

文章编号:1007-5321(2017)06-0065-09

DOI:10.13190/j.jbupt.2017-019

高斯过程回归补偿 ARIMA 的网络流量预测

田中大, 李树江, 王艳红, 王向东

(沈阳工业大学 信息科学与工程学院, 沈阳 110870)

摘要: 为提高网络流量时间序列的中期预测精度,提出一种高斯过程回归模型补偿自回归积分滑动平均(ARIMA)模型的网络流量预测模型. 首先通过 Brock-Dechert-Scheinkman 统计量检验方法确定网络流量时间序列包含线性特征与非线性特征;然后利用 ARIMA 模型对网络流量时间序列进行非平稳建模,得到符合网络流量序列线性变化规律的模型,并通过人工蜂群算法优化的高斯过程回归模型对具有非线性特性的预测误差序列进行建模与预测;最后将 ARIMA 模型的预测值与高斯过程回归模型的预测误差值进行相加得到最终的网络流量预测值. 仿真对比实验结果表明,提出的预测方法具有更高的预测精度和更小的预测误差.

关键词: 网络流量; 预测; 自回归积分滑动平均; 高斯过程回归; BDS 统计量

中图分类号: TP393

文献标志码: A

Network Traffic Prediction Based on ARIMA with Gaussian Process Regression Compensation

TIAN Zhong-da, LI Shu-jiang, WANG Yan-hong, WANG Xiang-dong

(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: In order to improve the medium-term prediction accuracy of network traffic, a network traffic prediction method based on auto regressive integrated moving average (ARIMA) with Gaussian process regression compensation was proposed. Firstly, the linear and nonlinear characteristics of network traffic can be determined by Brock-Dechert-Scheinkman statistics. Then, ARIMA model was used for modeling the non-stationary network traffic time series. The linear model of network traffic sequence was obtained. The artificial bee colony algorithm optimized Gaussian process regression model was used as the prediction model of predictive error sequences with the nonlinear characteristic. Finally, the final prediction value was obtained by adding predictive values of ARIMA model and predictive error values of Gaussian process regression model. Simulation comparison shows that the proposed prediction method has higher prediction accuracy with the smaller prediction error.

Key words: network traffic; prediction; auto regressive integrated moving average; Gaussian process regression; Brock-Dechert Scheinkan statistics

近年来,随着信息技术的进步,网络的资源分配成为一个重要的研究课题. 当网络过载或拥塞时,完善的资源分配机制能保证重要或高优先级的业务

量不会发生延迟或丢弃,同时保证网络的高效运行^[1]. 网络流量预测技术的发展与成熟使得建立一种基于流量精确预测的动态资源分配成为可能^[2].

收稿日期: 2017-03-02

基金项目: 辽宁省自然科学基金重点项目(20170540686); 辽宁省教育厅科学研究项目(LGD2016009)

作者简介: 田中大(1978—), 男, 讲师, E-mail: tianzhongda@126.com.

网络流量预测根据应用场合不同,通常可分为短期预测、中期预测及长期预测^[3-4]。一般而言,长期预测通常是以比较大的颗粒度即月、日的长时间段内的历史数据为参考来分析、建模、预测未来时刻流量变化,通常更注重趋势上的准确,而不是预测值的绝对准确。短期预测则更要求实时性,要求对未来秒级甚至更小尺度的网络流量进行预测。而中期预测一般是依据几分钟至几小时的历史数据,预测出未来几分钟或几小时内网络流量的变化,同时也要求具有较高的预测准确度。相对于长期预测,中短期预测更具有挑战性和难度。

目前,网络流量的中短期预测模型包括线性模型与非线性模型。线性预测模型有自回归滑动平均 (ARMA, auto regressive moving average) 模型^[5]、自回归积分滑动平均 (ARIMA, auto regressive integrated moving average) 模型^[6-7] 及差分自回归求和滑动平均模型^[8-9] 等。姜明等^[10] 利用以上线性模型对不同时间尺度下的网络流量预测进行了研究,仿真结果表明线性模型的预测精度有限。同时,随着现今网络模型的动态化和复杂化,网络流量特性已经偏离了相关学者认为的泊松或马尔可夫分布^[11],因此线性模型无法反映网络流量表现出的新特性。而非线性预测模型的典型代表包括支持向量机^[12-13]、最小二乘支持向量机^[14-15] (LSSVM, least square support vector machine)、人工神经网络^[16-18] 及灰色模型^[19] 等。相对于线性模型,非线性模型的预测精度有了一定程度的提高,但也存在着各自的缺陷。同时,进一步的研究也指出非线性模型对于时间序列中线性成分的预测精度有限。对于1个时间序列,因为无法确定其是否包含线性或者非线性成分,所以采用单一的线性或非线性预测模型是不适合的。最合适的方法是针对时间序列中的线性成分采用线性预测模型,而非线性成分则采用非线性预测模型。

笔者通过 BDS (Brock-Dechert-Scheinkman) 统计量检验方法对实际采集的中期网络流量时间序列进行分析表明,网络中期流量既含有线性成分又含有非线性成分。因此,利用 ARIMA 模型对原始网络中期流量进行预测,得到网络中期流量的线性趋势;然后利用具有良好非线性逼近能力的高斯过程回归模型对预测误差进行预测,并通过人工蜂群 (ABC, artificial bee colony) 算法对高斯过程回归模型的参数进行优化,以提高预测精度;最后将 ARIMA 模型的线性预测值与高斯过程回归模型的非线性预测误

