时间序列预测方法综述

杨海民[1] 潘志松[2] 白 玮[2]
(陆军工程大学研究生院 南京 210007)[1] (陆军工程大学指挥控制工程学院 南京 210007)[2]

摘 要 时间序列是按照时间排序的一组随机变量,它通常是在相等间隔的时间段内依照给定的采样率对某种潜在过程进行观测的结果。时间序列数据本质上反映的是某个或者某些随机变量随时间不断变化的趋势,而时间序列预测方法的核心就是从数据中挖掘出这种规律,并利用其对将来的数据做出估计。针对时间序列预测方法,着重介绍了传统的时间序列预测方法、基于机器学习的时间序列预测方法和基于参数模型的在线时间序列预测方法,并对未来的研究方向进行了进一步的展望。

关键词 时间序列,时间序列预测,机器学习,在线学习
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.004

Review of Time Series Prediction Methods

YANG Hai-min[1] PAN Zhi-song[2] BAI Wei[2]
(School of Graduate, Army Engineering University of PLA, Nanjing 210007, China)[1]
(College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China)[2]

Abstract Time series is a set of random variables ordered in timestamp. It is often the observation of an underlying process, in which values are collected from uniformly spaced time instants, according to a given sampling rate. Time series data essentially reflects the trend that one or some random variables change with time. The core of time series prediction is mining the rule from data and making use of it to estimate future data. This paper emphatically introduced a summary of time series prediction method, namely the traditional time series prediction method, machine learning based time series prediction method and online time series prediction method based on parameter model, and further prospected the future research direction.

Keywords Time series, Time series prediction, Machine learning, Online learning

1 引言

通俗地说,时间序列是按照时间排序的一组随机变量,它通常是在相等间隔的时间段内,依照给定的采样率对某种潜在过程进行观测的结果[1]。时间序列数据通常是一系列实值型数据,用 $X_1, X_2, X_3, \cdots, X_t, X_t \in \mathbb{R}\ (t \in Z)$ 表示时间。现实生活中,在一系列时间点上观测数据是司空见惯的活动,在农业、商业、气象、军事和医疗等研究领域都包含大量的时间序列数据。总之,目前时间序列数据正以不可预测的速度几乎产生于现实生活中的每一个应用领域。

时间序列数据的研究方法主要包括分类、聚类和回归预测等方面,本文重点讨论时间序列预测方法。现实生活中的时间序列数据预测问题有很多,包括语音分析、噪声消除[2]以及股票市场的分析[3-4]等,其本质主要是根据前 T 个时刻的观测数据推算出 $T+1$ 时刻的时间序列的值。

本文主要介绍时间序列预测方法,第2节介绍时间序列数据的特点以及相关的时间序列参数模型;第3节主要描述传统的时间序列预测方法;第4节主要介绍基于机器学习的时间序列预测方法;第5节介绍基于参数模型的在线时间序列预测方法;第6节介绍时间序列预测方法的进一步研究方向;最后总结全文。

2 时间序列数据的特点以及时间序列参数模型

时间序列数据本质上反映的是某个或者某些随机变量随时间不断变化的趋势,而时间序列预测问题的核心就是从数据中挖掘出这种规律,并利用其对将来数据做出估计。本节主要介绍时间序列数据的特点以及相关的时间序列参数模型,其对时间序列预测方法具有很好的指导作用。

2.1 时间序列数据的特点

时间序列数据被看作一种独特的数据来处理,其具有以下特点。

(1)时间序列数据与其他类型的数据的最大区别在于当

到稿日期:2018-02-04 返修日期:2018-04-03 本文受国家重点研发计划"网络空间安全"重点专项(2017YFB0802800)资助。
杨海民(1990-),男,博士生,主要研究方向为机器学习、时间序列预测,E-mail:haiminyang_nj@126.com;潘志松(1973-),男,教授,博士生导师,主要研究方向为机器学习、网络空间安全,E-mail:haiminyang_nj@126.com(通信作者);白 玮(1983-),男,讲师,主要研究方向为网络空间安全。

前时刻的数据值与之前时刻的数据值存在着联系，该特点表明过去的数据已经暗示了现在或者将来数据发展变化的规律，这种规律主要包括了趋势性、周期性和不规则性。趋势性反映的是时间序列在一个较长时间内的发展方向，它可以在一个相当长的时间内表现为一种近似直线的持续向上或持续向下或平稳的趋势。周期性反映的是时间序列受各种周期因素影响所形成的一种长度和幅度固定的周期波动。不规则性反映的是时间序列受各种突发事件、偶然因素的影响所形成的非趋势性和非周期性的不规则变动。

(2)时间序列的平稳性和非平稳性。时间序列的平稳性表明了时间序列的均值和方差在不同时间上没有系统的变化，而非平稳性意味着均值和方差随着时间推移会发生变化。也就是说，时间序列的平稳性保证了时间序列的本质特征不仅仅存在于当前时刻，还会延伸到未来。

(3)时间序列数据的规模不断变大。一方面，随着各方面硬件技术的不断发展，实际应用中数据的采样频率不断提高，因此时间序列的长度也不断变大，仅仅把时间序列看作单纯的一维向量数据来处理不可避免地会带来维数灾难等问题；另一方面，很多实际应用中的时间序列数据不仅仅是单纯的一维数据，往往包含了一组数值，这一组数值之间也存在着联系，多维时间序列对时间序列预测提出了新的要求。

实际上，在具体研究时间序列预测方法的过程中，时间序列数据的这些特点是需要首先考虑的，这是完成预测工作的难点和关键。结合这些特点进行时间序列预测，才能针对实际问题给出满意的结果。

2.2 相关的时间序列参数模型

经典的时间序列模型包括移动平均模型(Moving Average，MA)、自回归模型(Auto Regressive，AR)、自回归移动平均模型(Auto Regressive Moving Average，ARMA)。假设 x_t 表示 t 时刻的时间序列的值，p 和 q 表示时间窗的大小，ε_t 表示 t 时刻的白噪声，$\alpha_1,\cdots,\alpha_p$ 和 $\beta_1,\cdots,\beta_q$ 表示权重系数。

MA(q)表示为：

$$X_t=\sum_{i=1}^{q}\beta_i\varepsilon_{t-i}+\varepsilon_t$$

AR(p)和 ARMA(p,q)分别表示为：

$$X_t=\sum_{i=1}^{p}\alpha_i X_{t-i}+\varepsilon_t$$

$$X_t=\sum_{i=1}^{p}\alpha_i X_{t-i}+\sum_{i=1}^{q}\beta_i\varepsilon_{t-i}+\varepsilon_t$$

