基于DTW相似判定的周期性时间序列预测方法

李文海[1,2] 程佳宇[2] 谢晨阳[2]
(软件工程国家重点实验室 武汉 430072)[1] (武汉大学计算机学院 武汉 430072)[2]

摘 要 针对大样本下周期性时间序列预测的问题,文中给出了一种基于DTW距离的相似样本度量方法。首先,给出周期性时间序列预测问题的定义,并基于支持向量回归方法分析大量噪声点对预测误差的影响。然后,通过对时间序列周期分段来构建相似性度量,在给定预测样本容量下确定给定预测条件的相似样本子集。同时,基于误差调谐函数对SVM的核函数进行调整,以进一步提升预测精度。最后,基于常用的周期性时间序列,在预测精度上将所提方法与已有算法进行实验比较,并分析该模型的参数敏感性。实验结果验证了所提方法的有效性。

关键词 DTW,支持向量机,时间序列,相似性,预测,数据挖掘
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.05.024

Prediction Method of Cyclic Time Series Based on DTW Similarity

LI Wen-hai[1,2] CHENG Jia-yu[2] XIE Chen-yang[2]
(State Key Laboratory of Software Engineering of China, Wuhan 430072, China)[1]
(School of Computer Science, Wuhan University, Wuhan 430072, China)[2]

Abstract This paper presented a DTW distance-based sampling framework to effectively improve the accuracy of cyclic time series prediction in large-scale datasets. It addresses the problem of noisy identification for each given prediction condition, and formalizes the impact of noise with the SVR-based predicting method. On top of the DTW-based similarity measurement, this paper presented an end-to-end identification method to improve the quality of the training set. It also introduced a regularized function in the kernel function of SVR, such that the generalization error can be minimized based on the distances between each training instance and the prediction condition. The experiment conducts a series of widely a dopted cyclic time series to evaluate the precision and stability of the proposed method. The results demonstrate that in terms of high-quality training instances and the weighted regularization strategy, the proposed method remarkably outperforms its competitors in most of the datasets.

Keywords DTW, Support vector machine, Time series, Similarity, Prediction, Data mining

1 引言

作为一种常见的数据表达形式,时间序列(Time Series)广泛存在于工农业生产、金融预测和信息分析中。时间序列作为时间表征下的有序点集,根据时间采样或插值形成一定间隔的统计值,从而得到时间序列的特征集,对现实世界的观测结果进行了描述。在此基础上的时间序列模式挖掘和聚类分析则是通过时间序列的统计值或形状信息来提取序列及其子序列的相似特征。在这类应用中,如果一些子序列频繁出现一类相似的特征,那么对这些相似特征进行分析和提取是极具意义的。利用这类频繁出现的模式,在未来某时刻可实现对时间序列的预测或异常检测。

时间序列的应用广泛存在于各行各业中,常见于精准营销和目前蓬勃发展的线上服务、电力行业中的用电量预测需求、以提升商家收益为目标的商品销售量的精准预测等生活模式,以及土壤侵蚀、元器件损伤和物种数量变动等成因相关性分析和监控预测的社会问题。上述应用中的时间序列存在一些潜在的共性特征。首先,它们都是可供观测的序列,且具有长效性,如政府监测部门或企业的科研单位对一些基本资源和重点关心的外延环境信息进行长期监测,保存了大量的监测历史记录。同时,这类时间序列普遍还具有周期性[1],例如居民耗电的月度峰值普遍处于6月—8月的若干个时段内。企业用电一般依据商品的类型具有较强的周期性等。另外,部分时间序列在变化过程中还具有单调性,例如指定区域的耗电量在长期变化的过程中整体上呈逐年递增的趋势,多年来观测到的环境监测指标中CO_2气体和酸性物质的比重呈现总体增加的态势。综上所述,有效利用这些共性特征对时序进行分析预测具有重要意义。

目前已有的针对时间序列预测问题的研究中,研究热点主要包括以下几个方面。1)训练样本的选择:其核心在于如

到稿日期:2018-01-16 返修日期:2018-04-15 本文受国家自然科学基金(61572373,61472290,60903035),国家重点研发计划(2017YFC08038)资助。

李文海(1979—),男,博士,副教授,主要研究方向为数据库、并行计算和知识发现,E-mail:lwhaymail@zlcn.com(通信作者);**程佳宇**(1992—),女,硕士生,主要研究方向为数据挖掘;**谢晨阳**(1994—),女,硕士生,主要研究方向为文本处理和主题模型。

何建立历史样本集与预测序列之间的度量关系，可参照的方法包括空间上的欧氏距离计算[2]和 IIR 方法[3]等。2）训练模型的优化：针对时序预测应用得较多的最小二乘支持向量机（Least Squares Support Vector Machine，LSSVM），其出发点在于通过核调整降低偏离预测条件较远的噪声样本的权重，以减小模型的泛化误差（Generalization Error）[4]。该类研究在降低样本决策维方面，通过增量的 LSSVR 方法取得了一定的优化突破（本文 2.2 节将给出详细分析）。3）训练样本的相似性度量：应用熵值计算、DTW 距离等基于聚类的研究较多。而对于大规模时间序列的研究，上述方法中较为突出的问题集中在结合时间序列特征的相似性衡量和模型的参数选择上。

