小时间尺度网络流量混沌性分析及趋势预测

温祥西[1],孟相如[1],马志强[1],张永春[2]
(1.空军工程大学信息与导航学院,陕西西安 710077; 2.装甲兵学院,安徽蚌埠 233050)

摘 要: 小时间尺度的网络流量的混沌性被噪声掩盖难以预测,本文通过局部投影降噪得到可预测的混沌性流量趋势.针对网络流量存在的时变性和长周期性,提出一种最优样本子集在线模糊最小二乘支持向量机(Least Squares Support Vector Machine,LSSVM)预测方法:以与预测样本时间上以及欧式距离最近的样本点构成最优样本子集,并对其模糊化处理,最后采用模糊 LSSVM 训练获得预测模型.通过分块矩阵降低预测模型在线更新的运算复杂度.对真实网络流量的降噪以及预测的结果表明本文方法能够快速准确的预测网络流量趋势.

关键词: 网络流量; 趋势预测; 混沌理论; 最优样本子集; 最小二乘支持向量机
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2012) 08-1609-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.08.18

The Chaotic Analysis and Trend Prediction on Small-Time Scale Network Traffic

WEN Xiang-xi[1], MENG Xiang-ru[1], MA Zhi-qiang[1], ZHANG Yong-chun[2]
(1. *Institute of Information and Navigation, Air Force Engineering University, Xi'an, Shaanxi* 710077, *China*;
2. *Academy of Armoured Force, Bengbu, Anhui* 233050, *China*)

Abstract: The chaotic performance of small-time scale network traffic was covered by noise, which made the traffic unpredictable. This paper introduces the local projection to denosie network traffic; a chaotic and predictable traffic trend is obtained. As the network traffic series is long-period and time-varying, a new method named optimal training subset online fuzzy least squares support vector machines (OTSOF-LSSVM) is proposed. Samples temporal and distance nearest to prediction sample are chosen as optimal training subset, and the subset are fuzzified. On this basis, the prediction model is established by fuzzy LSSVM. The model update computational complexity is reduced by partitioned matrix calculation. The noise reduction and trend prediction on network traffic shows the proposed method can predict the trend quickly and exactly.

Key words: network traffic; trend prediction; chaotic theory; optimal training subset; least squares support vector machine (LSSVM)

1 引言

网络流量是记录和反映网络以及用户活动的重要载体,对它的预测能够指示未来网络的运行状况,为网络的带宽分配、流量控制、选路控制、故障管理等提供有效依据.网管人员可以依据预测结果采取相应措施避免故障发生,减小网络故障带来的破坏,增强网络的可生存性,降低维护成本.

目前网络流量的预测研究较多,其中很大一部分是以小时为单位的大时间尺度预测[1,2].虽然这些研究获得了较好的预测效果,但是在当前复杂多变的网络环境,它们只能从宏观上简单调整,相较而言小时间尺度的预测具有更重要的现实意义.然而小时间尺度网络流量数据由于其高维数、非线性等特点导致预测模型建立困难.传统的线性预测方法如 AR、ARMA、ARIMA 等[3~5]预测精度较差;基于神经网络、灰色模型、SVM 等非线性方法[6~9]能够较好的适应网络各状态参数高维、非线性等特点,提高了预测精度.随着网络流量数据的长周期性和混沌性的发现,结合混沌特性的研究得到了发展,文献[10]依据混沌序列的最大 Lyapunov 指数不变进行了预测,而结合神经网络、SVM 等智能方法和混沌性的研究获得了更好的预测效果[11,12].本文选择 SVM 作为流量预测方法,网络流量往往是时变的,传统的离线式 SVM 预测方法难以实时跟踪.文献[13]提出了在线 LSSVM 算法,实时更新预测模型,能够预测时变的混沌序列.文献[14]分析了在线 LSSVM 训练过程中样本的

收稿日期:2011-09-06;修回日期:2012-05-07
基金项目:国家自然科学基金(No.61003252);全军军事学研究生课题(No.2011JY002-524);空军工程大学研究生创新基金(No.201105)

更新方式,通过迭代修改训练函数,提高了训练效率.基于在线 LSSVM 的混沌时间序列预测本质上是一种机器学习方法,机器学习是通过对历史样本(先验知识)的学习来寻找解决当前问题的方法.机器学习效果的好坏往往取决于对先验知识的占有量,对类似问题学习的越充分(相当于对类似历史样本训练的越多),越容易获得问题解.但是,网络流量具有长周期性,要获得足够多的类似的样本需要较大的时间窗口,难以满足实时性要求.同时根据作者自己的工作和文献[2,5]中发现这些预测获得的结果均存在"一步延迟"现象,即预测结果同实际流量值之间存在一个单位的滞后.这些预测值在波形上与真实值吻合的非常好,但仔细观察可以发现看似很好的预测结果实际上并不正确,因为所得的预测结果总是落后真实值一个时间单位,也就是说此时获得的预测结果跟踪的是当前的值,而并非预测未来的值.作者认为这主要是由于流量中的高维噪声掩盖了它的混沌性,使得流量难以准确预测.

为解决上述问题,本文对网络流量数据进行降噪处理获得网络流量的趋势序列.网络流量的趋势序列去除了短暂的突发流量的影响,从本质上体现了网络的发展趋势,对其的预测研究更有利于网络的控制和故障管理.在此基础上针对网络流量趋势的特征提出 OTSOF-LSSVM 预测算法实现网络流量的精确预测.

