行 10 次测试以统计评价指标的平均值以及标准偏差.每个数据集上每个评估指标的最佳值以粗体显示，并采用后缀 •/◦标注 95%显著性水平的成对 t-检验意义下，本文方法明显优于/劣于对比方法的情况.图 2 展示了本文方法 RNN_α 在雅虎财经股票数据上预测误差的 PACF 图(400 轮训练迭代).通过图表，易观察到如下几个事实：(1)使用本文方法训练的模型都实现了独立的预测误差，这验证了本文方法的有效性；(2)对比原始的基础模型，本文方法基本上在所有数据集上取得了(四个评价指标上)显着的提升，这验证了本文方法的优越性；(3)本文方法可与其他两种实用的预处理技术兼容，其性能在大多数数据集和评价指标上未表现出下降；(4)指数光滑预处理和本文方法的结合方案在大多数数据集和评价指标上取得了最佳，而且其性能对基础模型鲁棒.综上，本节采用对比实验验证了本文方法的有效性和优越性，并展示了它与几种实用的预处理技术的兼容性.

讨论. 本文的实验部分考虑了数据集的多样性，即有高噪音非平稳的金融数据，低噪音非平稳的上牌量数据，以及高噪音近似平稳的天气数据(含周期性).在面对不同类型的时序数据时，混合模型往往需要“定制化”预处理过程以保证误差序列独立.

① https://github.com/blackeye735/Wind-Speed-Prediction
② https://tianchi.aliyun.com/competition/entrance/231641/information