针对大样本下周期性时间序列的预测问题，基于样本相似性的度量给出了一种基于 DTW 距离的改进方法，并通过对时序周期分段来度量相似性，从而在所需容量一定时确定给定预测条件的相似样本子集；同时结合支持向量回归分析时序问题的优势，分析此过程中大量噪声点产生的原因，并基于误差调谐函数对 SVM 的核函数进行调整以提升预测精度；最后通过实验分析所提方法的预测精度和参数敏感性，并验证有效性。

2 周期性时间序列预测

本节给出时间序列预测问题的定义，对相关研究方法进行概述，并对 SVR 方法的误差成因进行分析。

2.1 问题的基本描述

考虑一个由一组时间有序的观测值构成的典型观测结果，其中每个观测值由一个因变量在某一时刻上的原始观测值或一个时间段的统计值构成。

定义 1（时间序列） 对于一个长度为 m 的时间序列，形如 $\boldsymbol{x}=\{\langle x_1,t_1\rangle,\cdots,\langle x_m,t_m\rangle\}$，其元素由观测值 $x_i\,|\,1\leqslant i\leqslant m$ 和时刻 $t_i\,|\,1\leqslant i\leqslant m$ 两个一一对应的单元构成。

不失一般性，假定时间序列为时间有序的实值，且任意相邻元素在时间维上具有相等的时间间隔，并记作 $t(x[i])=x_i\cdot t\,|\,1\leqslant i\leqslant m$，每个元素的时间单元忽略不计。针对因样本缺失或统计差异造成的不等间隔问题，可考虑引入插值方法进行处理。基于此，下面给出子序列的定义。

定义 2（子序列） 一个长度为 m 的时间序列 $\boldsymbol{x}$ 的任意一个时间连续子集 $\boldsymbol{x}[i:j]=\{x_i,\cdots,x_j\}\,|\,1\leqslant i\leqslant j\leqslant m$ 即为时间序列的子序列。

基于时间序列及其子序列的定义可知，一个子序列为时间序列的任意时间连续子集。在数据流上下文中假定一个当前时刻 n，经常需要根据某时间序列历史观测全集 $\boldsymbol{x}[1:n]$ 的所有子集来预测未来观测值。

定义 3（时间序列预测） 给定时间序列观测历史 $\boldsymbol{x}[1:n]$，则将时刻 $n+p$ 的实值 x_{n+p} 称为 p 步预测。

近年来，这一时间序列预测问题在机器学习中被广泛关注，如线性回归、高斯过程、神经网络和马尔科夫模型等，均先后被引入并用于解决时间序列预测问题。一种常用的思路是通过对观测值构建时间连续函数，再基于最优化理论降低函数拟合的误差，得到模型参数，并实现对未来时刻的预测。此外，根据预测步长 p 取值的不同，此问题可分为长效（Long-term）、中程（Intermediate-term）和短程（Short-term）预测 3 类。本文聚集于时间序列的短程预测问题，特别地，当时间序列的模式周期性重复时，可依照一个给定的周期提取时间序列的子序列，实现对模型的求解。

定义 4（周期性时间子序列） 给定一个长度为 n 的时间序列 $\boldsymbol{x}$，称 $\{\boldsymbol{x}_i\in\mathbb{R}^d\,|\,1\leqslant i\leqslant n-d+1\}$ 为 $\boldsymbol{x}$ 的 d 周期时间子序列，其中 $\boldsymbol{x}_i=\{x_i,\cdots,x_{i+d-1}\}$。

可以看到，周期性子序列为原始时间序列的等周期连续子集。给定周期和原始时间序列，可以确定一个周期性时间子序列集合。面向预测问题，若序列长度为 n 且周期为 d，则通过 p 步预测可以得到 $n-d-p+1$ 个 d 维条件因向量和 1 维决策变量。在提取时间序列的周期性子序列的基础上，引入机器学习算法可以生成模型，完成对决策变量的预测。

2.2 研究方法概述

为了准确预测时间序列，大量机器学习方法被引入，以实现预测误差的可控。一些研究通过度量[5-6]时间序列的相似性实现时间序列聚类，以提高预测精度。依据研究所采用的模型划分，时间序列预测的相关研究主要沿着 3 种信息理论智能分析范例（基于统计回归的方法 ARIMA[7]、采用神经网络（Neural Network，NN）的权重训练方法[1]和基于核结构的支持向量回归（Support Vector Regression，SVR）方法[2-4,8-10]）展开。本研究将在实验环节对基于 ARIMA 和 NN 的方法进行比较分析，下面重点对基于相似性度量和 SVR 时间序列预测方法进行概述。

针对通用样例选择的问题，IIR 方法[3]通过样本点之间的 Minkowski 距离度量构建目标样本的 kNN 样本集合，并分别针对分类问题和回归问题求取目标模型的最优参数解。Suykens 等[8]提出的 LSSVM 方法是基于最小二乘方法来求解 SVM，通过对非线性核支持向量的不等式约束构建等价的等式约束，并通过最小二乘代价函数实现对模式参数的求解。在此基础上，文献[7]利用了欧氏空间特性，在 LSSVM 的基础上特别针对时间序列进行了对 kNN 子集的筛选，然后基于条件维的距离度量和样本决策维得到统计度量来调整核参数，以提高预测精度。其调整思想在于分别构建预测条件与 k 个训练样本的欧氏距离和训练样本的方差，实现预测模型的核参数调整。

