一种新的基于 ARIMA-SVM 网络流量预测研究

邵 忻
(天津外国语大学 教育技术与信息学院, 天津 300204)

摘 要: 研究网络流量预测精度问题,网络流量受多种因素的综合影响,其变化具有周期性、非线性和随机性等特点,将 ARIMA 模型和 SVM 模型相结合建立一种网络流量预测模型。采用 ARIMA 预测网络流量周期性和线性变化趋势;然后采用 SVM 对网络流量非线性和随机性趋势进行拟合;最后将两者结果再次输入 SVM 进行融合,得到网络流量最终预测结果。采用具体网络流量数据对模型性能进行测试,仿真结果表明,ARIMA-SVM 提高了网络流量预测精度,降低了预测误差,能更全面刻画网络流量变化规律。

关键词: 自回归滑动平均模型(ARIMA); 支持向量机(SVM); 网络流量; 预测
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)05-1901-03
doi:10.3969/j.issn.1001-3695.2012.05.080

Application of ARIMA-SVM model in network traffic prediction

SHAO Xin
(*College of Educational Technology & Information*, *Tianjin Foreign Studies University*, *Tianjin* 300204, *China*)

Abstract: Network flow is affected by many factors, its change is cyclical, nonlinear and stochastic characteristics. This paper used ARIMA model and SVM model to establish a network traffic prediction model. It used ARIMA forecasting network traffic cyclic and linear trend, and then used SVM to network flow nonlinear and stochastic trend fitting. At last two results again entered the SVM integration, network traffic had been the final prediction results. By using network flow data for model performance test, simulation results show that, ARIMA-SVM increases network traffic prediction accuracy, reduces the prediction error, and it can fully characterize network traffic variation.

Key words: ARIMA(autoregressive integrating moving average); SVM(support vector machine); network traffic; prediction

0 引言

随着 Internet 发展,网络用户增多,信息流量增大,网络规模呈指数级增长,网络管理工作日益繁重,管理员必须及时掌握网络运行状态,以便采取相应的管理策略。网络流量是衡量网络运行负荷和状态的重要参数,因此对网络流量进行建模和预测是网络控制的基础。网络流量预测准确性由预测算法决定,设计高准确率的预测算法是网络流量控制研究中的热点[1]。

为了对网络流量进行准确预测,许多学者对其进行了大量而深入的研究,提出许多网络流量的预测方法,并取得了不错的成果。传统网络流量预测方法有线性回归分析、泊松、隐马尔可夫和时间序列等,这些方法比较简单,均基于网络流量是线性变化规律进行建模,难以准确地刻画现代网络流量与其影响因素间的高度非线性关系,因而网络流量预测精度不高,预测结果不是很可靠[2~4]。近些年,随着人工智能技术的发展,为网络流量预测拓展了新的研究空间,出现神经网络、支持向量机等网络流量预测方法。相对于传统预测方法,人工智能预测方法能够对网络流量非线性和不确定性变化规律进行精确描述,能更准确刻画网络流量的变化规律,预测精度进一步得到了提高[5,6]。然而神经网络存在收敛速度慢、局部极小等不足,泛化能力比较差。网络流量受到节假日、周末、上网人行为等因素影响,具有一定周期性、随机性和多尺度变化规律,是一个变化十分复杂的系统,仍然采用单一预测模型进行预测,只能反映网络流量的部分变化规律,预测结果与实际流量之间差距比较大[7]。

针对当前网络流量预测准确率低的问题,在深入分析网络流量特性的基础上,提出一种自回归滑动平均模型(ARIMA)和支持向量机(SVM)相结合的网络流量预测模型(ARIMA-SVM)。

1 ARIMA-SVM 网络流量的预测思想

大量研究表明,网络流量具有多尺度、非线性、周期变化等特点,传统线性模型或者单一神经网络、支持向量机都不能全面刻画日趋复杂化的网络流量变化规律。由于网络受到环境、用户和网络本身等影响,具有多尺度、突变性等特点,其变化具有非平稳性,因此可以表示为

$$z(t) = s(t) + v(t) \tag{1}$$

其中:$z(t)$为网络流量原始序列;$s(t)$为线性部分;$v(t)$为非线性部分。

ARIMA-SVM 对网络流量的基本思想:首先利用 ARIMA 对网络流进行预测,获取网络流量的线性变化规律,利用 SVM 对网络流量进行预测,捕捉网络流量的非线性变化规律;然后将 ARIMA 和 SVM 的预测结果作为 SVM 的输入,采用 SVM 对两者结果进行融合;最后获得网络流量的最终预测结果。因此 ARIMA-SVM 网络流量的数学描述方式为

收稿日期: 2011-09-02; **修回日期**: 2011-10-10
作者简介: 邵忻(1979-),女,天津人,讲师,硕士,主要研究方向为计算机软件与理论(shaoxinls@163.com).

$$\hat{z}(t) = \mathrm{SVM}(\hat{s(t)} + \hat{v(t)}) \tag{2}$$

其中：$\hat{s(t)}$ 为 ARIMA 预测结果，$\hat{v(t)}$ 为 SVM 预测结果。其网络流量的预测流程如图 1 所示。

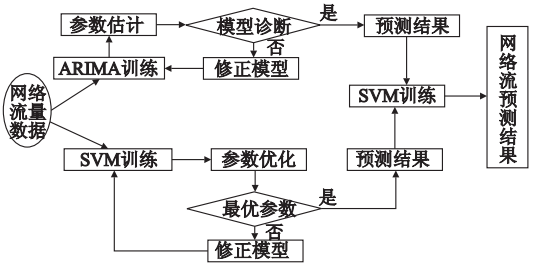

图1 ARIMA-SVM的网络流量预测流程

2 ARIMA-SVM 的网络流量预测模型

2.1 网络流量数据归一化处理

网络流量是多种影响的综合结果，变化范围比较大，需要进行归一化处理，本文采用式(1)对网络流量进行归一化处理。

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

最后对网络流量预测结果进行反归一化处理，恢复真实预测值。

$$x = x'(x_{\max} - x_{\min}) + x_{\min} \tag{4}$$

其中：x 表示网络流量原始数据，$x_{\max}$ 和 $x_{\min}$ 表示上界和下界。

2.2 ARIMA 预测模型

ARIMA 模型是自回归移动平均模型(ARMA)的推广 ARMA，由自回归(autoregression，AR)、I 差分(differencing)或结合(integration)和移动平均(moving-average，MA)三部分组成，比 ARMA 更加灵活。

设 $\{w_t, t = 0,1,\cdots\}$ 表示一个随机时间序列，如果存在一个非负整数 d，使得

$$\nabla^d w_t = x_t \tag{5}$$

其中：$\nabla = 1 - B$ 为差分算子，B 为后移算子。则该时间序列是一个平稳时间序列，否则要进行平稳化处理。

若是平稳时间序列，可得

$$\varphi(B) = 1 - \sum_{k=1}^{p} \varphi_k B^k \tag{6}$$

$$\phi(B) = 1 - \sum_{k=1}^{q} \phi_k B^k \tag{7}$$

其中：$\varphi(B)$ 和 $\phi(B)$ 互质，$\varphi_p \phi_q \neq 0$，$\{\alpha_t\}$ 为白噪声序列。

这样，$\{w_t, t = 0,1,\cdots\}$ 应该满足如下随机差分方程：

$$\varphi(B) x_t = \theta(B) \alpha_t \tag{8}$$

那么 $\{w_t, t = 0,1,\cdots\}$ 就为自回归求和滑动平均序列，记做 ARIMA(p,q,d)模型，d 为求和的阶数。

将式(5)代入式(8)可得：

$$\varphi(B)(1 - B)^d w_t = \theta(B) \alpha_t \tag{9}$$

那么称式(7)为广义自回归算子，这样式(7)就可转变成

$$\Phi(B) w_t = \theta(B) \alpha_t \tag{10}$$

2.3 ARIMA 模型参数估计

回归参数和平滑参数是 ARIMA 建模的关键，ARIMA(p,d,q)定阶方法通常由自相关函数 ACF 和偏自相关函数 PACF 决定，从最低阶 RIMA(1,1,1)模型慢慢递增尝试，并结合 Akaike 提出的使用广泛的最小信息准则(AIC)和具有相合性准则(SBC)的方法确定最适宜模型。AIC 和 SBC 函数定义为

$$\mathrm{AIC} = n\ln(\sigma_\varepsilon^2) + 2(p + q + 1) \tag{11}$$

$$\mathrm{SBC} = n\ln(\sigma_\varepsilon^2) + 2(p + q + 1) \tag{12}$$

2.4 支持向量机算法

根据给定 K 个样本数据集 $\{(x_i, y_i), i = 1,2,\cdots,k\}$，则有

$$f(x) = \omega^{\mathrm{T}} \Phi(x) + b \tag{13}$$

其中：ω 为权值向量，b 为偏置量。

一般支持向量机回归采用 ε 不敏感损失函数，ε 不敏感损失函数定义为

$$L_\varepsilon = \begin{cases} |f(x) - y| - \varepsilon & |f(x) - y| \geqslant \varepsilon \\ 0 & |f(x) - y| < \varepsilon \end{cases} \tag{14}$$

为了使回归模型获得更好的推广能力，需要同时考虑经验风险的最小化和模型复杂度。因此，SVM 回归实际上是一个优化求解问题：

$$\min_{w,b,\xi_i^*,i} = \frac{1}{2}\omega^{\mathrm{T}}\omega + c\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{15}$$

其中：ξ_i、ξ_i^* 为松弛变量，C 为惩罚参数。

最后，SVM 回归估计函数为

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*) k(x - x_i) + b \tag{16}$$

2.5 模型评价指标

本研究采用均方根误差(root mean squared predict error，RMSE)、平均绝对误差(mean absolute error，MAE)和平均绝对相对误差(mean absolute percent error，MAPE)对模型的性能进行评价，它们分别定义为

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{17}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|(y_t - \hat{y}_t)| \tag{18}$$

$$\mathrm{MAPE} = \left(\frac{1}{n}\sum_{t=1}^{n}\left|\frac{(y_t - \hat{y}_t)}{y_t}\right|\right) \times 100\% \tag{19}$$

其中：y_t 表示网络流量实际观测值，$\hat{y}_t$ 表示预测值。

3 ARIMA-SVM 在网络流量预测中的应用

3.1 数据来源

本研究对一个网站服务器的网络流量进行了采集，实时收集到 515 个数据，具体如图 2 所示。将数据分为两部分，前 465 个数据作为训练集，最后 50 个数据作为测试集，从图 2 可知，网络流量变化具有时变性、随机性，波动幅度比较大，因此首先采用式(1)对其进行归一化处理。

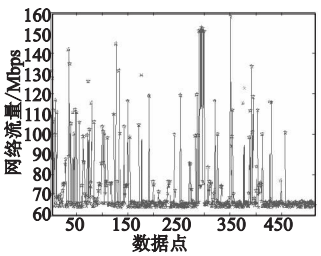

图2 网络流量原始数据

3.2　ARIMA 模型预测

3.2.1　ARIMA 模型参数估计

观察网络流量原始时间序列的 ACF 和 PACF(图 3)可知,ACF 值并未随延滞数增加而快速下降,因此该网络流量序列自我相关性极高,为非平稳的时间序列,必须对序列进行差分使其平稳化。对数据序列进行 2 阶差分的 ACF 和 PACF 如图 4 所示。从图 4 可发现,2 阶差分后的数据已经基本平稳化,所以设定 ARIMA 模型参数 $d=2$;然后采用从低阶到高阶的逐步试探法来识别模型的参数,经过比较分析,发现选择 ARIMA(2,1,2)模型的拟合效果较好,这样 ARIMA(2,1,2)就作为最优预测模型。

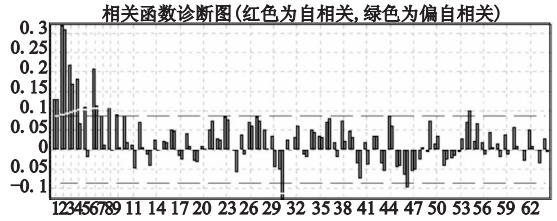

图3　原始网络流量的ACF和PACF值(参见电子版)

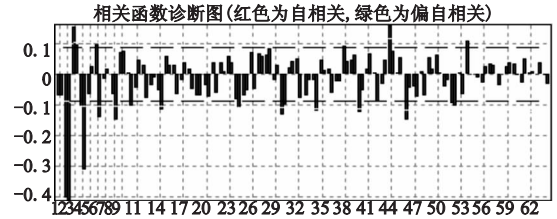

图4　2阶网络流量的ACF和PACF值(参见电子版)

3.2.2　ARIMA 的预测结果

采用 ARIMA(2,1,2)对网络流量测试集进行预测,得到图 5 所示的结果。

3.3　SVM 对网络流量进行预测

首先根据 ARIMA(2,1,2)确定 SVM 的样本阶数为 2;然后重构数据集;最后采用 SVM 对训练集进行学习,对测试集进行预测,得到的预测结果如图 6 所示。

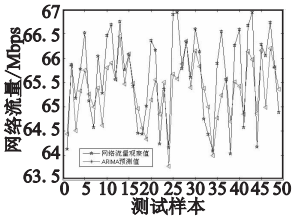

图5 ARIMA的网络流量预测结果

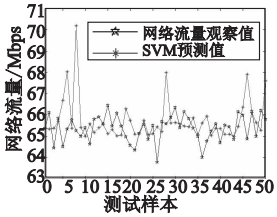

图6 SVM的网络流量预测结果

3.4　SVM 对两个预测结果融合

将 ARIMA(2,1,2)和 SVM 模型预测结果作为 SVM 的输入,网络流量实际观察值作为 SVM 期望输出,数据同样分成两部分,即训练集和测试集,从而得到最终网络流量的预测结果。预测结果如图 7 所示。

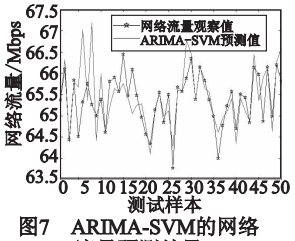

图7　ARIMA-SVM的网络流量预测结果

3.5　结果与分析

根据图 6 ~ 7 可知,ARIMA-SVM 模型的网络流量预测精度优于单一的 ARIMA 和 SVM 模型预测精度,主要因为单一预测模型只能反映网络流量部分信息;而 ARIMA-SVM 模型综合利用了 ARIMA 和 SVM 优势,达到优势互补,并克服了两者缺陷,更加全面地刻画了网络流量的变化规律,提高了网络流量预测精度,降低了预测误差。为了更具体地说明 ARIMA-SVM 模型预测结果的可靠性,将其预测结果和两个单一模型的误差进行对比分析,结果如表 1 所示。

表 1　几种预测模型的预测性能对比

模型	RMSE	MAE	MAPE
ARIMA	2.646	2.370	3.63
SVM	1.234	0.832	1.28
ARIMA-SVM	0.553	0.38	0.55

从表 1 评价指标值可知,ARIMA 和 SVM 预测误差高于本文提出的 ARIMA-SVM 预测结果,ARIMA-SVM 模型的整体预测性能最优。对比实验结果表明,本文提出的 ARIMA-SVM 模型能够精确描述网络流量变化的非线性和周期变化规律,是一种预测精度高、结果可靠的网络流量预测方法。

4　结束语

网络流量预测是网络控制的基础,其具有复杂性、非线性和高度自相似性,传统预测方法或单一预测方法难以全面刻画网络流量变化规律,预测结果有时不理想。针对网络流量变化的特点,本文提出了 ARIMA 和 SVM 相结合的网络流量预测模型。仿真结果表明,ARIMA-SVM 提高了网络流量预测精度,能够对网络流量变化规律进行很好的刻画,在网络流量等非线性预测问题中有着广泛的应用前景。

参考文献:

[1] 成兴. 网络流量预测方法和实际预测分析[J]. 计算机工程与应用,2008,42(23):129-130.

[2] 王升辉,裘正定. 结合多重分形的网络流量非线性预测[J]. 通信学报,2007,28(2):45-50.

[3] 邹柏贤,刘强. 基于 ARMA 模型的网络流量预测[J]. 计算机研究与发展,2011,39(12):1645-1652.

[4] 姜明,吴春明,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):2353-2359.

[5] LAHIRI S K, GHANTA K C. Prediction of pressure drop of slurry flow in pipeline by hybrid support vector regression and genetic algorithm model [J]. Chinese Journal of Chemical Engineering, 2008,16(6): 841-848.

[6] POURBASHEER E, RIAHI S, GANJALI M R, *et al*. Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity[J]. European Journal of Medicinal Chemistry,2009,44(12):5023-5028.

[7] HUANG Cheng-lung, WANG Chieh-jen. A GA-based feature selection and parameters optimizationfor support vector machines[J]. Expert Systems with Applications,2006,31(2):231-240.

[8] LIN Shih-wei, YING Kuo-ching, CHEN Shih-chieh, *et al*. Particle swarm optimization for parameter determination and feature selection of support vector machines[J]. Expert Systems with Applications, 2008,35(4):1817-1824.