混沌-支持向量机回归在流量预测中的应用研究

罗贤赛 夏靖波 王焕彬
(空军工程大学电讯工程学院 西安 710077)
(西安电子科技大学综合业务网理论及关键技术国家重点实验室 西安 710071)

摘 要 为了提高流量预测准确性,将混沌理论和支持向量机回归应用于网络流量预测。采用相空间重构理论计算实际流量的延时、嵌入维数和Lyapunov指数,证实网络流量存在混沌现象;据此建立混沌-支持向量机预测模型并确定训练样本对,对实际网络流量数据进行预测。结果表明,该方法能有效地进行流量预测,相对于BP神经网络方法,该方法具有更好的预测精度。

关键词 支持向量机,流量预测,回归,混沌
中图法分类号 TP393.07 **文献标识码** A

Application of Chaos-support Vector Machine Regression in Traffic Prediction

LUO Yun-qian XIA Jing-bo WANG Huan-bin
(The Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077, China)
(State Key Lab of Integrated Service Networks, Xidian University, Xi'an 710071, China)

Abstract A traffic forecasting model based on the support vector machine(SVM) and chaos was developed to improve the accuracy of the traffic prediction. Based on the phase space reconstruction, it calculates the real-time traffic's delay time, embedded dimension and Lyapunov exponent, and proves that the traffic chaos phenomena exists. That a chaos-SVM model was constructed and pairs of training samples was determined to forecast the real network traffic. The results show that the chaos-SVM model is able to predict network traffic effectively. In comparison with the BP neutral network, it has higher accuracy of prediction.

Keywords Support vector machine(SVM), Traffic prediction, Regression, Chaos

1 前言

网络流量与网络相关的活动联系在一起,是能够记录和反映网络及其用户活动的重要载体,网络流量预测能为网络中带宽分配、流量控制、选路控制、接纳控制、安全管理等提供有效依据,因此对网络流量的预测是目前研究的一个热点问题。众多学者先后提出了基于ARMA[1,2],FARIMA[3],指数平滑[4]等显示线性方法;基于神经网络[5]、灰色模型以及组合方式[6]的隐式非线性方法。显示方法建立线性模型实现预测,简单但精度差。隐式方法中神经网络得到广泛应用,神经网络能根据历史数据进行学习训练和经验积累,具有自适应能力特点,但是它得到的结果基于经验风险最小化,需要有足够大的样本数据数量,可能还会出现过学习问题导致网络推广能力较差的缺陷。因此需要更先进的预测算法,提高网络流量预测精度和效率。目前的研究结果表明网络流量存在着混沌特性[7,8],而混沌模型中相空间重构的方法以及近几年在人工智能领域中广受关注的统计学习理论里的支持向量机算法,可以将非线性序列映射到高维空间中去,把非线性序列中的动力学特性信息暴露出来。因此采用混沌理论的相空间重构理论对非线性数据进行相空间重构确定训练样本对,以解决训练样本对的确定问题[9],再应用支持向量机回归对训练样本对进行建模,实现对流量的预测。

2 网络流量序列的相空间重构方法

研究表明一个混沌系统产生的轨迹经过一定时期变化后,最终会做一种有规律的运动,产生一种规则的、有形的轨迹,这种轨迹在经过类似拉伸和折叠后转化成与时间相关的序列时,呈现出混乱的、复杂的特性。Packard等建议用原始系统中的某变量延迟坐标来重构相空间,Takens证明可以找到一个合适的嵌入维,即如果延迟坐标的维数 $m \geq 2d+1$,d 是动力系统的维数,在这个嵌入维空间里可以把有规律的轨迹恢复出来。这就是相空间重构理论[10]。

在重构相空间中,时间延迟 τ 和嵌入维数 d 的选取非常重要,研究表明如果 τ 太小,将不能展示系统的动力特征,τ 太大会使简单轨道变得复杂且会减少有效的数据点数;同样 d 太小嵌入空间无法包容动力系统的吸引子,从而无法全面展现系统的动力学特性;d 太大不仅会减少可用数据长度、增加计算工作量而且可能会增大预测误差。

到稿日期:2008-10-24 返修日期:2008-12-24 本文受综合业务网理论及关键技术重点实验室开放基金(ISN-9-08)资助。
罗贤赛(1981—),男,博士生,主要研究方向为军事通信网络运行质量评价、数据挖掘,E-mail:immortalluo@163.com;夏靖波(1963—),男,教授,博士生导师,主要研究方向为信息网络技术和通信网规划;王焕彬(1979—),男,博士生,主要研究方向为网格计算。