已有研究表明，通用回归[9]和支持向量回归[10]的泛化误差概率收敛于经验误差（Empirical Error）与一个上界函数之和，该上界函数与决策上界存在线性关系。换言之，针对可再生核 Hilbert 空间（RKHS），通过缩小样本决策维的值域空间，可以降低给定模型的预测误差。同时，值域空间的压缩不应过度牺牲样本容量[9]。基于核调整，Santos 等[10]给出了一种增量的 LSSVR 方法，通过在新增样本中挑选出与已有训练样本线性无关的部分加入模型，提高了模型的精度，减少了训练时间。类似地，度量样本之间相似性的还有熵[11]和 DTW[12]距离等，这些研究聚焦于样本聚类，未对核调整给出更为深入的解决方法。与此同时，围绕序列相似性问题，国内研究也沿着序列距离[13]和主成分分析方法[14]等方向开展。文献[15]综述了该问题近年来的研究现状和方法。近年来，面对支持向量预测问题，无核相关向量机[16]、增强学习方法[17]和多支持向量方法[18]逐渐被重视，相关研究也涉及到

核调整的思路，并通过引入新的模型来提高支持回归的精度。

2.3　支持向量回归

给定一组训练子序列及其 p 步预测值，对回归问题的目标进行定义。

定义 5（周期时间序列回归）　给定经由周期性变换及其 p 步预测的样本集 $D_n=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n)\}$，其中实例 $(\boldsymbol{x}_i,y_i)\in D_n\subset\mathcal{X}\times\mathcal{Y}$ 由 d 维条件向量 $\boldsymbol{x}_i\in\mathbb{R}^d$ 和决策值 $y_i\in\mathbb{R}$ 构成，回归问题可表示为如下函数：

$$f(\boldsymbol{x})=\langle\boldsymbol{w},\boldsymbol{x}\rangle+b \tag{1}$$

其中，符号 $\langle\boldsymbol{w},\boldsymbol{x}\rangle$ 为两个向量的内积。函数 $f(\boldsymbol{x})$ 的求解过程一般被定义为获取最后的权重系数向量 $\boldsymbol{w}$，以保证回归误差的最小化。针对非线性问题，SVR 引入高维特征空间来求解上述回归问题。通过构建一个映射函数 $\varphi:\mathbb{R}^d\to\Phi$ 向 h 维 Hilbert 空间 $\varphi:\mathbb{R}^d\to H$ 转换，则式(1)可被转换为一簇线性函数，如下：

$$\mathcal{F}_\Phi=\{f:f(\boldsymbol{x})=\langle\boldsymbol{w},\varphi(\boldsymbol{x})\rangle+b\,|\,\varphi:\mathbb{R}^d\to\Phi,\boldsymbol{w}\in\mathbb{R}^h\} \tag{2}$$

其中，映射函数 $\varphi:\mathbb{R}^d\to\Phi$ 可表达为满足 Mercer 条件的向量内积 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)\rangle$，以便基于风险最小化原则实现模型求解。常用核函数有：

1）线性核：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle$；

2）多项式核：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle+c)^k$；

3）RBF 核：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/h^2)$。

该问题的求解方法中，涉及到优化多项式的方法需要基于凸函数实现模型的最优化求解，具有较高的时间复杂度。近年来，具有更低代价的最小二乘支持向量回归方法[6]被广泛采用。通过定义最小二乘代价函数：

$$LS(f(\boldsymbol{x})-y)=(f(\boldsymbol{x})-y)^2 \tag{3}$$

LSSVR 模型的求解被转换为约束风险最小化问题：

$$\begin{aligned}&\min\ J(\boldsymbol{w},\boldsymbol{e})=\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{\gamma}{2}\sum_{i=1}^{n}\delta e_i^2\\&\text{s.t.}\ \ y_i=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b+e_i\end{aligned} \tag{4}$$

值得指出的是，式(4)中的误差变量 e_i 和权重因子 δ_i 为模型训练过程中得到的样本加权权重系数，核调整可根据一定的策略对它们进行加权，以优化基于预测条件与训练样本的距离计算。4.1 节将给出一个基于序列距离度量的调整框架。基于回归的稳定性定律[5]，可得如下定理。

定理 1　给定有界核 $\forall\boldsymbol{x}\in\mathcal{X},K(\boldsymbol{x},\boldsymbol{x})\leqslant h^2$，任意预测条件 $\boldsymbol{z}_i\in\mathbb{R}^d$ 在 LSSVR 一个学习算法 A 和训练样本子集 D 下的 Leave-One-Out 的风险满足：

$$E_D[\,|L(A_D,\boldsymbol{z}_i)-L(A_{D^{\setminus i}},\boldsymbol{z}_i)|\,]\leqslant\frac{\gamma\sigma^2h^2}{2} \tag{5}$$

该定理表明：当通过一个训练子集 D 在 LSSVR 算法 A 上生成模型来对预测条件 $\boldsymbol{z}_i$ 进行预测时，其误差期望具有上界。由于经验误差 $L(A_D,\boldsymbol{z}_i)$ 与样本集相对于 $\boldsymbol{z}_i$ 的距离呈线性关系，样本集的条件维与预测条件 $\boldsymbol{z}_i$ 越相似，Leave-One-Out 风险 $L(A_{D^{\setminus i}},\boldsymbol{z}_i)$ 越小，对应的预测误差就越小。换言之，若能够有效度量样本与预测条件的相似度，则基于相似度高的样本训练模型可以降低模型误差。同时，该样本容量为原始样本全集的一个子集，其效率相较于基于全集的训练过程有显著提高。

