计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400056

面向海量数据的网络流量混沌预测模型

向昌盛[1]　陈志刚[2]
1 湖南工程学院计算机与通信学院　湖南 湘潭 411104
2 中南大学计算机学院　长沙 410000
(13077331687@163.com)

摘　要　针对网络流量的混沌特性以及海量特性,为弥补网络流量预测模型存在的不足,以获得更优的网络流量预测结果,提出了面向海量数据的网络流量混沌预测模型。该模型首先采用小波分析对原始网络流量时间序列进行多尺度处理,得到不同特征的网络流量分量,然后对网络流量分量的混沌特性进行分析,分别进行重构,并采用机器学习算法中的极限学习机进行建模与预测,最后采用小波分析对网络流量分量的预测结果进行叠加,得到原始网络流量数据的预测值,并进行网络流量预测的仿真实验。实验结果表明,所提模型的网络流量预测精度超过90%,不仅预测精度结果远远超过其他网络流量预测模型的结果,而且其网络流量预测的结果更加稳定,因此是一种有效的网络流量建模与预测工具。

关键词:小波分析;网络流量;建模与预测;仿真测试;海量特征;极限学习机
中图法分类号　TP181

Chaotic Prediction Model of Network Traffic for Massive Data

XIANG Chang-sheng[1] and CHEN Zhi-gang[2]
1 School of Computer and Communication,Hunan Institute of Engineering,Xiangtan,Hunan 411104,China
2 School of Computer Science and Engineering,Central South University,Changsha 410000,China

Abstract　Aiming at the chaotic and massive characteristics of network traffic,in order to make up for the shortcomings of network traffic prediction model to obtain better network traffic prediction results,a chaotic network traffic prediction model for massive data is proposed. First,wavelet analysis is used to deal with the original network traffic time series in multi-scale to obtain network traffic components with different characteristics. Then,the chaotic characteristics of network traffic components are analyzed and reconstructed respectively. The extreme learning machine in machine learning algorithm is used to model and predict. Finally,wavelet analysis is used to overlay the prediction results of network traffic components to get the original network traffic data prediction value,and the network traffic prediction simulation experiment is carried out. Experimental results show that,compared with other network traffic prediction models,the network traffic prediction accuracy of the proposed model is more than 90%,and the network traffic prediction results are more stable. It is an effective tool for network traffic modeling and prediction.

Keywords　Wavelet analysis,Network traffic,Modeling and prediction,Simulation test,Massive features,Extreme learning machine

1　引言

随着网络应用的范围日益拓宽,每天都有大量的网络流量数据产生,网络系统出现拥塞的频率越来越高。网络流量预测可以对将来网络系统的状态可能性做出科学的评价,并对网络流量的变化趋势做出准确的预测,因此网络流量建模与预测分析研究是研究者们关注的焦点[1-3]。

目前存在大量的网络流量建模与预测模型,这些模型均取得了不错的应用效果[4]。尤其是近几年来,随着人工智能技术的发展,网络流量建模与预测得到前所未有的发展,其中最具代表性的网络流量建模技术有:灰色模型、人工神经网络、支持向量机等[5-7]。灰色模型属于线性建模技术,人工神经网络和支持向量机为非线性建模技术。然而网络流量与经济、网络用户、上网价格等多种因素有关,是一个非线性动力系统,具有一定的混沌特性,同时亦具有一定的多层次性,因此单一的灰色模型、人工神经网络、支持向量机等均难以对网络流量变化特性进行全面准确的预测[8-10]。小波分析可以对非线性信号进行多层次分解,能够更好地描述非线性信号的变化特性,因此有学者提出小波分析和神经网络、小波分析和支持向量机相融合的网络流量预测模型,该类模型使得网络

