

## 基于高斯过程回归的网络流量预测模型

李振刚\*

(天津城建大学 信息中心, 天津 300384)

(\* 通信作者电子邮箱 lzhg@tcu.edu.cn)

**摘要:** 针对传统网络流量预测精度低难题, 为了获得理想的网络流量预测结果, 提出一种基于高斯过程回归 (GPR) 的网络流量预测模型。该模型首先计算延迟时间和嵌入维数, 构建高斯过程回归的学习样本; 然后采用高斯过程回归对网络流训练集进行学习, 并采用入侵杂草优化对高斯过程回归的参数进行优化; 最后采用经典的网络流量测试集对该模型性能进行实验测试。实验结果表明, 高斯过程回归模型提高了网络流量的预测精度。

**关键词:** 网络流量; 高斯过程回归; 入侵杂草优化; 延迟时间; 嵌入维数

**中图分类号:** TP301.6 **文献标志码:** A

### Network traffic forecasting model based on Gaussian process regression

LI Zhengang\*

(Information Center, Tianjin Chengjian University, Tianjin 300384, China)

**Abstract:** To solve the defect of traditional network traffic prediction forecasting, and obtain good forecasting results of network traffic, a network traffic forecasting model based on Gaussian Process Regression (GPR) was proposed. Firstly, the time delay and embedding dimension of network traffic were calculated to construct the learning samples of GPR, and then training samples were input to Gaussian process to learn in which Invasive Weed Optimization (IWO) algorithm was used to optimize the parameters of Gaussian process, and finally, the forecasting model of network traffic was established based on the optimal parameters, and the performance was tested by network traffic data. The results show that the proposed model can improve the forecasting precision of network traffic and it has great practical application value.

**Key words:** network traffic; Gaussian Process Regression (GPR); Invasive Weed Optimization (IWO); delay time; embedding dimension

近年来随着网络业务多样化, 网络越来越拥挤, 同时用户对服务质量要求也相应提高, 提高预测精度成为网络领域中的研究重点和难点<sup>[1]</sup>。

针对网络流量随机性、突变性以及混沌性, 一些学者将非线性理论和混沌理论应用于网络流量预测, 提出基于贝叶斯网络、支持向量机、灰色理论、神经网络、相关向量机等网络流量预测模型<sup>[2-5]</sup>, 尤其是神经网络具有优异的非线性预测能力, 成为当前主要研究方向<sup>[6]</sup>。但是神经网络自身存在许多难以克服的缺陷, 如网络结构的构造要求训练样本数量大, 在小样本条件下, 预测结果不稳定, 易出现“过拟合”现象<sup>[7]</sup>。高斯过程回归 (Gaussian Process Regression, GPR) 是一种基于贝叶斯网络的新型机器学习算法, 不仅具有贝叶斯网络推理能力, 可解释性强, 同时具有了支持向量机的小样本、非线性、高维等问题的自适应处理能力, 是机器学习领域的研究热点<sup>[8]</sup>。大量研究实践表明, 相对于支持向量机和神经网络, GPR 模型具有易实现、泛化能力更强等优点, 可以获得较好的建模性能<sup>[9]</sup>。然而, GPR 在实际应用中, 参数对其性能影响至关重要, 目前常采用共轭梯度法确定最优参数, 但共轭梯度法存在对初始值敏感、易陷入局部最优等弊端, 对 GPR 预测性能产生不利影响<sup>[10]</sup>。

为了提高网络流量的预测精度, 提出一种高斯过程回归的网络流量预测模型 (Invasive Weed Optimization-Gaussian

process Regression, IWO-GPR), 采用入侵杂草优化 (Invasive Weed Optimization, IWO) 对参数进行优化, 并采用经典的网络流量测试集来验证 IWO-GPR 的性能。

### 1 高斯过程回归模型

#### 1.1 高斯过程回归算法

给定训练集  $D: \{X_i, t_i\}_{i=1}^N$ ,  $N$  为训练样本数,  $X_i$  为对应  $t_i$  时刻的向量, 高斯过程的随机变量联合概率分布函数为:

$$P(t | C(X_m, X_n; \Theta), \{X_n\}) = \frac{1}{Z} \exp\left(-\frac{1}{2}(t - \mu)^T \cdot C(X_m, X_n; \Theta)^{-1}(t - \mu)\right) \quad (1)$$

其中:  $X_m$  和  $X_n$  表示第  $m$  和  $n$  个向量;  $C(X_m, X_n; \Theta)$  为参数协方差函数;  $\mu$  为均值向量。