2 理论基础

2.1 相空间重构理论

本文进行网络流量趋势预测的基础是混沌理论,而混沌时间序列预测的基础是状态空间的重构理论,即把具有混沌特性的时间序列重建为一种低阶非线性动力学系统.通过相空间重构,可以找出隐藏的混沌吸引子的演化规律,使现有的数据纳入某种可描述的框架之下.根据 Takens 定理[15],网络流量趋势预测问题可以描述为在一定条件下对满足特定条件的 m(相空间的嵌入维数)和 τ(时延),存在一个光滑映射,使得 $f: R^m \to R$,即:

$$x_{n+1} = f(x_{n-(m-1)\tau}, x_{n-(m-2)\tau}, \cdots, x_{n-\tau}, x_n) \quad (1)$$

其中 x_{n+1} 是预测的目标值(未来时刻网络流量值),$\{x_{n-(m-1)\tau}, x_{n-(m-2)\tau}, \cdots, x_{n-\tau}, x_n\}$ 为待预测样本,f 为预测模型.对于长度为 n 的时间序列,经过相空间重构,可以得到学习样本:

输入:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_{n-1-(m-1)\tau} \end{bmatrix} = \begin{bmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(m-1)\tau} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n-1-(m-1)\tau} & x_{n-1-(m-2)\tau} & \cdots & x_{n-1} \end{bmatrix}$$

对应输出:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1-(m-1)\tau} \end{bmatrix} = \begin{bmatrix} x_{2+(m-1)\tau} \\ x_{3+(m-1)\tau} \\ \vdots \\ x_n \end{bmatrix} \quad (2)$$

预测样本:$\boldsymbol{x}_{n-(m-1)\tau} = [x_{n-(m-1)\tau}, \cdots, x_{n-\tau}, x_n]$.

将预测样本输入映射(1)中获得预测目标 x_{n+1} 时刻的值.

2.2 LSSVM 预测原理

LSSVM 是 SVM 的一种改进,引入最小二乘损失函数和等式化约束的方法,使问题的求解变为解线性方程,避免了解二次规划问题,所需的计算资源较少,具有较快的求解速度.在获得最优样本子集的基础上,本文选择 LSSVM[16] 作为学习机器构建预测模型 f.基于 LSSVM 网络流量回归预测可以描述为:

对重构后的样本集 $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^k$,$\boldsymbol{x}_i$ 为 m 维输入向量,y_i 为一维输出向量,k 为样本个数.由于 $\boldsymbol{x}_i$ 与 y_i 间为非线性关系,因此将 $\boldsymbol{x}_i$ 映射到高维特征空间中,LSSVM 的基本思想是在高维空间中对样本进行线性回归:

$$y = \boldsymbol{w}^{\mathrm{T}} \varphi(\boldsymbol{x}) + b \quad (3)$$

其中,$\varphi(\boldsymbol{x})$ 为非线性映射函数,$\boldsymbol{w}$ 为法向量,b 为偏置量,根据结构风险最小化原理,对以上问题的求解可描述如下:

$$\begin{aligned} &\min_{w,b,e} \quad J(\boldsymbol{w}, b, e) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2}\sum_{i=1}^{k} \xi_i^2 \\ &\text{s.t.} \quad y_i - \xi_i = \boldsymbol{w}^{\mathrm{T}} \varphi(\boldsymbol{x}_i) + b \ , \ i = 1, \cdots, k \end{aligned} \quad (4)$$

C 为惩罚因子,ξ_i 为训练误差.为求解此优化问题,可引入 Lagrange 函数:

$$L(\boldsymbol{w}, b, e, \alpha) = J(\boldsymbol{w}, b, e) + \sum_{i=1}^{k} \alpha_i [y_i - \xi_i - \boldsymbol{w}^{\mathrm{T}} \varphi(\boldsymbol{x}_i) - b] \quad (5)$$

其中 $\alpha_i, i = 1, \cdots, k$ 为 Lagrange 乘子,由 KKT 条件得到如下关系式:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}} = 0 \ \to \ \boldsymbol{w} = \sum\limits_{i=1}^{k} \alpha_i \varphi(\boldsymbol{x}_i) \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \ \to \ y_i - \xi_i - \boldsymbol{w}^{\mathrm{T}} \varphi(\boldsymbol{x}_i) - b = 0 \\ \dfrac{\partial L}{\partial b} = 0 \ \to \ \sum\limits_{i=1}^{k} \alpha_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = 0 \ \to \ \alpha_i = C\xi_i \end{cases} \quad (6)$$

核函数 $K_{ij}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)\rangle$,则式(6)的求解可形式化为:

$$\begin{bmatrix}0 & \boldsymbol{e}^{\mathrm{T}}\\ \boldsymbol{e} & Q+C^{-1}\boldsymbol{I}\end{bmatrix}\times\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix}\tag{7}$$

其中 Q 是元素为 K_{ij} 的 $k\times k$ 阶核矩阵,$\boldsymbol{I}$ 为单位矩阵,向量 $\boldsymbol{e}=[1,\cdots,1]^{\mathrm{T}}$,向量 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_k]^{\mathrm{T}}$,向量 $\boldsymbol{y}=[y_1,\cdots,y_k]^{\mathrm{T}}$.求解式(7)得到 α_i,b 代入式(3)中即可得到 LSSVM 的混沌时间序列回归模型为:

$$\begin{aligned}f(\boldsymbol{x})&=\sum_{i=1}^{k}\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x})+b\\ y&=f(\boldsymbol{x})\end{aligned}\tag{8}$$

对应的预测样本为 $\boldsymbol{x}_p$,则得到的预测值为 $y_p=f(\boldsymbol{x}_p)=\sum_{i=1}^{k}\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x}_p)+b$.

3　网络流量混沌性分析

混沌是一种介于确定性和随机性之间的随机敏感充分不规则的非线性动力现象,宏观上表现为无序无律的混乱运动,以及对初值十分敏感的蝴蝶效应;微观上呈现无穷嵌套几何自相似性.混沌时间序列区别于随机时间序列,它是有序和无序、确定性和随机性的中间态,具有短期的可预测性和长期的不可预测性.网络流量存在混沌性,但是关于网络流量混沌性的研究却很少,本节通过二维相图分析网络流量数据.

3.1　网络流量序列分析

本文分析实际流量数据 LBL-tcp-3.tcp(http://ita.ee.lbl.gov/html/contrib/LBL-TCP.html),此数据采样时间共 2 个小时,数据共 1789995 个.本文分析目标为小时间尺度预测,故以 1s 为间隔进行重采样,得到长度为 7199 的流量序列 $tr(i)$.对这些流量数据进行归一化,得到的原始的网络流量时间序列如图 1.