到稿日期:2020-04-14　返修日期:2020-09-02　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61672540);湖南省自然科学基金(2018JJ2082);2018年湖南省教育厅优秀青年项目(18B386)
This work was supported by the National Natural Science Foundation of China(61672540),Natural Science Foundation of Hunan Province,China(2018JJ2082) and Outstanding Young Scholars Program of Hunan Provincial Education Department,2018(18B386).
通信作者:陈志刚 (czg@csu.edu.cn)

流量的预测精度明显优于单一的神经网络和支持向量机的预测精度[11-13]。在实际应用中，神经网络和支持向量机存在明显的缺陷，如神经网络的收敛效率低、支持向量机的学习效率低等[14-15]。

极限学习机是一种现代机器学习算法，其收敛速度明显优于神经网络，且参数选取简单，学习速度快于支持向量机。因此，本文根据网络流量的变化特性，以提高网络流量预测精度为目标，提出了面向海量数据的网络流量混沌预测模型，并对其性能进行了验证性测试。

2 网络流量的混沌特性分析

2.1 网络流量的混沌特性

对网络流量历史数据序列进行混沌处理和分析之前，需要判定网络流量历史数据序列的混沌特性。如果没有混沌特性，进行混沌特性分析就是错误的，同时也会影响网络流量预测的建模效果。混沌系统对初值十分敏感，其在相空间会呈指数速率进行发散，这与 Lyapunov 指数的变化趋势相符，因此可以对网络流量历史数据序列的 Lyapunov 指数进行计算。如果其 Lyapunov 指数大于 0，则表示该网络流量历史数据序列具有混沌特性[16]。

2.2 网络流量时间序列的重构

当前混沌时间序列基本上采用相空间重构方法实现，该方法可以重构与原非线性系统具有相似变化规律的状态空间。设一个一维的网络流量历史数据序列为 $x_i, i=1,2,\cdots,n$，通过确定最合理的延迟时间 τ 和嵌入维数 m 来推导一个多维的网络流量历史数据序列，具体如下：

$$y_t=\{x_t, x_{t+\tau}, x_{t+2\tau}, \cdots, x_{t+(m-1)\tau}\} \tag{1}$$

其中，$t=1,2,\cdots,n-(m+1)\tau$。

由式(1)可知，最优 m 和 τ 是网络流量时间序列的相空间重构重点键。当前有许多确定 m 和 τ 的方法，如相关法、互信息法确定最优 τ，试算法、虚假邻近点法确定最优 m，本文选择互信息法和虚假邻近点法分别确定 τ 和 m。

3 面向海量数据的网络流量混沌预测模型

3.1 小波分析理论

小波分析是一种时频局部化分析方法，可以对复杂信号进行多尺度细化，能够对信号进行更细致的分析和处理。对于信号 $x(t)$，通过小波函数 $\varphi(t)$ 进行 a 尺度和 b 的平移后，可以得到时域表达式为：

$$f_x(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}x(t)\varphi\left(\frac{t-b}{a}\right)\mathrm{d}t, a>0 \tag{2}$$

首先采用小波分析的 Mallat 算法对网络流量历史数据的序列进行分解，得到高频部分和低频部分，同时可以对低频部分再进行细分。一般通过 3～4 层的细分分解，就可以得到比较理想的细分结果，再对网络流量细分结果进行建模与预测，最后通过小波分析的重构对预测结果进行融合。

3.2 极限学习机

相比传统人工神经网络，极限学习机不需要调整输入层的权值和隐含层的阈值，是一种新型的机器学习算法，其只需设置隐含层神经元数量(k)，训练过程十分简单，学习效率高。对于训练样本集合$\{x_i, t_i\}$，$i=1,2,\cdots,n$，O 表示网络输出值，则极限学习机可以表示为：

