基于非负矩阵分解的IP流量预测

高 茜 李广侠 胡 婧
(解放军理工大学通信工程学院 南京 210007)

摘 要 为解决宽带多媒体卫星通信系统中的IP流量预测问题,首先使用多用户的IP流量作为训练数据,通过非负矩阵分解迭代方法将其分解为基向量矩阵和编码矩阵,之后再通过ARIMA模型在时间维度上对编码矩阵中的各个行向量进行预测,最后依照预测结果和基向量矩阵合成出各个用户的IP流量预测结果。由于经非负矩阵分解后,编码矩阵中的行向量个数小于用户个数,因此相对于原始的单个用户独立预测方法,新方法可以降低运算的复杂度。仿真实验证实了本方法预测的准确性。

关键词 宽带,多用户,非负矩阵分解,预测
中图法分类号 TP927 **文献标识码** A

Nonnegative Matrix Factorization-based IP Traffic Prediction

GAO Qian LI Guang-xia HU Jing
(Institute of Communication Engineering, PLA University of Science & Technology, Nanjing 210007, China)

Abstract In the face of limited satellite bandwidth resources and users' increasing demands, it becomes a significant issue to realize reasonable and efficient bandwidth allocation among users for broadband multimedia satellite communications system. Traffic prediction plays an important role in system resource allocation and management. In the process, IP traffic for multiple users which is regarded as training data was decomposed into basis matrix and coding matrix based on NMF (Nonnegative Matrix Factorization). Then each row vector of encoding matrix was predicted on time dimension on the basis of ARIMA model. At the end, prediction results were combined with basis matrix to generate IP traffic prediction results of each user. After NMF decomposition, the native umber of row vector is fewer than the number of users. As a result, compared with traditional signal user prediction method, the method proposed in this paper can reduce computational complexity. Tests indicate the accuracy of this prediction method.

Keywords Broadband, Multi-user, Nonnegative matrix factorization (NMF), Prediction

1 引言

随着网络宽带化的进程,宽带多媒体卫星通信系统正经历着高速的发展,未来将有更多的用户通过卫星信道实现网络接入。而面对有限的卫星带宽资源和日益增长的用户需求,如何在用户间实现合理、高效的带宽分配就成为了宽带多媒体卫星通信系统研究领域的一个重要问题。对于带宽分配而言,最简单的方式是在每位用户的通信时间段内,为其分配固定的带宽。带宽定为其整个通信时段内,带宽需求的上限。这种带宽分配方式虽然降低了整个系统带宽分配的复杂度,但由于各用户实际的带宽需求是动态变化的,将导致大量珍贵的频谱资源被浪费,因而降低了整个卫星通信系统的资源利用效率。因此,研究具有更高利用率的动态带宽分配方案就变得非常重要。针对带宽分配,学者们提出了多种有效的解决方案[1-3]。但在这些方案中有一个十分重要又常常被简化的问题——网络IP流量的建模和预测。

几乎所有的带宽分配方案都是在某一特定网络IP流量预测模型的基础上提出的。例如,在Secchi R和Barsocchi等人提出的方案中[1],网络的流量建模为由独立随机过程所驱动的线性动态系统。而在Jiang等人的方案中[2],则采用了BP(Back Propagation)神经网络对流量进行建模。Delli等人使用基于Markov调制泊松过程模型对业务流量进行建模,并在此基础上设计了带宽分配机制[3]。

图1给出了宽带多媒体卫星通信系统的结构图。在系统中,各卫星终端(Satellite Terminal,ST)负责管理多个用户终端的连接。

在卫星系统上行信道中,用户终端(User Terminal,UT)的业务数据是经由ST聚集,并经由卫星转发,传输给信关站(Gateway Station,GS),最后接入地面骨干IP网。下行信道的传输过程与之相反。而在这个传输过程中,卫星系统的流量控制等功能都是由地面的网络控制中心(Network Control Centre,NCC)集中决策。这种集中决策过程与地面网的中心