注意到对于 ARMA 模型，当权重系数 $\beta_1,\cdots,\beta_q$ 全为 0 时，其可以被看作一个 AR 模型。因为 MA，AR 和 ARMA 都具有弱平稳性，其均值和协方差都不取决于 t，即 $E(X_t)=\mu$，$\mathrm{cov}(X_t,X_{t+k})=E(X_t-u)(X_{t+k}-u)=\gamma_k,k\in\mathbb{Z}$。

3 传统的时间序列预测方法

传统的时间序列预测方法主要是在确定时间序列参数模型的基础上求解出模型参数，并利用求解出的模型完成预测工作。Box 和 Jenkins[5] 提出的“Box-Jenkins 方法”非常流行。“Box-Jenkins 方法”总的策略主要包含 3 步：对于给定的时间序列，首先确定适当的 p,d,q 值；然后通过最有效的方法估计出模型中具体的参数值；最后检验拟合模型的适当性，并且适当地改进该模型。

对于 d 值的确定，可以通过对原始时间序列进行差分，然后检验差分后的时间序列的平稳性来确定 d 值的大小。p 值和 q 值的大小可以通过偏自相关函数 PACF(Partial Auto-Correlation Function)和自相关函数 ACF(Auto-Correlation Function)来确定。偏自相关函数主要用于在消除中间介入变量 $X_{t-1},X_{t-2},X_{t-3},\cdots,X_{t-k+1}$ 的影响后确定 X_t 和 X_{t-k} 的相关系数 ϕ_{kk}，这个系数称为 k 阶滞后偏自相关系数。自相关函数可以用于确定自相关系数 r_k。ϕ_{kk} 和 r_k 的计算公式如下：

$$\phi_{kk}=corr(X_t,X_{t-k}\mid X_{t-1},X_{t-2},\cdots,X_{t-k+1}),k=1,2,\cdots$$

$$r_k=\frac{\sum_{t=k+1}^{n}(X_t-\overline{X})(X_{t-k}-\overline{X})}{\sum_{t=1}^{n}(X_t-\overline{X})^2},k=1,2,\cdots$$

最大似然估计经常被用来进行时间序列模型参数的估计，根据观察到的样本值，构建关于模型参数的概率密度，求解出使其最大的模型参数值，该方法要求噪声符合一个具体分布。下面给出 AR(p)的似然函数求解法[6]，样本($x_1,x_2,\cdots,x_t$)中的前 p 个观测值($x_1,x_2,\cdots,x_p$)合成一个($p\times1$)向量 x_p，其均值为 u_p，σ^2V_p 表示($p\times p$)的协方差矩阵：

$$\sigma^2V_p=\begin{bmatrix}\gamma_0&\gamma_1&\gamma_2&\cdots&\gamma_{p-1}\\\gamma_1&\gamma_0&\gamma_1&\cdots&\gamma_{p-2}\\\gamma_2&\gamma_1&\gamma_0&\cdots&\gamma_{p-3}\\\vdots&\vdots&\vdots&\cdots&\vdots\\\gamma_{p-1}&\gamma_{p-2}&\gamma_{p-3}&\cdots&\gamma_0\end{bmatrix}$$

其中，γ_j 是 AR(p)的第 j 阶自协方差。前 p 个观察值的概率密度服从一个高斯分布 $N(u_p,\sigma^2V_p)$：

$$\begin{aligned}&f_{X_p,X_{p-1},\cdots,X_1}(x_p,x_{p-1},\cdots,x_1;\theta)\\&=(2\pi)^{-p/2}\mid\sigma^{-2}V_p^{-1})\mid^{1/2}\exp[-\frac{1}{2\sigma^2}(x_p-u_p)'V_p^{-1}(x_p-u_p)]\\&=(2\pi)^{-p/2}(\sigma^{-2})^{k/2}\mid V_p^{-1}\mid^{1/2}\exp[-\frac{1}{2\sigma^2}(x_p-u_p)'V_p^{-1}(x_p-u_p)]\end{aligned}$$

对于样本余下的观察值($x_{p+1},x_{p+2},\cdots,x_t$)，可应用预测误差分解。以前 $t-1$ 个观察值为条件，第 t 个观察值的概率密度是服从高斯分布的。对于 AR(p)来说，只有最近 p 个观察值会影响当前时刻的序列值，因此，当 $t>p$ 时：

$$\begin{aligned}&f_{X_t\mid X_{t-1},\cdots,X_1}(x_t\mid x_{t-1},\cdots,x_1;\theta)\\&=f_{X_t\mid X_{t-1},\cdots,X_{t-p}}(x_t\mid x_{t-1},\cdots,x_{t-p};\theta)\\&=\frac{1}{\sqrt{2\pi\sigma^2}}\exp[\frac{-(x_t-\alpha_1x_{t-1}-\alpha_2x_{t-2}-\cdots-\alpha_px_{t-p})^2}{2\sigma^2}]\end{aligned}$$

这样，全部样本的似然函数为：

$$\begin{aligned}&f_{X_t,X_{t-1},\cdots,X_1}(x_t,x_{t-1},\cdots,x_1;\theta)\\&=f_{X_p,X_{p-1},\cdots,X_1}(x_p,x_{p-1},\cdots,x_1;\theta)\times\\&\quad\prod_{i=p+1}^{t}f_{X_i\mid X_{i-1},\cdots,X_{i-p}}(x_i\mid x_{i-1},\cdots,x_{i-p};\theta)\end{aligned}$$

最大似然函数难以求解时，可以通过 EM 算法(Expectation Maximization Algorithm)求解。EM 经常用来求解含有隐变量的概率模型的最大似然估计函数。EM 算法由 Dempster 于 1977 年提出[7]，主要通过两个步骤进行循环交替计算。

E 步：计算期望，利用对隐变量的现有估计值计算其最大似然估计值；

M 步：最大化在 E 步求得的最大似然估计值来计算参数的值。

状态空间方程也可以用来实现时间序列的预测，其主要通过卡尔曼滤波来进行求解。卡尔曼滤波是一种连续修正系统的线性投影算法，其进行时间序列预测通常是把回归模型表示成动态系统方程形式[6,8-9]，然后应用卡尔曼滤子对方程进行处理，从而进行模型参数的估计，继而预测时间序列。