2.1 延时的计算方法

目前，延时 τ 的选择方法主要有自相关法、平均位移法、去偏复自相关法、互信息法等[10]。由于互信息法适用于大数组、非线性问题，本文采用互信息法计算延时，计算方法为：

$$I(\tau)=\sum_{n=1}^{N}P(x(n),x(n+\tau))\log\left[\frac{P(x(n),x(n+\tau))}{P(x(n))\cdot P(x(n+\tau))}\right] \tag{1}$$

$I(\tau)\geqslant 0$，$P(\cdot)$为概率。$I(\tau)$指出从一个序列得到多少关于另一个序列的信息；Fraer[11]建议 $I(\tau)$达到第一个极小点所对应的 τ 作为嵌入时间延迟。

2.2 嵌入维数的计算方法

求取嵌入维数的方法主要有关联指数饱和法、假近邻法、Cao 方法[12]等，本文选用 Cao 方法对 m 进行选取。定义

$$a(i,d)=\frac{\| y_i(d+1)-y_{n(i,d)}(d+1)\|}{\| y_i(d)-y_{n(i,d)}(d)\|} \tag{2}$$

其中，$i=1,2,\cdots,N-d\tau$，$\|\cdot\|$为欧式距离，满足：$\| y_k(m)-y_l(m)\|=\max|x_{k+j\tau}-x_{l+j\tau}|$，$0\leqslant j\leqslant m-1$。$y_i(d+1)$是 $d+1$ 次嵌入维重构空间的第 i 个向量。如果 d 是一个嵌入维，在 d 维相空间邻近的两个点在 $d+1$ 维相空间依然邻近，这样的点对被称为“真实邻近点”，反之被称为“假邻近点”。Cao 引入一个量

$$E1(d)=E(d+1)/E(d) \tag{3}$$

其中，$E(d)$为 $a(i,d)$的平均值。

$$E(d)=\frac{1}{N-d\tau}\sum_{i=1}^{N-d\tau}a(i,d) \tag{4}$$

Cao 发现，当 d 比某一个 d_0 大时，$E1(d)$停止变化，于是 d_0+1 便给出了序列的最小嵌入维。同时 Cao 还定义 $E2(d)$用于区分确定性混沌信号和随机信号，如果是随机信号 $E2(d)$对任何 d 为 1，对于混沌信号 $E2(d)$将不会始终为 1。

2.3 最大 Lyapunov 指数的计算方法

得到了延时和嵌入维数以后可以计算 Lyapunov 指数，通过 Lyapunov 指数可以检验流量的时间序列是否存在混沌现象，正的 Lyapunov 指数意味着混沌。其计算方法主要有 Jacobin 法、Wolf 法和小数据量法[5]。本文采用由 Michael. T. Rosenstein 提出的小数据量方法进行计算。由 Sato 等改进的估计表达式为：

$$\lambda_1(i,k)=\frac{1}{k\Delta t}\frac{1}{M-k}\sum_{j=1}^{M-k}\log\frac{d_j(i+k)}{d_j(k)} \tag{5}$$

其中 k 是常数，$d_j(k)$是基本轨道上第 j 对最邻近点对经过 i 个离散时间步长后的距离，Δt 为样本周期，M 为重构相空点的个数。

3 支持向量机回归原理

SVM 最初被用于模式识别等分类问题，后来被推广到非线性回归估计和曲线拟合中，得到用于曲线拟合的支持向量回归机(Support Vector Machine for Regression)[13]，也表现出了良好的效果。假设训练样本为 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，在线性条件下，SVM 回归机使用线性函数 $f(x,w)=(w\cdot\phi(x))+b$ 对样本点进行拟合。在非线性条件下，则将样本映射到高维特征空间，在高维特征空间中建立线性模型 $f(x,w)=(w\cdot\phi(x))+b$，其中 $\phi(x)$是将样本点映射到高维空间的非线性变换，SVM 回归机可以表示为：

$$\min\frac{1}{2}\|w\|^2+c\sum_{i=1}^{l}(\xi_i+\xi_i^*) \tag{6}$$

$$\text{St } f(x_i,w)-y_i\leqslant\varepsilon+\xi_i,\ i=1,\cdots,l$$

$$y_i-f(x_i,w)\leqslant\varepsilon+\xi_i^*,\ i=1,\cdots,l$$

$$\xi_i\geqslant 0,\xi_i^*\geqslant 0,i=1,\cdots,l$$

其中：$\|w\|^2$ 代表与模型复杂度相关的因素；模型采用 ε 不敏感损失函数，松弛变量 ξ_i,ξ_i^* 表示样本偏离 ε 不敏感区域的程度；c 为惩罚系数。