3　DTW-LSSVR 时间序列预测模型

本节给出 DTW 距离来度量周期子序列的相似性，进而构建参数化 kNN 样本子集。

3.1　基于 DTW 距离的 kNN 训练集的选择方法

时间序列的相似性度量由来已久，其中动态时间拉伸（Dynamic Time Warping）距离由于对时态平移和伸缩不敏感而被广泛用于时间序列聚类中。对于两条长度为 a 和 b 的序列 $\boldsymbol{u}=\{u_1,\cdots,u_a\}$ 和 $\boldsymbol{v}=\{v_1,\cdots,v_b\}$，其距离度量过程可表述为下述递归过程。

$$\begin{aligned}DTW(\boldsymbol{u}[1:a],\boldsymbol{v}[1:b])=\tau(u_a,u_b)+\\\begin{cases}DTW(\boldsymbol{u}[1:a-1],\boldsymbol{v}[1:b])\\DTW(\boldsymbol{u}[1:a-1],\boldsymbol{v}[1:b-1])\\DTW(\boldsymbol{u}[1:a],\boldsymbol{v}[1:b-1])\end{cases}\end{aligned} \tag{6}$$

即两条序列的 DTW 距离为两条序列之间逐点最小距离 $\tau(u_a,u_b)$ 的累积和。图 1 为逐点距离求和得到 DTW 距离的示意图。其中序列 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 为不等长的两条序列，对比欧氏距离，对于 $\boldsymbol{V}$ 序列中的任意一点 a，其到序列 $\boldsymbol{U}$ 的距离并不是到时刻对应点 b 的距离，而是离序列 $\boldsymbol{U}$ 中最近点 c 的距离。同理，对所有最近距离成对点的距离度量求和，得到序列之间的距离。基于 DTW 距离，可对不等长序列进行序列相似性判定，且相似度可以较好地反映序列的形状，因而被广泛采用。通过构建参数化 kNN 样本集，可以针对任意给定预测条件得到 k 个 DTW 距离最小的样本。由定理 1 可知，训练样本与预测条件之间的相似度越高且训练样本决策维的值域越小，则模型的泛化误差越小。图 2 引入了 Laser 数据集[10]，验证了 DTW 距离在周期性时间序列下对相似样本遴选的有效性。

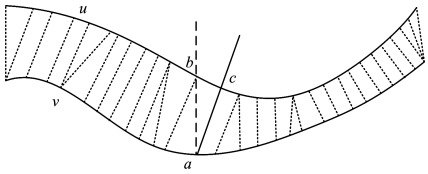
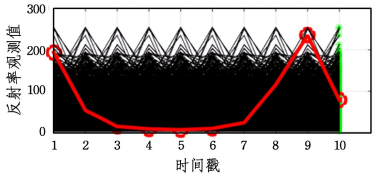
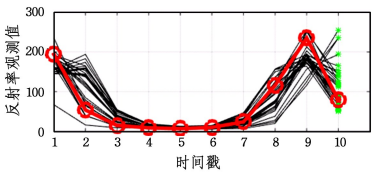


图 1　DTW 距离度量的示意图（向量长度分别为 31 和 30）

Fig. 1　Schematic diagram of DTW distance metrics (vector lengths are 31 and 30 respectively)



(a) Laser 样本全集



(b) $k=200$ 时得到的 DTW 相似样本

图 2　采用 kNN 对周期性时间子序列进行相似样本过滤的结果

Fig. 2　Similar sample filtering results for periodic time subsequences using kNN

从图 2(a)中可以看到，原始 Laser 数据集的周期性特征显著，给定的预测样本的条件维（图中加重序列）具有大量极其接近的训练样本。上例中，周期参数设置为 9，可以看到单

步预测条件下的值域空间约为 250。图 2(b)给出当 $k=200$ 时 DTW 距离最近样本的条件维及其决策维。从图中可以看到相似样本与预测条件较为接近,图 2(b)基于预测序列相似性挑选出来的子样本集的决策维值域约为[0,200],均值为 80(即该预测序列的决策维坐标期望值),同时方差明显比图 2(a)小。对于通用的时间序列,我们可以采用模糊集分段的方式来确定需要抽取的相似样本子集的容量 k。假定原始数据集的容量为 N,通过自相关性分析之后得到合适的周期取值为 q;决策维对应的最大取值为 d、最小取值为 s,将[s,d]的取值空间均分为 q 段之后,落在这 q 段中样本数量最多的空间样本容量为 n,则通过实验总结 k 的取值可通过公式 $k=N/n$ 进行计算。

3.2 面向 LSSVR 的 DTW 距离调整函数

针对式(4)中 LSSVR 给出的模型求解,可通过构建拉格朗日乘子进行参数估计以降低计算复杂度。该拉格朗日函数被定义为:

$$\mathcal{L}(\boldsymbol{w},b,\boldsymbol{e};\boldsymbol{\alpha})=J(\boldsymbol{w},\boldsymbol{e})-\sum_{i=1}^{n}\alpha_i\{\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b+e_i-y_i\} \tag{7}$$

经过最小二乘回归[6],参数 w 和 e 可被消元,从而得到:

$$\begin{bmatrix}0 & \mathbf{1}_n^T\\ \mathbf{1}_n & \boldsymbol{\Omega}+\gamma^{-1}\boldsymbol{\Delta}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix} \tag{8}$$

其中,$\boldsymbol{y}=(y_1,\cdots,y_n)^{\mathrm{T}}$ 为训练样本的决策维向量;调整基 $\boldsymbol{\Delta}=\mathrm{diag}(\delta_1^{-1},\cdots,\delta_n^{-1})$ 可依据预测条件进行调整;$\boldsymbol{\Omega}$ 表示训练实例之间的成对核距离,即:

$$\boldsymbol{\Omega}_{ij}=\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)=K(\boldsymbol{x}_i,\boldsymbol{x}_j),i,j=1,\cdots,n \tag{9}$$

考虑经由 DTW 距离下 kNN 挑选后的样本集中每个实例与预测条件 $\boldsymbol{x}$ 间均构建一个归一化条件维距离:

$$\delta_e(\boldsymbol{x}_i)=\frac{DTW(\boldsymbol{x},\boldsymbol{x}_i)-\min\limits_{\boldsymbol{x}'\in D(\boldsymbol{x})}(DTW(\boldsymbol{x},\boldsymbol{x}'))}{\max\limits_{\boldsymbol{x}'\in D(\boldsymbol{x})}(DTW(\boldsymbol{x},\boldsymbol{x}'))-\min\limits_{\boldsymbol{x}'\in D(\boldsymbol{x})}(DTW(\boldsymbol{x},\boldsymbol{x}'))} \tag{10}$$

和一个归一化决策维距离:

$$\delta_r=\frac{|y_i-\hat{D}(y)|}{\max\limits_{y'\in D(y)}|y'-\hat{D}(y)|} \tag{11}$$

基于该距离,可归一化条件维和决策维得到调整函数:

$$\rho_i=\frac{\ln(\delta_e(\boldsymbol{x}_i)/\delta_r(y_i))+\delta_r(y_i)}{\delta_e(\boldsymbol{x}_i)+\delta_r(y_i)} \tag{12}$$

通过替换式(8)中的调整基:

$$\delta_i=e^{-\rho_i\delta_e(\boldsymbol{x}_i)}+e^{-(1-\rho_i)\delta_r(y_i)} \tag{13}$$

式(4)可基于样本调整基求解,变量 e 为每个样本的偏差值。基于 LSSVR 的回归过程,该模型可通过在相关参数上求取偏导数后取极值进行求解,过程如下:

$$\begin{cases}\dfrac{\partial\mathcal{L}}{\partial\boldsymbol{w}}=0\rightarrow\boldsymbol{w}=\sum\limits_{i=1}^{n}\alpha_i\varphi(\boldsymbol{x}_i)=0\\ \dfrac{\partial\mathcal{L}}{\partial b}=0\rightarrow\sum\limits_{i=1}^{n}\alpha_i=0\\ \dfrac{\partial\mathcal{L}}{\partial e_i}=0\rightarrow\alpha_i=\gamma e_i\sum\limits_{i=1}^{n}(e^{-\rho_i\delta_e(\boldsymbol{x}_i)}+e^{-(1-\rho_i)\delta_r(y_i)})\\ \dfrac{\partial\mathcal{L}}{\partial\rho_i}=0\rightarrow-\delta_e(\boldsymbol{x}_i)e^{-\rho_i\delta_e(\boldsymbol{x}_i)}+\delta_r(y_i)e^{-(1-\rho_i)\delta_r(y_i)}=0\\ \dfrac{\partial\mathcal{L}}{\partial\alpha_i}=0\rightarrow\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b+e_i-y_i=0\end{cases} \tag{14}$$

其中,向量下标(如 e_i,ρ_i,α_i)为针对向量的每个元素分别求取偏导数。

从式(13)可以看到,该调整分别在条件维距离(式(10))和决策维归一化距离(式(11))上进行自适应得到,二者系数 ρ_i 的求解在式(14)中可解释为极值系数。由偏导数的性质可以得到,基于式(14)求解条件维和决策维系数 ρ_i 可保证模型(式(4))目标函数的经验误差最小化。在这一点上,文献[7]通过指定具体的经验值来确定条件维和决策维的影响因子。但相对于由式(14)进行最小二乘求解,指定参数的方式难以获取最优系数。此外,模型的样本误差 e_i 是采用指数加权方式通过求取偏导数得到的,在 α_i 的基础上可保证模型的经验误差最小化。基于式(14)依次求解参数 ρ_i,b,$\boldsymbol{w}$,e_i,α_i,可得到面向给定预测条件(式(10)中 $\boldsymbol{x}$)的调整模型。

3.3 预测样本模型的生成及其预测算法

由上述基于样本的训练集选择过程和模型求解方法可知,给定周期参数 d 和预测步长 p,预测任意时刻的观察值 x_n,均可基于预测条件 $\boldsymbol{x}=\{x_{n-d+1},\cdots,x_n\}$ 首先对所有 $n-d-p+1$ 个样本全集进行 kNN 相似子序列的选择,然后采用 LSSVR 求解并预测 $y=x_{n+p}$。其中,式(10)-式(13)在回归之前可根据所有 k 个基于 DTW 距离选择的训练样本及其与预测条件 $\boldsymbol{x}$ 间的距离直接获得,在此基础上的模型参数基于式(4)和式(14)进行最小二乘回归得到。考虑一个数据流环境下的周期性时间序列预测问题:伴随观测点的不断加入,当前预测条件的起始时间戳 $x_{n-d-p+1}\cdot t$ 和历史样本不断增长,因而该方法总可以根据最新加入的观测值构建预测条件,挑选训练样本子集,从而进行模型的生成和预测。不失一般性,算法描述假定一个无界样本集 $\mathcal{S}$,基于周期参数 d、预测步长 p 和样本子集容量 k,对最新观测值进行训练子集的选择和基于 LSSVR 的预测。