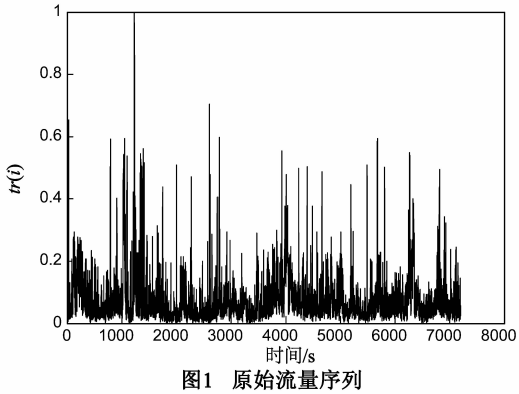


图1　原始流量序列

最大 Lyapunov 指数是目前常用的时间序列混沌特性检验方法,本文采用小数据量法[17]计算原始流量的最大 Lyapunov 指数为 0.31 > 0,可以判断小时间尺度下的网络流量是混沌的.依据原始的网络流量序列 $s(i)$ 获得其二维相图,如图 2.

从网络流量序列的二维相图可以看出小时间尺度的网络流量数据相图紊乱,并没有显示出混沌时间序列所特有的无穷嵌套几何自相似性.这主要是因为高维的噪声将网络流量的混沌性掩盖.以这个原始序列作为混沌预测的目标,采用 LSSVM、在线 LSSVM 等方法进行预测,作者同样获得了类似文献[5]的结果,即也存在"一步延迟"现象,这主要是由于网络流量中的噪声是随机的,无法预测的,所以叠加了这些噪声的网络流量预测模型的预测效果也比较差.

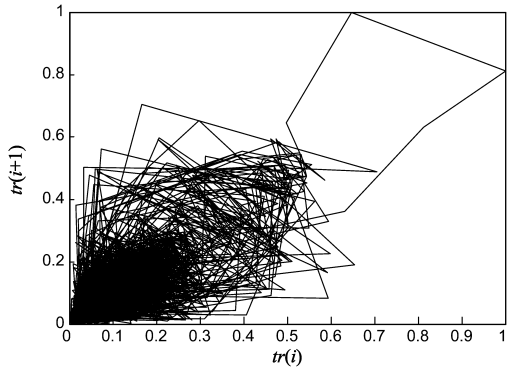


图2　原始流量的二维相图

3.2　网络流量降噪

从上一小节的分析可以看出网络流量的混沌性被高维的噪声所掩盖,对原始的流量数据预测困难.因此,需要对原始数据进行降噪处理,以恢复其内在的混沌性.现有的混沌去噪方法主要有:影子定理去噪、小波方法以及局部投影算法.本文采用局部投影算法对网络流量数据进行降噪处理[18].

局部投影算法实质上是一种子空间去噪算法,它采用局部投影策略来估计噪声,其中局部邻域的选取直接影响去噪的效果.由于受到噪声的影响,相空间中的相点都偏离了各自的初始位置,因此,在所选中的邻域点中势必会包含一些虚假邻域点,降低去噪性能.PCA 从原理上来说是子空间去噪的一种特殊形式,只是最终的目的不相同,本文首先对样本进行主成分分析(Principle Component Analysis, PCA)处理,在主元空间中寻找局部邻域以降低邻域虚假率.降噪的最终目的是为了实时的预测网络流量,因此降噪的过程也应是实时的,但是由于降噪过程涉及矩阵奇异值分解,计算量较大,本文提出一种在线的局部投影降噪方法,具体过程为:

步骤 1　选择合适的嵌入维数 m 和时延 τ(这里 $m=6,\tau=1$),对已知流量数据进行相空间重构,并存储.

步骤 2　对该相空间进行 PCA 处理,主成分分量占

90%,并在主成分空间中进行邻域选取(邻域点个数取10),在原相空间中找出对应的相点,采用局部投影降噪对已知的流量数据进行降噪处理,在噪声估计时引入语音增强方法中的频域约束(Spectrum Domain Constraint,SDC)[19]方法,获得降噪后完整的流量序列.

步骤 3 对于新到的样本数据,对其进行 PCA 降维处理,并在存储的历史样本中找出其邻域相点集(H).

步骤 4 对所获得的局部邻域进行投影降噪处理,获得最新时刻降噪后的流量值.

步骤 5 分别找出 H 中各个样本点的邻域集(H_i,i =1,2,…,10),并将此时的样本数据置于各个邻域集中,分别对这些进行降噪处理,获得最新时刻在各个邻域中降噪后对应的流量值.

步骤 6 取每次降噪后获得的流量平均值作为最终的降噪结果存储于趋势序列中.

步骤 7 滑动窗口,获得新样本数据,重复 3 ~ 6 步.

对原始的网络流量数据进行降噪后获得的流量趋势序列 $tr'(i)$如图 3.

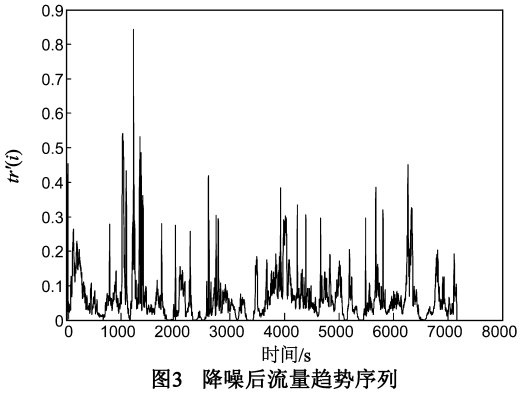


图3 降噪后流量趋势序列

其中,前 2000 个点作为已知的历史数据进行降噪,后面部分采用上述方法进行实时降噪处理.降噪过程将原始流量数据中高维的随机部分消除,保留其平滑的低维特征,体现了网络流量的变化趋势.由于每个降噪过程涉及的矩阵分解以及重构大小仅为 6 × 10,因此每一次的降噪过程所需的时间很少(平均为 0.024s),且最终的降噪结果为多次邻域降噪结果的均值,结果更加平滑准确.获得的降噪后序列的二维相图,如图 4.

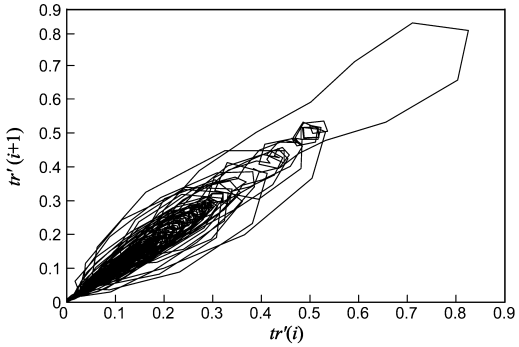


图4 降噪后流量二维相图

从图 4 可以看出,通过降噪原来的杂乱无序的网络流量相图显示出了自身内在的混沌性,但是这种混沌性不是固定不变的,又存在着时变性.

4 流量预测建立

从上面的降噪分析以及二维相图可以看出网络流量具有混沌性、时变性以及长周期性.单纯的采用 LSSVM 难以满足预测要求,这里提出一种解决方法:OTSOF-LSSVM.下面从最优样本的子集选取和算法的推导具体介绍 OTSOF-LSSVM 算法.

4.1 最优样本子集选择

由最大 Lyapunov 指数预测方法[10]可以得出:在一个混沌时间序列中与预测样本最相关的信息在与其欧式距离最近的样本点中.从几何上解释,与预测样本点欧式距离最近的样本点位于与预测样本点所在轨道最近的邻域轨道上,根据混沌时间序列的无穷嵌套自相似性,该样本点与其最近邻点的发展轨迹是相似的,因此该最近邻点包含了预测的最大的预测相关信息.类似的,与预测样本欧式距离最小的样本点集处于与预测样本最近的邻域轨道集上,同样也与预测样本有着很大的相关性,因此可以说在混沌时间序列中预测样本与训练样本之间是距离相关的.另外,由于网络流量数据是时变的,我们假设网络流量序列是时间相关的且时间上离预测样本最近的 k_1 个样本与预测样本完全相关.根据上面的分析,获得的关于预测样本的最优样本子集包括两部分:与其时间上最近的 k_1 个样本和距离上最近的 k_2 个样本.其中,时间相关的样本可以由时间窗确定,距离相关的可由预测样本和历史样本集联合确定.随着距离的增大,样本点与预测样本的相关性逐渐减弱,为了更好的描述 k_2 个样本点与预测样本之间的关系,这里引入模糊逻辑来表示它们之间的相关性,即用模糊隶属度表示样本点的重要性.模糊隶属度采用 ε 不敏感模型确定:

$$s_i=\begin{cases}1, & 0\leqslant d_i\leqslant\varepsilon_1\\ \dfrac{\varepsilon_2-d_i}{\varepsilon_2-\varepsilon_1}, & \varepsilon_1<d_i<\varepsilon_2\\ \eta, & d_i\geqslant\varepsilon_2\end{cases}\tag{9}$$

其中 d_i 表示第 i 个样本点到预测样本的欧式距离,η 是极小的常数.

4.2 OTSOF-LSSVM 算法

对于模糊 LSSVM,设样本总数为 k,对不同样本进行模糊化处理,即引入不同的惩罚系数 s_i,则式(5)变为:

$$\min_{w,b,e} \quad J(\boldsymbol{w},b,e)=\frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{k}s_i\xi_i^2 \qquad (10)$$

$$\text{s.t.} \quad y_i-\xi_i=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b\ ,\ i=1,\cdots,k$$

其中 s_i 为第 i 个样本模糊隶属度,同样经过 Lagrange 函数和 KKT 条件变换可得:

$$\begin{bmatrix}0 & \boldsymbol{e}^{\mathrm{T}}\\ \boldsymbol{e} & \boldsymbol{Q}+\boldsymbol{C}_s\end{bmatrix}\times\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix} \qquad (11)$$

其中 $\boldsymbol{C}_s=\mathrm{diag}(C^{-1}s_1^{-1},\cdots,C^{-1}s_k^{-1})$. 由上一小节可知选择的最优子集包括两部分:时间相关的的 k_1 个样本($S_{k_1}=\{\boldsymbol{x}_i,y_i\}_{i=1}^{k_1}$)和距离上最近的 k_2 个样本($S_{k_2}=\{\boldsymbol{x}_i,y_i\}_{i=1}^{k_2}$). S_{k_1}与待测样本是完全相关的,故 $s_i=1,(i=1,2\cdots,k_1)$, S_{k_2}与待测样本是距离相关的,设 S_{k_2}中各样本的模糊隶属度为 $s_i=1,(i=k_1+1,k_1+2\cdots,k)$,则:

$$\boldsymbol{C}_s=\begin{bmatrix}C^{-1}\boldsymbol{I}_{k1} & \boldsymbol{0}\\ \boldsymbol{0} & \mathrm{diag}\,(C^{-1}s_i^{-1})\end{bmatrix}\ ,\ i=k_1+1,\cdots,k \qquad (12)$$

设 $\boldsymbol{G}=[\boldsymbol{Q}+\boldsymbol{C}_s]$,代入式(11)求解得:

$$\begin{cases}b=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{y}/\boldsymbol{e}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{e}\\ \boldsymbol{\alpha}=\boldsymbol{G}^{-1}(\boldsymbol{y}-\boldsymbol{e}b)\end{cases} \qquad (13)$$

由式(13)可知,求解 $\boldsymbol{\alpha}$、b 的关键在于求解 $\boldsymbol{G}^{-1}$,这是一个矩阵求逆运算,矩阵规模为 $k\times k$,其计算复杂度为 $o(k^3)$. 根据分块矩阵的思想可将 $\boldsymbol{G}$ 重新写为:

$$\boldsymbol{G}=\begin{bmatrix}\boldsymbol{Q}_{k_1}+C^{-1}\boldsymbol{I}_{k_1} & \boldsymbol{R}^{\mathrm{T}}\\ \boldsymbol{R} & \boldsymbol{Q}_{k_2}+\mathrm{diag}(C^{-1}s_i^{-1})\end{bmatrix}=\begin{bmatrix}\boldsymbol{G}_{k_1} & \boldsymbol{R}^{\mathrm{T}}\\ \boldsymbol{R} & \boldsymbol{G}_{k_2}\end{bmatrix} \qquad (14)$$