表 2 对比方法在雅虎财经股票价格、盐城汽车上牌量、柏林天气预测任务上的性能

数据集	对比模型	*MSE*	*MAE*	*CA*	*IPE*?	本文方法	*MSE*	*MAE*	*CA*	*IPE*?	•/◦
雅虎财经	Naive	6.3518	0.9595	0.5210	0/50*						
	RNN	416.5±10.35	4.278±0.1450	0.5063±0.0014	0/50	RNN_{α}	6.545±0.0130	0.9589±0.0081	0.5127±0.0009	49/50	•
	RNN^{d}	6.295±0.0139	0.9609±0.0015	0.5415±0.0010	50/50	RNN_{α}^{d}	6.273±0.0051	0.9570±0.0005	0.5422±0.0010	50/50	•
	RNN^{es}	402.0±5.169	4.163±0.1120	0.5185±0.0421	0/50	RNN_{α}^{es}	6.273±0.0037	0.9573±0.0011	0.5365±0.0078	50/50	•
	GRU	1277±81.85	7.897±0.1040	0.5129±0.0024	0/50	GRU_{α}	6.334±0.0083	0.9562±0.0071	0.5160±0.0021	48/50	•
	GRU^{d}	6.278±0.0041	0.9600±0.0026	0.5386±0.0018	50/50	GRU_{α}^{d}	6.269±0.0121	0.9554±0.0005	0.5413±0.0012	50/50	
	GRU^{es}	1269±67.43	7.732±0.0972	0.5270±0.0083	0/50	GRU_{α}^{es}	6.938±0.0181	0.9883±0.0112	0.5333±0.0095	50/50	•
	LSTM	3506±118.2	17.95±1.428	0.5145±0.0026	0/50	$LSTM_{\alpha}$	6.327±0.0091	0.9570±0.009	0.5152±0.0027	47/50	•
	$LSTM^{d}$	6.286±0.008	0.9614±0.0006	0.5409±0.0009	50/50	$LSTM_{\alpha}^{d}$	6.254±0.0121	0.9559±0.0010	0.5410±0.0012	50/50	•
	$LSTM^{es}$	3499±86.46	16.88±1.317	0.5407±0.0016	0/50	$LSTM_{\alpha}^{es}$	6.940±0.0244	0.9551±0.0036	0.5202±0.0046	50/50	•
	FTNet	656.9±99.52	4.278±0.2021	0.5324±0.0014	2/50	$FTNet_{\alpha}$	6.989±0.0250	0.9585±0.0090	0.5331±0.0029	45/50	•
	$FTNet^{d}$	6.313±0.0041	0.9557±0.0006	0.5414±0.0011	50/50	$FTNet_{\alpha}^{d}$	6.321±0.012	0.9578±0.0007	0.5412±0.0013	50/50	◦
	$FTNet^{es}$	507.4±72.31	4.978±0.2133	0.5125±0.0014	4/50	$FTNet_{\alpha}^{es}$	**6.163±0.0171**	0.9727±0.0011	0.5372±0.0005	50/50	•
	TCN	3937±135.5	21.84±1.321	0.5086±0.0010	0/50	TCN_{α}	6.978±0.5760	0.9651±0.0050	0.5212±0.0032	49/50	•
	TCN^{d}	6.291±0.0168	**0.9546±0.0008**	0.5334±0.0015	50/50	TCN_{α}^{d}	6.331±0.0117	0.9635±0.0008	0.5353±0.0013	50/50	◦
	TCN^{es}	2874±97.27	17.33±1.107	0.5120±0.0009	0/50	TCN_{α}^{es}	0.6271±0.0109	0.9614±0.0011	0.5433±0.0012	50/50	•
	DSANet	3260±215.8	15.65±2.792	0.5238±0.0003	0/50	$DSANet_{\alpha}$	17.93±1.172	1.445±0.0930	**0.5603±0.0018**	27/50	•
	$DSANet^{d}$	6.849±0.065	0.9837±0.0019	0.5433±0.0012	50/50	$DSANet_{\alpha}^{d}$	6.699±0.042	0.9765±0.0039	0.5456±0.0008	50/50	
	$DSANet^{es}$	3194±224.1	15.47±2.152	0.5198±0.0013	0/50	$DSANet_{\alpha}^{es}$	13.79±0.9583	1.127±0.1045	0.5324±0.0013	50/50	•
数据集	对比模型	*MSE*(1e5)	*MAE*(1e2)	*CA*	*IPE*?	对比模型	*MSE*(1e5)	*MAE*(1e2)	*CA*	*IPE*?	
盐城汽车上牌量	Naive	43.04	16.49	0.5080	×						
	RNN	6.671±0.4752	5.031±0.1021	0.9245±0.0128	×	RNN_{α}	6.018±0.2810	4.932±0.0530	0.9287±0.0024	√	•
	RNN^{d}	9.817±1.168	5.600±0.0352	0.9011±0.0226	×	RNN_{α}^{d}	6.721±0.2620	5.104±0.0152	0.9287±0.0024	√	•
	RNN^{es}	7.692±0.0345	5.872±0.0099	0.8749±0.0091	×	RNN_{α}^{es}	6.659±0.0318	5.023±0.0073	0.9538±0.0045	√	•
	GRU	5.594±0.1710	4.836±0.0340	0.9315±0.0046	√	GRU_{α}	5.329±0.1973	4.721±0.0420	0.9385±0.0068	√	•
	GRU^{d}	6.129±0.1553	4.949±0.0350	0.9266±0.0041	√	GRU_{α}^{d}	6.1120±0.2624	4.828±0.0151	0.9477±0.0044	√	•
	GRU^{es}	5.646±0.4481	4.921±0.0029	0.9559±0.0129	√	GRU_{α}^{es}	5.6245±0.0546	**4.707±0.0134**	0.9547±0.0079	√	•
	LSTM	5.594±0.1710	4.801±0.0461	0.9251±0.0044	√	$LSTM_{\alpha}$	5.534±0.1724	4.713±0.0330	0.9301±0.0115	√	•
	$LSTM^{d}$	6.039±0.2853	4.888±0.0050	0.9333±0.0126	√	$LSTM_{\alpha}^{d}$	6.050±0.1986	4.783±0.0062	0.9344±0.0095	√	