差值进行相加而得到最终的预测值。通过仿真对比表明,所提出的网络流量中期预测方法具有较高的预测精度。

1 网络流量的线性与非线性判定

笔者利用 BDS 统计量检验法进行网络流量的线性/非线性判定。BDS 统计量检验法是基于 G-P 相关积分创建的一种统计量,用于判定时间序列的非线性特征。对于长度为 n 的时间序列,构造嵌入维数为 m 、时延为 τ 的嵌入向量

$$\mathbf{x}_t^m = (x_{1\tau}, x_{2\tau}, \dots, x_{(m-1)\tau}) \quad (1)$$

其相关积分为

$$C(m, n, r) = \frac{2}{M(M-1)} \sum_{t < s} H(r - \|\mathbf{x}_t^m - \mathbf{x}_s^m\|) \quad (2)$$

其中: $M = n - (m - 1)$; r 为包含时间序列点的 m 维球的半径; s 为参考距离; H 为 Heaviside 函数,即

$$H(z) = \begin{cases} 0, & z \leq 0 \\ 1, & z > 0 \end{cases} \quad (3)$$

用 BDS 进行检查前,需要消除原时间序列线性相关的成分,通常对原时间序列拟合自回归模型 $AR(p)$,在寻找到合适的阶数 p 后,计算 $AR(p)$ 的残差序列并对该残差序列进行 BDS 检验。BDS 检验的零假设:该残差序列是独立同分布的。如果结果拒绝零假设,则意味着原时间序列在某个显著水平下是非线性的。

采集了 250 组辽宁联通 169 核心网络接入路由器的网络中期流量数据,数据采集尺度为 10 min,数据单位为 MB。250 组网络流量数据如图 1 所示。

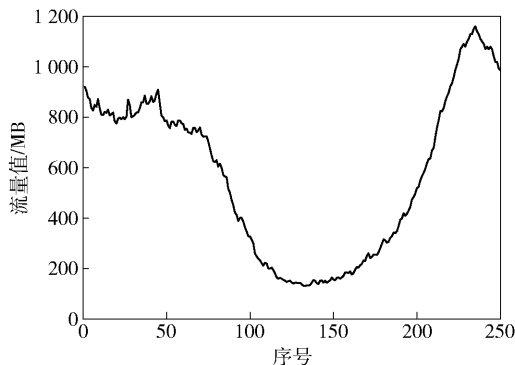


图1 网络中期流量样本序列

令 $AR(p)$ 模型的 p 为 2,得到残差序列,并对残差序列进行 BDS 检验,取嵌入维数 m 为 2,取 r 为序列标准差的 0.5、0.75、1、1.25 及 1.5 倍数,检验结

果如表 1 所示.

表 1 网络流量的 BDS 检验结果

标准差倍数	BDS 检验结果
0.50	136.890 1
0.75	85.156 8
1.00	56.261 2
1.25	43.836 3
1.50	37.987 6

标准正态分布 95% 的临界值为 1.96, 从表 1 中可知, 网络流量时间序列的 BDS 统计量检验结果远大于 1.96, 这就表明网络流量时间序列的残差序列不是独立同分布的时间序列, 具有强烈的非线性特征. 故而网络流量时间序列既含有线性又含有非线性成分. 因此对残差序列需要用非线性预测模型进行预测. 基于以上分析, 笔者提出了一种线性与非线性模型相结合的预测方法.

2 ARIMA 预测模型

ARIMA 模型的基本思想是通过多次差分使非平稳序列平稳化, 差分次数是参数 d , 然后用以 p, q 为参数的 ARMA 模型对该平稳序列建模, 之后经过反变换得到原序列. 以 p, d, q 为参数的 ARIMA 模型预测方程可表示为

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{4}$$

其中: y_t 为时间序列的样本值, $\varphi_i (i = 1, 2, \cdots, p)$ 和 $\theta_i (i = 1, 2, \cdots, q)$ 为模型参数, ε_t 为独立正态分布的白噪声.

平稳化时间序列后, 首先计算原始序列的自相关函数和偏自相关函数. 对于时间序列 y_t , 有自协方差为

$$\gamma_k = \frac{1}{N} \sum_{j=1}^{N-k} y_k y_{j+k} \tag{5}$$

自相关函数为

$$\rho = \frac{\gamma_k}{\gamma_0} \tag{6}$$

偏自相关函数为

$$\begin{aligned} \alpha_{11} &= \rho_1 \\ \alpha_{k+1, k+1} &= \left(\rho_{k+1} - \sum \rho_{k+1-j} \alpha_{kj} \right) \left(1 - \sum_{j=1}^k \rho_j \alpha_{kj} \right)^{-1} \\ \alpha_{k+1, j} &= \alpha_{kj} - \alpha_{k+1, k+1} \alpha_{k, k-j+1} \end{aligned} \tag{7}$$

可通过 ρ_k, α_k 的截尾性初步确定模型的阶数.

时间序列的参数辨识可通过最小二乘估计得到, 即估计参数 $\varphi_1, \varphi_2, \cdots, \varphi_p, \theta_1, \theta_2, \cdots, \theta_q$ 使式(8)最小.

$$\sum_{t=1}^N \alpha_t^2 = \sum_{t=1}^N (\theta_q^{-1}(Z) \varphi_p(Z) \nabla^d y_t)^2 \tag{8}$$

对不同的 p, d, q 参数进行组合, 通过赤池信息量准则得到最优的参数模型. 赤池信息量准则定义为

$$A = -2 \ln L + 2n \tag{9}$$

其中: L 为模型的极大似然参数, n 为模型的独立参数.

ARIMA 模型的实质就是差分运算与 ARMA 模型的结合, 即通过适当阶数的差分操作后, 可实现任何非平稳时间序列的平稳, 适合于线性时间序列的预测.