其中,$\boldsymbol{G}_{k_1}$由样本集 $S_{k_1}=\{\boldsymbol{x}_i,y_i\}_{i=1}^{k_1}$的核矩阵 $\boldsymbol{Q}_{k_1}$和模糊隶属度矩阵 $C^{-1}\boldsymbol{I}_{k_1}$确定; $\boldsymbol{G}_{k_2}$由样本集 $S_{k2}=\{\boldsymbol{x}_i,y_i\}_{i=k_1+1}^{k}$的核矩阵 $\boldsymbol{Q}_{k_2}$和模糊隶属度矩阵 $\mathrm{diag}(C^{-1}S_i^{-1}),(i=k_1+1,\cdots,k)$确定; $\boldsymbol{R}=K(\boldsymbol{x}_i,\boldsymbol{x}_j),(\boldsymbol{x}_i\in S_{k_2},\boldsymbol{x}_j\in S_{k_1})$. 对于文中提出的在线问题,若采取简单的每次样本更新则重新计算一遍 $\boldsymbol{G}^{-1}$,计算量大. 观察样本的更新过程,$\boldsymbol{G}_{k_2}$随着待测样本的改变完全改变,因此文献[14]提出的在线更新运算在此并不适用. 但是 $\boldsymbol{G}_{k_1}$部分的更新采用滑动时间窗的方式实现的可以采用迭代运算的方法实现快速计算 $\boldsymbol{G}_{k_1}^{-1}$. 为了降低运算量,这里引入分块矩阵思想,首先给出引理.

引理[20]　给定可逆矩阵 $\boldsymbol{A}$、$\boldsymbol{D}$ 以及矩阵 $\boldsymbol{U}$, $\boldsymbol{V}$ 则等式:

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{U}\\ \boldsymbol{V} & \boldsymbol{D}\end{bmatrix}^{-1}=$$

$$\begin{bmatrix}\boldsymbol{A}^{-1}+\boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{D}-\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{D}-\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\\ -(\boldsymbol{D}-\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1} & (\boldsymbol{D}-\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\end{bmatrix} \qquad (15)$$

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{U}\\ \boldsymbol{V} & \boldsymbol{D}\end{bmatrix}^{-1}=$$

$$\begin{bmatrix}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1} & -\boldsymbol{U}\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}\\ -\boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1} & \boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{U}\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}\end{bmatrix} \qquad (16)$$

成立.

由引理知,若知道 $\boldsymbol{G}_{k_1}^{-1}$或者 $\boldsymbol{G}_{k_2}^{-1}$则可以通过式(15)或者式(16)求出 $\boldsymbol{G}^{-1}$. S_{k_1}的更新是通过窗口滑动,可以通过迭代的方式求解 $\boldsymbol{G}_{k_1}^{-1}$. 为了推导 $\boldsymbol{G}_{k_1}^{-1}$的迭代算法,首先给出引理的推论.

推论　给定可逆矩阵 $\boldsymbol{D}$ 以及标量 $\boldsymbol{A}$,行向量 $\boldsymbol{U}$ 以及列向量 $\boldsymbol{V}$,若有$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{U}\\ \boldsymbol{V} & \boldsymbol{D}\end{bmatrix}^{-1}=\begin{bmatrix}\boldsymbol{B} & \boldsymbol{C}\\ \boldsymbol{E} & \boldsymbol{F}\end{bmatrix}$,其中 $\boldsymbol{B}$、$\boldsymbol{C}$、$\boldsymbol{E}$、$\boldsymbol{F}$ 分别与 $\boldsymbol{A}$、$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{D}$ 规模相同,则有: $\boldsymbol{D}^{-1}=\boldsymbol{F}-\boldsymbol{E}\boldsymbol{C}\boldsymbol{B}^{-1}$成立.

证明　由引理可知 $\boldsymbol{F}=\boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{U}\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}$

故有 $\boldsymbol{D}^{-1}=\boldsymbol{F}-\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{U}\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}$

而$\boldsymbol{C}=-\boldsymbol{U}\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}$, $\boldsymbol{E}=-\boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}$,

$\boldsymbol{B}=(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}$

$$\begin{aligned}\boldsymbol{E}\boldsymbol{C}\boldsymbol{B}^{-1}&=(-\boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1})(-\boldsymbol{U}\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1})(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})\\&=\boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}\boldsymbol{U}\boldsymbol{D}^{-1}\end{aligned}$$

由于 $\boldsymbol{A}$ 为标量, $\boldsymbol{U}$ 为行向量 $\boldsymbol{V}$ 为列向量故有$(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}$为标量

所以 $\boldsymbol{E}\boldsymbol{C}\boldsymbol{B}^{-1}=\boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}\boldsymbol{U}\boldsymbol{D}^{-1}=\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{U}\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V})^{-1}$

故: $\boldsymbol{D}^{-1}=\boldsymbol{F}-\boldsymbol{E}\boldsymbol{C}\boldsymbol{B}^{-1}$成立,推论得证.