	$LSTM^{es}$	5.628±0.1792	4.821±0.0210	0.9582±0.0120	√	$LSTM_{\alpha}^{es}$	5.638±0.0156	4.772±0.0213	**0.9614±0.0135**	√	•
	FTNet	5.928±0.0490	4.907±0.0071	0.9251±0.0090	√	$FTNet_{\alpha}$	5.855±0.1121	4.973±0.0091	0.9351±0.0075	√	•
	$FTNet^{d}$	6.537±0.3080	5.113±0.0580	0.9262±0.0147	√	$FTNet_{\alpha}^{d}$	6.475±0.2971	5.013±0.0970	0.9329±0.0035	√	•
	$FTNet^{es}$	5.724±0.0232	4.833±0.0283	0.9517±0.0037	√	$FTNet_{\alpha}^{es}$	5.671±0.0371	4.827±0.0391	0.9513±0.0028	√	•
	TCN	5.777±0.4781	5.145±0.0233	0.9259±0.0061	√	TCN_{α}	5.482±0.1511	4.956±0.0121	0.9259±0.0061	√	•
	TCN^{d}	6.111±0.2553	5.198±0.0140	0.9233±0.0146	√	TCN_{α}^{d}	6.046±0.2270	5.281±0.0154	0.9455±0.0088	√	•
	TCN^{es}	5.821±0.3412	5.023±0.0121	0.9344±0.0097	√	TCN_{α}^{es}	5.512±0.0837	4.920±0.0120	0.9415±0.0021	√	•
	DSANet	5.388±0.1690	4.936±0.0171	0.9296±0.0120	√	$DSANet_{\alpha}$	5.152±0.2590	4.753±0.0110	0.9396±0.0058	√	•
	$DSANet^{d}$	6.234±0.1809	4.911±0.0110	0.9222±0.0111	√	$DSANet_{\alpha}^{d}$	5.964±0.2470	4.917±0.0180	0.9400±0.0078	√	
	$DSANet^{es}$	5.211±0.1923	4.831±0.0092	0.9322±0.0087	√	$DSANet_{\alpha}^{es}$	**5.179±0.2310**	4.778±0.0109	0.9406±0.0037	√	•
数字集	对比模型	*MSE*	*MAE*	*CA*	*IPE*?	对比模型	*MSE*	*MAE*	*CA*	*IPE*?	
柏林天气	Naive	4.624	1.65	0.5104	×						
	RNN	4.618±0.1009	1.632±0.0078	0.5405±0.0103	√	RNN_{α}	4.462±0.2350	1.624±0.0060	0.5627±0.0165	√	
	RNN^{d}	4.323±0.0022	1.613±0.0020	0.5728±0.0036	√	RNN_{α}^{d}	4.317±0.0099	1.613±0.0040	0.5739±0.0040	√	•
	RNN^{es}	4.527±0.0847	1.620±0.0219	0.5499±0.0039	√	RNN_{α}^{es}	4.497±0.2147	1.632±0.0120	0.5534±0.0159	√	
	GRU	4.541±0.0670	1.604±0.0128	0.5361±0.0121	√	GRU_{α}	4.301±0.1180	**1.584±0.0102**	0.5712±0.0194	√	•
	GRU^{d}	4.339±0.0120	1.617±0.0029	0.5607±0.0031	√	GRU_{α}^{d}	4.339±0.0101	1.614±0.0040	0.5615±0.0041	√	•
	GRU^{es}	4.584±0.0914	1.622±0.0020	0.5422±0.0095	√	GRU_{α}^{es}	4.334±0.0246	1.619±0.0071	0.5600±0.0045	√	•
	LSTM	4.727±0.1277	1.632±0.0150	0.5391±0.0028	×	$LSTM_{\alpha}$	4.357±0.0350	1.619±0.0140	0.5594±0.0038	√	•
	$LSTM^{d}$	4.352±0.0221	1.618±0.0149	0.5631±0.0024	√	$LSTM_{\alpha}^{d}$	4.351±0.0230	1.618±0.0090	0.5624±0.0025	√	•
	$LSTM^{es}$	4.739±0.1239	1.647±0.0252	0.5353±0.0064	√	$LSTM_{\alpha}^{es}$	4.421±0.0805	1.620±0.0012	0.5556±0.0052	√	•
	FTNet	4.340±0.0060	1.618±0.0028	0.5876±0.0143	√	$FTNet_{\alpha}$	4.335±0.0310	1.616±0.0011	0.5863±0.0136	√	•
	$FTNet^{d}$	4.351±0.0269	1.619±0.0030	0.5744±0.0045	√	$FCNTNet_{\alpha}^{d}$	4.349±0.0330	1.620±0.0080	0.5753±0.0036	√	•
	$FTNet^{es}$	4.489±0.0958	1.612±0.0170	0.5932±0.0129	√	$FTNet_{\alpha}^{es}$	4.237±0.0011	1.622±0.0071	**0.6023±0.0011**	√	
	TCN	4.466±0.0709	1.629±0.0084	0.5473±0.0079	√	TCN_{α}	4.438±0.0492	1.626±0.0101	0.5692±0.0071	√	•
	TCN^{d}	4.298±0.0190	1.623±0.0074	0.5530±0.0096	√	TCN_{α}^{d}	4.287±0.0140	1.620±0.0048	0.5546±0.0054	√	•
	TCN^{es}	4.260±0.0228	1.633±0.0069	0.5318±0.0119	√	TCN_{α}^{es}	4.228±0.0460	1.598±0.0062	0.5513±0.0151	√	•
	DSANet	4.556±0.0820	1.664±0.0120	0.5630±0.0083	√	$DSANet_{\alpha}$	4.510±0.0560	1.665±0.0079	0.5634±0.0019	√	•
	$DSANet^{d}$	4.443±0.0940	1.639±0.0161	0.5458±0.0185	√	$DSANet_{\alpha}^{d}$	4.432±0.0698	1.638±0.0120	0.5618±0.0151	√	•
	$DSANet^{es}$	4.214±0.0491	1.617±0.0211	0.5620±0.0133	√	$DSANet_{\alpha}^{es}$	**4.200±0.0368**	1.596±0.0132	0.5606±0.0190	√	•