3 ABC 算法优化的高斯过程回归

高斯过程回归是近年发展起来的一种机器学习回归方法, 它有着严格的统计学习理论基础, 方便预测未来事物的发展. 其对处理高维数、小样本、非线性等复杂问题具有很好的适应性, 且泛化能力强^[20]. 它的基本原理是: 假设给定训练数据集 $D = \{(\mathbf{x}_i, y_i) | i = 1, 2, \cdots, n\}$, 其中 \mathbf{x}_i 表示训练数据集 D 中的第 i 个输入向量, y_i 表示训练数据集 D 中的第 i 个目标输出, n 表示训练数据集中样本的个数.

假定 f 是一个高斯过程, 即 $f \sim \text{GP}(m, k)$, f 是一个以 m 为均值函数、 k 为协方差函数的高斯过程. 高斯过程是一个随机过程, 与高斯分布类似, 高斯过程完全由其均值函数与协方差函数确定. 根据高斯过程的定义可知, $f(\mathbf{x}_1), f(\mathbf{x}_2), \cdots, f(\mathbf{x}_n)$ 服从多元高斯分布, 且该多元高斯分布的均值向量为 $m(\mathbf{x}_i)$, 协方差矩阵为 \mathbf{K} , 因此有

$$f(\mathbf{x}_i) \sim N[m(\mathbf{x}_i), \mathbf{K}], i = 1, 2, \cdots, n \tag{10}$$

$$D: y_i = f(\mathbf{x}_i), i = 1, 2, \cdots, n \tag{11}$$

实际目标输出 \mathbf{y} 往往会包含一些噪声:

$$\mathbf{y} = f(\mathbf{x}_i) + \varepsilon_i \tag{12}$$

其中 $\varepsilon \sim N(0, \sigma_n^2)$. 于是问题转化为已经观测到训练数据 $D: y_i = f(\mathbf{x}_i) + \varepsilon_i (i = 1, 2, \cdots, n)$, 需要在测试数据集 $D_* = \{(\mathbf{x}_i, y_i) | i = n+1, n+2, \cdots, n+n_*\}$ 预测对应的输出值 f_* . 训练数据集的输出向量 \mathbf{y} 和测试数据集的预测值 f_* 的多元高斯分布为

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim N \left(0, \begin{bmatrix} \mathbf{K} + \sigma_n^2 \mathbf{I} & \mathbf{K}_*^T \\ \mathbf{K}_* & \mathbf{K}_{**} \end{bmatrix} \right) \tag{13}$$

其中

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & k(\mathbf{x}_n, \mathbf{x}_2) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

$$\mathbf{K}_* = [k(\mathbf{x}_*, \mathbf{x}_1), k(\mathbf{x}_*, \mathbf{x}_2), \cdots, k(\mathbf{x}_*, \mathbf{x}_n)]$$

$$\mathbf{K}_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$$

根据多元高斯分布的条件分布形式,可得出高斯过程预测方程的关键式为

$$f_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \text{GP}[m(\mathbf{x}_*), \text{cov}(f_*)] \quad (14)$$

其中:矩阵 \mathbf{X} 由训练数据输入 \mathbf{x}_i 的列向量组成,矩阵 \mathbf{X}_* 由测试集的输入 \mathbf{x}_{i*} 的列向量组成.

$$m(\mathbf{x}_*) = \mathbf{K}_* (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \quad (15)$$

$$\text{cov}(f_*) = \mathbf{K}_{**} - \mathbf{K}_* (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K}_*^T \quad (16)$$

高斯过程回归中协方差函数是一个满足 Mercer 条件的对称函数,其在有限输入集上正定,因此协方差函数等价于核函数,将式(15)改写为

$$m(\mathbf{x}_*) = \sum_{i=1}^n \alpha_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}_*) \quad (17)$$

其中

$$\alpha_i = [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \delta_n^2 \mathbf{I}]^{-1} \mathbf{y}$$

这里选择平方指数函数作为核函数,即

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \delta_f^2 \exp \left(-\frac{1}{2l^2} (\mathbf{x}_i - \mathbf{x}_j)^2 \right) + \delta_n^2 \delta_{ij} \quad (18)$$

其中: l 为关联性测定超参数, δ_f^2 为核函数的信号方差, δ_n^2 为噪声方差, δ_{ij} 为 Kronecker 符号. 一般将 l 、 δ_f^2 、 δ_n^2 称为超参数. 超参数直接决定了高斯过程回归的预测性能. 目前一般通过共轭梯度法来确定高斯过程回归模型的参数,但是共轭梯度法通过梯度下降求取最优解,容易陷入局部最优,同时算法效果和收敛性依赖初始值.

ABC 算法来源于蜜蜂采蜜的行为^[21]. ABC 算法相对于遗传算法、差分进化算法以及粒子群优化算法具有极强的竞争力^[22]. 基于此,采用 ABC 算法对高斯过程回归的超参数进行优化.

ABC 算法中蜜源的位置抽象为解空间内的点,蜜源($i = 1, 2, \cdots, S$)的质量为解的适应度. 设求解问题的维数为 d ,对于第 t 次迭代蜜源 i 的位置可表示为 $\mathbf{X}_i^t = [x_{i1}^t \ x_{i2}^t \ \cdots \ x_{id}^t]$, $x_{id} \in (L_d, U_d)$,其中 L_d 为搜索空间下限, U_d 为搜索空间上限,蜜源的位置按式(19)在搜索空间随机产生.

$$x_{id} = L_d + \text{rand}(0, 1) (U_d - L_d) \quad (19)$$

搜索开始阶段,引领蜂在蜜源周围根据式(20)

搜索而产生一个新的蜜源.