目前使用最广泛的协方差函数为:

$$C(X_m, X_n; \Theta) = \theta_1 \exp\left(-\frac{1}{2} \sum_{l=1}^L \frac{(X_m^{(l)} - X_n^{(l)})^2}{\delta^2}\right) + \theta_2 \quad (2)$$

其中:  $L$  为向量  $X_i$  中元素个数;  $\delta$  为方差;  $\theta_1$  和  $\theta_2$  为参数<sup>[11]</sup>。

对于一个新输入向量  $X_{N+1}$ , 采用预测模型对其  $t_{N+1}$  的概率分布、期望和方差进行预测, 设模型函数表示形式为:

$$t_n = y(X_n) + v_n \quad (3)$$

其中:  $y(X_n)$  是模型函数;  $v_n$  是预测误差。

那么模型概率为:

$$P(T_N | \{X_n\}, A, B) = \int P(T_N | \{X_n\}, y, v) P(y | A) \cdot P(v | B) dy dv \quad (4)$$

其中:  $P(y | A)$  为  $y(x)$  的事先概率分布;  $A$  是  $P(y | A)$  的一组超参数;  $P(v | B)$  为预测误差的事先概率分布;  $B$  是表示误差  $v$  的参数。

令  $T_N = (t_1, t_2, \dots, t_N)$ ,  $T_{N+1} = (t_1, t_2, \dots, t_N, t_{N+1})$ , 那么  $t_{N+1}$  条件分布可以采用式(5)表示, 并通过它对  $t_{N+1}$  进行预测:

$$P(t_{N+1} | D, A, B, X_{N+1}) = \frac{P(T_{N+1} | \{X_n\}, A, B, X_{N+1})}{P(T_N | \{X_n\}, A, B)} \quad (5)$$

根据贝叶斯法可推断出  $t_{N+1}$  的分布函数为:

$$P(t_{N+1} | D, C(X_n, X_m; \Theta), X_{N+1}, \Theta) = \frac{P(T_{N+1} | C(X_n, X_m; \Theta), \Theta, X_{N+1}, \{X_n\})}{P(T_N | C(X_n, X_m; \Theta), \Theta, \{X_n\})} \quad (6)$$

代入高斯分布公式, 初步计算可以简化为:

$$P(t_{N+1} | D, C(X_n, X_m; \Theta), X_{N+1}, \Theta) = \frac{Z_N}{Z_{N+1}} \exp \left[ -\frac{1}{2} (t_{N+1}^T C_{N+1}^{-1} t_{N+1} - t_N^T C_N^{-1} t_N) \right] \quad (7)$$

其中:  $Z_N$  和  $Z_{N+1}$  是两个正规化常数。

经过进一步的简化, 最终可以得到下式:

$$P(t_{N+1} | D, C(X_n, X_m; \Theta), \Theta, X_{N+1}) = \frac{1}{Z} \exp \left( -\frac{(t_{N+1} - \hat{t}_{N+1})^2}{2\sigma_{N+1}^2} \right) \quad (8)$$

其中:  $\hat{t}_{N+1}$  为期望值。

因此, 根据其期望和方差可以进行预测。

## 1.2 高斯过程的超参数优化

高斯过程回归模型的参数  $\theta_1$  和  $\theta_2$  对预测结果影响甚大, 当前最优参数常采用共轭梯度法确定, 但共轭梯度法存在自身难以克服的缺陷, 难以获得全局最优的超参数, 不能建立整体性能最优的预测模型。入侵杂草优化(IWO)是一种模拟杂草种子生长、繁殖和竞争的群智能算法, 搜索能力强, 可以找到问题的全局最优解, 因此本文采用 IWO 对高斯过程回归参数  $\theta_1$  和  $\theta_2$  进行优化。

### 1.2.1 入侵杂草优化

入侵杂草优化(IWO)算法是一种模拟杂草入侵种子的空间扩散、生长、繁殖和竞争等过程的群智能优化算法。其执行步骤为:

- 1) 初始化种群。将一组初始解随机地散布在问题空间中。
- 2) 生长繁殖。每个杂草种子生长, 并根据其适应性(繁殖能力)产生种子。
- 3) 空间扩散。产生的种子随机地散布在整个搜索区域, 长成新种杂草。
- 4) 重复步骤2)、3), 直到杂草种子的最大数。
- 5) 竞争性生存。只有较好适应性的杂草个体能生存并产生种子, 其他则消亡。
- 6) 重复步骤2)~6), 直到最大代数为止<sup>[12]</sup>。

