Shai Ben-David
John Case
Akira Maruoka (Eds.)

LNAI 3244

Algorithmic Learning Theory

**15th International Conference, ALT 2004
Padova, Italy, October 2004
Proceedings**

Springer

TP301.6-53
A465
2004

Shai Ben-David John Case
Akira Maruoka (Eds.)

Algorithmic Learning Theory

15th International Conference, ALT 2004
Padova, Italy, October 2-5, 2004
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Shai Ben-David
University of Waterloo, School of Conputer Science
Waterloo, Ontario, Canada
E-mail: shai@cs.uwaterloo.cs

John Case
University of Delaware
Department of Computer and Information Sciences
Newark, DE 19716, USA
E-mail: case@cis.udel.edu

Akira Maruoka
Tohoku University
Graduate School of Information Sciences
Sendai 980-8579, Japan
E-mail: maruoka@ecei.tohoku.ac.jp

Library of Congress Control Number: 2004113282

CR Subject Classification (1998): I.2.6, I.2.3, F.1, F.2, F.4, I.7

ISSN 0302-9743
ISBN 3-540-23356-3 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2004
Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 11326540 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 3244

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004.

Vol. 3244: S. Ben-David, J. Case, A. Maruoka (Eds.), Algorithmic Learning Theory. XIV, 505 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004.

Vol. 3215: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. XXX, 900 pages. 2004.

Vol. 3214: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. XXX, 1200 pages. 2004.

Vol. 3213: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. XXX, 1200 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004.

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004.

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004.

Vol. 3171: A.L.C. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence – SBIA 2004. XVII, 548 pages. 2004.

Vol. 3159: U. Visser, Intelligent Information Integration for the Semantic Web. XIV, 150 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004.

Vol. 3071: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World. XIII, 409 pages. 2004.

Vol. 3070: L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Artificial Intelligence and Soft Computing - ICAISC 2004. XXV, 1208 pages. 2004.

Vol. 3068: E. André, L. Dybkjær, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems. XII, 324 pages. 2004.

Vol. 3067: M. Dastani, J. Dix, A. El Fallah-Seghrouchni (Eds.), Programming Multi-Agent Systems. X, 221 pages. 2004.

Vol. 3066: S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymała-Busse (Eds.), Rough Sets and Current Trends in Computing. XX, 853 pages. 2004.

Vol. 3065: A. Lomuscio, D. Nute (Eds.), Deontic Logic in Computer Science. X, 275 pages. 2004.

Vol. 3060: A.Y. Tawfik, S.D. Goodwin (Eds.), Advances in Artificial Intelligence. XIII, 582 pages. 2004.

Vol. 3056: H. Dai, R. Srikant, C. Zhang (Eds.), Advances in Knowledge Discovery and Data Mining. XIX, 713 pages. 2004.

Vol. 3055: H. Christiansen, M.-S. Hacid, T. Andreasen, H.L. Larsen (Eds.), Flexible Query Answering Systems. X, 500 pages. 2004.

Vol. 3040: R. Conejo, M. Urretavizcaya, J.-L. Pérez-de-la-Cruz (Eds.), Current Topics in Artificial Intelligence. XIV, 689 pages. 2004.

Vol. 3035: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XII, 326 pages. 2004.

Vol. 3034: J. Favela, E. Menasalvas, E. Chávez (Eds.), Advances in Web Intelligence. XIII, 227 pages. 2004.

Vol. 3030: P. Giorgini, B. Henderson-Sellers, M. Winikoff (Eds.), Agent-Oriented Information Systems. XIV, 207 pages. 2004.

此为试读,需要完整PDF请访问: www.ertongbook.com

Vol. 3029: B. Orchard, C. Yang, M. Ali (Eds.), Innovations in Applied Artificial Intelligence. XXI, 1272 pages. 2004.

Vol. 3025: G.A. Vouros, T. Panayiotopoulos (Eds.), Methods and Applications of Artificial Intelligence. XV, 546 pages. 2004.

Vol. 3020: D. Polani, B. Browning, A. Bonarini, K. Yoshida (Eds.), RoboCup 2003: Robot Soccer World Cup VII. XVI, 767 pages. 2004.

Vol. 3012: K. Kurumatani, S.-H. Chen, A. Ohuchi (Eds.), Multi-Agnets for Mass User Support. X, 217 pages. 2004.

Vol. 3010: K.R. Apt, F. Fages, F. Rossi, P. Szeredi, J. Váncza (Eds.), Recent Advances in Constraints. VIII, 285 pages. 2004.

Vol. 2990: J. Leite, A. Omicini, L. Sterling, P. Torroni (Eds.), Declarative Agent Languages and Technologies. XII, 281 pages. 2004.

Vol. 2980: A. Blackwell, K. Marriott, A. Shimojima (Eds.), Diagrammatic Representation and Inference. XV, 448 pages. 2004.

Vol. 2977: G. Di Marzo Serugendo, A. Karageorgos, O.F. Rana, F. Zambonelli (Eds.), Engineering Self-Organising Systems. X, 299 pages. 2004.

Vol. 2972: R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, H. Sossa (Eds.), MICAI 2004: Advances in Artificial Intelligence. XVII, 923 pages. 2004.

Vol. 2969: M. Nickles, M. Rovatsos, G. Weiss (Eds.), Agents and Computational Autonomy. X, 275 pages. 2004.

Vol. 2961: P. Eklund (Ed.), Concept Lattices. IX, 411 pages. 2004.

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. X, 301 pages. 2004.

Vol. 2930: F. Winkler (Ed.), Automated Deduction in Geometry. VII, 231 pages. 2004.

Vol. 2926: L. van Elst, V. Dignum, A. Abecker (Eds.), Agent-Mediated Knowledge Management. XI, 428 pages. 2004.

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004.

Vol. 2915: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction. XIII, 558 pages. 2004.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. XX, 512 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. XVI, 1075 pages. 2003.

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. XV, 504 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003.

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. X, 215 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003.

Vol. 2871: N. Zhong, Z.W. Raś, S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems. XV, 697 pages. 2003.

Vol. 2854: J. Hoffmann, Utilizing Problem Structure in Planing. XIII, 251 pages. 2003.

Vol. 2843: G. Grieser, Y. Tanaka, A. Yamamoto (Eds.), Discovery Science. XII, 504 pages. 2003.

Vol. 2842: R. Gavaldá, K.P. Jantke, E. Takimoto (Eds.), Algorithmic Learning Theory. XI, 313 pages. 2003.

Vol. 2838: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Knowledge Discovery in Databases: PKDD 2003. XVI, 508 pages. 2003.

Vol. 2837: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Machine Learning: ECML 2003. XVI, 504 pages. 2003.

Vol. 2835: T. Horváth, A. Yamamoto (Eds.), Inductive Logic Programming. X, 401 pages. 2003.

Vol. 2821: A. Günter, R. Kruse, B. Neumann (Eds.), KI 2003: Advances in Artificial Intelligence. XII, 662 pages. 2003.

Vol. 2807: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. XIII, 426 pages. 2003.

Vol. 2801: W. Banzhaf, J. Ziegler, T. Christaller, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life. XVI, 905 pages. 2003.

Vol. 2797: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. XII, 281 pages. 2003.

Vol. 2792: T. Rist, R.S. Aylett, D. Ballin, J. Rickel (Eds.), Intelligent Virtual Agents. XV, 364 pages. 2003.

Vol. 2782: M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (Eds.), Cooperative Information Agents VII. XI, 345 pages. 2003.

Vol. 2780: M. Dojat, E. Keravnou, P. Barahona (Eds.), Artificial Intelligence in Medicine. XIII, 388 pages. 2003.

Vol. 2777: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. XIV, 746 pages. 2003.

Vol. 2752: G.A. Kaminka, P.U. Lima, R. Rojas (Eds.), RoboCup 2002: Robot Soccer World Cup VI. XVI, 498 pages. 2003.

Vol. 2741: F. Baader (Ed.), Automated Deduction – CADE-19. XII, 503 pages. 2003.

Vol. 2705: S. Renals, G. Grefenstette (Eds.), Text- and Speech-Triggered Information Access. VII, 197 pages. 2003.

Vol. 2703: O.R. Zaïane, J. Srivastava, M. Spiliopoulou, B. Masand (Eds.), WEBKDD 2002 - MiningWeb Data for Discovering Usage Patterns and Profiles. IX, 181 pages. 2003.

Vol. 2700: M.T. Pazienza (Ed.), Extraction in the Web Era. XIII, 163 pages. 2003.

Vol. 2699: M.G. Hinchey, J.L. Rash, W.F. Truszkowski, C.A. Rouff, D.F. Gordon-Spears (Eds.), Formal Approaches to Agent-Based Systems. IX, 297 pages. 2002.

