Mach Learn
DOI 10.1007/s10994-013-5408-x

Adaptively learning probabilistic deterministic automata from data streams

Borja Balle · Jorge Castro · Ricard Gavaldà

Received: 7 December 2012 / Accepted: 9 August 2013

Abstract Markovian models with hidden state are widely-used formalisms for modeling sequential phenomena. Learnability of these models has been well studied when the sample is given in batch mode, and algorithms with PAC-like learning guarantees exist for specific classes of models such as Probabilistic Deterministic Finite Automata (PDFA). Here we focus on PDFA and give an algorithm for inferring models in this class in the restrictive *data stream* scenario: Unlike existing methods, our algorithm works incrementally and in one pass, uses memory sublinear in the stream length, and processes input items in amortized constant time. We also present extensions of the algorithm that (1) reduce to a minimum the need for guessing parameters of the target distribution and (2) are able to adapt to changes in the input distribution, relearning new models when needed. We provide rigorous PAC-like bounds for all of the above. Our algorithm makes a key usage of stream sketching techniques for reducing memory and processing time, and is modular in that it can use different tests for state equivalence and for change detection in the stream.

Keywords PAC learning · Data streams · Probabilistic automata · PDFA · Stream sketches

1 Introduction

Data streams are a widely accepted computational model for algorithmic problems that have to deal with vast amounts of data in real-time and where feasible solutions must use very little time and memory per example. Over the last ten years, the model has gained popularity

Editors: Jeffrey Heinz, Colin de la Higuera, and Tim Oates.

B. Balle · J. Castro · R. Gavaldà (✉)
LARCA research group, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain
e-mail: gavalda@lsi.upc.edu

B. Balle
e-mail: bballe@lsi.upc.edu

J. Castro
e-mail: castro@lsi.upc.edu

Published online: 03 October 2013