

一种基于 EMD 和 RVM 的自相似网络流量预测模型

柏 骏¹ 夏靖波¹ 赵小欢²

(空军工程大学信息与导航学院 西安 710077)¹ (95034 部队 百色 533616)²

摘 要 针对自相似网络流量提出了一种基于 EMD(经验模态分解)和 RVM(相关向量机)的自相似时间序列预测模型。该模型利用 EMD 将滑动窗口内的小时间尺度网络流量序列分解为多个 IMF(固有模态函数)分量,以去除流量序列长相关性;然后采用 RVM 对其中的高频分量进行拟合,而对低频分量则使用 ARMA 构建预测模型;最后合成各分量的预测结果。实验表明,该模型能准确地预测流量时间序列的幅值及其趋势,与同类型预测方法相比,其预测性能更好。

关键词 网络流量,预测模型,经验模态分解,相关向量机

中图分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.029

Prediction Model of Network Traffic Based on EMD and RVM

BAI Jun¹ XIA Jing-bo¹ ZHAO Xiao-huan²

(Institute of Information and Navigation, Air Force Engineering University, Xi'an 710077, China)¹

(Unit 95034, Baise 533616, China)²

Abstract A prediction model was proposed for self-similar network traffic based on EMD (empirical mode decomposition) and RVM(relevant vectors machine). Firstly, the network traffic in slipping window is decomposed into multiple IMF(intrinsic mode function) using EMD, and then RVM is applied to fit high frequent components while ARMA is used to structure prediction model for low frequent components, lastly, all the components' forecasting result is composed. The experiment indicates that the proposed model can accurately predict traffic time series' amplitude and trend, and compared to other method, achieves higher prediction accuracy than that of other similar prediction methods.

Keywords Network traffic, Prediction model, EMD, RVM

1 引言

互联网流量建模与预测对于研究网络特性、提升运维管理水平以及设计和优化网络协议具有重要意义^[1,2]。网络流量具有普遍的自相似特性,主要表现为长相关性^[3,4],且其自相似程度直接影响网络节点队列大小、等待时延和丢包率等。因此在研究网络流量的建模和预测过程中,必须考虑其自相似性。

研究人员主要从 3 方面入手展开相关研究:一是通过短相关(short range dependent, SRD)模型来拟合长相关性^[5,6],但该类模型对网络流量的预测精度较差;二是提出具有长相关特性的预测模型,如自回归滑动平均(fractional auto-regressive integrated moving average, FARIMA)^[7]、分数布朗运动(fractional Brownian motion, FBM)等,然而该类模型并未显著提高预测准确率,反而增加了复杂度,致使实用性较差;三是将具有自相似性的流量数据转化为短相关数据^[8,9],以降低计算复杂度。研究表明经小波变换或 EMD^[9]分解后的时间序列不具有长相关性,再利用短相关模型或非线性预

测模型进行预测能够取得更好的预测精度。

虽然经过 EMD 分解后得到的 IMF 不再具有自相似性,但其中的高频分量仍表现出明显的非线性特性,若使用线性模型进行预测,精度较低,难以满足实际需要。为此,本文采用非线性预测 RVM 回归模型对高频分量进行预测,提出一种基于 EMD 和 RVM 的网络流量预测模型,以提高自相似网络流量时间序列的预测精度。

2 预测原理及方法

2.1 EMD

EMD 算法基于一项简单假设:任何时间序列都由一系列不同时间尺度的简单固有模态函数(IMF)组成。这些 IMF 满足以下条件:1)信号极值点的数量与过零点的数量必须相等或最多相差一个;2)在任一时间点上,信号的局部最大值与局部最小值定义的包络的均值为零。

EMD 经过一系列分解和筛选处理^[10],可以从原始时间序列中分解出 n 个 IMF 分量,即 C_1, C_2, \dots, C_n 和一个残余分量 r_n ,而原始的时间序列 $X(t)$ 可以表示成它们的和,即:

到稿日期:2014-03-06 返修日期:2014-05-13 本文受国家自然科学基金面上项目(61272486),陕西省科技计划自然科学基金重点项目(2012JZ8005),全军军事学研究生课题(2010XXXX-488)资助。