令 y_t 表示 t 时刻观察到的变量的一个 $(n\times 1)$ 向量，y_t 的动态模型可以以可能观察到的 $(r\times 1)$ 向量 ξ_t 来表示，ξ_t 称为状态变量。y_t 的动态系统方程如下：

$$\xi_{t+1}=F\xi_t+v_{t+1}$$

$$y_t=A'x_t+H'\xi_t+w_t$$

其中，F，A'，H' 分别为 $(r\times r)$，$(n\times p)$，$(n\times r)$ 矩阵，x_t 是外生或前定变量的 $(p\times 1)$ 向量。上述两个方程分别称为状态方程和观察方程。

AR(p) 的状态方程和观察方程表示如下：

$$X_{t+1}-u=\sum_{i=1}^{p}\alpha_i(X_{t-i}-u)+\varepsilon_{t+1}$$

$$X_t=u+[1,0,\cdots,0][X_t-u,X_{t-1}-u,\cdots,X_{t-p+1}-u]^{\mathrm{T}}$$

为充分利用测量值和预测值，卡尔曼滤波并不是取其中一个作为输出，也不是简单地求平均值。在卡尔曼滤波方法中，计算输出分为预测过程和修正过程，其假设预测过程噪声 $w(n)\sim N(0,Q)$，测量噪声 $v(n)\sim N(0,R)$。

预测过程中的预测值为：

$$x(n|n-1)=Ax(n-1|n-1)+Bu(n)$$

最小均方误差矩阵为：

$$P(n|n-1)=AP(n-1|n-1)A^{\mathrm{T}}+Q$$

修正过程中的误差增益为：

$$K(n)=P(n|n-1)H^{\mathrm{T}}(n)[R(n)+H(n)P(n|n-1)H^{\mathrm{T}}(n)]^{-1}$$

修正值为：

$$x(n|n)=Ax(n|n-1)+K(n)[z(n)-H(n)x(n|n-1)]$$

最小均方误差矩阵为：

$$P(n|n)=[I-K(n)H(n)]P(n|n-1)]$$

其中，$x(n)$ 为状态向量，$u(n)$ 为输入向量，$z(n)$ 为观测向量，A 是状态转移矩阵，B 是控制输入矩阵，$x(n|n-1)$ 为用 n 时刻以前的数据对 n 时刻进行估计的结果，$x(n|n)$ 为用 n 时刻及 n 时刻以前的数据对 n 时刻进行估计的结果，右侧公式为最小预测均方误差，

$$P(n|n-1)=E\{[x(n)-x(n|n-1)][x(n)-x(n|n-1)]^{\mathrm{T}}\}$$

其中，$P(n|n)$ 为修正后的最小均方误差矩阵，$K(n)$ 为误差增益，$H(n)$ 为观测向量。通过卡尔曼滤波进行时间序列预测涉及到自回归模型，所以不可避免要对数据进行分布假设。

对于求得参数后的模型，通过残差分析来检验拟合模型的适当性并适当改进模型。残差在数值上等于 X_t 真实值与预测值之差，由于 ε_t 是白噪声，合理的残差应该大致满足均值类似于为零、标准差为 1 的正态分布[10-11]。

传统的时间序列预测方法非常依赖参数模型的选择，能否正确选择参数模型在很大程度上决定了预测结果的准确率。

4　基于机器学习的时间序列预测方法

时间序列数据预测工作本质上与机器学习方法分类中的回归分析之间存在着紧密的联系。经典的支持向量机 SVM、贝叶斯网络 BN、矩阵分解 MF 和高斯过程 GP 在时间序列预测方面均取得了不错的效果。早期的人工神经网络 ANN 也被用来获取时间序列中长期的趋势。随着深度学习的崛起，其也可以被看作实现时间序列预测的有效工具。

4.1　基于支持向量机的时间序列预测方法

支持向量机(Support Vector Machine，SVM)是以统计学习理论为基础的机器学习方法[12-13]，其主要以 VC 维理论和结构风险最小化原理为基础[14-15]，同时也是建立在几何距离[16]基础上的首个学习算法。支持向量机坚实的理论基础保证了其在解决小样本、高维数据和非线性问题方面展现出特有的优势。

Kim[17]直接用支持向量机来预测股票价格指数，并通过对比实验检验了该方法的可行性。Gestel 等[18]提出把贝叶斯证据框架[19-20]应用到最小二乘支持向量机[21-22]来推断关于金融时间序列的数据的模型参数和相关的波动性，从而实现时间序列的预测。该方法主要分为 3 个推断步骤，第 1 步利用最小二乘支持向量机推断时间序列模型参数；第 2 步主要负责推断与正则化和噪声方差相关的超参数；第 3 步对时间序列模型证据进行评估，其主要目的是选择核函数的调节参数和最重要的输入集合。除了把 SVM 应用到金融时间序列预测外，Mellit 等[23]还利用 SVM 对气象方面的时间序列数据进行预测。

Tay 和 Cao 在把 SVM 应用到时间序列预测方面做了很多工作。2001 年，他们直接使用 SVM 来预测金融时间序列数据，并通过对比实验检验了方法的可行性[24]；在此基础上，他们于 2003 年发现 SVM 预测性能的变化受到自由参数的影响，并提出一种自适应的参数选择方法，该方法主要是把金融时间序列的不稳定性应用到 SVM 中；同时通过实验证明了自由参数对预测性能的影响和自适应参数选择方法对性能的提升。Cao 于 2003 年又提出把混合专家系统结构(Mixture of Experts，ME)[25-28]应用到 SVM 来进行时间序列预测。该方法主要分为 2 个阶段，第 1 阶段利用自组织特征映射(Self-Organizing Feature Map，SOM)[29]作为聚类算法，把输入空间划分成几个不相交的区域；第 2 阶段通过多个 SVM 也就是 SVM 专家系统对每个区域进行预测。

4.2　基于贝叶斯网络的时间序列预测方法

贝叶斯网络[30-32](Bayesian Network，BN)本质上是一个有向无环图，使用概率网络进行不确定性推理。这个有向无环图中的节点表示随机变量，节点之间的有向弧代表变量之间的直接依赖关系。利用贝叶斯网络进行时间序列预测工作，主要是针对给定数据集学习出贝叶斯网络结构[33]，这部分工作是构建整个预测模型的基础，其目的就是找出一个最适合数据的网络结构。

Das 和 Ghosh 在把 BN 应用到气象时间序列数据预测方

面做了很多工作，他们主要是把时空信息融入到现有网络结构中。2011 年，Das 和 Ghosh[34] 在 BN 的基础上提出了一种新的网络结构，该结构首次考虑了气候变量间的时空相互关系。在天气状况的预测过程中，首先对气候变量间的时空关系进行预测，预测完成后再把这种关系应用到最终的天气状况预测中。2017 年，他们还提出了基于语义贝叶斯网络的多元气象时间序列预测网络 semBnet[35]，其本质也是把时空信息作为领域知识具体化到语义贝叶斯网络中。