对于问题式(6)，通常通过求解模型的 Lagrange 对偶问题获得原问题的最优解：

$$f(x)=\sum_{i=1}^{l}(a_i^*-a_i)k(x_i,x)+b \tag{7}$$

其中 $k(x_i,x_j)$称为核函数，满足 Mercer 条件且 $k(x_i,x_j)=(\phi(x_i)\cdot\phi(x_j))$。其中惩罚系数 c、不敏感系数 ε、核函数的选择与模型的成功存在一定关系[14]，对 SVM 参数的选择还没有一种有效的方法，只是凭借经验、试验对比、大范围搜寻或者利用软件包提供的交互检验功能进行寻优。

4 混沌-支持向量机回归预测模型

对于给定的流量时间序列 $x_1,x_2,\cdots,x_{N-1},x_N$，采用相空间重构法，将其转换成维数为 m，延时为 τ 的新数据空间，即：

$$Y(n)=[x(n-(m-1)\tau),\cdots,x(n-\tau),x(n)]$$

其中 $n\in[(m-1)\tau,N]$，$Y(n)$为重构后的相点。利用重构后的状态矢量对流量测量值进行预测，可以构造映射(回归估计函数)$f:R^m\to R$，使得：

$$x(n+1)=f(Y(n)) \tag{8}$$

设当前时刻为 n，训练数据数量为 N，则训练数据可以表示为：

$$(Y(n),x(n+1)),n=(m-1)\tau,\cdots,N-1$$

根据已知样本序列确定训练数据，应用支持向量机回归进行训练求得最佳模型 f；对未来时刻 $\hat{x}(t+1)$的预测值，以其重构相空间中前 $(m-1)\tau$ 变量作为输入，应用训练得到的支持向量机模型进行预报。

5 混沌-SVM 回归的流量预测方法

5.1 预测步骤

(1)根据混沌系统的相空间重构理论求出训练历史数据的最佳嵌入维数 d 和时延 τ，生成训练样本。

(2)对生成的训练样本进行归一化处理，以便提高收敛速度、缩短训练时间。

(3)以生成的训练样本利用支持向量机回归进行训练，得到预测模型。

(4)利用训练好的支持向量机回归模型预测未来流量值。

5.2 实例分析

5.2.1 预测模型的建立

本文采用来自 Http://newsfeed.ntcu.net/~news/2006 的流量文件对模型进行测试，介绍模型的应用方法及效果。该流量文件收集了主节点路由器 news 从 2006 年 7 月 21 日到 2006 年 9 月 3 日共 45 天网络每小时访问流量(流量曲线如图 1 所示)，得到 $45\times 24=1080$ 个数据；用前 40 天的 960 个数据作为已知数据训练模型；后 5 天的 120 个数据作为预测数据用以校验模型的预测效果。

采用互信息法得到前 40 天流量序列的延时为 3，利用 Cao 方法得到嵌入维数为 10；利用小数据量法得到最大 Lya-

punov 指数为 $\lambda=0.0.167$；说明该时间序列为混沌时间序列。利用相空间重构理论得到 932 个训练样本，利用支持向量机回归对这些训练样本进行训练。

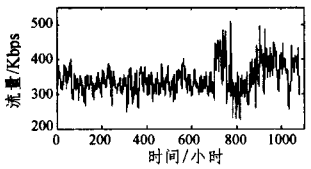

图 1　采集的原始流量

在使用 SVM 时，最重要的是核函数的选择，目前常用的核函数有线性核函数、多项式核函数、径向基核函数和 sigmoid 核函数，其中实验证明径向基核函数相比其他核函数不仅具有较少的参数还具有良好的性能[15]，因此本文选择径向基核函数。支持向量机回归模型中模型参数确定目前在理论上还没有统一的标准，本文采用 Libsvm 工具箱实现回归预测模型，在 SVM 回归模型中采用网格搜索法确定最佳模型参数。采用网格搜索时，给定惩罚系数 c、不敏感系数 ε 和宽度系数 δ 的取值范围以及步进长度，然后对三者取值并进行组合训练，最后选择误差最小一组的参数作为最优的 c，ε 和 δ。如果结果不理想，可以重新设定三者取值范围和步长进行训练。通过网格搜索最终确定 $c=64$，$\varepsilon=0.5$ 和 $\delta=0.0625$ 作为模型参数值。最后利用训练好的模型对测试数据进行预测，其预测曲线如图 2 所示。从图中曲线可以看到模型较好地预测了流量变化。