柏 骏(1985-),男,博士生,主要研究方向为网络管理与信息安全;夏靖波(1963-),男,博士后,教授,博士生导师,主要研究方向为通信网络管理与评估方面;赵小欢(1984-),博士,工程师,主要研究方向为军事信息网络管理与安全。

$$X(t) = \sum_{i=1}^n C_i + r_n \quad (1)$$

文献[9]证明了自相似网络流量经过 EMD 分解后得到的各个 IMF 分量是短相关的。

2.2 RVM

RVM^[11]是一种新的机器学习方法,其决策形式与 SVM 相同,通过引入稀疏贝叶斯学习理论,不仅使其具备了 SVM 避免过学习的优点,且极大地减少了核函数的计算量,还弥补了 SVM 存在的稀疏性不强、计算量大、核函数必须满足 Mercer 条件以及需人为凭经验确定参数等不足。因此,与 SVM 相比,RVM 更稀疏,计算时间更短,更适用于在线分析处理数据。目前 RVM 已被广泛应用于故障预报^[12]、网络流量分类^[13]等领域,取得了良好的效果。

与 SVM 不同,RVM 通过最大化后验概率(MAP)求解相关向量的权重。对于给定的训练样本集 $\{t_n, x_n\}$,RVM 的模型输出定义为

$$t_n = \sum_{n=1}^N \omega_n K(x, x_n) + \omega_0 + \epsilon_n \quad (2)$$

其中, $K(x, x_n)$ 是选用的核函数, $\{\omega_n\}_{n=0}^N$ 代表不同的权重。对于 $t_n = y(x_n, \omega) + \epsilon_n$,假设噪声 ϵ_n 服从均值为0,方差为 σ^2 的高斯分布,则 $\rho(t_n | \omega, \sigma^2) = N(y(x_n, \omega), \sigma^2)$ 。为了避免过学习,为 ω 加上先决条件,使 ω 服从期望值为0的正态分布:

$$p(\omega_i | \alpha_i) = N(\omega_i | 0, \alpha_i^{-1}) \quad (3)$$

于是将对 ω 的求解转化为对 α 的求解,通过不断迭代,其中大部分 α_i 趋近无限大,对应的 ω_i 则为零,少量的 ω_i 稳定趋于有限值,对应的 x_i 即为相关向量(Relevant Vectors)。

2.3 基于 EMD 和 RVM 的自相关时间序列预测模型

本文采用滑动窗口方法,利用 EMD 将滑动窗口内的已知序列进行分解,通过分解的 IMF 分量预测后续序列。

首先确定滑动窗口长度 w 以及预测步长 k ,并利用 EMD 将滑动窗口内的流量时间序列分解为 n 个 IMF 和 1 个残余分量;然后对其中的低频和低频分量分别建立预测模型,最后将各分量预测结果进行叠加合成。预测流程如图 1 所示。

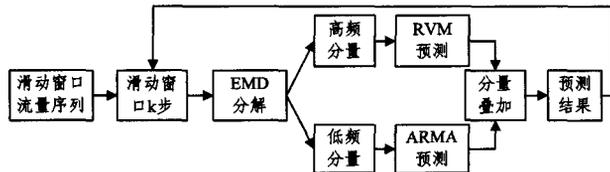


图 1 基于 EMD 和 RVM 的预测方法流程

对窗口内流量序列 $X_i: \{X_1, X_2, \dots, X_w\}$ 进行 EMD 分解,通过各分量的频谱分析发现,高频分量 IMF1 规律性及周期性较弱,难以使用线性预测模型直接预测,故采用 RVM 回归模型进行预测;其余低频分量则采用 ARMA 预测模型。

利用 RVM 对高频分量 IMF1: $\{H_1, H_2, \dots, H_w\}$ 训练预测模型,预测后续分量 $\{H_{w+1}, \dots, H_{w+k}\}$ 。预测步骤如下:

步骤 1 将 IMF1 序列按照如下公式

$$\hat{H}_i = 2 \cdot \frac{H_i - \min(H_i)}{\max(x_i) - \min(x_i)} - 1 \quad (4)$$