### 1.2.2 入侵杂草优化优化高斯过程回归参数

- 1) 初始 IWO 算法参数, 根据相关研究设置高斯过程回归参数  $\theta_1$  和  $\theta_2$  的范围。
- 2) 随机产生  $m$  粒种子, 每一个粒种子由  $\theta_1$  和  $\theta_2$  两部分组成。
- 3) 对每一粒种子进行解码, 得到高斯过程回归  $\theta_1$  和  $\theta_2$

值, 然后采用 GPR 对训练集进行学习, 并根据 10 折交叉验证得到作为每一粒种子适应度值。适应度函数定义如下:

$$f = \sum_{i=1}^n \frac{|\hat{x}_i - x_i|}{x_i} \times 100\% \quad (9)$$

其中:  $x_i$  表示第  $i$  个样本实际值;  $\hat{x}_i$  表示第  $i$  个样本预测值。

4) 杂草种子根据其适应度值来产生新的种子, 具体为:

$$w_n = \frac{f - f_{\min}}{f_{\max} - f_{\min}} (s_{\max} - s_{\min}) + s_{\min} \quad (10)$$

其中:  $f$  适应度值;  $f_{\max}$ 、 $f_{\min}$  分别为当前杂草的最大和最小适应度值;  $s_{\max}$ 、 $s_{\min}$  分别一个杂草能产生最大和最小值种子数。

5) 将种子按  $N(0, \sigma_i)$  分布在杂草周围; 并根据式(11)更新  $\sigma_i$  的值, 并将产生的种子添加到解集中。

$$\sigma_{\text{iter}} = \frac{(\text{iter}_{\max} - \text{iter})^n}{(\text{iter}_{\max})^n} (\sigma_{\text{initial}} - \sigma_{\text{final}}) + \sigma_{\text{final}} \quad (11)$$

其中:  $\text{iter}_{\max}$  是最大代数,  $\sigma_{\text{iter}}$  是当代标准差,  $\sigma_{\text{initial}}$ 、 $\sigma_{\text{final}}$  分别是标准差的初值和终值;  $n$  是非线性调和指数。

6) 按照适应度值, 对种群中种子排序, 选择优秀种子。

7) 如果迭代次数大于最大迭代数, 则终止算法。

8) 对最优种子进行解码, 得到最优高斯过程回归参数  $\theta_1$  和  $\theta_2$  值。

## 2 高斯过程回归的网络流模型

1) 收集若干网络流量的历史时间序列数据, 得到  $X_i (i = 1, 2, \dots)$ 。

2) 由于网络流量数值变化较大时, 对高斯过程学习产生不利影响, 因此对网络流量进行归一化处理, 具体为:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (12)$$

式中:  $x_{\max}$  和  $x_{\min}$  为最大值和最小值;  $x_i$  为原始值。

3) 由于网络流量数据常具有混沌特性, 因此采用自相关法和假近邻法分别计算网络流量的延迟时间和嵌入维数, 并对网络流量数据进行相空间重构, 构建高斯过程的学习样本, 并将其分为训练和测试集。

4) 将训练集输入到高斯过程回归中进行学习, 并采用 IWO 对参数  $\theta_1$  和  $\theta_2$  进行优化, 建立网络流量预测模型。

5) 对网络流量的预测精度和预测误差进行分析。

## 3 网络流量预测的实例分析

### 3.1 数据来源

为了测试 IWO-GPR 模型的网络流量预测性能, 数据采用网络流量文库 (<http://newsfeed.ntcu.net/~news/2013/>) 主节点路由器 2013 年 7 月 1 日到 7 月 30 日的每小时访问流量, 得到 720 个样本, 具体如图 1 所示。仿真程序使用 Java 语言编写, 在 JDK1.4.2 环境下运行, 前 620 个样本作为训练集, 其余样本作为测试集。