$$\sum_{i=1}^{k}\beta_i g(\alpha_i \cdot x_i+b_i)=t_j \tag{3}$$

其中，$g()$表示激励函数，α 表示输入和隐含层神经元之间的连接权值，β 表示隐含层和输出层神经元之间的连接权值，b 表示隐含层的阈值。

因此，式(3)可以简化为：

$$\boldsymbol{H}\beta=\boldsymbol{T} \tag{4}$$

其中，$\boldsymbol{H}$ 表示输出矩阵，具体如下：

$$\boldsymbol{H}=\begin{bmatrix}h(x_1)\\ \vdots \\ h(x_n)\end{bmatrix}=\begin{bmatrix}g(\alpha_1\cdot x_1+b_1) & \cdots & g(\alpha_1\cdot x_1+b_k)\\ \vdots & \vdots & \vdots \\ g(\alpha_1\cdot x_n+b_1) & \cdots & g(\alpha_1\cdot x_n+b_k)\end{bmatrix} \tag{5}$$

由于输入层的权值和隐含层的阈值均采用随机方式确定，极限学习机的训练过程与式(6)的求解等价，即：

$$\min_{\beta}\|\boldsymbol{H}\beta-T\| \tag{6}$$

那么可以得到：

$$\hat{\beta}=\boldsymbol{H}^{+}Y \tag{7}$$

在极限学习机训练过程中，连接权值 α、β 和阈值 b 会影响其性能，采用随机方式确定难以得到最优参数 α,β,b 值，因此选择粒子群算法来确定最优的 α,β,b 值。

3.3 粒子群算法优化极限学习机的参数

第 i 个粒子的位置和速度向量分别为 $\boldsymbol{X}_i=(\boldsymbol{X}_{i1},\boldsymbol{X}_{i2},\cdots,\boldsymbol{X}_{iD})^{\mathrm{T}}$ 和 $\boldsymbol{V}_i=(\boldsymbol{V}_{i1},\boldsymbol{V}_{i2},\cdots,\boldsymbol{V}_{iD})^{\mathrm{T}}$，其中，$D$ 表示解搜索空间的维数，其极值为 $P_i=(P_{i1},P_{i2},\cdots,P_{iD})^{\mathrm{T}}$，整个数子群的极值为 $P_g=(P_{g1},P_{g2},\cdots,P_{gD})^{\mathrm{T}}$，那么在 $t+1$ 次迭代过程中，速度和位置更新公式为：

$$\boldsymbol{V}_{iD}^{t+1}=\omega V_{iD}^{t}+c_1 r_1(P_{iD}^{t}-X_{iD}^{t})+c_2 r_2(P_{gD}^{t}-X_{iD}^{t}) \tag{8}$$

$$X_{iD}^{t+1}=X_{iD}^{t}+V_{iD}^{t+1} \tag{9}$$

其中，c_1 和 c_2 为常数，r_1 和 r_2 为随机数，ω 为惯性权值。

粒子群算法优化极限学习机的参数步骤如下。

(1)设置粒子群算法参数，如粒子的数量、最大迭代次数、常数 c_1 和 c_2 等；

(2)产生初始粒子群，粒子采用十进制编码方式，用一个粒子位置向量表示极限学习机参数 α,β,b 的一组组合；

(3)设迭代次数为 $t=1$；

(4)对每一个粒子的位置进行解码，并将其作为极限学习机参数 α,β,b，极限学习机根据参数 α,β,b 进行训练，计算网络流量的预测值与实际值之间的偏差，并作为粒子群算法的适应度函数，适应度函数如下：

$$F_{fit}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{10}$$

其中，$\hat{y}_i$ 表示第 i 个网络流量训练样本的预测值，N 表示网络流量训练样本的数量；

(5)采用适应度函数值对粒子位置的优劣进行评价，将每一个粒子位置与当前个体最优位置和粒子群最优位置进行比较，如果粒子位置优于它们的位置，就用粒子位置替换当前个体最优位置和粒子群最优位置，这就使得粒子群朝着参数最优组合方向搜索；