归一化处理,有 $\hat{H}_i \in [-1, 1]$ 。

步骤 2 从 X_i 中确定输入变量 $X_{m,l+1} =$

$$\begin{bmatrix} X_1 & X_2 & \dots & X_m \\ X_{1+k} & X_{2+k} & \dots & X_{m+k} \\ \vdots & \vdots & \dots & \vdots \\ X_{1+l \cdot k} & X_{2+l \cdot k} & \dots & X_{m+k \cdot l} \end{bmatrix} \text{ 以及输出变量 } Y_{k,l+1} =$$

$$\begin{bmatrix} X_{m+1} & X_{m+2} & \dots & X_{m+k} \\ X_{m+k+1} & X_{m+k+2} & \dots & X_{m+2k} \\ \vdots & \vdots & \dots & \vdots \\ X_{m+l \cdot k+1} & X_{m+l \cdot k+2} & \dots & X_{m+k \cdot (l+1)} \end{bmatrix}, \text{ 其中 } m+k(l+1) \leq w, \text{ 训练 RVM 预测模型。}$$

步骤 3 以 $X_p = [X_{w-m+1} \ X_{w-m+2} \ \dots \ X_w]$ 作为预测模型的输入,输出 $Y_p = [X_{w+1} \ X_{w+2} \ \dots \ X_{w+k}]$ 作为预测值,并将预测结果逆归一化。

步骤 4 最后将该预测值与 ARMA 预测的各分量预测值叠加,得到窗口流量序列的预测结果,并将窗口向前滑动 k 个时间序列,重复步骤 1。

3 网络流量预测

3.1 预测性能指标

预测效果的分析评价大多采用平均相对误差(MAPE)、均方根误差(RMSE)或归一化均方差(NMSE)等指标来衡量预测性能,其中 $MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - y'_i}{y_i} \right|$, $RMSE =$

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2}, NMSE = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2}{\text{var}(y_i)}$$

。但上述指标仅反映预测值与真实值偏差的绝对值,并不能反映其对序列趋势预测的准确性,难以准确描述网络流量预测中出现的“一步延迟”现象。为此,引入趋势预测准确率(Accuracy of Trend Prediction, ATP)的概念,旨在反映其预测序列趋势的能力。

对于预测真实值 y_i 及其预测值 \hat{y}_i (其中, $i=1, 2, \dots, N$),当 $(y_i - y_{i-1}) * (\hat{y}_i - \hat{y}_{i-1}) \geq 0$ 时,有 $T_i=1$,否则 $T_i=0$,于是有趋势预测准确率

$$APT = \frac{\sum_{i=1}^N T_i}{N} \quad (5)$$

本文采用 MAPE 以及 APT 两个指标对预测算法的性能进行衡量。

3.2 实验与分析

网络流量预测实验所采用的历史流量数据 201007301400.dump 来自于 MAWI 工作组,数据产生时间为 2010 年 7 月 30 日 14:00 至 14:15,900s 内共采集数据包 33900444 个,平均流量为 24.33Mb/s。以 1s 为时间间隔,记录该时间段内的字节数,形成小时间尺度流量时间序列,记为 TTS。

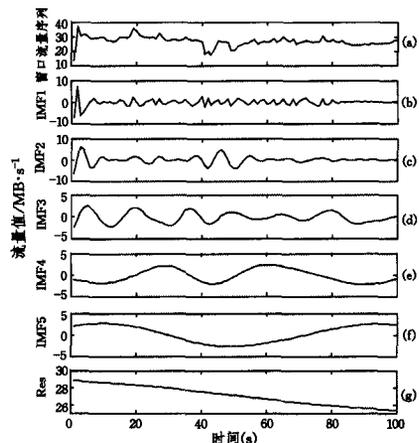


图 2 窗口流量序列及 EMD 分解所得分量

初始化滑动窗口 $w=100$, 预测步长 $k=1$, 将滑动窗口内流量序列用 EMD 进行分解, 如图 2 所示。

对于高频分量 $IMF1$, 利用 RVM 回归模型进行预测。首先需要确定输入变量 X 的维数 (m, l) 。以 $IMF1(1;100)$ 作为训练数据, $IMF1(101;200)$ 作为测试数据, 遍历 $(m+l+1)=100$, 其中 $m>9, l>9$ (m, l 不宜过小)。由图 3 可知, 在 $m>27$ 后, 随着 m 增大, MAPE 呈下降趋势, 在 $m=50$ 左右开始趋于平稳。当 $(m, l)=(48, 51)$ 时, MAPE 最小, 为 5.39%。故选择 $m=48, l=51$ 作为输入变量的维数。

接着, 利用 RVM 对 $[X_{48.52}, Y_{1.52}]$ 进行训练并建立预测模型, 将 $X_p = [IMF1_{53} \quad IMF1_{54} \quad \dots \quad IMF1_{100}]$ 作为预测模型的输入, 输出预测结果 $\hat{IMF1}_{101}$ 。然后, 向前滑动时间窗口, 重复上述步骤依次求得 $\hat{IMF1}_{102}, \hat{IMF1}_{103}, \dots, \hat{IMF1}_{900}$ 。

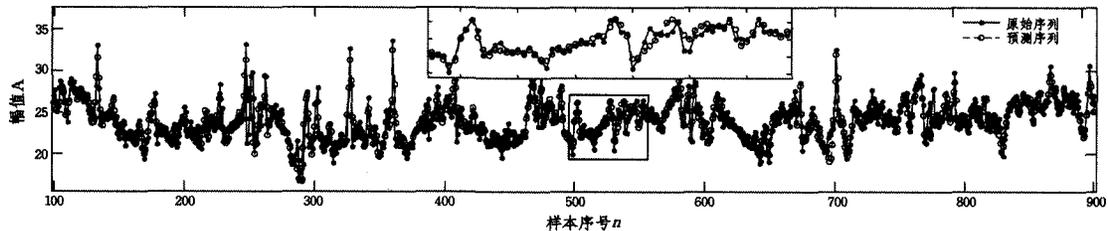


图 4 原始序列及其预测

一步预测需要在 1s 内完成上述步骤, 这对硬件速度, 尤其是在高速网络环境下, 提出了很高的要求, 因此考虑多步预测。按照上述方法进行多步预测, 其 MAPE 及 APT 如图 5 所示。可以发现, MAPE 在 $k=[1, 9]$ 区间线性上升, 随后趋缓, 预测步长越大, 预测误差越大; APT 随着步长的增加呈下降趋势, 下降至 50% 左右后趋于平稳, 但对于趋势预测而言, 50% 的准确率已经失去了预测的意义。

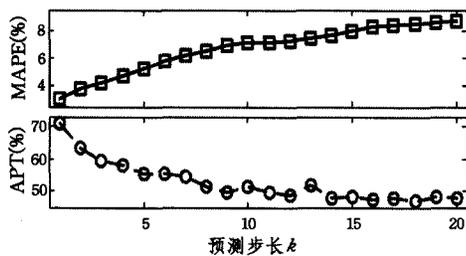


图 5 多步预测

3.3 实验方法比较

对 3 组不同的数据进行了预测, 结果如表 1 所列。可以发现 3 组数据的预测都取得了比较理想的结果, 表明该方法具有较好的预测准确性以及扩展性; 在很多文献中都使用小波与 SVM 或 LS-SVM^[14] 相结合的方法, 本文同样利用窗口滑动的方法将小波+SVM 用于预测上述 3 组数据, 结果表明本文所提方法在 MAPE 上略好于小波+SVM, 在 APT 方面与之不相伯仲; 使用 ARMA 预测, 虽然 MAPE 也能达到 5% 左右, 仅仅稍低于前两种方法, 但其趋势预测概率 APT 则远远低于本文所提方法, 其预测结果在图上直观地表现为“一步延时”。这表明仅仅用 MAPE 衡量预测结果的好坏是片面的, 不足以准确描述预测模型性能, 需结合 APT 共同描述预测结果。这也是本文提出 APT 概念的意义所在。

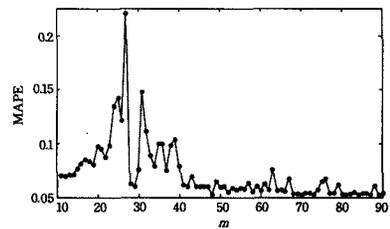


图 3 输入变量维数与 MAPE 的关系

与此同时, 利用 ARMA 对 $IMF2, IMF3, IMF4, IMF5$ 以及 Res 分量按照窗口滑动方式进行一步预测。按照式(1)将上述预测结果相加, 最终预测结果如图 4 所示。从图中可以发现, 该方法的一步预测能较为准确地预测流量时间序列的幅值及其趋势。

表 1 与其他方法的比较

数据	本文方法		小波+SVM		ARMA	
	MAPE	APT	MAPE	APT	MAPE	APT
MAWI ¹	3.07%	71.00%	3.77%	71.87%	5.04%	45.12%
MAWI ²	3.56%	77.17%	4.48%	75.77%	5.81%	49.06%
LBNL	3.91%	80.95%	4.26%	74.88%	4.82%	40.00%

注: MAWI¹ 为前文实验所用数据, MAWI² 数据产生时间为 2010 年 8 月 30 日 14:00 至 14:15, LBNL 数据来自于劳伦斯伯克利国家实验室 (<http://www.icir.org/enterprise-tracing>)。

结束语 为了克服在预测上网络流量的自相似性所带来的困难, 本文采取滑动窗口的方式, 在对已知序列进行 EMD 分解的基础上利用 RVM 和 ARMA 对各分量进行预测。利用 EMD 可有效降低自相似程度, 且 RVM 对非线性高频分量具有很好的拟合效果。实验结果表明该方法对于自相似的网络流量时间序列具有很好的预测能力。此外, 为了反映模型预测序列趋势的能力, 引入了趋势预测准确率 (APT) 的概念, 该指标对于股市、期货预测以及网络拥塞预警等具有很好的指导意义。

参考文献

- [1] 杨双懋, 郭伟, 唐伟. 基于 FARIMA-GARCH 模型的网络业务预测算法[J]. 通信学报, 2013, 34(3): 23-31
- [2] 郭通, 兰巨龙, 李玉峰, 等. 基于量子自适应粒子群优化径向基函数神经网络的网络流量预测[J]. 电子与信息学报, 2013, 35(9): 2220-2226
- [3] Shiravi Y G K A, Min P S. Congestion Prediction of Self-Similar Network through Parameter Estimation[C]// 2006 IEEE/IFIP Network Operations and Management Symposium (S1542-1201). Vancouver, BC: IEEE/IFIP, 2006: 1-4
- [4] 石江涛, 王永纲, 戴雪龙, 等. 自相似网络业务流量的研究与实现[J]. 通信学报, 2005, 26(6): 115-120

[5] JI Q J. Can Multifractal traffic burstiness be approximated by markov modulated poisson processes? [C]// Proceedings 12th IEEE International Conference Networks, 2004 (ICON 2004), 2004; 26-30

[6] 邹柏贤, 姚志强. 一种网络流量平稳化方法[J]. 通信学报, 2004, 25(8): 14-23

[7] Liu J K, Shu Y T, Zhang L F, et al. Traffic modeling based on FARIMA models [C]// IEEE Canadian Conference Electrical and Computer Engineering. Canndian, 1999; 162-167

[8] 高茜, 冯琦, 李广侠. 基于组合模型的自相似业务流量预测[J]. 计算机科学, 2012, 39(4): 123-126

[9] 高波, 张钦宇, 梁永生, 等. 基于 EMD 及 ARMA 的自相似网络流量预测[J]. 通信学报, 2011, 32(4): 47-56

[10] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis [C]// Proceedings of the Royal Society of London, Series A, 1998, 454: 903-995

[11] Tipping M. Sparse Bayesian Learning and the Relevance Vector Machine [J]. Journal of Machine Learning Research, 2001, 1(1): 211-244

[12] 胡昌华, 王兆强, 周志杰, 等. 一种 RVM 模糊模型辨识方法及在故障预报中的应用[J]. 自动化学报, 2011, 37(4): 503-512

[13] 夏靖波, 柏骏, 赵小欢, 等. 基于相关向量机的在线网络流量分类方法[J]. 吉林大学学报: 工学版, 2014, 44(2): 459-464

[14] 王晓兰, 张万宏, 王慧中. 基于小波变换和 AR-LSSVM 的非平稳时间序列预测[J]. 控制与决策, 2008, 23(3): 357-360

(上接第 112 页)

此时的自由度为

$$d_{sum}^{\infty} = d_1 + \min(M_2, N_1 - d_1, N_2 - d_1) \quad (15)$$

比较发现接收端协作并不能提高自由度, 因为 $M_1 > N_1$, 主网络的接收端没有多余的天线, 与 $N_1 < M_1 + M_2$ 情况类似, 并且为了彻底消除干扰, 将抑制次网络发送端发送消息的天线数。