Vol. 2691: V. Mařík, J.P. Müller, M. Pechoucek (Eds.), Multi-Agent Systems and Applications III. XIV, 660 pages. 2003.

Foreword

Algorithmic learning theory is mathematics about computer programs which learn from experience. This involves considerable interaction between various mathematical disciplines including theory of computation, statistics, and combinatorics. There is also considerable interaction with the practical, empirical fields of machine and statistical learning in which a principal aim is to predict, from past data about phenomena, useful features of future data from the same phenomena.

The papers in this volume cover a broad range of topics of current research in the field of algorithmic learning theory. We have divided the 29 technical, contributed papers in this volume into eight categories (corresponding to eight sessions) reflecting this broad range. The categories featured are Inductive Inference, Approximate Optimization Algorithms, Online Sequence Prediction, Statistical Analysis of Unlabeled Data, PAC Learning & Boosting, Statistical Supervised Learning, Logic Based Learning, and Query & Reinforcement Learning.

Below we give a brief overview of the field, placing each of these topics in the general context of the field. Formal models of automated learning reflect various facets of the wide range of activities that can be viewed as *learning*.

A first dichotomy is between viewing learning as an indefinite process and viewing it as a finite activity with a defined termination. Inductive Inference models focus on indefinite learning processes, requiring only eventual success of the learner to converge to a satisfactory conclusion.

When one wishes to predict future data, success can be enhanced by making some restrictive but true assumptions about the nature (or regularities) of the data stream. In the learning theory community, this problem is addressed in two different ways. The first is by assuming that the data to be predicted is generated by an operator that belongs to a restricted set of operators that is known to the learner a priori. The PAC model and some of the work under the Inductive Inference framework follow this path. Alternatively, one could manage without any such prior assumptions by relaxing the success requirements of the learner: rather than opting for some absolute degree of accuracy, the learner is only required to perform as well as any learner in some fixed family of learners. Thus, if the data is erratic or otherwise hard to predict, the learner can ignore poor accuracy as long as no member of the fixed reference family of learners can do no better. This is the approach taken by some Online Sequence Prediction models in the indefinite learning setting and, also, by most of the models of Statistical Learning in the finite horizon framework.

Boosting is a general technique that applies a given type of learner iteratively to improve its performance. Boosting approaches have been shown to be effective for a wide range of learning algorithms and have been implemented by a variety of methods.

A second dichotomy is between *Supervised* and *Un-Supervised* learning. The latter we refer to as *learning from Unlabeled Data*. In the first scenario, the data has the form of an example-label pairs. The learner is trained on a set of such pairs and then, upon seeing some fresh examples, has to predict their labels. In the latter model, the data points lack any such labeling, and the learner has to find some persistent regularities in the data, on the basis of the examples it has seen. Such regularities often take the form of partitioning the data into clusters of similar points, but in some cases take other forms, such as locating the boundaries of the support of the data generating distribution.

Many learning algorithms can be viewed as searching for an object that fits the given training data best. Such *optimization* tasks are often computationally infeasible. To overcome such computational hurdles, it is useful to apply algorithms that search for approximations to the optimal objects. The study of such algorithms, in the context of learning tasks, is the subject of our Approximate Optimization Algorithms session.

There is a large body of research that examines different *representations* of data and of learners' conclusions. This research direction is the focus of our Logic Based Learning session.

A final important dichotomy separates models of interactive learning from those that model passive learners. In the first type of learning scenarios the actions of the learner affect the training data available to it. In the Query Learning model this interaction takes the form of queries of certain (pre-indicated) type(s) that the learner can pose. Then the data upon which the learner bases its conclusions are the responses to these queries. The other model that addresses interactive learning is Reinforcement Learning, a model that assumes that the learner takes actions and receives *rewards* that are a function of these actions. These rewards in turn are used by the learner to determine its future actions.

August 2004

Shai Ben-David
John Case
Akira Marouka

Preface

This volume contains the papers presented at the 15th Annual International Conference on Algorithmic Learning Theory (ALT 2004), which was held in Padova (Italy) October 2–5, 2004. The main objective of the conference was to provide an interdisciplinary forum for discussing the theoretical foundations of machine learning as well as their relevance to practical applications. The conference was co-located with the 7th International Conference on Discovery Science (DS 2004) and the 11th Conference on String Processing and Information Retrieval (SPIRE 2004) under the general title "The Padova Dialogues 2004".