算法 1 基于 kNN 的样本选择算法

输入:当前历史样本集合 $D_{n-d-p+1}$,预测条件 $y=x_{n+p}$

输出:训练样本子集 $D(\mathbf{x})\subset D_{n-d-p+1}$

1. 基于预测条件 $y=x_{n+p}$,计算所有 $D_{n-d-p+1}$ 中样本条件维 $\mathbf{x}'$ 与 $y=x_{n+p}$ 的 DTW 距离,形成距离集合 $dist_{n-d-p+1}$;
2. 对 $dist_{n-d-p+1}$ 增序排列,取 $D_{n-d-p+1}$ 中对应于前 k 个距离的训练样本,构成 $D'(\mathbf{x})$;
3. 保留 $D'(\mathbf{x})$ 中每个周期内与预测条件的 DTW 距离最小的样本子集,并将其输出到 $D(\mathbf{x})$。

算法 1 对不断增加的样本集对时间戳递增的预测条件 $\boldsymbol{x}$ 进行 kNN 子样本选择。随着新观测值加入到序列中,预测样本全集不断产生新的周期子序列,当前时刻的所有观测值 $x_1,\cdots,x_n$ 可以形成周期子序列训练全集 $D_{n-d-p+1}$。通过逐一计算其中每个子序列与预测条件之间的 DTW 距离,该算法在预测一个新的条件时具有 $O(n)$ 的 DTW 距离计算复杂度。此外,虽然时态拉伸会增加最小二乘的风险误差,但 DTW 距离对时态拉伸不敏感。如图 2(a)所示,DTW 距离下的最近样本子集将包含 9 个时刻的 kNN 样本,事实上峰值在 1 到 8 的样本子集与预测条件尽管形状极相似,但是其时态拉升(峰值 1 到 8 相对于预测条件 9 分别时态拉伸了 8 到 1)会增加模型的风险误差。在实际应用这些样本前,依据时态距离对样本子集 $D'(\boldsymbol{x})$ 进行删除,仅保留时态间隔 d 内的候选样本中与 $\boldsymbol{x}$ 的 DTW 最小的样本,并最终构成训练集 $D(\boldsymbol{x})$。如图 2(b)所示,基于算法 2 的第 2-3 行,保留的样本均与预测条件 $\boldsymbol{x}$ 具有相同的峰值时间戳。

给定上述 3 个参数 d,p,k，对于每个预测条件 $\boldsymbol{x}$，可以得到一个训练集 $D(\boldsymbol{x})$。该训练集中每个样本的条件维是 $\boldsymbol{x}$ 的全局 kNN 样本，且其时态距离总体上与 $\boldsymbol{x}$ 等周期间隔。假定依据 3.2 节得到相关系数后，可以基于下式完成对 $\boldsymbol{x}$ 的预测：

$$y=f(\boldsymbol{x})=\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i)+b \tag{15}$$

基于上述分析，依据选择后的样本集 $D(\boldsymbol{x})$，模型生成和针对 $\boldsymbol{x}$ 的预测如算法 2 所示。

算法 2　加权调整 DTW-LSSVR 预测算法

输入：训练样本子集 D(**x**)，预测条件 **x**

输出：预测值 $y=x_{n+p}$

1. 基于式(13)计算训练样本与预测条件的调整基 δ_i；
2. 采用交叉验证求取式(4)在调整基(式(13))上的模型参数 b，**w**，e_i；
3. 基于式(14)得到样本权重系数 α_i 并完成模型训练；
4. 依据式(15)在 D(**x**)上的训练，基于权重系 α_i 得到 **x** 的预测值 $y=x_{n+p}$。

由前文的分析可知，算法 2 第 1 行的调整可以直接根据确定后的训练样本子集 $D(\boldsymbol{x})$ 的条件维和决策维分别计算。第 2 行按照交叉验证方式对式(4)给出的约束最优化问题进行求解，并确定其截距等参数。实现过程中，所有 LSSVR 可采用 5-fold 交叉验证进行求解[4]。由于算法 1 仅保留 $\boldsymbol{x}$ 的 kNN 样本中同等周期间隔的训练样本，因此相较于原始样本集 $D_{n-d-p+1}$，训练样本集 $D(\boldsymbol{x})$ 显著较小，有效减少了训练时间。值得注意的是，原始 LSSVR 需对每个新加入样本更新训练全集，尽管可采用增量方式[10]更新模型参数，但其更新代价仍然较在 $D(\boldsymbol{x})$ 上训练更大(实际应用中大样本环境下 $D(\boldsymbol{x})$ 的样本容量一般较 $D_{n-d-p+1}$ 小 1～2 个数量级)。第 3—4 行通过常数代价分别计算权重系数和预测值 $y=x_{n+p}$。对于任意增量数据流 $\mathcal{S}$，对最近时刻的预测可在当前样本全集的基础上依次调用算法 1 和算法 2 实现，并基于新加入的观测值产生新的预测样本，更新样本全集，以实现联机(Online)时间序列预测。由式(14)可知，模型的调整与条件和决策距离有关，即该求解过程可以有效保证条件维接近的历史样本能够更大程度地影响决策函数。在周期性时间序列中，相同曲线的子样本集逐点趋于一致。当样本空间增加时，通过 k 的设定，测试样本容量随之增加，可以保证测试样本容量与预测条件之间的相似性，从而保证了模型的健壮性。