(6)更新粒子的位置和速度；

(7)将迭代次数增加到 $t=t+1$；

(8)如果迭代次数超过最初设置的最大迭代次数，那么根据粒子群的最优位置向量得到极限学习机的最优参数，否则就返回步骤(5)继续执行，具体如图 1 所示。

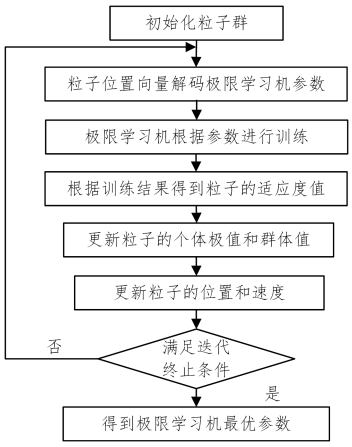

图 1 粒子群算法优化极限学习机参数的流程

Fig. 1 Process of optimizing parameters of extreme learning machine by particle swarm algorithm

3.4 面向海量数据的网络流量混沌预测步骤

(1)针对一个具体的网络服务器端口，采用相关设备采集网络流量历史数据，组成一个时间序列数据。

(2)受到多种因素的影响，采集的原始网络流量历史数据可能存在错误的数据点，因此通过加权平均来替换错误的数据点，并采用式(11)对其进行归一化处理。

$$z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{11}$$

(3)设置小波分析的分解层次，采用小波分析对网络流量时间序列进行分解，得到高频的网络流量子序列和低频的网络流量子序列。

(4)选择互信息法和虚假邻近点法分别确定网络流量各子序列的 τ 和 m。

(5)根据 τ 和 m 对网络流量子序列进行相空间重构，得到相应的网络流量子序列学习样本。

(6)对于每一个网络流量的各子序列，用极限学习机对其进行建模与预测，得到各自的预测结果。

(7)通过小波分析的重构将各网络流量子序列的预测结果进行叠加，得到网络流量的预测结果，具体流程如图 2 所示。

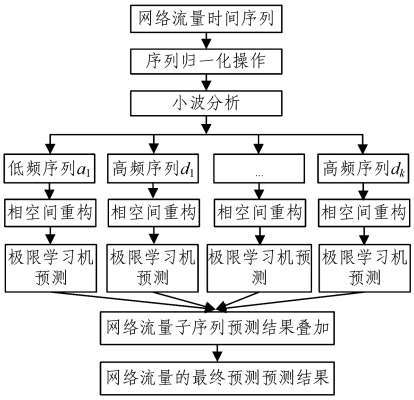

图 2 面向海量数据的网络流量混沌预测模型流程

Fig. 2 Flow chart of chaotic network traffic prediction model for massive data

4 网络流量预测的实例分析

4.1 网络流量的数据

为了分析面向海量数据的网络流量混沌预测模型(WA-PHR-ELM)的性能，本文选择了一个网络流量历史数据作为研究对象，共得到 50000 个样本数据，为了使仿真实验结果更加可信，进行了 5 次仿真实验，每一次实验选择不同比例的样本数据组成预测检验样本集合，其余部分作为训练样本集合，具体如表 1 所列。

表 1 5 次仿真实验的训练样本和测试样本数量分布

Table 1 Distribution of training samples and test samples in 5 simulation experiments

实验编号	训练样本数量	测试样本数量
1	10000	4000
2	15000	4500
3	20000	5000
4	30000	6000
5	40000	10000

为了使本文模型的网络流量预测结果具有可比性，本文选择了 4 种模型进行对比实验，具体设计如下。

(1)小波分析-支持向量机(WA-SVM)。该模型与本文建模的思想相同，但是其网络流量预测算法采用支持向量机。

(2)小波分析-BP 神经网络(WA-BPNN)。该模型与本文建模的思想相同，但是其网络流量预测算法采用 BP 神经网络。

(3)相空间重构-极限学习机(PHR-ELM)。该模型对归一化后的网络流量直接进行混沌分析和相空间重构，其网络流量预测算法采用极限学习机，没有对网络流量序列进行小波分析的细化处理。