$$v_{id} = x_{id} + \varphi (x_{id} - x_{jd}) \quad (20)$$

其中: $j \in \{1, 2, \cdots, S\}$, $j \neq i$,代表在蜜源中随机选择一个不同于 i 的蜜源; $\varphi \in [-1, 1]$. 当新蜜源 $\mathbf{V}_i = [v_{i1} \ v_{i2} \ \cdots \ v_{id}]$ 的适应度优于 \mathbf{X}_i 时,利用贪婪算法将 \mathbf{V}_i 代替 \mathbf{X}_i ,否则保留 \mathbf{X}_i . 然后跟随蜂根据所有引领蜂的蜜源分享信息,按概率

$$p_i = \frac{\text{fit}_i}{\sum_{i=1}^S \text{fit}_i} \text{fit}_i \quad (21)$$

进行跟随. 其中 fit_i 为第 i 个蜜源的适应度值. 也即跟随蜂产生一个 $[0, 1]$ 间的随机数并与 p_i 进行比较,如果该随机数小于 p_i 则按式(20)产生一个新蜜源. 搜索中,蜜源 \mathbf{X}_i 经过若干次迭代到达阈值而没有找到更好的蜜源,则 \mathbf{X}_i 将被抛弃,与之对应的引领蜂转变为侦查蜂,并按式(22)随机产生一个新的蜜源替代 \mathbf{X}_i .

$$\mathbf{X}_i^{t+1} =$$

$$\begin{cases} L_d + \text{rand}(0, 1)(U_d - L_d), & \text{尝试次数} \geq \text{最大开采次数} \\ \mathbf{X}_i^t, & \text{尝试次数} < \text{最大开采次数} \end{cases} \quad (22)$$

为了获取最小的均方根误差(RMSE, root mean square error)以及和 ABC 算法保持一致,采用式(23)作为 ABC 优化高斯过程回归参数的适应度函数.

$$\text{fit}_i = \frac{1}{1 + r_i} \quad (23)$$

其中 r_i 为实际值与预测值的 RMSE:

$$r_i = \sqrt{\frac{1}{S} \sum_{j=1}^S (y_j - \hat{y}_j)^2} \quad (24)$$

ABC 算法优化高斯模型回归预测模型的步骤如下:

步骤1 生成训练数据样本集,同时参数初始化,包括蜜源的个数 S ,最大迭代次数 M ,蜜源最大开采次数为 L ,待优化的超参数 l 、 δ_f^2 、 δ_n^2 的取值范围,令 $t = 1$;

步骤2 为蜜源分配一只引领蜂,按照式(20)进行搜索,产生新蜜源 \mathbf{V}_i ;

步骤3 代入样本数据,根据高斯过程回归模型计算输出预测值,并根据式(23)计算适应度值,根据贪婪算法保留蜜源;

步骤4 按式(21)计算蜜源被跟随的概率,跟随蜂搜索,根据贪婪算法保留蜜源;

步骤 5 判断蜜源 X_i 是否应抛弃,若是,则引领蜂转变为侦查蜂,否则转步骤 7;

步骤 6 侦查蜂按式(22)产生新的蜜源;

步骤 7 令 $t = t + 1$,如果满足终止条件,输出最优超参数,否则转到步骤 2 继续执行。

4 网络流量预测模型

设网络流量时间序列为 $T(k)$, k 为数据采集时刻,高斯过程回归补偿 ARIMA 的网络流量预测模型如图 2 所示. 通过单步预测迭代循环的方式实现网络流量多步预测,即对 $k+1$ 时刻的网络流量进行预测,然后将预测值看作实际值,通过滑窗机制代入网络流量序列,并剔除最旧的数据,然后按照预测算法进行 $k+2$ 时刻网络流量的预测,直到满足最大预测步数. 单步迭代预测方法较直接多步预测方法减少了模型训练时间,同时可利用拉伊达准则等算法对预测值数据进行合理性检测,结合插值操作,减少预测误差。

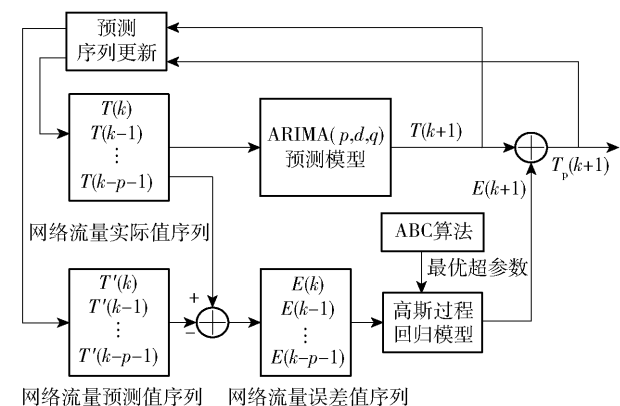


图 2 网络流量预测模型

预测步骤描述如下:

步骤 1 预测模型训练,利用网络流量样本数据建立 ARIMA 预测模型,通过 ARIMA 的预测流量值与实际流量值误差建立高斯过程回归预测模型,模型中的超参数通过 ABC 算法得到;

步骤 2 利用 $T(k), T(k-1), \dots, T(k-1-p)$ 通过建立的 ARIMA 模型预测下一时刻的网络流量值 $T'(k+1)$;

步骤 3 利用网络流量实际值与 ARIMA 预测值生成预测误差序列 $E(k), E(k-1), \dots, E(k-1-p)$,通过高斯过程回归模型生成网络流量误差的预测值 $E'(k+1)$;

步骤 4 将 $T'(k+1)$ 与 $E'(k+1)$ 相加,得到最

终的预测值 $T_p(k+1)$;

步骤 5 更新预测序列,将 $T'(k+1)$ 放入网络流量预测值队列中,将 $T_p(k+1)$ 放入网络流量实际值序列中,同时均剔除 2 个序列中最旧的数据,重复步骤 2,直到预测结束。

5 仿真

仿真数据利用上文介绍的 250 组辽宁联通 169 核心网接入路由器的网络中期流量数据(数据采集尺度为 10 min),通过前 200 组数据进行 ARIMA 与高斯过程回归预测模型的建立,后 50 组数据用来进行预测模型精度的验证。

首先利用网络流量数据进行 ARIMA 预测模型的建立,训练完毕其模型为 $ARIMA(10, 1, 7)$,利用该模型对 50 组网络流量测试样本进行了预测. 图 3 为 ARIMA 模型的网络流量预测值与实际值的对比曲线. 从图中可看出 ARIMA 模型的网络流量预测值在趋势上可以较好地拟合实际值,预测出了网络流量时间序列中的线性特征,但是存在预测误差过大的问题,同时该预测误差是非线性的,因此需要对其预测误差进行补偿修正。

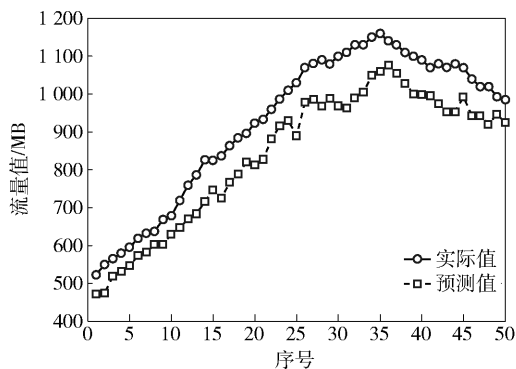


图 3 ARIMA 模型的网络流量实际值与预测值对比

利用得到的 ARIMA 预测模型对 200 组网络流量训练样本数据进行预测,得到 200 组网络流量样本预测误差,其曲线如图 4 所示. 从图 4 可见,样本预测误差时间序列具有强烈的非线性特征,因此选择高斯过程回归模型是合适的. 利用预测误差序列完成高斯过程回归预测模型的建立。

ABC 算法参数设置:蜜源的个数 S 为 20,最大迭代次数 M 为 100,蜜源最大开采次数 L 为 50. 由于待优化的参数为 3 个,则最优解是 3 维向量. 待优化的超参数取值范围为 $l \in [0, 10], \delta_f^2 \in [0, 10\ 000], \delta_n^2 \in [0, 100]$. 图 5 为 ABC 算法优化高斯

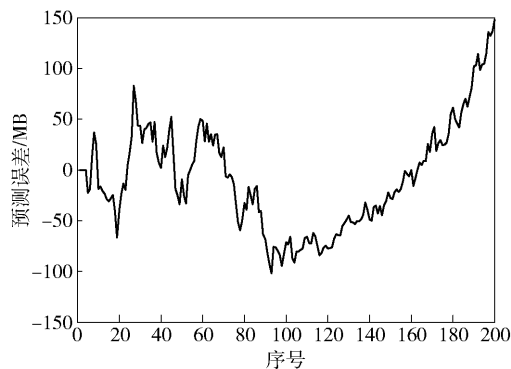


图4 网络流量训练样本预测误差

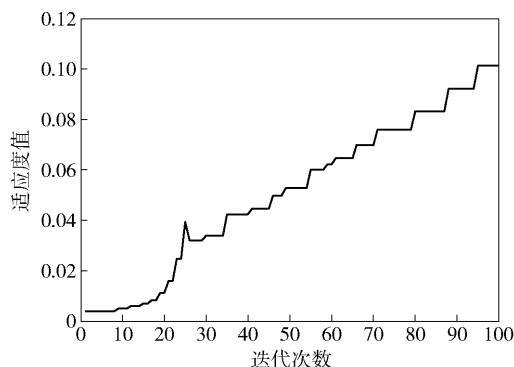


图5 ABC算法适应度变化曲线

过程回归适应度曲线,寻优之后的结果为 $l = 5.61$, $\delta_f^2 = 73.27$, $\delta_n^2 = 13.26$. 利用 ABC 算法优化后的超参数,高斯过程回归模型对 ARIMA 模型网络流量预测误差的预测对比曲线如图 6 所示,从图中可观察到高斯过程回归模型对于 ARIMA 模型的预测误差具有很好的预测能力. 将 ARIMA 预测模型的网络流量预测值与高斯过程回归模型的网络流量误差预测值进行相加,即可得到最终的预测值,图 7 为最后叠加后得到的预测值与实际值的对比曲线.