注：带*表示存在独立误差的股票数量占股票总数(50)的数量比例.

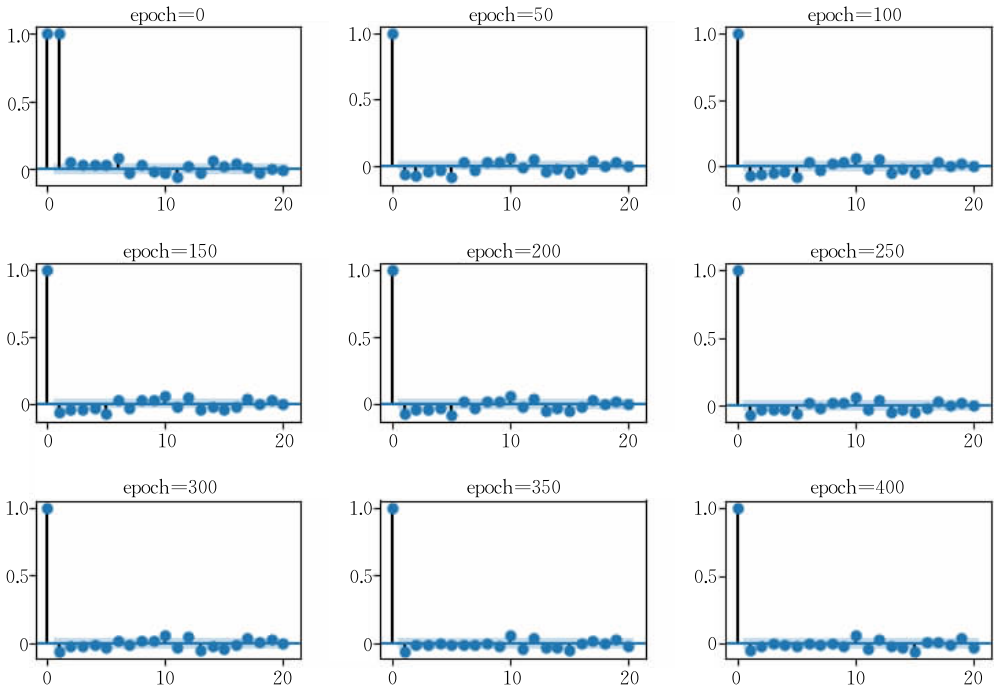

图 2　本文方法 RNN_α 的误差 PACF 图

差分等预处理技术可以有效去噪，但是通常会损失一部分有用的信息，在统计学习中往往需要各种检测方法以确定差分阶数. 而本文所提方法可自适应地应对误差自相关且截尾的情况. 以式(11)为例，$\alpha=0$即可处理误差独立的情况；$\alpha=1$，即等价于结合了一阶差分技术的预测模型；当自适应学习参数 α 时，在功能上可以视为自适应学习对原始序列的预处理，参数 α 表示了预处理中的光滑程度. 此外，式(11)中的预测函数仅考虑了自相关误差一阶截尾的情况，本文所提算法也可以推广到高阶截尾的情况. 较大的阶数往往意味着误差序列存在较长的相关性和拖尾性，而在实际应用中，使用者通常无法提前获知该信息. 我们在做了大量实验后，建议阶数取 1 或者 2 即可. 当然，也可以将阶数 p 被设定为超参数，然后采用零阶优化的方法，比如网格搜索(grid search)或者贝叶斯估计(Bayes Estimation)等，来学习超参数 p，但是这些工作并非本文的重点.