4.3 基于矩阵分解的时间序列预测方法

矩阵分解(Matrix Factorization，MF)也是机器学习中的一项重要工作，其在协同过滤[36-40]、协作排序[41-42]和社会网络分析[43]等领域都发挥了重要作用。MF 本质就是针对原始矩阵找出两个小规模矩阵，使得这两个小矩阵的乘积最大程度地近似拟合原始矩阵。高维时间序列的表示可以采用矩阵形式，其每一列对应的是时间点，每一行对应的是一维的时间序列特征。

MF 在处理拥有缺失数据的时间序列数据预测中取得了很好的效果。网络流量数据本质上是一个时间序列数据，其中存在缺失值的情况特别常见。Zhang 等于 2009 年利用矩阵分解对网络流量数据进行估计，提出一种新的稀疏正则矩阵分解(Sparsity Regularized Matrix Factorization，SRMF)[44]框架，该框架是在原有 MF 的基础上充分利用网络流量数据所蕴含的时空信息，进一步提高网络流量预测的准确性。Zhang 等[45]于 2010 年对手机无线网中的定位技术进行研究，真实的手机轨迹本质上也是一个时间序列数据，这些轨迹数据具有低秩特征，从而选择 MF 来进行过处理。在现有定位技术研究中，结合时间稳定性和数据低秩特征的新机制确实减小了定位误差。

Yu 等[46]在 2016 年的 NIPS 会议上提出一种时间正则矩阵分解(Temporal Regularized Matrix Factorization，TRMF)技术来处理高维时间序列的预测问题，其不仅能很好地处理高维数据中的缺失数据，还展现出了对噪声数据的鲁棒性。TRMF 使用一种新的正则化算子来表达时间序列数据在时间维度上的这种依赖性，不同于以往的用基于图的正则化算子[44-45,47]来处理这种依赖性的方法，该正则化算子克服了基于图的正则化算子不能表示时间序列数据之间负相关依赖性的缺陷。另外，基于图的正则化算子表示的是时间序列数据之间明确的时间依赖关系，在实际应用中这些明确的依赖关系往往不容易获取。

4.4 基于高斯过程的时间序列预测方法

高斯过程(Gauss Process，GP)最早于 20 世纪 90 年代末被提出。在 2006 年，Rasmussen 和 Williams 在 *Gaussian Processes for Machine Learning*[48]一书中从机器学习核方法角度对高斯过程重新加以阐述，并对回归和分类两类问题做了系统的理论和数值实验分析。核函数的采用保证了高斯过程具有很强的处理非线性问题的能力，其还具有其他很多优点[49-53]。高斯过程通常也可以被看作是一个采用概率分布进行描述的正态随机过程，可以对复杂的时变非参数函数进行通用简易求解，具有很灵活的非参数性质，能够对时间序列数据进行有效的估计。

对于过程 $X_1, X_2, X_3 \cdots, X_t, X_t \in \mathbb{R}, t \in \mathbb{Z}$，其任意的有限序列都符合多元高斯分布，这个过程就是 GP。GP 是通过均值函数和协方差核函数对数据进行非参数形式的表示，其中协方差核函数代表了数据之间的依赖关系，因此找到一个合适的核是 GP 对数据进行建模的关键[54-57]。采用 GP 进行时间序列预测时，通常都会假设噪声服从高斯分布。文献[58]介绍了高斯过程卷积模(Gaussian Process Convolution Model，GPCM)，其通过两个阶段的非参数生成过程对稳定时间序列进行建模，改进了 GP 在对时间序列数据进行建模过程中的不足。GPCM 是一个连续时间下的非参数窗口移动平均过程，在某些条件下，其本质上是一个带有以概率形式定义的非参数核的高斯过程。GPCM 通过变分推断过程对协方差核进行封闭形式的概率估计，从而实现对时间序列数据的预测。自动贝叶斯协方差发现(Automatic Bayesian Covariance Discovery，ABCD)对时间序列数据构建自然语言描述，使用带有复合协方差核函数的 GP 对未知的时间序列进行描述，但利用单一的时间序列数据学习到的核对数据的描述往往不够充分。对此，Hwang 等[59]提出使用两个相关联的核学习方法对多个时间序列进行建模，实验表明该方法的准确性更高。

4.5 基于深度学习的时间序列预测方法

深度学习在语音识别、计算机视觉和自然语言处理等领域取得了重大突破，也可以被看作实现时间序列预测的有效工具。深度学习在本质上是多层神经网络，其主要模型包括卷积神经网络(Convolutional Neural Network，CNN)和循环神经网络(Recurrent Neural Networks，RNN)。利用深度模型来预测时间序列在本质上与其他领域没有太大区别，只需对模型的输入和输出做好严格定义。

文献[60]用带有受限玻尔兹曼机(Restricted Boltzmann Machines，RBM)的深度置信网(Deep Belief Network，DBN)来处理时间序列预测问题；文献[61]比较了 DBN 和集成去噪自编码器(Stacked Denoising Auto-encoders)在时间序列预测问题上的性能差异；文献[62]用深度网络来预测室内温度；文献[63]用深度网络来预测交通流量。这些文献中对深度学习模型都只是简单直接的运用，没有结合时间序列数据的性质。文献[64-66]对利用深度学习来完成各种时间序列预测的问题做了很好的总结。

除了利用数据来训练新的深度学习网络结构以完成预测任务外，还可以对已有的深度学习模型进行修改，以实现预测目标。Borovykh 等[67]对现有的深度学习模型 WaveNet[68]进行修改，从而实现对大范围历史数据和序列之间关系的使用。

4.6 基于混合模型的时间序列预测方法

在利用深度学习来预测时间序列时，上述工作对深度学习模型都只是简单的直接应用，没有结合时间序列数据的特点。在时间序列预测问题中，我们更加倾向于利用结合时间序列数据特点的混合模型[69-74]来完成预测工作。文献[69-71]结合 ARIMA 模型和多层感知机(Multilayer Perceptron，MLP)来预测时间序列数据，随着深度网络结构的发展，CNN 和 RNN 的混合模型也被用来完成预测工作。下面主要介绍文献[72-74]的工作。