(4)小波分析-极限学习机(WA-ELM)。该模型对网络流量进行了小波分析和混沌分析，其网络流量预测算法采用极限学习机，但是其极限学习机参数 α，β，b 采用随机方式确定。

4.2 网络流量历史数据的混沌特性识别

采用小数据量法计算得到网络流量历史数据的最大 Lyapunov 指数，具体如图 3 所示。由图 3 可知，横坐标 i 表示离散步时间，$y(i)$ 表示指数发散率，采用小数据量法计算出该网络流量历史数据的最大 Lyapunov 指数为 0.091 53>0，这表明本文收集的网络流量历史数据具有一定的混沌特性。

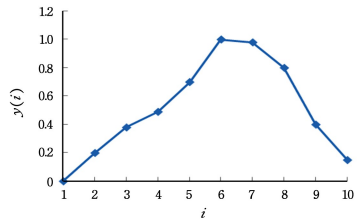

图 3 用小数据量法计算最大 Lyapunov 指数

Fig. 3 Calculating the maximum Lyapunov exponent with small amount of data

4.3 确定网络流量历史数据的延迟时间和嵌入维数

由于网络流量历史数据具有混沌性，本文采用相关系数法来确定网络流量历史数据的延迟时间，结果如图 4 所示。由图 4 可知，延迟时间 $\tau=6$。利用虚假邻近点法确定网络流

量历史数据的维数 m，结果如图 5 所示。由图 5 可知，随着 m 不断增加，虚假最邻近点的比例相应地增加。当 $m=6$ 时，虚假最邻近点的比例不再减少，因此可以确定网络流量历史数据的最优维数 $m=6$。本文将确定的延迟时间 $\tau=6$ 和嵌入维数 $m=6$ 用于 PHR-ELM 的网络流量预测建模。

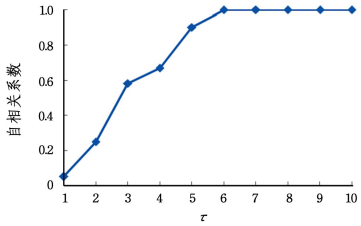

图 4 网络流量历史数据的延迟时间确定

Fig. 4 Determination of delay time of network traffic history data

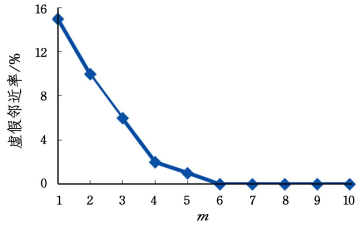

图 5 网络流量历史数据的维数确定

Fig. 5 Determination of dimension of network traffic history data

本文采用 Mallat 算法对表 1 所列的网络流量序列进行分解，共进行了 4 个层次的分解，网络流量的子序列分别为低频序列 a_4 和高频序列 d_4，d_3，d_2，d_1，其中，a_4 子序列描述网络流量的整体变化趋势，d_4，d_3 子序列描述网络流量的周期性变化规律，d_2，d_1 子序列描述网络流量的随机性、非线性变化特点。利用相关系数法、虚假邻近点法确定子序列的延迟时间和维数 m，结果如表 2 所列。

表 2 网络流量子序列的延迟时间和嵌入维数

Table 2 Delay time and embedding dimension of network traffic subsequence

网络流量子序列的名称	τ	m
a_4	2	7
d_4	4	6
d_3	3	6
d_2	1	3
d_1	2	5

4.4 与其他极限学习机的网络流量预测性能的对比

采用 WA-PHR-ELM，PHR-ELM，WA-ELM 对表 1 列出的网络流量数据进行建模与预测，统计它们的网络流量的预测精度，结果如图 6 所示。由图 6 可知，相比其他极限学习机的网络流量预测模型，PHR-ELM，WA-ELM，WA-PHR-ELM 的网络流量预测精度明显提升，这是因为 PHR-ELM 和 WA-ELM 只单独考虑网络流量的混沌性或多特征性，没有同时考虑网络流量的混沌性和多特征性，无法准确对网络流量变化特性进行全面的建模与描述，所以对比模型的网络流量预测误差较大。而 WA-PHR-ELM 不仅可以更全面、准确地描述网络流量变化特性，还引入了粒子群算法确定了极限学习机参数 α，β，b 值，改善了网络流量的预测效果，因此 WA-PHR-ELM 是一种行之有效的网络流量预测模型。