设更新前后得到的 $\boldsymbol{G}_{k_1}$矩阵分别表示为 $\boldsymbol{G}_{k_1}^{old}$、$\boldsymbol{G}_{k_1}^{new}$,根据更新方法,可将两矩阵的差别形式化为:

$$\boldsymbol{G}_{k_1}^{old}=\begin{bmatrix}G_{11} & \boldsymbol{\rho}_{old}^{\mathrm{T}}\\ \boldsymbol{\rho}_{old} & \boldsymbol{\Phi}\end{bmatrix}\rightarrow\boldsymbol{G}_{k_1}^{new}=\begin{bmatrix}\boldsymbol{\Phi} & \boldsymbol{\rho}_{new}^{\mathrm{T}}\\ \boldsymbol{\rho}_{new} & G_{kk}\end{bmatrix} \qquad (17)$$

将$(\boldsymbol{G}_{k_1}^{old})^{-1}$写成分块矩阵形式: $(\boldsymbol{G}_{k_1}^{old})^{-1}=\begin{bmatrix}\gamma & \boldsymbol{M}\\ \boldsymbol{N} & \boldsymbol{Z}\end{bmatrix}$,则由推论可得: $\boldsymbol{\Phi}^{-1}=\boldsymbol{Z}-\boldsymbol{N}\boldsymbol{M}/\gamma$. 再将 $\boldsymbol{\Phi}^{-1}$代入引理中式(15)可得出$(\boldsymbol{G}_{k_1}^{new})^{-1}$,这样就将求$(\boldsymbol{G}_{k_1}^{new})^{-1}$的计算复杂度由 $o(k_1^3)$降为 $o(k_1^2)$,这就实现了由$(\boldsymbol{G}_{k_1}^{old})^{-1}$向$(\boldsymbol{G}_{k_1}^{new})^{-1}$的快速迭代运算. 获得$(\boldsymbol{G}_{k_1})^{-1}$后,计算$(\boldsymbol{G}_{k_2}-\boldsymbol{R}\boldsymbol{G}_{k_1}^{-1}\boldsymbol{R}^{\mathrm{T}})^{-1}$,再根据式(15)可得$(\boldsymbol{G}_{k_1}^{new})^{-1}$. 求解$(\boldsymbol{G}_{k_2}-\boldsymbol{R}\boldsymbol{G}_{k_1}^{-1}\boldsymbol{R}^{\mathrm{T}})^{-1}$的计算复杂度为 $o(k_2{}^3)$,所以采用这种方式计算$(\boldsymbol{G}_{k_1}^{new})^{-1}$的总的计算复杂度为 $o(k_1^2+k_2^3)$,由于 $k_1>>k_2$,故 $k_1^2+k_2^3<<k^3=(k_1+k_2)^3$ 大大降低了计算

复杂度,这样就实现了 $\boldsymbol{G}^{-1}$快速计算,代入式(13)获得更新的预测模型.

4.3 OTOSF-LSSVM 预测模型

根据上面的分析,建立 OTSOF-LSSVM 网络流量预测模型,预测流程如图 5 所示.

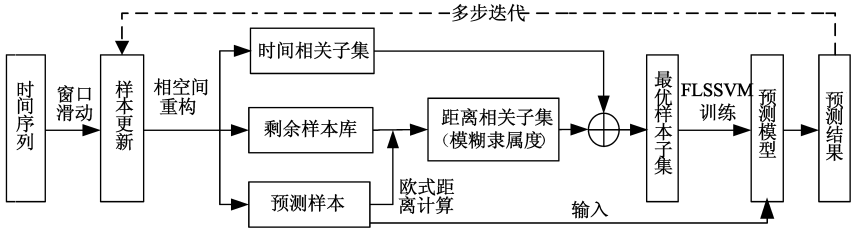


图5 OTSOF-LSSVM网络流量趋势预测模型

具体的步骤如下:

步骤 1 样本更新,随着时间窗口的滑动对样本进行更新,纳入新样本的同时去掉最老的样本(此时的样本存储的降噪后的流量趋势序列).

步骤 2 相空间重构,设置延时 τ 和嵌入维数 m,对窗口内时间序列进行相空间重构,获得训练样本库和预测样本,选择时间上距预测样本最近的 k_1 个样本作为时间相关子集.

步骤 3 距离相关最优样本子集选择,计算当前样本库中的样本与预测样本的欧式距离,选择最近的前 k_2 个样本组成距离相关最优样本子集,根据距离确定各个样本的模糊隶属度.

步骤 4 预测模型建立,选择恰当的核参数和惩罚因子,对训练样本进行迭代方法的 FLSSVM 训练获得预测模型,在线迭代过程如前一小节描述.

步骤 5 预测输出,将预测样本输入到预测模型中得到单步的预测结果.

步骤 6 窗口滑动,重复 1~5 步.

此外,若要对网络流量趋势进行多步预测,可以将预测的结果迭代回时间序列中获得新的预测样本,然后重复 3~5 步.

5 实验

为了验证本文方法的有效性,本文对 LBL-tcp-3.tcp 降噪后获得的网络流量趋势进行单步预测和多步预测实验.

5.1 单步预测实验

分别采用离线 LSSVM、文献[14]的在线 LSSVM 以及 OTSOF-LSSVM 对网络流量趋势进行单步预测.训练使用的核函数为 RBF 核,核带宽 $\sigma^2 = 0.05$, $C = 2000$.为了比较方便,离线预测选择前 1874-2100 个样本为样本集;在线预测时间窗口长度为 226,初始窗口为 1874-2100;OTSOF-LSSVM 的窗口长度为 2100,时间相关样本选择的窗长为 206,距离最优样本子集选择 $k_2 = 20$.这样就保证了 3 种方法的训练初始数据集相同,且训练集大小一直保持不变(均为 220 个).分别对后面 200 个的数据进行单步预测(2101-2300 这段数据较具代表性,因此选择其为预测目标),实验环境为 Pentium(R) Dual-Core 2.7G CPU,2G 内存,Windows XP 系统,Matlab7.4.0,预测结果如下:

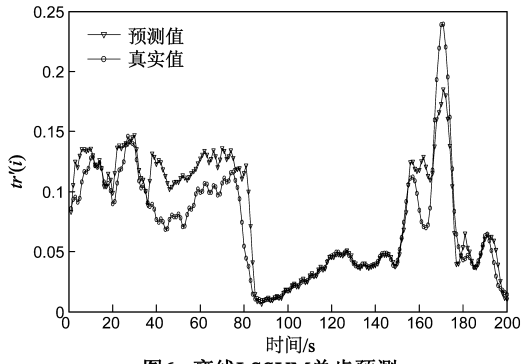


图6 离线LSSVM单步预测

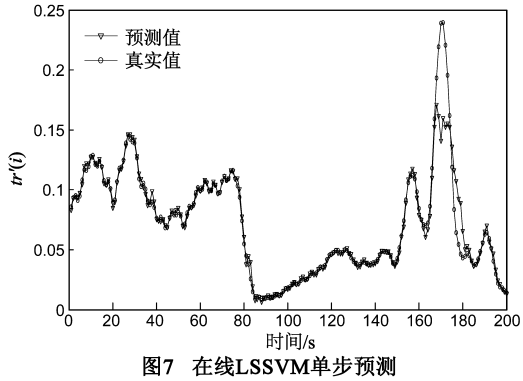


图7 在线LSSVM单步预测

从图 6 可以看出,离线 LSSVM 预测的结果同真实值出现了较大的偏差,这是因为离线 LSSVM 方法训练样本固定,获得的预测模型是固定不变的,而网络流量是一个时变的,所以离线 LSSVM 在大多数情况下无法预测未来流量.在线 LSSVM 方法能够较好的适应网络流量的时变性,大部分情况下都能很好的预测出未来的流量,但是在突变的情况下(图 7 中 160-180 点)难以预测,这是由于在线预测的历史样本获取窗口较短,获得的训练样本中没有与该段变化特征相似的历史样本,在该段数据中没有充分学习,预测效果较差.OTSOF-LSSVM 方法获得了最好的预测结果,从图 8 可以看出即使在突变的情况下依然能够准确预测出未来时刻的流量,这主要是因为虽然训练样本大小相同,但是经过最优样本的选取以及模糊化处理后这些样本包含了历史样本库中(前 1994 个样本)与待预测样本最相关的信息,在这个历史样本库中存在与该突变类似的样本信息,所以训练得到的预测模型能够准确的预测.从单

步预测的模型更新时间分析,文献[13]的在线预测方法更新时间为 0.129s,文献[14]为 0.012s,而本文方法为 0.016s,这是因为文献[13]的更新方法是每次更新均重新计算一遍$(\boldsymbol{G}^{new})^{-1}$,计算复杂度为 $o(220^3)$,文献[14]的迭代算法计算复杂度为 $o(220^2)$远远小于文献[13]的方法,本文的算法为 $o(200^2+20^3)$,略小于文献[14].但是本文方法还涉及距离相关样本的选取和模糊隶属度的确定,因此最终时间略高于文献[14].最后,从图 8 可以看出预测的结果并没有"一步延迟"现象,每一步都非常准确的预测出了未来时刻的值,这是由于通过降噪,去掉了高维的噪声部分,剩下的是流量的低维趋势部分,容易建立预测模型.统计总体单步预测的时间平均为 0.04s,可以满足一秒内预测下一秒网络流量趋势的要求,可看出本文提出的方法有着较强的现实意义.

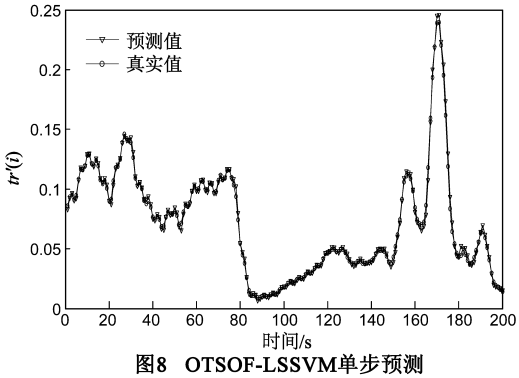


图8　OTSOF-LSSVM单步预测

5.2　多步预测实验

在现实中,对未来网络流量的预测能够预测的时间是越长越有利于网络的调整和控制,因此对网络流量的多步预测研究是非常有意义的.这里对网络流量进行了多步预测实验:已知量为前 2100 个时刻的网络流量值,预测目标点为 2101-2020 段,即向后预测 20 步.实验分别采用在线 LSSVM 和 OTSOF-LSSVM 进行多步预测,仍选用 RBF 核,$\sigma^2=0.05$, $C=2000$.训练样本集的获取仍采用单步预测的窗口获取,实验的结果如图:

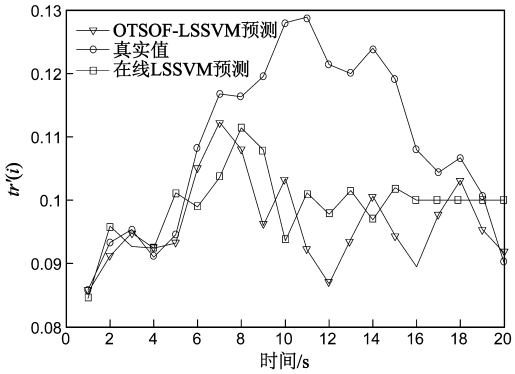


图9　多步迭代预测

可以看出在线 LSSVM 预测的方法进行多步预测只能准确的向前预测 4 步,而 OTSOF-LSSVM 方法可以向前准确预测 7 步.虽然初始时刻预测获得的值与真实值很接近,但是后面的预测结果偏差较大,这也从另一方面体现了网络流量的混沌性(初值敏感性).同时,网络流量的最大 Lyapunov 指数为 0.31,采用最大 Lyapunov 指数预测方法在理论上能够获得的最大预测步数是 3 步,可见本文提出的方法相较在线 LSSVM 和最大 Lyapunov 指数预测方法可以获得更好的预测效果以及更大的向前预测步数.

6　结论

小时间尺度的网络流量存在混沌性,但噪声的影响将其混沌性掩盖,导致难以对其进行预测.本文通过对网络流量的降噪处理去除了其中的高维噪声分量,保留了网络流量的趋势部分,恢复了网络流量的混沌性,使得网络流量趋势可以预测.在此基础上,本文从机器学习的本质出发,结合混沌时间序列本身的无穷嵌套自相似性和最大 Lyapunov 指数固定不变等特性提出 OTSOF-LSSVM 预测方法,解决由网络流量的混沌性和长周期性带来的问题,并通过分块矩阵思想降低预测模型的更新时间.通过对实际网络流量数据 LBL-tcp-3.tcp 的实验表明本文提出的方法相较 LSSVM,在线 LSSVM 方法能够获得较快的速度以及更高的精度.本文提出的方法能够准确的预测小时间尺度网络流量的变化趋势,为更好的控制和管理网络提供了新的思路和方法.