4　实验验证

本节引入了几种常用的周期性时间序列进行实验比较与验证，对模型的精度和参数影响进行了验证性研究。文献[1]和文献[7]的实验数据 Mackey-Glass(MG)、Lorenz(LZ)等长时间序列在序列预测的研究中均被广泛采用。实验环境统一为 Matlab 2014b，所有参与比较的算法均采用 Matlab 语言实现：LSSVR 来源于文献[6]，Neural Network 的 ANN 算法来源于文献[19]，ARIMA 来源于文献[4]，通过在 LSSVR 中加入算法 1 和调整过程得到 DTW-LSSVR。

4.1　时间序列预测模型的比较

通过对不同时间序列预测模型进行比较，来验证算法的有效性。采用 Sales 数据预测样本的最后 35 个观测值，比较 LSSVR-mean(对样本观测值求取周期内的观测均值)，NN 和 DTW-LSSVR 3 类方法，结果如图 3 所示。

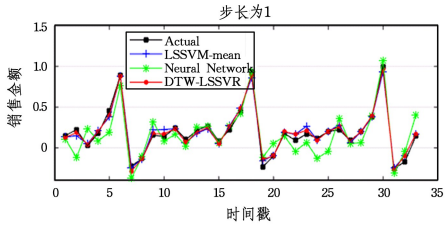


图 3　销售数据单步预测的比较结果

Fig. 3　Comparison results for 1-step forecast on sales data

从图 3 中可以看到，DTW-LSSVR 相对于其他方法具有明显优势：在时刻 2，10 和 24 等预测点上，其绝对误差比 LSSVR-mean 的小一半以上；同时，在大多数观测点上比 NN，DTW-LSSVR 和 LSSVR-mean 的预测误差更小。

图 4 为取 CO_2 数据的 60 个观测值做 3 步预测所得到的结果与原始观测值(Actual)的对比图。从图中可以看到，相对于原始 LSSVR 模型和 LSSVR-mean 模型，基于全部样本训练的 LSSVR 模型表现出显著更高的误差，主要原因是 CO_2 数据随着时间变化总体递增，而该方法用历史样本进行训练，导致预测值趋近于历史均值(如式(15)所示)。考虑到该序列各周期内的变化相对平稳，采用均值方法对数据进行转换，使得 DTW-LSSVR 较 LSSVR-mean 的提升并不是特别显著，前者在波动时刻点 4 和 27 的精度较后者有一定提升。

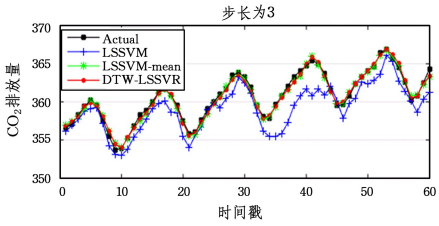


图 4　采用 CO_2 数据进行 3 步预测的比较结果

Fig. 4　Comparison results for 3-step forecast on CO_2 data

采用 Laser 数据预测样本的最后 100 个观测值，共做 5 步预测，比较 NN 和 DTW-LSSVR 两类方法，结果如图 5 所示。由于步长较大，在突变的波动点容易产生较大的预测误差。例如，在时刻 21－26 这几个预测点上，相对于 NN，DTW-LSSVR 的结果更加接近真实值，绝对误差小一半以上；同时，在预测点 5，12，13 上，DTW-LSSVR 也比 NN 在预测精度上更有优势。

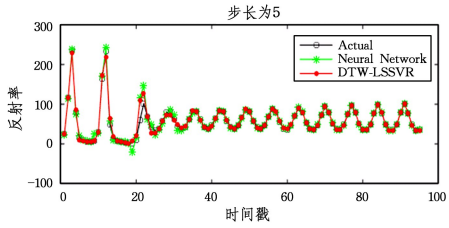


图 5　采用 Laser 数据进行 5 步预测的比较结果

Fig. 5　Comparison results for 5-step forecast on Laser data

4.2　模型参数影响的验证与分析

以平均绝对误差(Mean Absolute Error，MAE)为衡量指标，采用 8 种长时间序列测试数据集进行实验对比。通过将 k 置为 100 和 200，对 DTW-LSSVR 进行样本选择。实验结果如表 1 所列，其中 ARIMA 和 NN 作为对比方法来验证所

提模型的精度，黑体部分突出了同一数据下的最小误差方法（其中 MG 和 LZ 后附生成数据集的参数，DL 为 DTW-LSSVR 的首字母缩写，PE 为 Porland Electronic 的首字母缩写）。

表 1 长事件序列预测的 MAE 对比

Table 1 MAE comparison of long event sequence prediction

数据集	*MAE*			
	ARIMA	NN	DL-100	DL-200
MG-50	0.0031	0.0013	0.001	**0.0009608**
LZ-9/3	0.1769	0.0596	**0.0555**	0.0629
Laser	13.27	1.5019	1.0976	**1.0816**
PE	0.0393	0.057	0.028	**0.0274**
Traffic	0.0607	0.1166	0.0606	**0.0507**
Taylor	280.4	**136.29**	143.69	175.84
Sales	0.1759	0.1187	0.0407	**0.0339**
CO_2	**0.3283**	1.6072	0.6989	0.7809