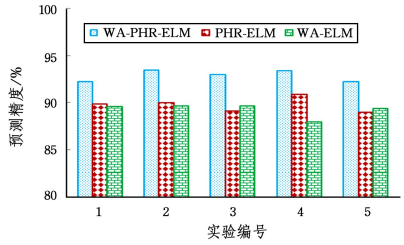

图 6 与其他极限学习机的网络流量预测精度的对比

Fig. 6 Comparison of network traffic prediction accuracy with other extreme learning machines

4.5 与典型模型的网络流量预测性能的对比

将 WA-PHR-ELM 与当前的经典网络流量预测模型 WA-SVM，WA-BPNN 进行对比，网络流量预测精度如图 7 所示。图 7 给出了网络流量预测精度，其中，WA-PHR-ELM 的网络流量数据超过 90%，相比 WA-SVM 和 WA-BPNN，WA-PHR-ELM 的网络流量预测精度得到了不同程度的提升，这是因为 WA-PHR-ELM 采用极限学习机对网络流量进行训练，更好地描述了网络流量的变化态势，打破了支持向量机和神经网络的局限性，获得了理想的网络流量预测效果，预测结果具有较好的通用性与稳定性，这验证了 WA-PHR-ELM 的优越性。

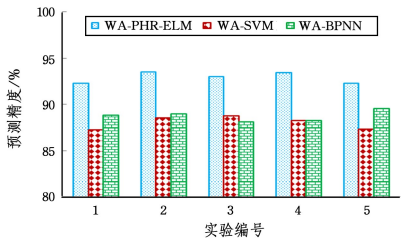

图 7 与典型模型的网络流量预测精度的对比

Fig. 7 Comparison of network traffic prediction accuracy with typical models

结束语 网络流量具有比较明显的时变性和混沌性，针对当前网络流量预测误差大的缺陷，本文提出了基于小波分析、混沌分析和极限学习机相融合的网络流量预测模型。首先该模型采用小波分析对原始网络流量数据序列进行多尺度分解，得到不同变化特征的网络流量子序列，然后对每一个网络流量的子序列进行混沌分析和重构，并采用极限学习机对重构后的网络流量的子序列进行建模与预测，最后应用实例的预测结果表明，本文模型具有预测精度高、通用性强、预测结果稳定等优点，为时变性的网络流量预测提供了一种新的建模思路。

网络流量与网民的上网习惯、网络上业务类型等多种因素相关，具有一定的周期性变化特点，同时具有随机性、混沌性等时变性变化特点，是一种具有多重变化特征的数据序列。因此采用单一建模方法无法全面、准确地描述其变化特点，使得网络流量的预测结果不准确。本文引入小波分析能够从网络流量序列中提取不同的变化特征，从而对网络流量进行更精确的建模与预测，大幅降低了网络流量的预测误差。

人工神经网络等传统机器学习算法存在收敛速度慢等缺

陷，因此本文引入极限学习机对网络流量进行建模，并引入粒子群算法解决极限学习机参数优化难的问题，减少了人为确定参数的不利影响。结果表明，极限学习机的网络流量预测值与实测值十分吻合，具有明显的优越性。

本文模型以网络流量一维时间序列为基础，利用混沌理论对网络流量时间序列进行重构，以解决随机性、非平稳性对网络流量建模与预测的干扰。在没有考虑外界因素的条件下，能够准确预测出网络流量的变化趋势，且网络流量的预测效果得到了明显改善，简化了网络流量的建模过程，为网络流量预测提供了一种新的思路，具有十分广泛的应用前景。