文献[72]提出了一种 SOCNN(Significance-Offset Convolutional Neural Network)模型来解决多元非同步时间序列数据的预测问题，这个模型受到标准的 AR 模型和 RNN 中门机制思路的启发。门机制的原理可以用 $f(x)=c(x)\otimes\sigma(x)$ 表示，f 表示输出函数，c 表示候选输出，σ 是元素值在[0,1]之间的矩阵，$\otimes$运算为矩阵的 Hadamard 乘积，即对应元素相乘。门机制的本质其实是对原始输出赋予不同的权重后再输出。文中研究的金融数据往往来自不同的途径和不同的时刻，因此这些数据是非同步的。对于这些多元非同步数据，通过增加一维时间间隔特征，可以将其转化为标准的同步时间序列数据。多元时间序列数据$\{x_1,x_2,\cdots,x_T\}(x_t\in\mathbb{R}^n)$展开后如矩阵 $X_{n\times T}$所示：

$$X_{n\times T}=\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1T} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2T} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3T} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nT} \end{bmatrix}$$

SOCNN 由两个卷积神经网络 CNN 构成，分别是 offset 和 significance。offset 网络负责对矩阵 $X_{n\times T}$的列向量进行处理，也就是在特征维度上对多元时间序列数据进行评估，其本质上是一个特征选择过程；而 significance 网络对矩阵 $X_{n\times T}$的行向量进行处理，其本质上类似于一个自回归 AR 过程，主要在时间维度上对多元时间序列进行选择。最后通过权重向量与经过 offset 和 significance 处理过的数据的 Hadamard 乘积来获得预测结果。文中通过实验将 SOCNN 与 VAR，CNN 和 LSTM 做了对比，结果表明，SOCNN 对异步时间序列数据取得最小的均方误差(Mean Squared Error，MSE)。

文献[73]提出了一种混合模型 R2N2 (Residual RNN)来对多元时间序列问题进行预测。R2N2 首先通过一个线性模型 f，例如 VAR 等，对数据进行预测，在此基础上通过 RNN 对线性模型预测得到的残差进行预测。最终的预测结果通过 f 与 RNN 预测的残差相加获得。文中实验只是简单对比了 R2N2 和单一的 RNN，VAR 的效果差距，证明了 R2N2 能取得更好的预测效果。

文献[74]提出一种新的深度学习框架 LSTNet(Long- and Short-term Time-series Network)来解决多元时间序列数据的预测问题。LSTNet 使用 CNN 来提取数据间短期的依赖关系，通过 RNN 来发现数据长期的模式和趋势，充分利用了 CNN 和 RNN 的优点。LSTNet 由 4 个单元构成，即卷积单元、循环单元、循环跳跃单元和自回归单元。卷积单元和循环单元分别实现了 CNN 和 RNN 功能。循环跳跃单元主要是在循环单元上做出修改，从而能够成功获得数据的周期模式。相对于一般的 RNN，循环跳跃单元中隐含层与一个周期之前时刻的隐含层建立联系，而不是前一时刻的隐含层。增加自回归单元主要是为了防止神经网络对输入数据尺寸变化不敏感而影响预测结果的精度。

5　基于参数模型的在线时间序列预测方法

随着计算机网络技术的广泛应用和普及，数据规模的急剧增长给传统的批处理机器学习预测方法带来了严峻的挑战，也严重影响了预测方法的效率，利用在线学习方法来进行时间序列数据预测成为了新的趋势。相对于传统的批处理学习方法，在面对新样本到来时，在线学习方法不需要处理整个数据集，仅需要处理这个新的样本，大大提高了方法的效率。目前利用在线学习对时间序列数据进行预测的相关工作还比较少，Anava 等[75]提出基于参数模型的在线时间序列预测算法 ARMA-ONS，其本质主要是通过在线算法对 AR 模型参数进行求解，并随新时间序列数据的到来更新模型参数的过程，这样就把传统的时间序列模型和在线学习有效结合起来。

ARMA-ONS 算法利用牛顿法 ONS 来求解 AR($p+m$)的权重系数，此方法的主要贡献是证明了在在线学习中，可以用 AR($p+m$)($m\in\mathbb{Z}$)来预测 ARMA(p,q)。相对于 ARMA(p,q)模型，AR($p+m$)模型只涉及到权重系数 $\alpha_1,\cdots,\alpha_{p+m}$，其求解更加容易。另外，在线算法 ARMA-ONS 对时间序列进行分析的优势还体现在不需要对噪声分布进行假设，而传统时间序列预测算法通常会假设噪声服从高斯分布。ARMA-ONS 的具体流程如算法 1 所示。

算法 1　ARMA-ONS(p,q)

1. 输入：ARMA 模型参数 p,q；步长 η；(m+p)×(m+p) 矩阵 A_0
2. 设置 $m=q\cdot\log_{1-\epsilon}((TLM_{max})^{-1})$
3. 任意选择 $\gamma_1\in\kappa$
4. for t=1 to (T−1) do
5. 　预测 $\widetilde{X}_t(\gamma^t)=\sum_{i=1}^{m+p}\gamma_i^t X_{t-i}=\gamma_t^T X_t$。
6. 　观察 X_t 的值并计算损失 $l_t(\gamma^t)$。
7. 　$\nabla_t=\nabla l_t(\gamma^t)$，更新 $A_t\leftarrow A_{t-1}+\nabla_t\nabla_t^T$
8. 　$\gamma_{t+1}\leftarrow\prod_{\kappa}^{A_t}(\gamma_t-\eta A_t^{-1}\cdot\nabla_t)$
9. end for

下面主要以算法 1 为例，形式化在线时间序列预测问题。Shwartz[76]于 2011 年对在线学习做了定义：在线学习是指在给定之前全部或者部分问题的正确答案，可能还有些额外信息的基础上，学习者对现有问题进行回答的过程，这个过程会一直持续下去。根据在线学习的定义，在线时间序列预测的主要任务就是在 t 时刻对当前时刻的时间序列值 X_t 进行估计，在得到 X_t 的真实值后，根据损失对模型参数进行调整，在 $t+1$ 时刻到来时继续进行估计并一直持续下去。对于 ARMA-ONS 算法，定义 t 时刻的损失函数为：

$$l_t(\gamma^t)=l_t(X_t,\widetilde{X}_t(\gamma^t))=l_t(X_t,(\sum_{i=1}^{m+p}\gamma_i^t X_{t-i}))$$

其中，$\widetilde{X}_t(\gamma^t)$是对 X_t 的预测值，$\gamma_1^t,\gamma_2^t,\cdots,\gamma_{m+p}^t$是 t 时刻 AR 的权重系数，$m+p$ 和 p 分别表示 AR 和 ARMA 时间窗的大小。为了保证预测误差最小，在 t 时刻，ARMA-ONS 算法的优化目标函数是：

$$\gamma_t=\arg\min_{\gamma}\|X_t-\gamma^T X_t\|_2^2$$

$$\text{s.t. } t\geqslant m+p$$

其中，向量 γ_t 表示$[\gamma_1^t,\gamma_2^t,\cdots,\gamma_{m+p}^t]^T$，$X_t$ 表示$[X_{t-1},X_{t-2},\cdots,X_{t-(m+p)}]^T$。在 t 时刻，求解出 γ_t 并利用其对 X_{t+1}做出预测。

在线算法中定义了 regret[77]来衡量在线优化算法的优劣，其定量描述了在线算法在不同时刻损失的和与批处理最

优解获得的损失之间的差异。一种好的在线算法需要保证 regret 与时间长度 T 之间服从次线性关系，即 $\lim_{T\to\infty}\text{regret}/T=0$，从而可以说明在线算法与批处理算法之间获得的损失差异可以忽略。对于在线时间序列预测方法，同样使用 regret 来评价权重系数求解算法的合理性。对于 ARMA-ONS 算法，用 $R_T(\gamma)$ 表示，其定义如下：

$$R_T(\gamma)=\sum_{t=1}^{T}l_t(\gamma_t)-\min_{\gamma}\sum_{t=1}^{T}l_t(\gamma)$$

ARMA-ONS 算法利用在线牛顿法 ONS[78] 来求解 AR 的权重系数。ONS 基于著名的 Newton Raphson 离线优化方法，并利用了损失函数的二阶信息，它的 regret 为 $O(\log T)$，与时间长度 T 之间是次线性关系。因此，ARMA-ONS 算法的权重系数调整策略即算法 1 中的步骤 7 和步骤 8 是合理的，其中：

$$\prod_{\kappa}^{A_t}(y)=\arg\min_{x\in\kappa}(y-x)^{\mathrm{T}}A_t(y-x)$$

下面明确算法 1 中的几个参数，这些参数的定义具体参照文献[75]。向量 γ_t 的选择范围用 κ 表示：

$$\kappa=\{\gamma\in\mathbb{R}^{m+p},|\gamma_i|<1,m,p\in\mathbb{N},i\in[1,m+p]\}$$

κ 中任意两个向量之间的最大距离用 D 表示：

$$D=\max_{\gamma_1,\gamma_2\in\kappa}\|\gamma_1-\gamma_2\|_2\leqslant 2\max_{\gamma\in\kappa}\|\gamma\|_2=2\sqrt{(m+p)}$$

G 是损失函数导数 $\|\nabla l_t(\gamma_t)\|$ 的上确界，$A_0=\varepsilon I_{m+p}$，$\varepsilon=\frac{1}{\eta^2D^2}$，$\eta=\frac{1}{2}\min\{4GD,\lambda\}$，$I_{m+p}$ 表示 $m+p$ 阶的单位矩阵，λ 和 L 分别表示损失函数 l_t 的 exp-concavity 常数和 Lipschitz 常数，$M_{\max}$ 是 ARMA 中噪声期望的上界。

6 进一步的研究方向

时间序列预测方法虽然经历了很长时间的发展，但数据规模的急剧增长给传统的批处理机器学习预测方法带来了严峻的挑战，也严重影响了预测方法的效率。利用在线学习方法来进行时间序列数据预测成为新的趋势，这也为我们进一步的研究指明了方向。

(1)将基于参数模型的在线时间序列预测方法与现实世界中时间序列数据本身的特点相结合。Anava 等提出的方法主要是针对传统的时间序列模型，往往忽视了现实世界中时间序列数据本身的特点和性质，特别是大数据环境下以流形式存在的时间序列数据的特点。研究者可以考虑提出一种新的基于平稳稀疏 AR 模型的在线时间序列预测算法，在现有算法 ARMA-ONS 优化目标的基础上增加模型平稳项和数据稀疏项，从而既保留现有参数模型 AR 本身的性质，又兼顾时间序列数据自身的特点。

(2)将在线时间序列预测方法与多元时间序列模型相结合。在传统的多元统计分析中，通常是以批处理的方法对多元时间序列模型的参数进行估计。这些方法包括条件似然方法和确切似然方法，这两种方法通常会假设噪声服从高斯分布，而在现实世界中，噪声往往不可能总是正好服从标准的正态分布。目前利用在线学习对时间序列数据进行预测的相关工作还比较少，研究者可以考虑如何在在线时间序列预测中更方便地求解多元时间序列模型的参数。

(3)时间序列数据往往包含异常数据点，在时间序列中，如果一个数据点与大部分数据点的行为明显不一致，其通常被确定为异常点。换句话说，如果时间序列中的大部分数据是系统或者程序正常运行时产生的，那么异常点就是这些系统或者程序以不正常的方式运行产生的。研究者可以考虑如何通过在线算法对时间序列中的异常点进行检测，从而减小预测方法的误差。

结束语 本文首先介绍了时间序列的概念，然后阐述了时间序列数据的特点，在此基础上分 3 个方面介绍了时间序列预测方法，包括传统的时间序列预测方法、基于机器学习的时间序列预测方法以及基于参数模型的在线时间序列预测方法。时间序列数据规模的急剧增长为我们进一步的研究指明了方向，包括如何将基于参数模型的在线时间序列预测方法与现实世界中时间序列数据本身的特点相结合，如何将在线时间序列预测方法与多元时间序列模型相结合，如何处理时间序列数据中的异常点，这些都需要研究者进一步努力。

参 考 文 献

[1] YUAN J D, WANG Z H. Review of Time Series Representation and Classification Techniques [J]. Computer Science, 2015, 42(3): 1-7. (in Chinese)
原继东，王志海. 时间序列的表示与分类算法综述[J]. 计算机科学，2015，42(3)：1-7.

[2] GAO J, SULTAN H, HU J, et al. Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison [J]. IEEE Signal Processing Letters, 2010, 17(3): 237-240.

[3] ROJO-ALVAREZ J L, MARTINEZ-RAMON M, PRADO-CUMPLIDO M, et al. Support Vector Method for Robust ARMA System Identification [J]. IEEE Transactions on Signal Processing, 2004, 52(1): 155-164.

[4] GRANGER C W J, NEWBOLD P. Forecasting Economic Time Series [M]. New York: Academic Press, 1986.

[5] BOX G, JENKINS G. Time Series Analysis, Forecasting and Control [M]. Holden-Day, 1990.

[6] HAMILTON J. Time Series Analysis [M]. Princeton: Princeton University Press, 1994.

[7] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum Likelihood from Incomplete Data via the EM Algorithm [J]. Journal of the Royal Statistical Society, Series B (Methodological), 1977, 39(1): 1-38.

[8] DURBIN J, KOOPMAN S J. Time Series Analysis by State Space Methods [M]. Oxford: Oxford University Press, 2012.

[9] KALMAN R E. A New Approach to Linear Filtering and Prediction Problems [J]. Journal of Fluids Engineering, 1960, 82(1): 35-45.

[10] ALQUIER P, LI X, WINTENBERGER O. Prediction of Time Series by Statistical Learning: General Losses and Fast Rates [J]. Dependence Modelling, 2014, 1: 65-93.

[11] KUZNETSOV V, MOHRI M. Generalization Bounds for Time Series Prediction with Non-Stationary Processes [M]// Algorithmic Learning Theory. Springer International Publishing, 2014: 260-274.

[12] CRISTIANINI N, TAYLOR J S. Introduction to Support Vector

Machines [M]. 李国正，王猛，曾华军，译. 北京：电子工业出版社，2004.

[13] ZHANG X G. Introduction to Statistical Learning Theory and Support Vector Machines [J]. Acta Automatica Sinica，2000，26(1)：32-41. (in Chinese)
张学工. 关于统计学习理论与支持向量机 [J]. 自动化学报，2000，26(1)：32-41.

[14] VAPNIK V N. The Nature of Statistical Learning Theory [M]. 张学工，译. 北京：清华大学出版社，2000.

[15] VAPNIK V N. Statistical Learning Theory [M]. 许建华，张学工，译. 北京：电子工业出版社，2004.

[16] CRISTIANINI N，TAYLOR J S. An Introduction to Support Vector Machines [M]. Cambridge：Cambridge University Press，2000.

[17] KIM K. Financial Time Series Forecasting Using Support Vector Machines [J]. Neurocomputing，2003，55(1)：307-319.

[18] GESTEL T V，SUYKENS J A K，BAESTAENS D E，et al. Financial Time Series Prediction Using Least Squares Support Vector Machines within The Evidence Framework [J]. IEEE Transactions on Neural Networks，2001，12(4)，809-821.

[19] MACKAY D J C. Bayesian Interpolation [J]. Neural Computation，1992，4(3)：415-447.

[20] MACKAY D J C. Probable Networks and Plausible Predictions—A Review of Practical Bayesian Methods for Supervised Neural Networks [J]. Network Computation in Neural Systems，1995，6(3)：469-505.

[21] SUYKENS J A K，VANDEWALLE J. Least Squares Support Vector Machine Classifiers [J]. Neural Processing Letters，1999，9(3)：293-300.

[22] SUYKENS J A K. Least Squares Support Vector Machines for Classification and Nonlinear Modeling [J]. Neural Network World，2000，10(1)：29-48.

[23] MELLIT A，PAVAN A M，BENGHANEM M. Least Squares Support Vector Machine For Short-Term Prediction of Meteorological Time Series [J]. Theoretical Applied Climatology，2013，111(1)：297-307.

[24] TAY F E H，CAO L. Application of Support Vector Machines in Financial Time Series Forecasting [J]. Omega，2001，29(4)：309-317.

[25] JACOBS R A，JORDAN M A，NOWLAN S J，et al. Adaptive Mixtures of Local Experts [J]. Neural Computation，1991，3(1)：79-87.

[26] JORDAN M I，JACOBS R A. Hierarchical Mixtures of Experts and the EM Algorithm [J]. Neural Computation，1994，6(2)：181-214.

[27] WEIGEND A S，MANAGEAS M. Analysis and Prediction of Multi-Stationary Time Series [C]// Proceedings of the Third International Conference on Neural Networks in the Capital Markets. 1995.

[28] WEIGEND A S，MANAGEAS M，SRIVASTAVA A N. Nonlinear Gated Experts for Time Series：Discovering Regimes and Avoiding Over-fitting [J]. International Journal of Neural Systems，1995，6(4)：373-399.

[29] KOHONEN T. Self-Organization and Associative Memory [M]. Springer，1989.

[30] PEARL J. Fusion，Propagation and Structuring In Belief Networks [J]. Artificial Intelligence，1986，2(3)：241-288.

[31] COOPER G F，HERSKOVITS E. A Bayesian Method for the Induction of Probabilistic Networks from Data [J]. Machine Learning，2008，9(4)：309-347.

[32] ZHANG M，ZHOU Z. Multi-Label Neural Networks with Application to Function Genomics and Text Categorization [J]. IEEE Transactions on Knowledge and Data Engineering，2006，18(10)：1338-1351.

[33] MEZ J，MATEO J L，PUERTA J. Learning Bayesian Networks by Hill Climbing：Efficient Methods Based on Progressive Restriction of the Neighborhood [J]. Data Mining Knowledge Discovery，2011，22(1-2)：106-148.

[34] DAS M，GHOSH S K. A Probabilistic Approach for Weather Forecast Using Spatio-temporal Inter-relationship among Climate Variables [C]// International Conference on Industrial and Information Systems. IEEE，2015.

[35] DAS M，GHOSH S K. SemBnet：A Semantic Bayesian Network for Multivariate Prediction of Meteorological Time Series Data [J]. Pattern Recognition Letters，2017，93：192-201.

[36] KOREN Y，BELL R，VOLINSKY C. Matrix Factorization Techniques for Recommender Systems [J]. Computer，2009，42(8)：30-37.

[37] KOREN K. Factorization Meets The Neighborhood：A Multifaceted Collaborative Filtering Model [C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2008：426-434.

[38] SALAKHUTDINOV R，MNIH A. Probabilistic Matrix Factorization [C]// International Conference on Neural Information Processing Systems. Curran Associates Inc.，2007：1257-1264.

[39] XU M，ZHU J，ZHANG B. Bayesian Nonparametric Maximum Margin Matrix Factorization for Collaborative Prediction [C]// Advances in Neural Information Processing Systems. 2012.

[40] SREBRO N. Maximum margin matrix factorization[J]. Advances in Nips，2005，37(2)：1329-1336.

[41] BALAKRISHNAN S，CHOPRA S. Collaborative Ranking[C]// ACM International Conference on Web Search and Data Mining. ACM，2012：143-152.

[42] WEIMER M，KARATZOGLOU A，LE Q，et al. CoFiRank—Maximum Margin Matrix Factorization for Collaborative Ranking [C]// Neural Information Processing Systems. 2007：1593-1600.

[43] KROHN-GRIMBERGHE A，DRUMOND L，FREUDENTHALER C. Multi-Relational Matrix Factorization Using Bayesian Personalized Ranking for Social Network Data [C]// Proceedings of the Fifth International Conference on Web Search and Web Data Mining(WSDM 2012). Seattle，WA，USA，2012.

[44] ZHANG Y，ROUGHAN M，WILLINGER W，et al. Spatio-temporal Compressive Sensing and Internet Traffic Matrices [C]// ACM Sigcomm Conference on Data Communication. 2009.

[45] RALLAPALLI S，QIU L，ZHANG Y，et al. Exploiting Temporal Stability and Low-Rank Structure for Localization in Mobile Networks [C]// International Conference on Mobile Computing and Networking(Mobicom'10). 2010.

[46] YU H F，RAO N，DHILLON I S. Temporal Regularized Matrix

Factorization for High-Dimensional Time Series Prediction [C]// NIPS. 2016.

[47] XIONG L, CHEN X, HUANG T K, et al. Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization [C]// Siam International Conference on Data Mining (SDM 2010). Columbus, Ohio, USA, DBLP, 2010: 211-222.

[48] RASMUSSEN C E, WILLIAMS C K I. Gaussian Processes for Machine Learning [M]. The MIT Press, 2006.

[49] ALEXE B, DESELAERS T, FERRARI V. What is An Object [C]// Computer Vision and Pattern Recognition. IEEE, 2010: 73-80.

[50] EVERINGHAM M, GOOL L V, WILLIAMS C K I, et al. The Pascal, Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[51] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object Detection with Discriminatively Trained Part-Based Models [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(9): 1627-1645.

[52] KAPOOR A, GRAUMAN K, URTASUNR, et al. Gaussian Processes for Object Categorization[J]. International Journal of Computer Vision, 2010, 88(2): 169-188.

[53] CHUM O, ZISSERMAN A. An Exemplar Model for Learning Object Classes [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2007(CVPR'07). IEEE, 2007: 1-8.

[54] DIOSAN L, ROGOZAN A, PECUCHET J P. Evolving Kernel Functions for SVMs by Genetic Programming [C]// International Conference on Machine Learning and Applications. IEEE, 2007: 19-24.

[55] WU B, ZHANG W, CHEN L, et al. A GP-based Kernel Construction and Optimization Method for RVM [C]// The International Conference on Computer and Automation Engineering. IEEE, 2010: 419-423.

[56] KLENSKE E D, ZEILINGER M N, SCHOLKOPF B, et al. Nonparametric Dynamics Estimation for Time Periodic Systems [C]// Communication, Control, and Computing. IEEE, 2014: 486-493.

[57] LLOYD J R, DUVENAUD D, GROSSE R, et al. Automatic Construction and Natural-Language Description of Nonparametric Regression Models [C]// Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014: 1242-1250.

[58] TOBAR F, BUI T D, TURNER R E. Learning Stationary Time Series Using Gaussian Processes with Nonparametric Kernels [C]// Advances in Neural Information Processing Systems 28 (NIPS 2015). 2015.

[59] HWANG Y, TONG A, CHOI J. Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series [C]// International Conference on International Conference on Machine Learning. JMLR. org, 2016: 3030-3039.

[60] KUREMOTO T, KIMURA S, KOBAYASHI K, et al. Time Series Forecasting Using a Deep Belief Network with Restricted Boltzman Machines [J]. Neurocomputing, 2014, 137(15): 47-56.

[61] TURNER J T. Time Series Analysis Using Deep Feed Forward Neural Networks [D]. Baltimore: University of Maryland, 2014.

[62] ROMEU P, ZAMORA-MARTINEZ F, BOTELLA-ROCAMORA P, et al. Time-Series Forecasting of Indoor Temperature Using Pre-trained Deep Neural Networks [C]// International Conference on Artificial Neural Networks. Berlin: Springer, 2013: 451-458.

[63] LV Y, DUAN Y, KANG W, et al. Traffic Flow Prediction with Big Data: A Deep Learning Approach [J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2): 865-873.

[64] LANGKVIST M, KARLSSON L, LOUTFI A. A Review of Unsupervised Feature Learning and Deep Learning for Time-Series Modeling [J]. Pattern Recognition Letters, 2014, 42(1): 11-24.

[65] GAMBOA J C B. Deep Learning for Time-Series Analysis [J]. Arxiv: 1701.01887, 2017.

[66] HEATON J B, POLSON N G, WITTE J H. Deep Learning in Finance [J]. Arxiv: 1602.06561, 2016.

[67] BOROVYKH A, BOHTE S, OOSTERLEE C W. Conditional Time Series Forecasting with Convolutional Neural Networks [J]. Arxiv: 1703.04691, 2017.

[68] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A Generative Model for Raw Audio [J]. Arxiv: 1609.03499, 2016.

[69] JAIN A, KUMAR A M. Hybrid Neural Network Models for Hydrologic Time Series Forecasting [J]. Applied Soft Computing, 2007, 7(2): 585-592.

[70] ZHANG G, PATUWO B E, HU M. Forecasting with Artificial Neural Networks: The State of The Art [J]. International Journal of Forecasting, 1998, 14(1): 35-62.

[71] ZHANG G. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model [J]. Neurocomputing, 2003, 50(1): 159-175.

[72] BINKOWSKI M, MARTI G, DONNAT P. Autoregressive Convolutional Neural Networks for Asynchronous Time Series [J]. Arxiv: 1703.04122, 2017.

[73] GOEL H, MELNYK I, BANERJEE A. R2N2: Residual Recurrent Neural Networks for Multivariate Time Series Forecasting [J]. Arxiv: 1709.03159, 2017.

[74] LAI G, CHANG W, YANG Y, et al. Modeling Long- And Short-Term Temporal Patterns with Deep Neural Networks [J]. Arxiv: 1703.07015, 2017.

[75] ANAVA O, HAZAN E, MANNOR S, et al. Online Learning for Time Series Prediction [J]. Journal of Machine Learning Research, 2013, 30: 172-184.

[76] SHWARTZ S S. Online Learning and Online Convex Optimization [J]. Foundations and Trends in Machine Learning, 2011, 4(2): 107-194.

[77] ZINKEVICH M. Online Convex Programming and Generalized Infinitesimal Gradient Ascent [C]// Proceedings of the Twentieth International Conference on Machine Learning (ICML). 2003: 928-936.

[78] HAZAN E, AGARWAL A, KALE S. Logarithmic regret algorithms for online convex optimization [J]. Machine Learning, 2007, 69(2-3): 169-192.