6　结　语

本文基于误差截尾的假设，对时间序列的预测可学习性提出了新的理论分析，并给出了数据依赖情形下的泛化误差界. 该界限包含一个序列复杂度度量和一个差异度量，前者描述了序列数据的非平稳性，后者可在适当假设下从数据中估计得到. 该误差界并不依赖于预测函数或者假设空间的显示表达，具有较强的普适性. 根据该理论结果，本文提出一种基于自回归模型的交替优化算法用于预测非平稳的时序数据. 通过在多个真实数据集上的实验，验证了该算法的有效性.

如何从理论上刻画时间序列的预测可学习性质是时间序列分析和机器学习理论中一个重要的研究课题. 本文在仅考虑了误差截尾的情形下，讨论如何设计可有效提升预测模型性能的算法以及计算其与数据分布相关的泛化误差界. 但仍有许多问题亟待进一步研究. 未来，我们可能会关心误差序列拖尾的情况以及算法相关的泛化性等问题.

参　考　文　献

[1] Chimmula V, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 2020, 135: 109864

[2] Shi Z, Zuo W, Liang S, et al. IDDSAM: An integrated disease diagnosis and severity assessment model for intensive care units. IEEE Access, 2020, 8: 15423-15435

[3] Song H, Rajan D, Thiagarajan J, Spanias A. Attend and diagnose: Clinical time series analysis using attention models// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, USA, 2018: 4091-4098

[4] Angryk R, Martens P, Aydin B, et al. Multivariate time series dataset for space weather data analytics. Scientific Data, 2020, 7(1): 1-13

[5] Che Z, Purushotham S, Cho K, et al. Recurrent neural

networks for multivariate time series with missing values. Scientific Reports, 2018, 8(1): 1-12

[6] Tang X, Yao H, Sun Y, et al. Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 5956-5963

[7] Sezer O B, Gudelek, M U, Ozbayoglu A M. Financial time series forecasting with deep learning: A systematic literature review: 2005—2019. Applied Soft Computing, 2020, 90: 106181

[8] Xue J, Zhou S, Liu Q, et al. Financial time series prediction using l2, 1RF-ELM. Neurocomputing, 2018, 277: 176-186

[9] Taylor S, Letham B. Forecasting at scale. The American Statistician, 2018, 72(1): 37-45

[10] Zhang S-Q, Zhou Z-H. ARISE: ApeRIodic SEmi-parametric process for efficient markets without periodogram and Gaussianity assumptions. arXiv:2111.06222, 2021

[11] Sun F-K, Chris L, Duane B. Adjusting for autocorrelated errors in neural networks for time series//Advances in Neural Information Processing Systems 34, Online. 2021: 29806-29819

[12] Beach C M, MacKinnon J G. A maximum likelihood procedure for regression with autocorrelated errors. Econometrica, 1978, 46: 51-58

[13] Frydman R. A proof of the consistency of maximum likelihood estimators of nonlinear regression models with autocorrelated errors. Econometrica, 1980, 48: 853-860

[14] Xiao Z, Oliver B L, Raymond J C, Enno M. More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. Journal of the American Statistical Association, 2003, 98(464): 980-992

[15] Cramér H. On some classes of nonstationary stochastic processes//Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 1961: 57-78

[16] Bates J, Granger C. The combination of forecasts. Journal of the Operational Research Society, 1969, 20(4): 451-468

[17] Lichtendahl K, Winkler R. Why do some combinations perform better than others? International Journal of Forecasting, 2020, 36(1): 142-149

[18] Godahewa R, Bergmeir C, Webb G I, et al. Monash time series forecasting archive. arXiv:2105.06643, 2021

[19] Isaac T. IEEE-CIS technical challenge on energy prediction from smart meter data. https://dx.doi.org/10.21227/2npg-c280, 2020

[20] Huang N, Shen Z, Long S, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903-995

[21] Zamparo M, Baldovin F, Caraglio M, Stella A. Scaling symmetry, renormalization, and time series modeling: The case of financial assets dynamics. Physical Review E, 2013, 88(6): 062808

[22] Durbin J, Watson G S. Testing for serial correlation in least squares regression. Biometrika, 1951, 38: 159-177

[23] Box G E P, Gwilym M J, Gregory C R, Greta M L. Time Series Analysis: Forecasting and Control. New York, USA: John Wiley & Sons, 2015

[24] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning. Chapter 3.5. Cambridge, USA: MIT Press, 2018

