Mach Learn (2014) 96:1–3
DOI 10.1007/s10994-013-5428-6

EDITORIAL

Introduction to the Special Issue on Grammatical Inference

Jeffrey Heinz · C. de la Higuera · Tim Oates

Received: 20 November 2013 / Accepted: 29 November 2013 / Published online: 14 December 2013

Abstract As a field, Grammatical Inference addresses both theoretical and empirical learning problems, and the collection of papers within this special issue attests both to the diversity of these problems as well as the advances and insights that are being made by the researchers working within it. Thus we hope this special issue is of interest to the readership of Machine Learning.

1 Introduction to the special issue

Grammatical Inference is concerned with learning automata, grammars, and other objects that make it possible to generate, derive, represent, or recognize formal languages when given some information about the languages. The algorithms, techniques, and theoretical results which have been obtained in the field are used in a variety of very different applications, including natural language, bioinformatics, psychology, software engineering, and many others. Grammatical Inference may not be familiar to the readers of this journal, which may be primarily due to the somewhat unconventional nature of the training data (strings) and the hypothesis space (grammars). However, Grammatical Inference shares many algorithmic principles, theoretical approaches, and empirical methods with Machine Learning, which we believe will become apparent as you read the articles herein.

This special issue presents seven articles from a total of 26 submissions received, nine of which were extended submissions from the *11th International Conference on Grammatical*

J. Heinz
University of Delaware, Newark, USA
e-mail: heinz@udel.edu

C. de la Higuera (✉)
Laboratoire d'Informatique de Nantes Atlantique, University of Nantes, Nantes, France
e-mail: cdlh@univ-nantes.fr

T. Oates
University of Maryland Baltimore County, Baltimore, USA
e-mail: oates@cs.umbc.edu

Inference held at the University of Maryland in September, 2012. Each accepted paper received two rounds of reviews. On the first round, each paper received at least three reviews. On the second round, they were reviewed by either one referee and a guest editor or by two referees. Among the 26 submissions, 6 were co-authored by guest editors of the special issue. These 6 papers were handled separately by the guest editors who were not co-authors of the papers.

The special issue includes papers on a variety of topics, including distributional learning, which is leading to better understanding of what learning context-free grammars (or languages) is about. The non-terminals are associated with sets of strings and contexts in a natural way. This approach can be extended beyond the context-free world, as is shown by Alexander Clark and Ryo Yoshinaka in their article titled *Distributional learning of Parallel Multiple Context-free Grammars*.

Spectral methods offer an interesting alternative to Expectation-Maximization or state merging when it comes to learning probabilistic finite state machines. One advantage is that they learn the weights in a global way if an initial architecture is proposed. This approach is followed and described in the article titled *Spectral Learning of Weighted Automata* by Borja Balle, Xavier Carreras, Franco Luque and Ariadna Quattoni.

The article titled *Learning Register Automata: From Languages to Program Structures* by Malte Isberner, Falk Howar and Bernhard Steffen concerns the use of Dana Angluin's well known L^* algorithm for model checking. This can appear as a surprising application of what was essentially designed as a theoretical and hypothetical algorithm, chiefly relying on an oracle. Today, the oracle can be simulated by the program one is to test!

Grammatical Inference is typically concerned with learning from batches of strings, the problem of learning from data streams introduces a new set of questions. How do we forget the strings seen so far yet keep enough knowledge to build the hypothesis? How do we update our hypothesis when we detect concept drift? These are some of the questions addressed by Borja Balle, Jorge Castro and Ricard Gavaldà in their article titled *Adaptively Learning Probabilistic Deterministic Automata from Data Streams*.

Competitions and challenges are an important part of Machine Learning, as they allow us to evaluate competing algorithms on shared tasks. In 2012 the PAutomaC competition took place, with the goal of learning probabilistic finite automata and hidden Markov models. The results of this competition, with an analysis of the problem and an assessment of the state of the art are presented in the article *PAutomaC: a Probabilistic Automata and Hidden Markov Models Learning Competition* by Sicco Verwer, Rémi Eyraud and Colin de la Higuera.

The winners of the competition, Chihiro Shibata and Ryo Yoshinaka, present in detail their method in the paper titled *A Comparison of Collapsed Bayesian Methods for Probabilistic Finite Automata*. This method is based on collapsed Gibbs sampling, a variant of Gibbs sampling which treats hidden variables differently. The methods is compared to several other statistical approaches to the induction of probabilistic finite automata.

The article titled *Improving Active Mealy Machine Learning for Protocol Conformance Testing* by Sicco Verwer, Fides Aarts, Harco Kuppens, Frits Vaandrager, and Jan Trestmans presents a novel application of Grammatical Inference to software verification, in this case the verification of the implementation of a standard protocol by comparing it to a reference implementation. Powerful active learning methods are used because the oracle is the protocol implementation itself, which can respond to (produce output for) millions of inputs.

We find this particular collection of papers is representative of the field of Grammatical Inference, and we hope it is of interest to the general readership of Machine Learning as well as to specialists in Grammatical Inference.

Lastly, for this special issue, we required the help of more than 70 reviewers which we really would like to thank today. We also are grateful to Sudha Subramanian, Melissa Fearon, and Peter Flach for their technical help and appreciated their advice during the handling of this work.