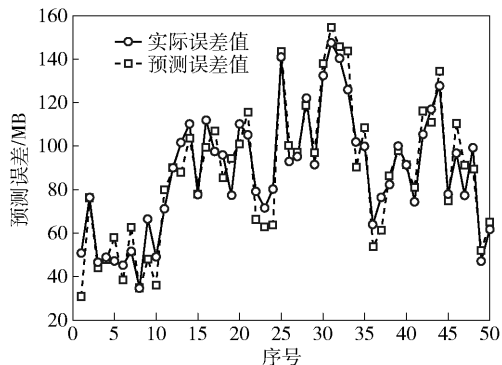


图6 高斯过程回归的网络流量误差预测曲线

为了对比所提出预测方法的效果,图 8 给出了

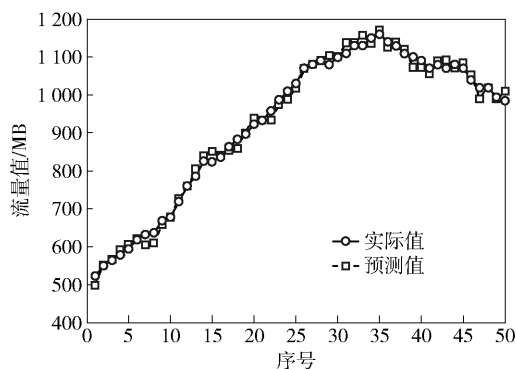


图7 叠加后的网络流量实际值与预测值的对比曲线

多核 LSSVM (优化后参数为 $\rho = 0.89$, $\gamma = 2.35$, $\sigma^2 = 1.05$, $q = 4.26$)^[15]、遗传算法优化回声状态网络(参数 SR 为 0.617 3, $N = 58$, IS 为 0.021 2, SD 为 0.419)^[16]、Elman 神经网络(输入层数为 30,最大迭代次数为 3 000,迭代目标为 0.000 1,隐层个数为 15,输出层个数为 1,利用迭代完成多步预测)^[17]、单一高斯过程回归(利用共轭梯度法确定模型的超参数为 $l = 12.68$, $\delta_f^2 = 27.23$, $\delta_n^2 = 8.67$) 4 种预测模型预测效果曲线对比. 由图 7 与图 8 可观察到,在数据拟合程度上所提出的组合预测方法要优于其他几种方法.

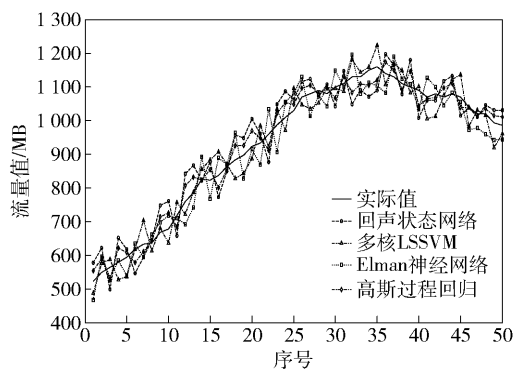


图8 其他方法实际值与预测值的对比曲线

图 9 为几种预测方法的误差分布图,从图中可观察到所提出的预测方法在预测误差上要小于其他预测方法,同时误差分布更加均匀,即预测误差随着预测步长的增长变化相对平稳,受预测步长的影响较小.

为了进行预测效果的对比,引入如下 4 种性能指标.

1) RMSE e_{RMSE}

$$e_{\text{RMSE}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (T_i - \hat{T}_i)^2} \quad (25)$$

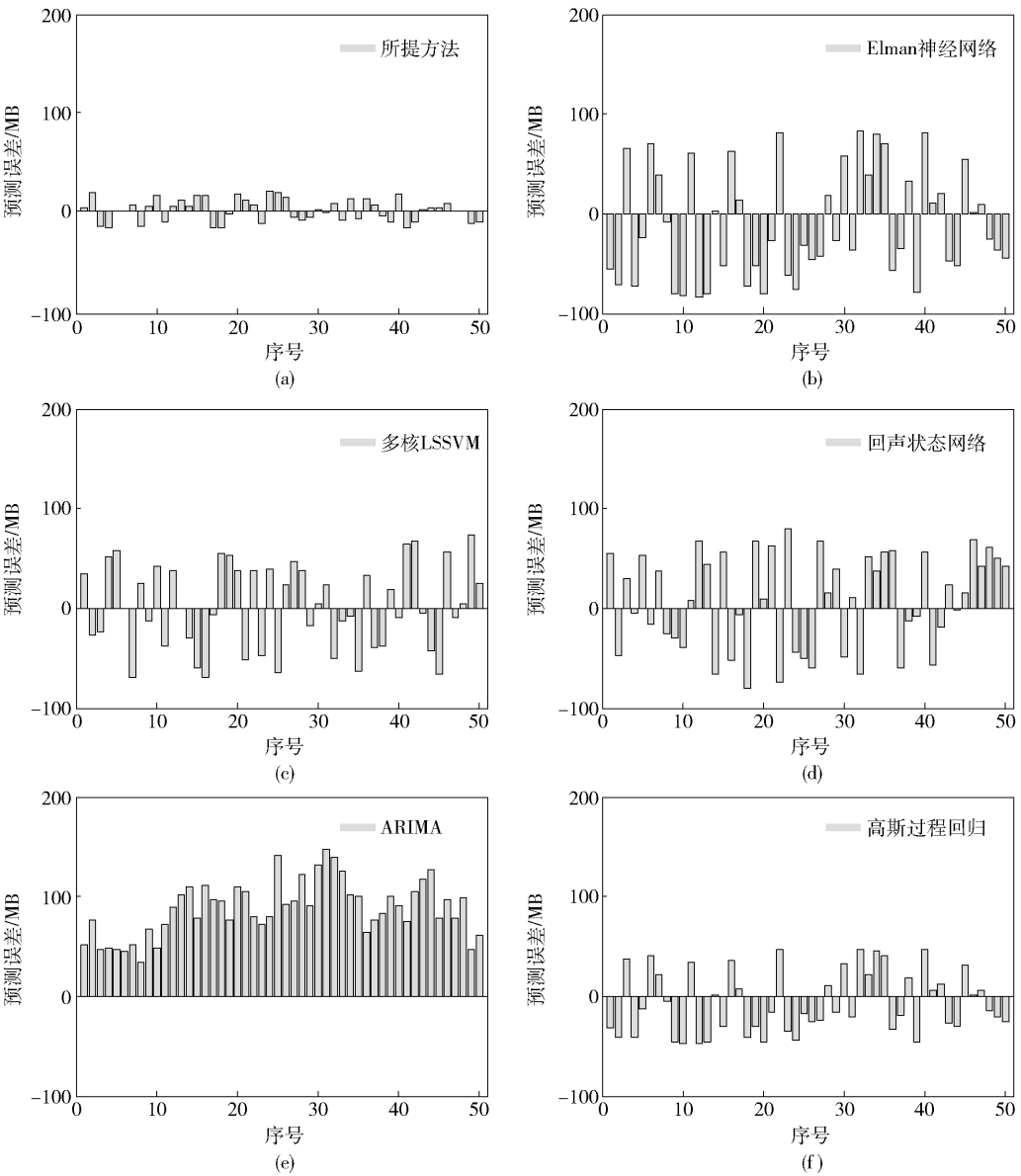


图9 预测误差分布

其中: N 为样本数量, T_i 为网络流量实际值, \hat{T}_i 为预测模型的网络流量预测值。

2) 平均绝对误差 (MAE, mean absolute error) e_{MAE}

$$e_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^N |T_i - \hat{T}_i|$$

(26)