[25] Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Chapter 27. Cambridge, UK: Cambridge University Press, 2014

[26] Doukhan P. Mixing: Properties and Examples. Germany: Springer Science & Business Media, 2012

[27] Zhang S-Q, Zhou Z-H. Harmonic recurrent process for time series forecasting//Proceedings of the 24th European Conference on Artificial Intelligence. Santiago de Compostela, Spain, 2020: 1714-1721

[28] Kuznetsov V, Mohri M. Generalization bounds for time series prediction with non-stationary processes//Proceedings of the 25th International Conference on Algorithmic Learning Theory. Ljubljana, Slovenia, 2014: 260-274

[29] Kuznetsov V, Mohri M. Learning theory and algorithms for forecasting non-stationary time series//Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 541-549

[30] Kuznetsov V, Mohri M. Forecasting non-stationary time series: From theory to algorithms//Advances in Neural Information Processing Systems 29, NIPS Tutoria. Barcelona, Spain, 2016

[31] Le-Thi H, Dinh T. DC programming and DCA: Thirty years of developments. Mathematical Programming, 2018, 169(1): 5-68

[32] Barron A. Complexity regularization with application to artificial neural networks. Nonparametric Functional Estimation and Related Topics, 1991: 561-576

[33] Girosi F, Jones M, Poggio T. Regularization theory and neural networks architectures. Neural Computation, 1995, 7(2): 219-269

[34] Yu H-F, Rao N, Dhillon I. Temporal regularized matrix factorization for high-dimensional time series prediction. Advances in Neural Information Processing Systems 29, Barcelona, Spain, 2016: 847-855

[35] Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014

[36] Zhang L, Aggarwal C, Qi G-J. Stock price prediction via discovering multi-frequency trading patterns//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 2141-2149

[37] Bai S, Kolter J, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271, 2018

[38] Cho K, Van-Merrienboer B, Gülçehre Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734

[39] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[40] Huang S, Wang D, Wu X, Tang A. DSANet: Dual self-attention network for multivariate time series forecasting// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China, 2019: 2129-2132

[41] Zhang S-Q, Zhou Z-H. Flexible transmitter network. Neural Computation, 2021, 33(11): 2951-2970

[42] Wu J-H, Zhang S-Q, Jiang Y, Zhou Z-H. Towards theoretical understanding of flexible transmitter networks via approximation and local minima. arXiv:2111.06027, 2021

[43] Smyl S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting, 2020, 36(1): 75-85

[44] Chen J, Yuan Z, Peng J, et al. DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 1194-1206

ZHANG Shao-Qun, Ph.D., assistant professor. His current research interests mainly include machine learning and data mining.

ZHANG Zhao-Yu, M. S. His research interests include machine learning and data mining.

JIANG Yuan, Ph. D., professor. Her research interests mainly include machine learning and data mining.

ZHOU Zhi-Hua, Ph. D., professor. His research interests mainly include machine learning and data mining.

Background

Theoretically investigating the predictability or predictable learnability of machine learning models for handling time series is an important research topic confronted in current learning theory. Conventional approaches often assume that errors at different times are independent of each other and then solve such forecasting problems using the maximum likelihood estimation framework. However, the errors sometimes are auto-correlated and not independent, which dissatisfies the prior usage of the traditional Gauss-Markov theorem. Taking the least-squares method as an example, when the errors are not independent, the variance of the estimated coefficients increases, and the standard error is underestimated. Therefore, its prediction accuracy will be reduced, and many statistics are no longer significant in practice. Therefore, when the errors are auto-correlated, how to design a time series prediction algorithm with good performance and theoretical guarantee is a significant research topic in machine learning.

Some efforts on this issue often rely on the explicit form of the hypothesis space corresponding to the forecasting algorithm. In contrast, there has been no systematic paradigm and guarantee for some models with ambiguous hypothesis space, such as neural networks, to analyze their predictive ability on non-stationary and error auto-correlated time series data. Based on the assumption that errors are auto-correlated and truncated, this paper proposes the predictable PAC learning theory and correspondingly presents the data-dependent learning bound. The bound contains a measure of sequence complexity and a discrepancy; the former indicates the inherent non-stationarity of the concerned time series, and the latter can be estimated from the data under mild assumptions. We also propose an auto-regressive model-based alternating optimization algorithm for forecasting non-stationary time series data. The experiments conducted on several real-world data sets confirm the effectiveness of our proposed algorithm.