从表 1 中不难看出，6 种数据的结果在 DL 下表现得更优，其中 5 种在 k 为 200 时候最优。对于递增时间序列 CO_2，ARIMA 的滑动平均测试表现得更好，而在变化值域较小的实验数据 Taylor 中，NN 的误差更小。对于长周期序列 Sales，Laser 和 LZ-9/3，由于其中观测值域在周期间的差异显著，ARIMA 比其他竞争对手的精度显著更低。同时也可以看到，对于 Sales 和 CO_2 这类非连续或周期间统计差异显著的子序列集，NN 的误差要显著高于 DL 模型。综合图 3—图 5 中 DL 模型与其他几种模型的对比不难看出，DL 模型对于周期统计差异显著的序列和波动较大的序列具有较好的健壮性，这可以通过其 kNN 的处理思路给予解释：由于 DL 模型总是依据每个预测条件维挑选形状最相似的样本进行预测，因此在大样本环境下其经验误差比 LSSVR 低；同时，LSSVR 基于历史样本进行回归预测，对于观测值不连续的数据，该方法总体上比 ARIMA 这类基于滑动平均思想的模型更有优势。比较 DL-100 和 DL-200 的精度也可以看到，模型对于参数的设置有一定依赖，但总体上在一定范围内对 k 的取值并不敏感。

结束语 本文基于 SVM 函数拟合方法，分析了产生大量噪声点的原因，通过对时间序列周期分段构建相似性度量，抽取预测序列对应的样本子集进行建模训练，同时基于误差调谐函数对 SVM 的核函数进行调整。另一方面，基于常用周期性时间序列进行了相关的实验论证，并将所提算法与已有算法进行指标比较，分析了模型参数的敏感性。实验结果证明，所提方法提升了最终的预测精度，并降低了时间成本，在大样本周期性时间序列的预测问题上有一定的应用价值。

参 考 文 献

[1] GHAREHBAGHI A, ASK P, BABIC A. A pattern recognition framework for detecting dynamic changes on cyclic time series [J]. Pattern Recognition, 2015, 48(3): 696-708.

[2] CHEN T T, LEE S J. A weighted ls-svm based learning system for time series forecasting [J]. Information Sciences, 2015, 299(1): 99-116.

[3] KANG P, CHO S. Locally linear reconstruction for instance-based learning [J]. Pattern Recognition, 2008, 41(1): 3507-3518.

[4] BOUSQUET O, ELOSSEEFF A. Stability and generalization [J]. Journal of Machine Learning Research, 2001, 3(2): 499-526.

[5] PETITJEAN F, KETTERLIN A, GANCARSKI P. A global averaging method for dynamic time warping, with applications to clustering [J]. Pattern Recognition, 2011, 44(1): 678-693.

[6] VIKJORD V V, JESSEN R. Information theoretic clustering using a k-nearest neighbors approach [J]. Pattern Recognition, 2014, 47(9): 3070-3081.

[7] ROJAS I, VALENZUELA O, ROJAS F, et al. Soft-computing techniques and arma model for time series prediction [J]. Neurocomputing, 2008, 71(4): 519-537.

[8] SUYKENS J A K, BRABANTER J D, LUKAS L, et al. Weighted least squares support vector machines: robustness andsparse approximation [J]. Neurocomputing, 2002, 48(1-4): 85-105.

[9] XU H. Robustness and regularization of support vector machines [J]. Journal of Machine Learning Research, 2009, 10(3): 1485-1510.

[10] SANTOS J D A, BARRETO G A. A regularized estimation framework for online sparse lssvr models [J]. Neurocomputing, 2017, 238(1): 1-12.

[11] CLÁUDIO R D S, SOARES C, KNOBBE A. Entropy-based discretization methods for ranking data [J]. Information Sciences, 2016, 329(1): 921-936.

[12] PARK H J, PARK W J, JUNG T, et al. On general purpose time series similarity measures andtheir use as kernel functions in support vector machines [J]. Information Sciences, 2014, 281(4): 478-495.

[13] WU H H, LIU G H, WANG W. The problem of similarity matching for uncertain time series[J]. Computer Research and Development, 2014, 51(8): 1802-1810. (in Chinese)
吴红花，刘国华，王伟. 不确定时间序列的相似性匹配问题[J]. 计算机研究与发展，2014，51(8)：1802-1810.

[14] HAN Z M. Research on Effective Clustering Algorithm for Hot Topic Time Series[J]. Journal of Computer, 2012, 35(11): 2337-2347. (in Chinese)
韩忠明. 面向热点话题时间序列的有效聚类算法研究[J]. 计算机学报，2012，35(11)：2337-2347.

[15] YUAN J D, WNAG Z H. Time Series Representation in Classification Algorithms[J]. Computer Science, 2015, 42(3): 1-7. (in Chinese)
原继东，王志海. 时间序列的表示于分类算法综述[J]. 计算机科学，2015，42(3)：1-7.

[16] HAN M, XU M L, MU D Y. Application of Non-nuclear Correlation Vector Machine in Time Series Prediction[J]. Journal of Computer, 2014, 37(12): 2427-2432. (in Chinese)
韩敏，徐美玲，穆大芸. 无核相关向量机在时间序列预测中的应用[J]. 计算机学报，2014，37(12)：2427-2432.

[17] CHEN Y, SHI Z H. Mixed Financial Time Series Forecasting Model Based on Adaboost and Regularized ELM and Its Application [J]. Mathematical Statistics and Management, 2017, 36(1): 112-124. (in Chinese)
陈艳，石智慧. 基于 Adaboost 和正则化 ELM 的混合金融时间序列预测模型及其应用[J]. 数理统计与管理，2017，36(1)：112-124.

[18] HAO Y H, ZHANG H F. Incremental Learning Method Based on Double Support Vector Regression Machine[J]. Computer Science, 2016, 43(2): 230-235. (in Chinese)
郝运河，张浩峰. 基于双支持向量回归机的增量学习方法[J]. 计算机科学，2016，43(2)：230-235.