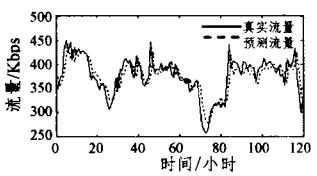

图 2　支持向量机回归流量预测结果

5.2.2　预测效果分析

采用神经网络的回归建模是当前研究最多和发展最快的预测模型，因此将 BP 神经网络和 SVM 回归模型预测值作为评估样本进行比较。采用均方误差（Mean Square Error，MSE）和平均相对误差（Mean Absolute Percentage Error，MAPE）评价模型的预测性能。

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_i-y_i')^2 \tag{9}$$

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-y_i'}{y_i}\right|\times 100\% \tag{10}$$

y_i 和 y_i' 分别为某时刻的实际和预测流量。使用 Matlab 工具箱建立 BP 神经网络预测模型，比较两种模型的预测效果。BP 神经网络中所选参数通过寻优确定。两种模型每天预测的指标结果如表 1 所列。

表 1　两种预测方法的性能比较

日期	BP 神经网络模型		混沌-支持向量机回归模型	
	MAPE	MSE	MAPE	MSE
2006-8-30	4.39%	449.0830	3.93%	407.4933
2006-8-31	4.90%	637.5117	5.04%	667.1241
2006-9-01	4.80%	479.0499	4.70%	437.6006
2006-9-02	4.54%	492.9286	4.41%	535.3157
2006-9-03	5.50%	662.7091	5.10%	606.2450
平均值	4.83%	544.2565	4.63%	530.7557

由表 1 可知，混沌-支持向量机预测方法与 BP 神经网络预测方法相比，在整体上预测误差更小，更适合网络流量预测。

结束语　本文研究了利用网络流量的混沌特性和支持向量机回归对网络流量进行预测的方法。根据相空间重构理论和支持向量机回归理论，通过相空间重构将流量序列映射到 m 维特征空间形成相点构成训练样本对，解决了样本对确定的问题。再利用训练样本对使用支持向量机回归方法构建预测模型，建立了混沌-支持向量机回归流量预测模型，对网络流量进行预测。研究结果表明混沌-支持向量机回归模型能够有效地预测网络流量，与 BP 神经网络方法相比具有更好的预测性能，更适合于网络流量预测。

参 考 文 献

[1] Sang A, Li S. Predictability analysis of network traffic [C]// Proceedings of INFOCOM 2000. Tel Aviv: IEEE, 2000: 342-351

[2] 邹伯贤，刘强. 基于 ARMA 模型的网络流量预测[J]. 计算机研究与发展，2002，39(12)：1645-1652

[3] Shu Y, Jin Z, Zhang L, et al. Traffic prediction using FARIMA models[C]// Proceedings of IEEE international Conference on communications. Vancouver: IEEE, 1999: 891-895

[4] 刘勇，靳新. 动态指数平滑模型在网络流量预测中的研究[J]. 火力与指挥控制，2008，33(3)：100-102

[5] 刘杰，黄亚楼. 基于 BP 神经网络的非线性网络流量预测[J]. 计算机应用，2007，27(7)：1770-1772

[6] 白燕. 基于灰色神经网络组合模型的流量预测与评估方法研究[D]. 西安：西安建筑科技大学，2007

[7] 赵其刚. 基于流量预测的下一代网络动态 Qos 研究[D]. 成都：西南交通大学，2006

[8] 陆锦军，王执铨. 基于混沌理论的网络数据流 RBF 神经网络预测[J]. 计算机工程，2006，32(23)：100-103

[9] 张颖路. 基于遗传算法优化支持向量机的网络流量预测[J]. 计算机科学，2008，35(5)：177-179

[10] 吕金虎，陆军安，陈士华. 混沌时间序列分析及应用[M]. 武汉：武汉大学出版社，2002

[11] Fraser A M. Independent Coordinates for Strange Attractors from Mutual Information [J]. Physical Review A, 1986, 33(2): 1134-1140

[12] Cao Lianyue. Practical method for determining the minimum embedding dimension of a scalar time series[J]. Physica D, 1997, 110(5): 43-50

[13] 邓乃扬，田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京：科学出版社，2004

[14] 刘靖旭，蔡怀平，谭跃进. 支持向量机回归参数调整的一种启发式算法[J]. 系统仿真学报，2007，19(7)：1540-1543

[15] Hsu CHih-Wei, Chang Chih-Chung, Lin Chih-Jen. A practical guide to SVM classification [EB/OL]. [2008-07-03]. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf