Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 318 (2015) 109–127

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Adapting Hidden Markov Models for Online Learning

Tiberiu Chis[1,2] Peter G. Harrison[3]

Department of Computing
Imperial College London
London, UK

Abstract

In modern computer systems, the intermittent behaviour of infrequent, additional loads affects performance. Often, representative traces of storage disks or remote servers can be scarce and obtaining real data is sometimes expensive. Therefore, stochastic models, through simulation and profiling, provide cheaper, effective solutions, where input model parameters are obtained. A typical example is the Markov-modulated Poisson process (MMPP), which can have its time index discretised to form a hidden Markov model (HMM). These models have been successful in capturing bursty behaviour and cyclic patterns of I/O operations and Internet traffic, using underlying properties of the discrete (or continuous) Markov chain. However, learning on such models can be cumbersome in terms of complexity through re-training on data sets. Thus, we provide an online learning HMM (OnlineHMM), which is composed of two existing variations of HMMs: first, a sliding HMM using a moving average technique to update its parameters "on-the-fly" and, secondly, a multi-input HMM capable of training on multiple discrete traces simultaneously. The OnlineHMM reduces data processing times significantly and thence synthetic workloads become computationally more cost effective. We measure the accuracy of reproducing representative traces through comparisons of moments and autocorrelation on original data points and HMM-generated synthetic traces. We present, analytically, the training steps saved through the OnlineHMM's adapted Baum-Welch algorithm and obtain, through simulation, mean waiting times of a queueing model. Finally, we conclude our work and offer model extensions for the future.

Keywords: HMM, online learning, adapted Baum-Welch, autocorrelation, MMPP.

1 Introduction

Over the last decade, the performance and reliability of networks and storage systems have been key issues for international enterprises with a global online presence. On a large scale, businesses increasingly face the technical challenges that network performance may impose on their critical IP applications, remote desktops or video conferencing such as a lag in waiting times, availability and congestion with varying real-time Internet conditions in each country. Additionally, cost of storing data is

[1] Thank you to Nigel Thomas and the UKPEW Committee

[2] Email: tiberiu.chis07@imperial.ac.uk

[3] Email: p.harrison@imperial.ac.uk

http://dx.doi.org/10.1016/j.entcs.2015.10.022

increasing as the ratio of users to servers is growing, leading to competition amongst top cloud providers (i.e. Amazon EC2). Users demand applications to be available on tablets and smartphones and expect reliable performance of these mobile devices. Therefore, devices, servers and networks must all meet the required quality of service (QoS) standards determined by realistic service level agreements (SLAs) from respective clients and vendors. However, maintaining a consistently smooth network performance, by continuously upgrading infrastructure or increasing bandwidth, is expensive to match a large user base with heavy demand; a similar problem faces cloud providers w.r.t. storage.

In an attempt to solve these challenges, engineers and researchers rely on a combination of methods. Geographically, large-scale systems are spread apart to minimise communication (and therefore minimise delays) between servers located near end users and data centers. With different locations experiencing variable connectivity, a challenge is to maintain consistent service for all users, irrespective of location. At packet level, methods such as *traffic classification* helps to regulate packet transmission and bandwidth [13]. However, simply classifying IP traffic is not enough (i.e. without also modelling system load, waiting time, packet type, arrival burstiness, etc.) because non-deterministic events dictate traffic behaviour and should be modelled to improve storage systems and network performance and reliability. Workload models allow for experimenting with new storage system designs, where production systems provide key characteristics of their applications [3]. Consequently, it is desirable to extract representative workload parameters from storage traces.

Simple models, such as Poisson processes, no longer provide realistic tools for modelling Internet traffic and storage access, as they fail to account for long-range dependence (LRD) or burstiness of packets and I/O commands. Therefore, to improve this, researchers are turning their attention to more complex models, such as Markov-modulated Poission processes (MMPPs) and hidden Markov models (HMMs). In fact, the MMPP can be viewed as a discretely indexed HMM by observing intervals between events as a sequence of random variables [17]. The HMM is a bivariate Markov chain of states and transitions composed of a hidden chain (with unknown states) and an observable chain. Such models have been more successful in accounting for LRD, self-similarity, burstiness in jobs and switching modes, where we turn the reader's attention to [15,16] for reference. Similarly, Harrison et al use HMMs to obtain input to performance models of Flash memory [3]. Despite the success of these models in classifying properties of Internet traffic and storage access, there exist inefficiencies in static learning and (unnecessary) repetitive training of data.

The need to learn data in an online manner (i.e. "on-the-fly") is particularly useful for live systems where latency has a significant impact for users. Some models in the literature have adapted learning algorithms to avoid re-training on data sets, where model parameters are updated with new data [6,11]. The incremental HMM [11], in particular, provides parsimony, and portability through its adapted expectation maximisation algorithm, which trains strictly on new data points. This