3) MAPE 平均绝对百分误差 (MAPE, mean absolute percentage error) e_{MAPE}

$$e_{\text{MAPE}} = \frac{1}{N} \sum_{i=1}^N \frac{|T_i - \hat{T}_i|}{T_i} \times 100$$

(27)

4) 可靠性

$$R^{(1-a)} = \left[\frac{\xi^{(1-a)}}{N} - (1-a) \right] \times 100\%$$

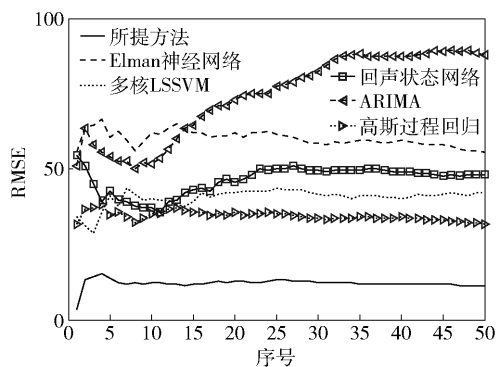
(28)

其中 $\xi^{(1-a)}$ 为在置信度 $(1-a)$ 下实际值落入预测置信区间的个数。

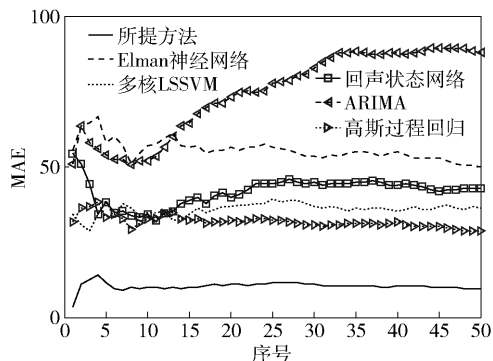
表 2 给出了几种预测方法的 RMSE、MAE、MAPE 的性能指标对比。图 10 是这些预测方法的 RMSE、MAE 以及 MAPE 随着预测步数增加的变化

表 2 几种方法误差性能指标对比

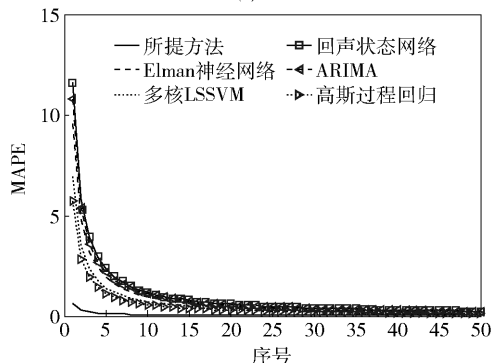
预测方法	RMSE	MAE	MAPE
所提方法	11.251 7	9.538 3	0.012 7
多核 LSSVM	24.341 3	36.245 5	0.139 3
Elman 神经网络	55.683 6	49.968 8	0.215 6
回声状态网络	47.959 6	42.721 2	0.191 6
ARIMA	87.725 9	84.135 9	0.231 7
高斯过程回归	31.828 3	28.561 7	0.114 2



(a) RMSE



(b) MAE



(c) MAPE

图 10 RMSE、MAE 以及 MAPE 变化对比曲线

曲线. 表 2 与图 10 中的结果同样显示, 所提出的预测方法在误差的性能指标上也优于其他预测方法.

图 11 为几种预测方法的可靠性值与置信度的分布图, 从图中可看出所提出的方法在可靠性上要优于其他几种方法.

6 结束语

提出了一种高斯过程回归补偿 ARIMA 的网络流量中期预测方法. 利用 ARIMA 模型进行网络流量线性成分的预测, 通过高斯过程回归模型进行具有非线性特征的网络流量预测误差的预测, 同时采用 ABC 算法对高斯过程回归模型的参数进行优化. 两种方法预测值累加得到最终的预测值, 通过仿真

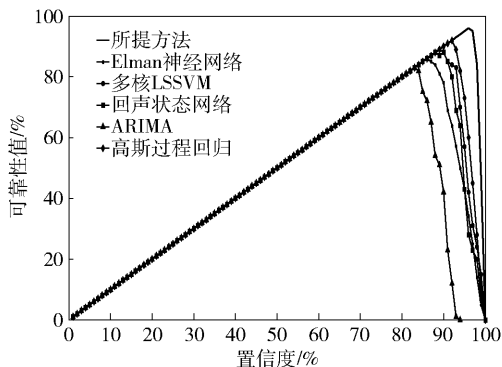


图 11 预测方法的可靠性值对比

实验对比表明所提出的方法具有很好的预测效果, 对解决未来设计网络资源分配原则、拥塞控制策略、网络管理优化等问题提供了良好的基础.

致谢 中国联合网络通信集团有限公司辽宁省分公司集团客户响应中心的顾斌工程师为笔者提供了网络流量数据, 在此表示感谢!

