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

2Electrical 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 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 an 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 λT. 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-time plot[16]. In this method the estimated variance of $X^{(m)}$ is
plotted against $m$ on log-log scale. A straight line with $(-\beta)$ slope indicates self-similarity. The
discovery of self-similarity in computer networks led researcher to use of fractional Brownian
motion (fBm) - a more general class of self-similar processes of which Brownian motion is a
special case - and fractional ARIMA process for modeling the self-similarity in the network data
traffic. In general the ARIMA trace is often obtained by generating a fractional Gaussian noise
trace with suitable $H$ and filtering this noise with the ARMA coefficients. Although the above-
mentioned models have been employed by several groups but the capability of these models for
predicting self-similar processes is not clear and prediction of Network data seems inevitable
[15]. For example Bates and McLaughlin [5] showed some evidence, both qualitatively and
quantitatively, to suggest that Ethernet data does not conform to popular self-similar models. The
evidence suggests that Ethernet is more impulsive than the Gaussian case, which these models
assume. There are also some other methods for self-similar data modeling which are less popular
than the above models [16][21]. These models are M/G/$\infty$ process, aggregated AR process;
heavy tailed on-off process, stable self-similar process, fractal shot noise and renewal process
and stochastic difference equations.

3 Self-Similar Layered Hidden Markov Model (SSLHMM)

Hidden Markov Models (HMM) have proven to be useful in a variety of real world applications
where considerations for uncertainty are crucial. Such an advantage can be more leveraged if
HMM can be scaled up to deal with complex problems. However, despite the broad range of
application areas shown for HMM, they do have limitations and do not easily handle problems
with certain characteristics. For example Markov Model has not reported as a successful
alternative for analyze the Network behavior. To extend HMM we only focus on complexity of
HMM for a certain category of problems with the following characteristics: 1) The uncertainty
and complexity embedded in these application makes it difficult and impractical to construct the
model in one step. 2) Systems are self-similar, contain self-similar structures and have been generated through recurrent processes. In the modeling of complex processes, when the number of states goes high, the maximization process gets more difficult. A solution provided in other literature is to use of a Layered HMM instead [2, 12]. Layered HMM has the capability to model more than one process. Hence, it provides an easier platform for modeling complex processes. Layered HMM is a combination of two or more HMM processes in a hierarchy. Fig. 1(b) shows a Layered HMM with 9 states and 3 super-states, or macro-states, which we refer to them as phases. As we can see, each phase is a collection of states bounded to each other. The real model transition happens among the states. However, there is another transition process in upper layer among phases. The comprehensive transition model is a function of transition among states and transition among phases.

SSLHMM is a special form of Layered HMM in which there are some constraints on state layer transition, phase layer transition, and observation distribution. There are also some extensions to HMM such as [12], [11] which through limited space we only refer to them for further study. The advantage of SSLHMM is that like any other self-similar model it is possible to learn the whole model having any part of the model. Although there are a couple of assumptions to hold such properties but fortunately for a large group of systems in nature self-similarity is one of their characteristics.

4 Experimental Result

We have applied SSLHMM approach to a real Network database collected during 18 weeks, from October 16 1994 to February 12 1995, on Cabletron corporate network. There are 16849 entries, representing measurements roughly every 10 minutes for 18 weeks. This network has a router with 16 ports connected to 16 links. The packet traffic of each port is investigated independently. There are 16 ports router that connect to 16 links, which in turn connect to 16 Ethernet subnets. Note that the traffic has to flow through the router ports in order to reach the 16 subnets. There are three independent variables:

- **Bandwidth**: the percentage of bandwidth utilization of a port during a 10 minute period.
- **Packet Rate**: the rate at which packets are moving through a port per minute.
- **Collision Rate**: the number of collided packets during a 10 minute period.

Fig. 2 shows the variance-time plot for the data. Here we measured the Hurst parameter by fitting the least square line through the resulting points in the plane, ignoring the small values of m. Values of the estimate $\beta$ of the asymptotic slope between $-1$ and $0$ suggest self similarity is given by $H = 1 - \beta / 2$. Fig.2 shows the data is indeed self-similar.
There are some amazing observations, which we found out through our experiments. The average fractal dimension for all three variables: Bandwidth, Packet and Collision is from 1.5 to 1.65. This may indicates that all three variables might have similar characteristics. We generated a synthetic data similar to Network data. User had the capability to define the number of sequence in experimental pool, length of each sequence, number of layers, number of states in each phase, number of phases and observation set for discrete environment or a range for continuous observation. We observed that the fractal dimension for the synthetic data for a SSLHMM with N=4 and n = 4 (number of states in each phase) is close to 1.5-1.6 in different experiments. This may indicates that this model, N and n are the good estimates for a SSLHMM or at least a good initial point to start. The number of states in HMM is hidden and has been a recurrent issue in past decade. The brute forth way to find the best number of states may not be suitable in some cases in which a model desired to be obtained in real time. In addition, a better starting point in a blind search decrease the search time in general.

