Georgios Paliouras Yasubumi Sakakibara (Eds.)

Grammatical Inference: Algorithms and Applications

7th International Colloquium, ICGI 2004
Athens, Greece, October 11-13, 2004
Proceedings

Volume Editors

Georgios Paliouras
Institute of Informatics and Telecommunications,
National Center of Scientific Research (NCSR), "Demokritos",
PO Box. 60228, Ag. Paraskevi, Attiki, 15310, Greece,
E-mail: paliourg@iit.demokritos.gr

Yasubumi Sakakibara
Department of Biosciences and Informatics,
Keio University,
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan,
E-mail: yasu@bio.keio.ac.jp

Preface

The Seventh International Colloquium on Grammatical Inference (ICGI-2004) was held in the National Centre for Scientific Research "Demokritos", Athens, Greece on October 11-13th, 2004. ICGI-2004 was the seventh in a series of successful biennial international conferences in the area of grammatical inference. Previous meetings were held in Essex, U.K.; Alicante, Spain; Montpellier, France; Ames, Iowa, USA; Lisbon, Portugal; Amsterdam, The Netherlands. This series of conferences seeks to provide a forum for presentation and discussion of original research papers on all aspects of grammatical inference.

Grammatical Inference, the study of learning grammars from data, is an established research field in Artificial Intelligence, dating back to the 60s and has been extensively addressed by researchers in automata theory, language acquisition, computational linguistics, machine learning, pattern recognition, computational learning theory and neural networks. ICGI-2004 has emphasized on the multi-disciplinary nature of the research field and the diverse domains in which grammatical inference is being applied, such as natural language acquisition, computational biology, structural pattern recognition, information retrieval, Web mining, text processing, data compression and adaptive intelligent agents.

We have received 45 high-quality papers from 19 countries. The papers were reviewed by at least two - in most cases three - reviewers. In addition to the 20 full papers, 8 short papers that received positive comments from the reviewers were accepted to appear in a separate section of this volume. The topics of the accepted papers vary from theoretical results of learning algorithms to innovative applications of grammatical inference and from learning several interesting classes of formal grammars to estimations of probabilistic grammars.

In conjunction to ICGI-2004, a context-free grammar learning competition, named Omphalos, took place. In an invited paper in this volume, the organisers of the competition report on the peculiarities of such an endeavor and some interesting theoretical findings. Last but not least, we are honoured by the contributions of our invited speakers Prof. Dana Angluin, from Yale University, USA, and Prof. Enrique Vidal, from Universidade Politecnica de Valencia, Spain.

The editors would like to acknowledge the contribution of the Program Committee and the Additional Reviewers in reviewing the submitted papers and thank the Organizing Committee for their invaluable help in organizing the conference. Particularly, we would like to thank Colin de la Higuera, Menno van Zaannen, Georgios Petasis, Georgios Sigletos and Evangelia Alexopoulou for their additional voluntary service to the grammatical inference community, through this conference. We would also like to acknowledge the use of the Cyberchair software, by Borbala Online Conference Services, in the submission and reviewing process. Finally, we are grateful for the generous support and sponsorship of the conference by NCSR "Demokritos", the PASCAL and KDnet European Networks of Excellence, and Biovista: Corporate Intelligence in Biotechnology.

October 2004

Georgios Paliouras
Yasubumi Sakakibara

Contents

Invited Papers

Learning and Mathematics 1
Dana Angluin

Learning Finite-State Models for Machine Translation 3
Enrique Vidal and Francisco Casacuberta

The Omphalos Context-Free Grammar Learning Competition 15
Bradford Starkie, François Coste and Menno van Zaanen

Regular Papers

Mutually compatible and incompatible merges for the search of the smallest consistent DFA 27
John Abela, François Coste and Sandro Spina

Faster gradient descent training of Hidden Markov Models, using individual learning rate adaptation 39
Pantelis Bagos, Theodore Liakopoulos and Stavros Hamodrakas

Learning Mild Context-Sensitiveness: Toward understanding children's language learning 51
Leonor Becerra-Bonache and Takashi Yokomori

Learnability of Pregroup Grammars 63
Denis Béchet, Annie Foret and Isabelle Tellier

A Markovian Approach to the Induction of Regular String Distributions ... 75
Jérôme Callut and Pierre Dupont

Learning Node Selecting Tree Transducer from Completely Annotated Examples 88
Julien Carme, Aurélien Lemay and Joachim Niehren

Identifying Clusters from Positive Data 100
John Case, Sanjay Jain, Eric Martin, Arun Sharma and Frank Stephan

Introducing Domain and Typing Bias in Automata Inference 112
François Coste, Daniel Fredouille, Christopher Kermorvant and Colin de la Higuera

Analogical Equations in Sequences: Definition and Resolution 124

Arnaud Delhay and Laurent Miclet

Representing Languages by Learnable Rewriting Systems 136
Rémi Eyraud, Colin de la Higuera and Jean-Christophe Janodet

A Divide-And-Conquer Approach to Acquire Syntactic Categories 148
Pablo Gamallo, Gabriel P. Lopes and Joaquim F. Da Silva

Grammatical Inference using Suffix Trees 160
Jeroen Geertzen and Menno van Zaanen

Learning Probabilistic Finite Automata 172
Colin de la Higuera and Jose Oncina

Navigation pattern discovery using grammatical inference 184
Nikolaos Karampatziakis, Georgios Paliouras, Dimitrios Pierrakos and Panagiotis Stamatopoulos

A Corpus-Driven Context-Free Approximation of Head-Driven Phrase Structure Grammar 196
Hans-Ulrich Krieger

Partial learning using link grammars data 208
Erwan Moreau

eg-GRIDS: Context-Free Grammatical Inference from Positive Examples using Genetic Search 220
Georgios Petasis, Georgios Paliouras, Constantine Spyropoulos and Constantine Halatsis

The Boisdale Algorithm Identifying a Class of Unification Grammar in the Limit from Positive Data 232
Bradford Starkie and Henning Fernau

Learning Stochastic Deterministic Regular Languages 244
Franck Thollard and Alexander Clark

Polynomial Time Identification of Strict Deterministic Restricted One-Counter Automata in Some Class from Positive Data 256
Mitsuo Wakatsuki, Kiyoshi Teraguchi and Etsuji Tomita

Poster Papers

Learning Syntax From Function Words 268
David Brooks and Mark Lee

Running FCRPNI in efficient time for Piecewise and Right Piecewise Testable Languages 270

Antonio Cano, José Ruiz and Pedro García

Extracting minimum length Document Type Definitions is NP-hard 272
Henning Fernau

Learning distinguishable linear grammars from positive data 274
J.A. Laxminarayana, José M. Sempere and G. Nagaraja

Extending Incremental Learning of Context Free Grammars in Synapse ... 276
Katsuhiko Nakamura

Identifying Left-Right Deterministic Linear Languages 278
Jorge Calera-Rubio and Jose Oncina

Efficient Learning of k-reversible Context Free Grammars from Positive Structural Examples 280
Shinnosuke Seki and Satoshi Kobayashi

An analysis of examples and a search space for PAC learning of simple deterministic languages with membership queries 282
Yasuhiro Tajima, Yoshiyuki Kotani and Matsuaki Terada

Author Index 285