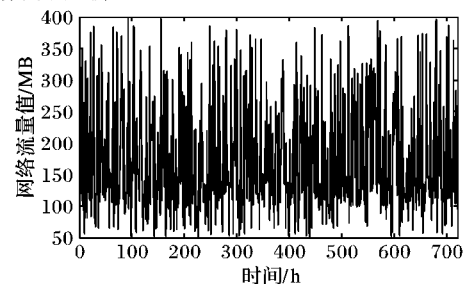


图1 网络流量

### 3.2 对比模型及评价标准

为了使 IWO-GPR 的预测结果具有可比性,选择共轭梯度法的 GRP(Gradient Ronjugate Particle, GRP)、支持向量机(Support Vector Machine, SVM)和 RBF 神经网络(RBF Neural Network, RBFNN)作为对比模型,模型性能采用均方根误差(Root Mean Square Error, RMSE)和平均相对百分比误差(Average Relative Percentage Error, MPAE)进行衡量,它们定义为:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{x}(i) - x(i))^2} \quad (13)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{x}(i) - x(i)}{x(i)} \right| \times 100\% \quad (14)$$

其中: $n$  表示样本数。

### 3.3 学习样本的构造

网络流量的自相关函数变化曲线如图 2 所示,从图 2 可知,网络流量的最佳延迟时间  $\tau = 5$ 。虚假近邻数和嵌入维数之间的变化关系如图 3 所示,从图 3 可知,随着嵌入维数的增加,虚假近邻数逐渐变小,当  $m = 5$  时,虚假近邻数不再变化,即网络流量的最优嵌入维数  $m = 5$ ,采用  $\tau = 5, m = 5$  对网络流量数据重构。

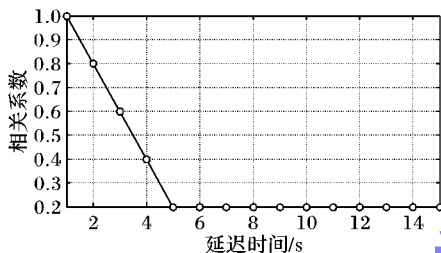


图2 网络流量的延迟时间( $\tau$ )

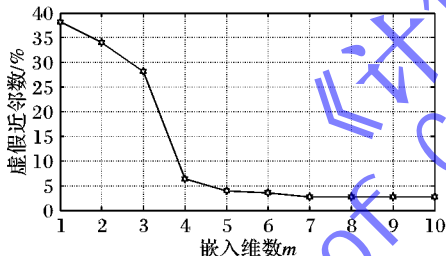


图3 网络流量的嵌入维数

### 3.4 结果与分析

#### 3.4.1 网络流量的预测结果

对于图1的网络流量, IWO-GPR 的预测结果和预测误差变化曲线分别如图4和5所示。从图4可知, IWO-GPR 可以较好地跟踪网络流量变化趋势, 从而使网络流量的实际值与预测比较接近, 偏差较小; 同时从图5可知, IWO-GPR 预测误差小, 预测误差变化范围波动小。实验结果表明, IWO-GPR 是一预测精度高、有效的网络流量预测模型。

GPR、IWO-GPR、SVM、RBFNN 的预测误差见表1。对表1结果进行分析, 可以得到如下结论:

1) 相对于 SVM、RBFNN, GPR、IWO-GPR 的预测误差均有所下降, 预测精度得以提高, 这表明 GPR 模型较好地克服了神经网络和支持向量机的缺陷, 可以建立性能更优的网络流量预测模型。

2) 相对于 GPR, IWO-GPR 预测误差大幅度下降, 这主要是由于采用 IWO 算法对 GRP 参数  $\theta_1$  和  $\theta_2$  进行优化, 获得了全局最优的参数, 较好地克服了共轭梯度法存在的初始值敏感、易陷入局最优等弊端, 提高了网络流量的预测精度。