在实际应用中，网络流量与诸多影响因素有关，但本文模型没有考虑这些影响因素对网络流量的影响，这使得本文模型的网络流量预测精度的提升程度有限。因此下一步工作将深入考虑网络流量预测模型的各种影响因素，使网络流量预测结果的可解释性更好，网络流量的预测模型具有更明确的物理意义。

参 考 文 献

[1] VELAN P, CERMAK M, CELEDA P, et al. A survey of methods for encrypted traffic classification and analysis[J]. International Journal of Network Management, 2015, 25(5): 355-374.

[2] LI W X, QI H, XU R H, et al. Research progress and trend of data center network traffic scheduling [J]. Chinese Journal of Computers, 2020, 43(4): 600-617.

[3] LV N, ZHOU J X, FENG X, et al. A time enhanced airborne network traffic identification method [J]. Journal of Northwest Polytechnic University, 2020, 38(2): 341-350.

[4] LOTFOLLAHI M, SIAVOSHANI M J, ZADE R S H, et al. Deep packet: A novel approach for encrypted traffic classification using deep learning [J]. Soft Computing, 2017, 28(9): 1-14.

[5] ZHANG J, BAI G W, SHA X L, et al. Mobile network traffic prediction model based on spatiotemporal characteristics [J]. Computer Science, 2019, 46(12): 108-113.

[6] GUO J, YU Y B, YANG C Y. Multi step network traffic prediction based on total attention mechanism [J]. Signal Processing, 2019, 35(5): 758-767.

[7] LI S, ZHOU Y T, CHI Y, et al. Application of Gaussian process mixture model to network traffic prediction [J]. Computer Engineering and Applications, 2020, 56(5): 186-193.

[8] LI X L, WU T. Efficient network traffic prediction method based on PF-LSTM network [J]. Computer Application Research, 2019, 36(12): 3833-3836.

[9] ZHANG C, PAUL P. Long-term mobile traffic forecasting using deep spatio-temporal neural networks[C]// Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing. Angeles: ACM, 2018: 231-240.

[10] ZHAO J H, WANG M X, QU H, et al. A traffic prediction algorithm for satellite networks based on adaptive klms [J]. Journal of Beijing University of Posts and Telecommunications, 2018, 41(3): 51-55.

[11] NAREJO S, PASERO E. An Application of Internet Traffic Prediction with Deep Neural Network [J]. Multidisciplinary Approaches to Neural Computing, 2018, 69(1): 139-149.

[12] HAN Y, JING Y W, JIN J Y, et al. Short term prediction of network traffic based on improved black hole algorithm optimized ESN [J]. Journal of Northeast University (Natural Science Edition), 2018, 39(3): 311-315.

[13] TIAN Z D, LI S J, WANG Y H, et al. Network traffic prediction based on Gaussian process regression compensation Arima [J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40(6): 65-73.

[14] CHEN Z, LIU Z, PENG L, et al. A novel semi-supervised learning method for Internet application identification [J]. Soft Computing, 2017, 21(8): 1963-1975.

[15] CHEN X, TANG J Y. Internet of Things traffic prediction model based on Bayesian and causal ridge regression [J]. Journal of Sichuan University (Natural Science Edition), 2018, 55(5): 965-970.

[16] LONG Z Y, AI J Q, ZOU H, et al. Network traffic prediction model based on improved gray wolf optimization algorithm [J]. Computer Application Research, 2018, 35(6): 1845-1848.

XIANG Chang-sheng, born in 1971, Ph.D, associate professor. His main research interests include artificial intelligence, data mining and machine learning.

CHEN Zhi-gang, born in 1964, professor, Ph.D supervisor, is a senior member of China Computer Federation. His main research interests include cluster computing, computer security, wireless networks, parallel and distributed system, etc.