参考文献:

- [1] 赵建龙, 曲桦, 赵季红, 等. 基于 Morlet-SVR 和 ARIMA 组合模型的网络流量预测[J]. 北京邮电大学学报, 2016, 39(2): 53-57.
Zhao Jianlong, Qu Hua, Zhao Jihong, et al. A comprehensive forecasting model for network traffic based on Morlet-SVR and ARIMA[J]. Journal of Beijing University of Posts and Telecommunications, 2016, 39(2): 53-57.
- [2] 段华琼, 唐宾徽. 基于线性多尺度模型的计算机网络数据流量预测[J]. 沈阳工业大学学报, 2017, 39(3): 322-327.
Duan Huaqiong, Tang Binhui. Prediction of data flow in computer network based on linear multi-scale model[J]. Journal of Shenyang University of Technology, 2017, 39(3): 322-327.
- [3] Tian Zhongda, Li Shuijiang, Wang Yanhong, et al. A network traffic hybrid prediction model optimized by improved harmony search algorithm[J]. Neural Network World, 2015, 25(6): 669-686.
- [4] Qu Hua, Ma Wentao, Zhao Jihong, et al. Prediction method for network traffic based on maximum correntropy criterion[J]. China Communications, 2013, 10(1): 134-145.
- [5] Laner M, Svoboda P, Rupp M. Parsimonious fitting of long-range dependent network traffic using ARMA models[J]. IEEE Communications Letters, 2013, 17(12): 2368-2371.

- [6] Yadav R K, Balakrishnan M. Comparative evaluation of ARIMA and ANFIS for modeling of wireless network traffic time series[J]. *EURASIP Journal on Wireless Communications and Networking*, 2014, 2014(1): 15.
- [7] Wang Jin. A process level network traffic prediction algorithm based on ARIMA model in smart substation[C]//2013 IEEE International Conference on Signal Processing, Communication and Computing. Kunming: IEEE Press, 2013: 1-5.
- [8] Ren Xunyi, Yang Yu, Zhang Junfeng, et al. Parameter estimation and application of time-varying FARIMA model[J]. *International Journal of Advancements in Computing Technology*, 2011, 3(3): 89-94.
- [9] Katris C, Daskalaki S. Comparing forecasting approaches for Internet traffic[J]. *Expert Systems with Applications*, 2015, 42(21): 8172-8183.
- [10] 姜明, 吴春明, 张旻, 等. 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, Wu Chunming, Zhang Min, et al. Research on the comparison of time series models for network traffic prediction[J]. *Acta Electronica Sinica*, 2009, 37(11): 2353-2358.
- [11] 马静, 沈来信, 盛文婷. 在线开放通信网络信道分配算法优化[J]. *沈阳工业大学学报*, 2017, 39(2): 193-197.
- Ma Jing, Shen Laixin, Sheng Wenting. Optimization for online open communication network channel allocation algorithm[J]. *Journal of Shenyang University of Technology*, 2017, 39(2): 193-197.
- [12] Liang Yonglin, Qiu Lirong. Network traffic prediction based on SVR improved by chaos theory and ant colony optimization[J]. *International Journal of Future Generation Communication and Networking*, 2015, 8(1): 69-78.
- [13] Liu Xingwei, Li Hua, Chen Lei, et al. An improved forecasting algorithm for wireless network traffic[J]. *IEICE Electronics Express*, 2011, 8(12): 916-922.
- [14] Tian Zhongda, Li Shuijiang. A network traffic prediction method based on IFS algorithm optimised LSSVM[J]. *International Journal of Engineering Systems Modelling and Simulation*, 2017, 19(4): 200-213.
- [15] 田中大, 高宪文, 石彤. 用于混沌时间序列预测的组合核函数最小二乘支持向量机[J]. *物理学报*, 2014, 63(16): 160508.
- Tian Zhongda, Gao Xianwen, Shi Tong. Combination kernel function least squares support vector machine for chaotic time series prediction[J]. *Acta Physica Sinica*, 2014, 63(16): 160508.
- [16] 田中大, 高宪文, 李树江, 等. 遗传算法优化回声状态网络的网络流量预测[J]. *计算机研究与发展*, 2015, 52(5): 1137-1145.
- Tian Zhongda, Gao Xianwen, Li Shuijiang, et al. Prediction method for network traffic based on genetic algorithm optimized echo state network[J]. *Journal of Computer Research and Development*, 2015, 52(5): 1137-1145.
- [17] Wang Junsong, Wang Jiukun, Zeng Maohua, et al. Prediction of Internet traffic based on Elman neural network[C]//2009 Chinese Control and Decision Conference. Guilin: IEEE Press, 2009: 1248-1252.
- [18] Qian Feng. A extreme learning machines approach for accurate estimation of large-scale IP network traffic matrix[J]. *Journal of Computational Information Systems*, 2012, 8(2): 755-762.
- [19] Cao Jianhua, Liu Yuan, Dai Yue. Network traffic prediction based on error advanced DGM(1, 1) model[C]//International Conference on Wireless Communication Networking and Mobile Computing. Shanghai: IEEE Press, 2007: 6353-6356.
- [20] 何志昆, 刘光斌, 赵曦晶, 等. 高斯过程回归方法综述[J]. *控制与决策*, 2013, 28(8): 1121-1129.
- He Zhikun, Liu Guangbin, Zhao Xijing, et al. Overview of Gaussian process regression[J]. *Control and Decision*, 2013, 28(8): 1121-1129.
- [21] Karaboga D, Basturk B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm[J]. *Journal of Global Optimization*, 2007, 39(3): 459-471.
- [22] Karaboga D, Gorkemli B, Ozturk C, et al. A comprehensive survey: artificial bee colony (ABC) algorithm and applications[J]. *Artificial Intelligence Review*, 2014, 42(1): 21-57.