到稿日期:2011-05-03 返修日期:2011-07-19 本文受国家自然科学基金项目(60972061,61032004),国家高技术研究发展计划("863"计划)项目(2008AA12A204,2008AA12Z307)资助。

高 茜(1984—),女,博士生,主要研究方向为宽带卫星通信,E-mail:gaoxiongmao1234@163.com;李广侠(1964—),男,教授,主要研究方向为卫星通信、卫星导航等;胡 婧(1981—),女,讲师,主要研究方向为卫星通信、通信网络等。

控制器相比，最大的不同在于卫星信道中存在较大的传输延时。以地球静止轨道卫星(Geostationary Earth Orbit，GEO)为例，从 ST 到 NCC 的单程传输时延约为 250ms。延时导致的直接结果是：NCC 对下一时段的带宽分配决策很大程度上是基于对下一时段网络流量的预测进行的。这种体系结构也促成了对网络流量预测的需求。

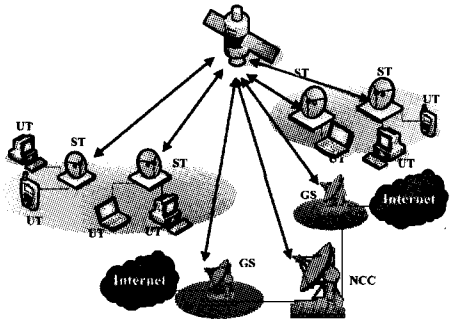

图 1　宽带多媒体卫星通信系统示意图

事实上网络流量预测并不是一个全新的问题，在网络研究领域已有多种预测模型被提出，主要有 Markov 模型、自回归滑动平均模型(Auto Regressive Moving Average，ARMA)、自回归求和滑动平均模型(Auto Regressive Integrated Moving Average，ARIMA)以及分数差分自回归求和滑动平均模型(Fractional Auto Regressive Integrated Moving Avera-ge，FARIMA)等。王升辉等[4]利用多重分形预测模型，将难以预测分析的长相关流量序列转化为可以用短相关线性模型预测的序列组。

此外，人工神经网络也因其良好的非线性预测能力，得以广泛应用，如基于神经网络或模糊神经网络的预测模型[5,6]。径向基函数(Radial Basis Function，RBF)神经网络是近年较为常用的网络流量预测工具，天津大学的王俊松和高志伟[7]利用 RBF 神经网络对网络流量建模，并进行了预测。虽然取得了较高的精度，但是回避了网络流量通常具有自相似性的特点，未对自相似流量的预测进行研究。

目前还有一些研究人员利用经验模式分解(Empirical Mode Decomposition，EMD)方法对流量的建模和预测进行研究，如单佩韦等[8]对 EMD 在网络流量的自相似参数估计方面应用的研究，高波等[9]提出的基于 EMD 的自相似网络流量预测方法。

上述这些方法考虑的都是单路网络预测问题，而在卫星上行信道中各个 ST 都负责多个 UT 的流量处理。此时如果使用单个用户分别处理的方式，将会给系统带来较大的运算开销。此外，虽然看似多用户的流量数据间是相互独立的，但其实在同一时间段内，这些用户间的流量是具有一定的相关性的。例如，在数字视频广播(Digital Video Broadcast，DVB)应用中，可能大多数个人用户收看视频的时间都集中在晚上的休息时间。因此这种相关性可通过对多用户流量数据的子空间投影方法加以分析和利用。

本文针对宽带多媒体卫星系统上行信道中，单个 ST 处理的多个 UT 流量的预测问题，在充分挖掘流量数据中的空间分布特性基础上，提出了一种基于非负矩阵分解的多用户预测方法。

本文第 2 节简要介绍非负矩阵分解的基本知识及加入连续性限制的非负矩阵分解；第 3 节提出基于非负矩阵分解的多用户 IP 流量预测方法；第 4 节通过基于实测数据的实验仿真验证本方法的有效性。

2　预备知识

已有的矩阵分解方法有 PCA(主成分分析)、ICA(独立成分分析)、SVD(奇异值分解)、VQ(矢量量化)、FA(因子分析)等，这些方法通常是在一定的限制下对数据进行线性变换或分解。不同的方法施加在其上的限制是不同的，这些方法在矩阵进行分解的时候允许分解的对象和结果为负。与之前的分解方法不同，非负矩阵分解(Nonnegative Matrix Factorization，NMF)是一种新的变换方法，能使分解的对象和分解的结果均为非负值。姜伟等[10]在 NMF 的基础上提出了一种局部敏感非负矩阵分解降维算法。

本节首先简要介绍非负矩阵分解的基本知识，之后介绍一种加入时域连续性限制的非负矩阵分解方法。

2.1　非负矩阵分解

非负矩阵分解是一种由 Lee 等人提出的线性子空间矩阵分解方法[11]。该方法相较于其它矩阵分解方法，最大的不同在于加入了分解结果非负性的限制。其分解形式如式(1)所示。

$$V \approx WH \tag{1}$$

也可表示为：

$$v_{ij} \approx (WH)_{ij} = \sum_{r=1}^{R} w_{ir} h_{rj} \tag{2}$$

式中，V 为 $M \times N$ 的非负矩阵。由 R 个列向量构成的矩阵 W 称为基矩阵，而 H 则表示 V 在 W 上的投影值，也被称为编码矩阵，且 $w_{ir} \geqslant 0, h_{rj} \geqslant 0$。此外，$R$ 满足：

$$(M+N)R < MN \tag{3}$$

由于具有非负性限制，V 中的列向量是由各个非负的基向量通过叠加方式进行重构的。因此，各个基向量可看作是对 V 的部分分解，也可认为每个基向量都承载了 V 的部分特征。也正是由于这种由整体到部分的分解，使得基矩阵和编码矩阵都具有了一定的稀疏特性[12]。

利用非负矩阵分解观测矩阵由整体到部分的分解特性，可以将多用户的流量矩阵投影到可表征其部分特征的基矩阵构成的子空间中去。由于特征子空间的维度远小于原流量数据的维度，使得流量矩阵得以压缩，这也就降低了预测步骤的运算量。此外，编码矩阵中的各个行向量所表征的是原流量矩阵各部分特征的时序变化情况，具有时域的连续性，因此对其预测的复杂度并不会比对原始流量数据预测的复杂度高。

对于非负矩阵分解的具体分解，Lee 等人给出了两种方法：基于欧氏距离的误差最小化方法，如式(4)所示，以及基于 Kullback-Leibler 散度的误差最小化方法，如式(5)所示[12]。

$$D_E(V \| WH) = \sum_{i,j} (v_{ij} - (WH)_{ij})^2 \tag{4}$$

$$D_{KL}(V \| WH) = \sum_{i,j} \left[v_{ij} \log \frac{v_{ij}}{(WH)_{ij}} - v_{ij} + (WH)_{ij} \right] \tag{5}$$

基于欧式距离的误差最小化方法是在加性高斯噪声条件下对 W 和 H 的极大似然估计，而基于 Kullback-Leibler 散度的误差最小化方法则是在观测数据由均值为 $(WH)_{ij}$ 的 Poisson 过程生成时的极大似然估计。且当使用 Kullback-Leibler 散度的误差最小方法进行非负矩阵分解时，分解结果对于观

测数据中的低频信息具有更好的重构效果。而在业务预测中，我们可认为主要由低频信息承载了业务的变换趋势，其对预测准确度的贡献要高于高频部分的信息。因而在后续的预测方法中，将采用基于 Kullback-Leibler 散度误差最小化的非负矩阵分解方法。

2.2 加入连续性限制的非负矩阵分解

在最初的非负矩阵分解算法中，并没有考虑分解得到的编码矩阵中各个行向量在时间维度上的连续性问题。但在许多新的应用场合，例如在对语音信号频谱矩阵的分析中，观测矩阵 V 的各个列向量在时间维度上是存在连续性的，并且在分解得到的编码矩阵中保持这种时序的连续性对后续处理也具有价值。因此，Tuomas 等人提出了一种加入连续性限制的非负矩阵分解方法[13]。

新方法在散度误差最小化的基础上，加入了编码矩阵时域连续性限制条件，如式(6)所示。

$$c(W,H)=c_r(W,H)+\alpha c_t(H) \tag{6}$$

式中，$c_r(W,H)$为由式(5)表示的重构误差项，$c_t(H)$为针对编码矩阵的时域连续性限制项，其具体形式如式(7)所示。α 为加权因子。

$$c_t(H)=\sum_{r=1}^{R}\frac{1}{\sigma_r^2}\sum_{j=2}^{N}(h_{r,j}-h_{r,j-1})^2 \tag{7}$$

式中，σ_r 为编码矩阵中第 r 行数据的标准差。$c_t(H)$是编码矩阵中相邻列向量间的归一化差异的平方和。

通过加入时域连续性限制项，使非负矩阵分解得到的编码矩阵在时间维度上具有更好的连续性。当应用于数据预测时，这种连续性将有助于提升预测结果的准确度。下一节将介绍一种基于存在时域连续性限制非负矩阵分解的多用户流量预测方法。

3 IP 流量预测

采用非负矩阵分解方法进行多用户网络流量预测的流程如图 2 所示。

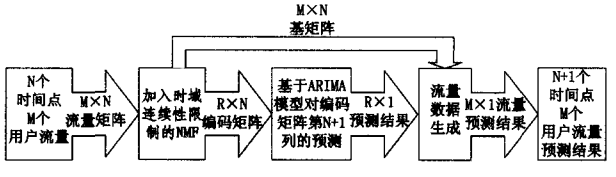

图 2 网络流量预测流程图

预测过程分为两个阶段：训练阶段和预测阶段。

3.1 训练阶段

1)假设存在 M 个用户的流量需要进行预测，首先获取 M 个用户前 N 个时刻的流量值，生成 $M\times N$ 的流量矩阵；

2)采用存在时域连续性限制的非负矩阵分解方法对流量矩阵进行分解，得到 $M\times R$ 的基矩阵和 $R\times N$ 的编码矩阵；

3)采用 ARIMA 模型对编码矩阵中各个行向量进行建模。

3.2 预测阶段

根据前 k 个时刻的数据对第 $k+1$ 时刻的数据进行预测：

1)通过建立的 ARIMA 模型以及之前编码矩阵中各行向量数据，得到第 $k+1$ 时刻编码矩阵的预测结果$\hat{h}_{k+1}$。

2)通过将基矩阵与编码矩阵第 $k+1$ 时刻的预测结果相乘得到 M 个用户第 $k+1$ 时刻的流量预测结果：

$$\hat{v}_{k+1}=W\hat{h}_{k+1} \tag{8}$$

式中，$\hat{v}_{k+1}$是 $k+1$ 时刻的预测值。

3)在第 $k+1$ 个时刻，通过 M 个用户的流量观测数据，并结合基矩阵分解得到的第 $k+1$ 时刻的编码矩阵值，可更新编码矩阵结果：

$$h_{k+1}=\frac{v_{k+1}}{W} \tag{9}$$

式中，v_{k+1}是 M 个用户在 $k+1$ 时刻的最新流量数据。

4)回到步骤 1)进行下一时刻的流量预测。

在对编码矩阵中各个行向量的预测中，采用的是 ARIMA 模型。这是因为相较于 ARMA 模型，ARIMA 模型能够更好地对信号非平稳特性进行建模，且相比于 FARIMA 模型，具有较低的复杂度。

4 仿真实验分析

本实验中采用的是实际的 Abilene III 网络流量数据。数据于 2004 年 6 月 1 日在 Abilene 网络的路由器节点处采集得到。该流量数据的基本情况如表 1 所列。

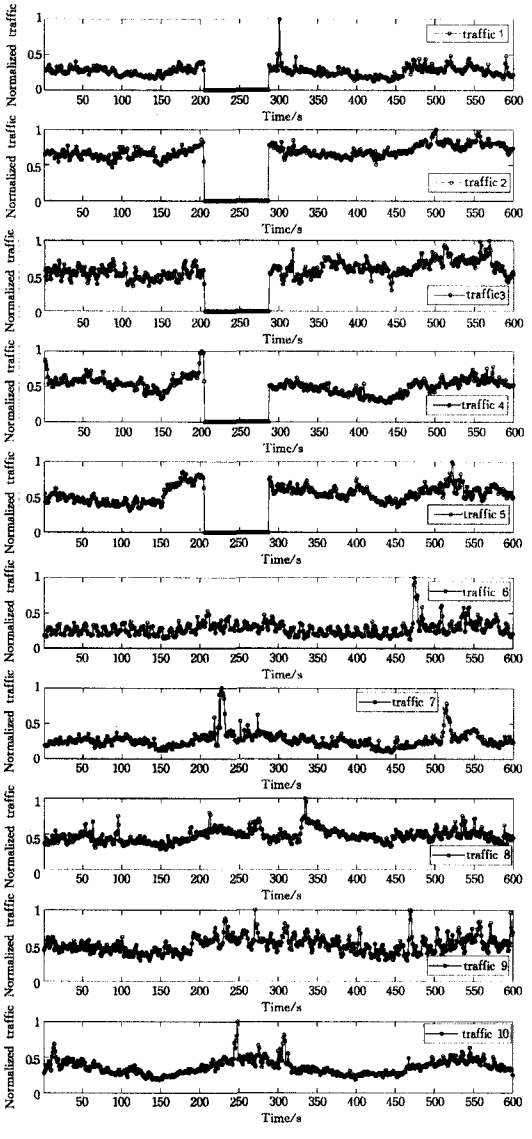

图 3 取自 Abilene III 的 10 路流量数据

表 1　Abilene III 流量数据的基本情况

Bandwidth	Total packets	Total flows	Utilization	Flows in progress	MTU
10Gbps	156M	683k	19%	62000	9000

注：MTU 为最大传输单元。

实验前首先对数据进行预处理。具体来说，随机选取其中 10 对节点间的流量进行统计，时间间隔为 5s，统计时长为 3000s。之后再对这些数据进行归一化处理。经处理后的这 10 路流量数据如图 3 所示。在实验中，取各路流量数据的前 500 点作为训练数据，后 100 点作为测试数据。

依照第 3 节所述的预测方法对流量进行预测时，首先取出各路流量数据中的前 500 点，构成 10×500 的流量矩阵 V。通过加入连续性限制的非负矩阵分解方法对 V 进行迭代分析，从而得到相应的基矩阵 W 和编码矩阵 H。此处，基矩阵大小设定为 10×6，编码矩阵大小为 6×500，连续性加权因子 α 取经验值 1.2。

然后，通过 ARIMA(p,d,q)模型对分解得到的 6 路编码矩阵分别进行建模。各 ARIMA(p,d,q)模型的阶数通过多次实验，以取拟合误差最小的方式确定。最终选定的阶数如表 2 所列。

表 2　编码矩阵各路 ARIMA 模型的阶数

第一路	第二路	第三路	第四路	第五路	第六路
(2,0,3)	(2,1,3)	(2,0,3)	(1,0,2)	(2,1,3)	(1,0,3)

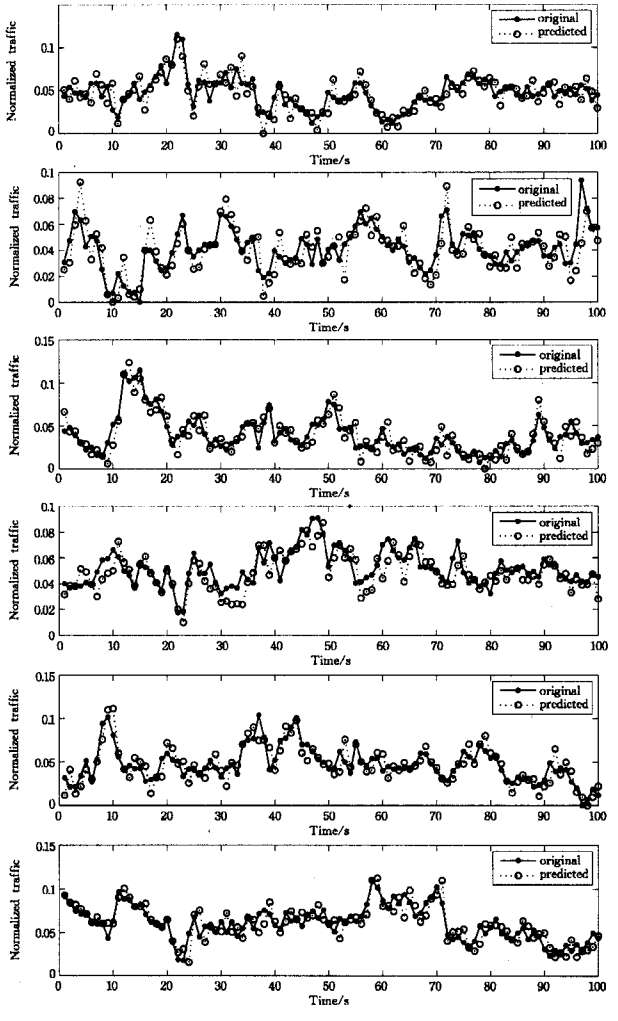

图 4　通过 ARIMA 模型对编码矩阵各路数据的预测结果

通过训练阶段得到的基矩阵，对由测试流量数据构成的 10×100 测试矩阵 V_t 进行分析，得到对应的编码矩阵 H_t：

$$H_t=\frac{V_t}{W} \tag{10}$$

其大小为 6×100。基于训练阶段得到的编码矩阵 ARIMA 模型，对测试编码矩阵 H_t' 中各路数据进行预测。预测结果如图 4 所示。

通过预测得到的编码矩阵 H_t' 及基矩阵 W，可得到各路流量数据的预测结果，如图 5 所示。

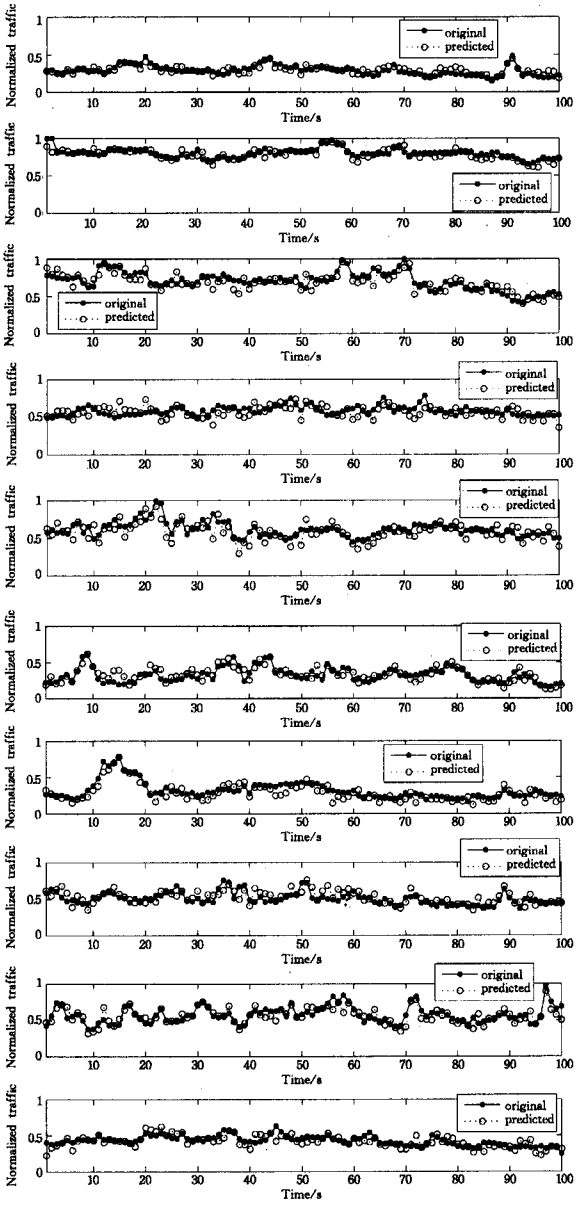

图 5　各路流量数据的预测结果

对于预测结果的准确度，本文采用各路流量的均方误差(Mean Squared Error，MSE)之和作为预测效果的评价指标，即：

$$e=\frac{1}{K}\sum_{i=1}^{10}\sum_{t=1}^{K}(\hat{v}_t^i-v_t^i)^2 \tag{11}$$

式中，e 代表各路预测结果的 MSE 之和，K 表示实际预测的流量数据个数，$\hat{v}_t^i$ 为第 k 路流量在 t 时刻的预测值，v_t^i 为第 k 路流量在 t 时刻的实际值。

为了验证本文提出的方法的有效性，将与直接采用 ARIMA 模型预测流量进行对比，并计算预测误差。基于本文方

法的预测误差和直接使用 ARIMA 模型所得到的预测误差如表 3 所列。

表 3 预测误差的比较

不同方法	e
基于 NMF 的预测	0.0651
直接预测	0.069

从表 3 中可以看出，基于 NMF 的预测方法和直接采用 ARIMA 模型的预测方法，其预测误差十分接近，但本文中的方法相对于直接预测方法来说省去了 4 路数据的建模开销，具有更低的计算复杂度。

结束语 本文针对宽带多媒体卫星通信系统中多路流量的预测问题，提出了一种基于非负矩阵分解的流量预测方法。通过非负矩阵分解，将原始多路流量数据投影到一个低维子空间，可降低预测数据量，提高整个预测系统的预测效率。通过实验表明，基于非负矩阵分解的流量预测方法与直接流量预测方法相比，预测误差十分接近，但预测计算复杂度更低。

参 考 文 献

[1] Secchi R, Barsocchi P, Davoli F. Linear quadratic control of service rate allocation in a satellite network[J]. Communications, IET, 2010, 4(13): 1580-1593

[2] Yueqiu J, Hang L. Study on bandwidth allocation of broadband multimedia communication satellite system[C]// Proceedings of the 3rd International Conference on Intelligent Networks and Intelligent Systems. 2010: 506-508

[3] Delli Priscoli F, Pompili D. A demand-assignment algorithm based on a Markov modulated chain prediction model for satellite bandwidth allocation[J]. Wirel. Netw., 2009, 15: 999-1012

[4] 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

[5] Satsri S, Ardhan S, Chutchavong V, et al. ANN based NGN IP traffic prediction in Thailand[C]// Proceedings of the International Conference on Control, Automation and Systems. 2007: 2154-2157

[6] Li R, Chen J, Liu Y, et al. WPANFIS: combine fuzzy neural network with multiresolution for network traffic prediction[J]. The Journal of China Universities of Posts and Telecommunications, 2010, 17(4): 88-93

[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)：6-11

[8] 单佩韦，李明. 基于 EMD 的自相似流量 Hurst 指数估计[J]. 计算机工程，2008，34(23)：128-129

[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] 姜伟，杨炳儒，隋海峰. 局部敏感非负矩阵分解[J]. 计算机科学，2010，27(12)：211-214

[11] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791

[12] Daniel D, Lee H. Sebastian Seung. Algorithms for non-negative matrix factorization[C]// Proceedings of the Conference on Advances in Neural Information Processing Systems. 2000: 556-562

[13] Tuomas V. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 1066-1074

(上接第 47 页)

$$e(P,V)=e(Q,U)e(P_{pub},\sum_{i=1}^{n}h_iQ_{ID_i})e(U_A,\sum_{i=1}^{n}h_iQ)e(P_{pub},\sum_{i=1}^{n}h_ih_AQ_{ID_A})$$

可知，我们只需要计算 $V=\sum_{i=1}^{n}V_i$，$U=\sum_{i=1}^{n}U_i$，而不需要单独验证各个签名 $\sigma_i=(W,U_A,U_i,V_i)$。签名长度的压缩率为 $1/n$，即与单个签名的长度相当，同时，整个签名验证过程只需要 5 个双线性对计算，因此具有较高的效率。

结束语 本文利用双线性对提出了一个基于身份的代理聚合签名方案。方案中聚合签名的长度仅与单个签名的长度相当，而签名的验证仅需要 5 个双线性对和 $3n$ 个群元素的标量乘计算，因而具有较高的效率。最后利用计算 CDH 问题的困难性证明了方案在随机预言模型下的不可伪造性。

参 考 文 献

[1] Shamir A. Identity-based cryptosystems and signature schemes [C]// Blakley G, Chaum D, eds. Proceedings of Crypto 1984, volume 196 of LNCS. 1984: 47-53

[2] Boneh D, Franklin M. Identity-based encryption from the Weil pairing[C]// Joe Kilian, ed. Proceedings of Crypto 2001. volume 2139 of LNCS. 2001: 213-229

[3] Hess F. Efficient identity based signature schemes based on pairings[C]// Kaisa Nyberg, Howard M, eds. Proceedings of SAC 2002. volume 2595 of LNCS. 2002: 310-324

[4] Paterson K G, Schuldt J C N. Efficient identity-based signatures secure in the standard model[C]// Proceedings of ACISP 2006. volume 4058 of LNCS. 2006: 207-222

[5] Mambo M, Usuda K, Okamoto E. Proxy signatures for delegating signing operation[C]// Proceedings of the 3rd ACM Conference on Computer and Communications Security. New York: ACM, 1996: 48-57

[6] Zhang F, Kim K. Efficient ID-based blind signature and proxy signature from bilinear pairings[C]// Proceedings of the 8th Australasian Conference on Information Security and Privacy. volume 2727 of LNCS. 2003: 312-323

[7] Wu W, Mu Y, Susilo W, et al. Identity-based proxy signature from pairings[C]// Proceedings of the 4th International Conference on Autonomic and Trusted Computing. volume 4610 of LNCS. 2007: 22-31

[8] 李明祥，韩伯涛，朱建勇，等. 在标准模型下安全的基于身份的代理签名方案[J]. 华南理工大学学报：自然科学版，2009，37(5)：118-122

[9] Boneh D, Gentry C. Aggregate and verifiably encrypted signatures from bilinear maps[C]// Advances in Cryptography-Eurocrypt 2003. volume 2656 of LNCS. 2003: 416-432

[10] Cheon J H, Kim Y, Yoon H J. A new ID-based aggregate signature with batch verification[OL]. http://eprint.iacr.org/2004/131

[11] Pointcheval D, Stern J. Security arguments for digital signatures and blind signatures[J]. Journal of Cryptology, 2000, 13(3): 361-396

[12] Herranz J. Deterministic identity-based signatures for partial aggregation[J]. Computer Journal, 2006, 49(3): 322-330

[13] Camenisch J, Hohenberger S, Pedersen M O. Batch Verification of short signatures[C]// Advances in Cryptography- Eurocrypt 2007. volume 4515 of LNCS. 2007: 246-263