参考文献

[1] Jun Jiang, Symeon Papavassiliou. Enhancing network traffic prediction and anomaly detection via statistical network traffic separation and combination strategies[J]. Computer Communications, 2006, 29(10): 1627 – 1638.

[2] 罗 ▮ 骞,夏靖波,王焕彬.混沌-支持向量机回归在流量预测中的应用研究[J].计算机科学,2009,36(7):244 – 246.
Luo Yun-qian, Xia Jing-bo, Wang Huan-bin. Application of Chaos-support Vector Machine Regression in Traffic Prediction [J]. Computer Science, 2009, 36(7): 244 – 246. (in Chinese)

[3] He Yu-jun, Zhu Youchan, Duan Dong-xing. Research on hybrid ARIMA and support vector machine model in short term load forecasting[A]. Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)[C]. Jinan P.R. China, 2006, Vol.1: 804 – 809.

[4] Chen Bor-Sen, Peng Sen-Chueh, Wang Ku-Chen. Traffic Modeling, prediction and congestion control for high-speed networks: a fuzzy AR approach[J]. IEEE Tans On Fuzzy Systems, 2000, 8(5): 491 – 508.

[5] 姜明,吴春明,张旻,胡大民.网络流量预测中的时间序列模型比较研究[J].电子学报,2009,37(11):2353 – 2358.
Jiang Ming, et al. Research on the comparison of time series models for network traffic prediction[J]. Acta Electronica Sinica,2009,37(11):2353 – 2358.(in Chinese)

[6] 姚奇富,李翠凤,马华林,张森.灰色系统理论和马尔柯夫链相结合的网络流量预测方法[J].浙江大学学报(理学版),2007,34(4):396 – 400.
Yao Qi-fu,Li Cui-feng,Ma Hua-lin,Zhang Sen. Novel network traffic forecasting algorithm based on grey model and Markov chain[J]. Journal of Zhejiang University (Science Edition), 2007,34(4):396 – 400.(in Chinese)

[7] Chen Y, Yang B. Small-time scale network traffic prediction based on flexible neural tree[J]. Applied Soft Computing Journal 2012,12(1):274 – 279.

[8] Dong-Chul Park. Prediction of network traffic using dynamic bilinear recurrent neural network[A]. Fifth International Conference on Natural Computation[C]. Tianjin China,2009, Vol. 2:419 – 423.

[9] 陈晓天,张顺颐,田婷婷.基于 BP 神经网络的 IP 网络流量预测[J].南京邮电大学学报(自然科学版),2010,30(2):16 – 21.
Chen Xiao-tian, Zhang Shun-yi, Tian Ting-ting. Internet traffic forecasting based on BP neural network[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science),2010,30(2):16 – 21.(in Chinese)

[10] 陆锦军,王执铨.基于混沌特性的网络流量预测[J].南京航空航天大学学报,2006,38(2):217 – 221.
Lu Jin-jun, et al. Prediction of network traffic flow based on chaos characteristics[J]. Journal of Nanjing University of Aeronautics &Astronautic, 2006, 38(2): 217 – 221.(in Chinese)

[11] Bao Rong Chang, Hsiu Fen Tsai. Improving network traffic analysis by foreseeing data-packet-flow with hybrid fuzzy-based model prediction[J]. Expert Systems with Applications 2009,36(3):6960 – 6965.

[12] Wei-Chiang Hong. Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm[J]. Neurocomputing 2011,74(12 – 13):2096 – 2107.

[13] 叶美盈,汪晓东,张浩然.基于在线最小二乘支持向量机回归的混沌时间序列预测[J].物理学报,2005,54(6):2568 – 2573.
Ye Mei-Ying, Wang Xiao-Dong, Zhang Hao-Ran. Chaotic time series forecasting using online least squares support vector machine regression[J]. Acta Physica Sinica, 2005, 54(6): 2568 – 2573.(in Chinese)

[14] 肖支才,王杰,等.基于在线 LSSVM 算法的变参数混沌时间序列预测[J].航空计算技术,2010,40(3):29 – 33.
Xiao Zhi-cai, Wang Jie, et al.. Predict the time series of the parameter-varying chaotic system based on recursive lease square support vector machine(RLS-SVM)[J]. Aeronautical Computing Technique,2010,40(3):29 – 33.(in Chinese)

[15] Takens F. Detecting strange attractors in turbulence[J]. Lecture Notes in Math,1987,898(8):175 – 198.

[16] J A K Suykens, J Vandewalt. Least squares support vector machine classifiers[J]. Neural Processing letters,1999,9(3):293 – 300.

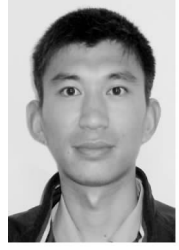
[17] M T Rosenstein, J J Collins, C J Deluca. A practical method for calculating largest Lyapunov exponents from small data sets[J]. Physica D,1993,65(1 – 2):117 – 134.

[18] 韩敏,项牧.局部投影去噪的一种改进的邻域选取方法[J].系统工程学报,2009,24(8):392 – 398.
HAN Min, XIANG Mu. An improved neighborhood selection method for local projection noise reduction[J]. Journal of Systems Engineering,2009,24(8):392 – 398.(in Chinese)

[19] Ephraim Y, Trees H L V. A signal subspace approach for speech enhancement[J]. IEEE Trans on Speech and Audio Processing,1995,3(7):251 – 261.

[20] Stoer J, Bulirsch R. Introduction to Numerical Analysis[M]. New York: Springer-Verlag,1993.

作者简介



温祥西 男,1984 年生于江苏连云港,博士研究生.主要研究方向为网络故障预测与健康管理.
E-mail:wxxajy@163.com

孟相如 男,1963 年生于陕西西安,教授,博士生导师.主要研究方向为宽带通信网络技术,网络故障诊断等.
E-mail:xrmeng@126.com

马志强 男,1968 年生于山东烟台,副教授,硕士生导师.主要研究方向为视频网络通信.
E-mail:mzqxa123@sina.com