To compare SSLHMM with HMM and conventional techniques, we employed a SSLHMM model in which user has the capability to define the number of states and number of phases. In this paper we only report the comparison of Baum-Welch forward algorithm for HMM with $n_{hmm}$ states and a 2-layer strong SSLHMM with $N$ phases and $n$ states. In addition we assume a one-to-one relationship among states and phases in two layers. In addition we ran a synthetic data generator and we found a SSLHMM with 4 phases and 4 states in each phase generate a sequence with fractal dimension about 0.6. We did use such number as an estimation for number of phases and states. The main purpose of this experiment is built on the following chain of principles:

- Assume there is a sequence of observation $O = \{o_1, o_2, \ldots, o_T\}$ as Network data.
- We would like to construct a model $\lambda$ for such data.
- We would like to illustrate that for $O = \{o_1, o_2, \ldots, o_T\}$, $P(O \mid \text{SSLHMM})$ is higher than $P(O \mid \text{HMM})$, and other techniques (the probability of the observation given each model).
We applied the HMM and SSLHMM to a given port of database with the purpose of modeling the Network data. We did test our technique through cross validation and in each round we trained the data with a random half of the data and test over the rest. We repeat the procedure for Bandwidth, Packet Rate and Collision Rate. Table 2 illustrates the comparison of HMM and SSLHMM for Bandwidth, Packet Rate and Collision Rate on 3 randomly selected ports. Respectively, we ran HMM with prior number of states equal to 2, 3, 4, 9 and 16, and SSLHMM with number of phases equal to 2, 3 and 4 (shown as 2-s, 3-s and 4-s in the Table 2). As it shows in Table 2 the SSLHMM model with \( N=4 \) outperforms other competitors in all series of experiments. -log(likelihood) increases by increasing the number of states more than 16 as it over fits the data. The best SSLHMM performance beats the best HMM by 23%, 41% and 38% for Collision Rate, Bandwidth and Packets Rate respectively. For ARMA, AR and RW techniques we used the error as the measure for likelihood with the assumption of a normal distribution for the trained data. Hence, the likelihood became the probability of error.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Bandwidth</th>
<th>Packet</th>
<th>Collision</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM-2</td>
<td>382.05</td>
<td>403.32</td>
<td>284.71</td>
</tr>
<tr>
<td>HMM-3</td>
<td>206.86</td>
<td>228.68</td>
<td>183.74</td>
</tr>
<tr>
<td>HMM-4</td>
<td>213.05</td>
<td>220.17</td>
<td>176.86</td>
</tr>
<tr>
<td>HMM-9</td>
<td>169.23</td>
<td>176.86</td>
<td>176.61</td>
</tr>
<tr>
<td>HMM-16</td>
<td>151.72</td>
<td>170.12</td>
<td>163.86</td>
</tr>
<tr>
<td>SSLHMM - 2-S</td>
<td>355.77</td>
<td>381.93</td>
<td>294.08</td>
</tr>
<tr>
<td>SSLHMM - 3-S</td>
<td>97.49</td>
<td>166.78</td>
<td>130.12</td>
</tr>
<tr>
<td>SSLHMM - 4-S</td>
<td>105.19</td>
<td>102.90</td>
<td>100.20</td>
</tr>
<tr>
<td>AR</td>
<td>227.74</td>
<td>300.50</td>
<td>267.35</td>
</tr>
<tr>
<td>ARMA</td>
<td>230.74</td>
<td>297.78</td>
<td>239.25</td>
</tr>
<tr>
<td>Random Walk</td>
<td>232.99</td>
<td>285.51</td>
<td>197.92</td>
</tr>
</tbody>
</table>

Table 2: Negative log likelihood of different techniques for Network data

Our experiments show SSLHMM approach behave properly and does not perform worse than HMM even when the data is not self similar or when we do not have enough information. However due to space limitation we do not report the result in this paper. The SSLHMM provides a more satisfactory model of the network data from three point of views. First, the time complexity is such that it is possible to consider model with a large number of states in a hierarchy. Second, these larger number of states do not require excessively large numbers of parameters relative to the number of states. Learning a certain part of the whole structure is enough to extend to the rest of the structure. Finally SSLHMM resulted in significantly better predictors; the test set likelihood for the best SSLHMM was about 100 percent better than the best ARIMA, ARMA and Random Walk.

While the SSLHMM is clearly better predictor than HMM, it is easily interpretable than an HMM as well. The notion of phase may be considered as a collection of locally connected sets, groups, levels, categories, objects, states or behaviors and it comes with the idea of granularity, organization and hierarchy. As it mentioned before in Network application domain a phase could define as “congestion” or “stable”. This characteristics is the main advantage of SSLHMM over other approaches such as FHMM [12]. SSLHMM is designed toward better interpretation as one the main goal of data mining approaches in general.

5 Acknowledgement

This work was partially supported by the National Science Foundation Grant: 9529615.
6 References

2. Adibi, J., Shen, W-M. General structure of mining through layered phases. in International Conference on Data Engineering. (2002). San Jose, California, USA: IEEE.