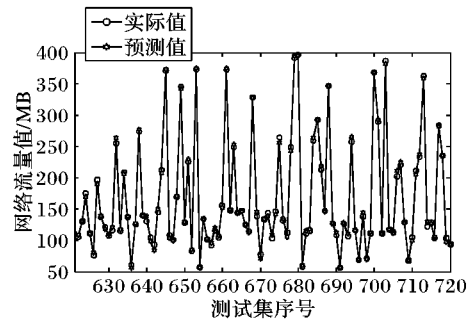


图4 IWO-GPR 的预测结果

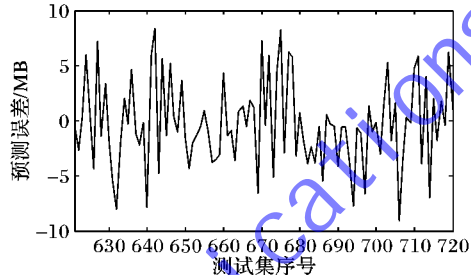


图5 IWO-GPR 的预测误差变化曲线

表1 不同模型的预测误差对比

模型	RMSE	MAPE/%	模型	RMSE	MAPE/%
RBFNN	12.14	8.83	GPR	5.992	3.74
SVM	10.00	7.54	IWO-GP	3.875	2.25

#### 3.4.2 含有噪的网络流量预测性能分析

为了测试 IWO-GPR 的适应性, 采用一个含噪网络流量数据进行测试实验, 具体如图6所示。最后100个样本作为测试集, 其余为训练集, 网络流量的最佳延迟时间和嵌入维数求解曲线如图7和8所示。从图7和8可知, 含噪网络流量的最优延迟时间和嵌入维数分别为  $\tau = 8, m = 6$ , 然后对含噪网络流量重构, IWO-GPR 的网络流量预测结果和预测误差如图9和10所示。从图9和10可知, 对于含噪网络流量, IWO-GPR 获得了较好的预测结果, 预测误差控制在有效的范围内, 这主要是 IWO-GPR 不仅利用 IWO 全局、局部的搜索能力, 而且利用 GRP 的强大非线性预测能力, 抗干扰能力强, 具有较强的适应性。

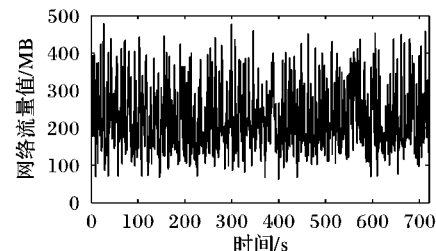


图6 含噪的网络流量

## 4 结语

针对网络流量的复杂性, 提出一种高斯过程回归的网络流量预测模型。首先根据网络流量混沌性构建高斯过程的学习样本, 然后将搜索能力强的人侵杂草优化引入到高斯过程回归参数优化中, 克服共轭梯度法存在的不足, 最后采用网络流量预测实例对模型性能进行分析。结果表明, IWO-GPR 可以对网络流量进行准确预测, 预测结果以为网络管理员提供有益参考意见, 降低网络拥塞频率, 具有较强的实用价值。

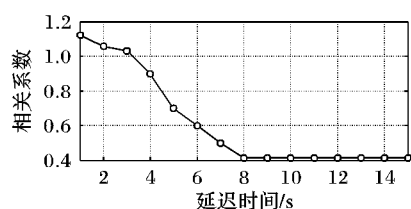


图7 含噪网络流量的延迟时间

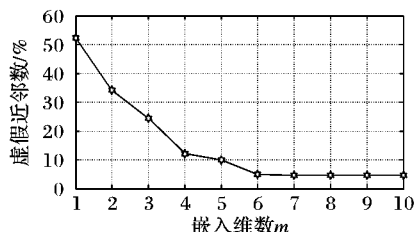


图8 含噪网络流量的嵌入维数

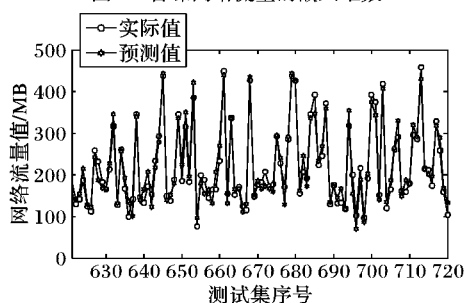


图9 IWO-GPR的含噪网络流量预测结果

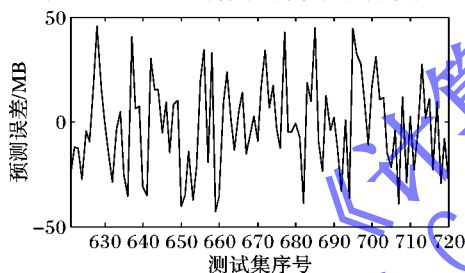


图10 含噪网络流量 IWO-GPR 的预测误差变化曲线

## 参考文献:

- [1] WEN X, MENG X, MA Z, *et al.* The chaotic analysis and trend prediction on small-time scale network traffic [J]. *Acta Electronica Sinica*, 2012, 40(8): 1609–1616. (温祥西, 孟相如, 马志强, 等. 小时间尺度网络流量混沌性分析及趋势预测[J]. *电子学报*, 2012, 40(8): 1609–1616.)
- [2] CHANG B R, TSAI H F. 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.
- [3] LUO Y, XIA J, WANG H. Application of chaos-support vector machine regression in traffic prediction [J]. *Computer Science*, 2009, 36(7): 244–246. (罗赞, 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.)
- [4] YAO Q, LI C, MA H, *et al.* Novel network traffic forecasting algorithm based on grey model and Markov chain [J]. *Journal of Zhejiang University: Sciences Edition*, 2007, 34(4): 396–400. (姚奇富, 李翠凤, 马华林, 等. 灰色系统理论和马尔可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.)
- [5] WANG J, GAO Z. Network traffic modeling and prediction based on RBF neural network [J]. *Computer Engineering and Applications*, 2008, 44(13): 6–7. (王俊松, 高志伟. 基于 RBF 神经网络的网络流量建模及预测[J]. *计算机工程与应用*, 2008, 44(13): 6–7.)
- [6] CHEN Y, YANG B, MENG Q. Small-time scale network traffic prediction based on flexible neural tree [J]. *Applied Soft Computing*, 2012, 12(1): 274–279.
- [7] XIONG N, LIU B. Online prediction of network traffic based on adaptive particle swarm optimisation and LSSVM [J]. *Computer Applications and Software*, 2013, 30(9): 21–24. (熊南, 刘百芬. 基于自适应粒子群优化 LSSVM 的网络流量在线预测[J]. *计算机应用与软件*, 2013, 30(9): 21–24.)
- [8] ZHOU X, WANG W, CHEN W. Network traffic prediction model based on wavelet transform and optimised support vector machine [J]. *Computer Applications and Software*, 2011, 28(2): 34–36. (周晓蕾, 王万良, 陈伟杰. 基于小波变换和优化的 SVM 的网络流量预测模型[J]. *计算机应用与软件*, 2011, 28(2): 34–36.)
- [9] LI J, ZHANG Y. Single-step and multiple-step prediction of chaotic time series using Gaussian process model [J]. *Acta Physica Sinica*, 2011, 60(7): 143–152. (李军, 张友鹏. 基于高斯过程的混沌时间序列单步与多步预测[J]. *物理学报*, 2011, 60(7): 143–152.)
- [10] SUN B, YAO H, LIU T. Short-term wind speed forecasting based on Gaussian process regression model [J]. *Proceedings of the Chinese Society for Electrical Engineering*, 2012, 32(29): 104–109. (孙斌, 姚海涛, 刘婷. 基于高斯过程回归的短期风速预测[J]. *中国电机工程学报*, 2012, 32(29): 104–109.)
- [11] SEEGER M. Gaussian processes for machine learning [J]. *International Journal of Neural Systems*, 2004, 14(2): 69–106.
- [12] SONG X, HU C. Discrete invasive weed optimization algorithm for 0/1 knapsack problem [J]. *Computer Engineering and Applications*, 2012, 48(30): 239–242. (宋晓萍, 胡常安. 离散杂草优化算法在 0/1 背包问题中的应用[J]. *计算机工程与应用*, 2012, 48(30): 239–242.)
- [12] ZAHIR T, ARSHAD K, KO Y, *et al.* A downlink power control scheme for interference avoidance in femtocells [C]// *Proceedings of the 2011 7th International Wireless Communications and Mobile Computing Conference*. Piscataway: IEEE Press, 2011: 1222–1226.
- [13] EFFROS M, GOLDSMITH A, LIANG Y. Generalizing capacity: new definitions and capacity theorems for composite channels [J]. *IEEE Transactions on Information Theory*, 2010, 56(7): 3069–3087.
- [14] LEE P, LEE T, JEONG J, *et al.* Interference management in LTE femtocell systems using fractional frequency reuse [C]// *Proceedings of the 2010 12th International Conference on Advanced Communication and Technology*. Piscataway: IEEE Press, 2010, 2: 1047–1051.
- [15] ETSI. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); FDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis (3GPP TR 36.921 version 11.0.0 Release 11) [S/OL]. [2013-08-09]. [http://www.etsi.org/deliver/etsi\\_tr/136900\\_136999/136921/11.00.00\\_60/tr\\_136921v110000p.pdf](http://www.etsi.org/deliver/etsi_tr/136900_136999/136921/11.00.00_60/tr_136921v110000p.pdf).
- [16] ETSI. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); TDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis (3GPP TR 36.922 version 11.0.0 Release 11) [S/OL]. [2013-08-09]. [http://www.etsi.org/deliver/etsi\\_tr/136900\\_136999/136922/11.00.00\\_60/tr\\_136922v110000p.pdf](http://www.etsi.org/deliver/etsi_tr/136900_136999/136922/11.00.00_60/tr_136922v110000p.pdf).

(上接第1242页)