将上述分析结果归纳于表 1, 经过比较发现只有当 $M_1 < N_1$ 或 $N_1 \geq M_1 + M_2$ 时接收端的协作能增加自由度。

表 1 $N_1 \geq M_2$ 时两个 PTP 网络 (M_1, M_2, N_1, N_2) 并存时的自由度

	$N_1 \geq M_1 + M_2$	$d_1 = M_1$ $d_{sum} = M_1 + \min(M_2, N_2 - M_1)$	$d_1 < M_1$ $d_{sum} = d_1 + \min(M_2, [N_2 - d_1]^+)$
	接收端有协作	$d_{sum}^{\infty} = d_1 + \min(M_2, [(N_2 + N_1 - 2d_1)/2]^+)$	
$M_1 \leq N_1$ $d_1 \in [0, M_1]$	$N_1 < M_1 + M_2$	$d_1 = M_1$ $d_{sum} = M_1 + \min(N_1 - M_1, [N_2 - M_1]^+)$	$d_1 < M_1$ $d_{sum} = d_1 + \min(M_2, N_1 - d_1, [N_2 - d_1]^+)$
	接收端有协作	$d_{sum}^{\infty} = d_{sum}$	
$M_1 > N_1, d_1 \in [0, N_1]$		$d_1 = N_1$ $d_{sum} = N_1$	
	接收端有协作	$d_1 < N_1$ $d_{sum} = d_1 + \min(M_2, N_1 - d_1, N_2 - d_1)$	$d_{sum}^{\infty} = d_{sum}$

$N_1 < M_2$ 的情况相当于将 $N_1 \geq M_2$ 情况下的发送端和接收端互换, 自由度推导方法和 $N_1 \geq M_2$ 完全一样, 只要将 $N_1 \geq M_2$ 时对应条件公式中的 M_1, M_2 分别换成 N_1, N_2 即可, 此时 d_{sum}^{∞} 代表在发送端进行协作时的自由度。受版面限制, 具体情况不再赘述。

结束语 本文讨论了两个 PTP 认知网络并存时的自由度, 给出了在主网络的自由度资源完全利用和部分利用两种情况下, 次网络利用剩余的空间资源进行通信时网络自由度的表达式; 此外还分析了两个网络在接收端以部分天线借用的协作方式进行通信时网络的总自由度。比较发现, 当满足一定条件时, 两网络接收端或发送端之间的协作能提高网络的总自由度。

参考文献

[1] Viswanath P, Tse D N C. Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality [J]. IEEE Trans.

Inf. Theory, 2003, 49(8): 1912-1921

[2] Viswanath S, Jindal N, Goldsmith A. Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels [J]. IEEE Trans. Inf. Theory, 2003, 49(10): 2658-2668

[3] Yu W, Cioffi J M. Sum capacity of Gaussian vector broadcast channels [J]. IEEE Trans. Inf. Theory, 2004, 50(9): 1875-1892

[4] Huang Chia-chi, Jafar S A. Degrees of freedom of the MIMO interference channel with cooperation and cognition [J]. IEEE Trans on Information Theory, 2009, 55(9): 4211-4220

[5] Vaze C S, Varanasi M K. A new outer bound via interference localization and degree of freedom regions of MIMO interference networks with no CSIT [J]. IEEE Trans. Inf. Theory, 2012, 58(11): 6853-6869

[6] Zhu Yan, Guo Dong-ning. The degrees of freedom of isotropic MIMO interference channels without state information at the transmitters [J]. IEEE Trans. Inf. Theory, 2012, 58(1): 341-352

[7] Vaze C S, Varanasi M K. On completing the degrees of freedom characterization of MIMO interference networks with CSIT [C]// IEEE International Symposium on Information Theory Proceedings, 2011

[8] Chen Guo-xing, Xiang Zheng-zheng, Xu Chang-qing, et al. On Degree of Freedom of Cognitive Networks with User Cooperation [J]. IEEE Wireless Communication Letters, 2012, 1(6): 617-620

[9] Park S H, Lee I. Degree of Freedom for Mutually Interfering Broadcast Channels [J]. IEEE Trans on Information Theory, 2012, 58(1): 393-401

[10] Fette B A. Cognitive Radio Technology [M]. Academic Press, 2006

[11] Haykin S. Cognitive radio: brain-empowered wireless communications [J]. IEEE J. Sel. Areas Commun., 2005, 23(2): 201-202

[12] Amir M, El K A, Nafie M. Constrained interference alignment and the spatial degrees of freedom of MIMO cognitive networks [J]. IEEE Trans. Inf. Theory, 2011, 57(5): 2994-3004

[13] Jafar S A. Interference alignment: a new look at signal dimensions in a communication network [M]. Foundations and Trends in Commun. And Inf. Theory, 2011

[14] Jafar S A, Fakhereddin M J. Degrees of Freedom for the MIMO Interference Channel [J]. IEEE Trans on Information Theory, 2012, 58(1): 2737-2641