Self-Similarity for Data Mining and Predictive Modeling
A Case Study for Network Data

Jafar Adibi¹, Wei-Min Shen¹, Eaman Noorbakhsh²

¹Information Sciences Institute, Computer Science Department,
University of Southern California
4676 Admiralty Way, Marina del Ray, CA 90292
{adibi,shen}@isi.edu

² Electrical Engineering Department, University of Southern California
University Campus Park, Los Angeles, CA 90089
noorbakh@usc.edu

Abstract. Recently there are a handful study and research on observing self-
similarity and fractals in natural structures and scientific database such as traffic
data from networks. However, there are few works on employing such
information for predictive modeling, data mining and knowledge discovery. In
this paper we study, analyze our experiments and observation of self-similar
structure embedded in Network data for prediction through Self Similar
Layered Hidden Markov Model (SSLHMM). SSLHMM is a novel alternative
of Hidden Markov Models (HMM) which proven to be useful in a variety of
real world applications. SSLHMM leverage HMM power and extend such
capability to self-similar structures and exploit this property to reduce the
complexity of predictive modeling process. We show that SSLHMM approach
can captures self-similar information and provides more accurate and
interpretable model comparing to conventional techniques.

1 Introduction

In recent years there have been interest in research and development for traffic
modeling and forecasting the Network data. Recent measurements of local-area and
wide-area traffic [16, 22, 23] have shown that network traffic exhibits variability at a
wide range of time scales. For instance, analysis of traffic data from networks and
services such as ISDN traffic, Ethernet LAN’s, Common Channel Signaling Network
(CCNS) and Variable Bit Rate (VBR) video have all convincingly demonstrated the
presence of features such as self-similarity, long range dependence, slowly decaying
variances, heavy-tailed distributions and fractal dimensions [23]. Through past
decade, different models suggested for Network behavior analysis. Although several
works attempt to analyze, model and predict Network data such as Markov Model,
but there are only a few remarkable works, which have shown successful, result and
the problem of modeling Network data still is an open issue [5, 9, 15, 19]. Recently
there have been several attempts to apply data mining techniques through fractal
dimensions and self-similarity. Among those, using fractal dimension, using fractal
dimension for dimension reduction [21], learning association rules[3] and application
in spatial joint [10] are considerable.
Self-Similarity for Data Mining and Predictive Modeling

Self-Similar Layered Hidden Markov Model (SSLHMM) has been introduced by Adibi et al in [1] with application in Network data. In this paper we only study and analyze our experiments and observations of such data and the relation among SSLHMM structure, fractal dimension and self-similarity. We would like to show if we can use the fractal dimension and self-similarity of a given data for a better estimation of SSLHMM structure.

2 Related Work and Background

The convectional methods for analyzing the network data was Poisson arrivals, in which the number of arrivals in the time interval $T$ follows the exponential distribution with parameter $\lambda T$. This model works well in case of traditional telephone network, but not for the internet traffic data. The failure of the Poisson model is explained in [19] by Paxson and Floyd. The second method which used for traffic modeling was Autoregressive type traffic models. These models define the next variates in the sequence as an explicit function of previous variates from the same time series within a time window stretching from the present into the past. These models are Auto Regressive (AR), Moving Average model (MA), ARMA model and also ARIMA model. This method worked well in the early years of developing Internet, because the change in the amount of traffic was not abrupt in those days. But recent studies have indicated that this method also fails for highly volatile traffic[4].

Recent measurements of local-area and wide-area traffic have shown that network traffic exhibits variability at a wide range of time scales. What is striking is the ubiquitousness of the phenomenon which has been observed in diverse networking contexts, from Ethernet to ATM, compressed VBR video, and HTTP-based WWW traffic [5, 9, 16, 22-24]. A number of performance studies have shown that self-similarity can have a detrimental on network performance leading to increased queuing delay and packet loss rate which implied that they also exhibited long range dependency (LRD). Recent research suggests that not only packet traffic, but also the TCP session arrival pattern is self-similar. This means that if a switched virtual circuit network is to be substituted for a connectionless TCP/IP, the network will have to cope with periodical overloads of control units of its switches. Since then, this feature has been discovered in many other traffic traces, such as Transmission Control Protocol (TCP), Motion Pictures Experts Group (MPEG) video, World Wide Web, and Signaling System traffic[6, 9, 15, 22, 24]. The importance of this discovery becomes apparent when it is observed that Poisson; ARMA and Markov processes are unable to exhibit LRD. In fact they are short-range dependent (SRD) processes. The major flaw with the traditional traffic models is that they do not model the burstiness of the Internet traffic correctly. The burstiness exists in every time-scale while with traditional models it disappears in the long time intervals.

The mathematical study of self-similar shapes and their relationship to natural shapes was first presented by Benoit Mandelbrot. Self-similar stochastic processes were introduced by Kolmogorov in a theoretical context and brought to the attention of probabilists and statisticians by Mandelbrot and his co-workers and have been used in hydrology, geophysics, biophysics, and biology and communication systems [17]. Among different alternative to test the self-similarity of a sequence we used variance-