The volume includes 29 technical contributions that were selected by the program committee from 91 submissions. It also contains the invited lecture for ALT and DS 2004 presented by Ayumi Shinohara (Kyushu University, Fukuoka, Japan) on "String Pattern Discovery". Furthermore, this volume contains the ALT 2004 invited talks presented by Nicolò Cesa-Bianchi (Università degli Studi di Milano, Italy) on "Applications of Regularized Least Squares in Classification Problems", and by Luc De Raedt (Universität Freiburg, Germany) on "Probabilistic Inductive Logic Programming".

Additionally, it contains the invited lecture presented by Esko Ukkonen (University of Helsinki, Finland) on "Hidden Markov Modelling Techniques for Haplotype Analysis" (joint invited talk with DS 2004). Moreover, this volume includes the abstract of the joint invited lecture with DS 2004 presented by Pedro Domingos (University of Washington, Seattle, USA) on "Learning, Logic, and Probability: A Unified View".

Finally, this volume contains the papers of the research tutorials on *Statistical Mechanical Methods in Learning* by Toshiyuki Tanaka (Tokyo Metropolitan University, Japan) on "Statistical Learning in Digital Wireless Communications", by Yoshiyuki Kabashima and Shinsuke Uda (Tokyo Institute of Technology, Japan), on "A BP-Based Algorithm for Performing Bayesian Inference in Large Perceptron-like Networks", and by Manfred Opper and Ole Winther on "Approximate Inference in Probabilistic Models".

ALT has been awarding the *E. Mark Gold Award* for the most outstanding paper by a student author since 1999. This year the award was given to Hubie Chen for his paper "Learnability of Relatively Quantified Generalized Formulas", co-authored by Andrei Bulatov and Victor Dalmau.

This conference was the 15th in a series of annual conferences established in 1990. Continuation of the ALT series is supervised by its steering committee consisting of: Thomas Zeugmann (Hokkaido Univ., Sapporo, Japan), Chair, Arun Sharma (Queensland Univ.of Technology, Australia), Co-chair, Naoki Abe (IBM T.J. Watson Research Center, USA), Klaus Peter Jantke (DFKI, Germany), Roni Khardon (Tufts Univ., USA), Phil Long (National Univ. of Singapore), Hiroshi Motoda (Osaka University, Japan), Akira Maruoka (Tohoku Univ., Japan), Luc De Raedt (Albert-Ludwigs-Univ., Germany), Takeshi Shi-

nohara (Kyushu Institute of Technology, Japan), and Osamu Watanabe (Tokyo Institute of Technology, Japan).

We would like to thank all individuals and institutions who contributed to the success of the conference: the authors for submitting papers, the invited speakers for accepting our invitation and lending us their insight into recent developments in their research areas, as well as the sponsors for their generous financial and logistical support.

We would also like to thank Thomas Zeugmann for assisting us via his experience in the publication of previous ALT proceedings, for providing the ALT 2004 logo, and for managing the ALT 2004 Web site. We are very grateful to Frank Balbach who developed the ALT 2004 electronic submission page.

Furthermore, we would like to express our gratitude to all program committee members for their hard work in reviewing the submitted papers and participating in online discussions. We are also grateful to the external referees whose reviews made a considerable contribution to this process.

We are also grateful to the DS 2004 chairs Einoshin Suzuki (PC Chair, Yokohama National University, Japan) and Setsuo Arikawa (Conference Chair, Kyushu University, Japan) for their effort in coordinating with ALT 2004, and to Massimo Melucci (University of Padova, Italy) for his excellent work as the local arrangements chair. Last but not least, Springer provided excellent support in preparing this volume.

August 2004

Shai Ben-David
John Case
Akira Maruoka

Organization

Conference Chair

Akira Marouka	Tohoku University, Sendai, Japan

Program Committee

Shai Ben-David	Univ. of Waterloo, Canada, Co-chair
John Case	Univ. of Delaware, USA, Co-chair
Nader Bshouty	Technion, Israel
Michael Collins	MIT, USA
Sanjoy Dasgupta	UC San Diego, USA
Peter Flach	Univ. of Bristol, UK
Steffen Lange	DFKI, Saarbrücken, Germany)
Jon Langford	TTI Chicago, USA
Gabor Lugosi	UPF, Barcelona, Spain
Rüdiger Reischuk	Univ. at Lübeck, Germany
Rocco Servedio	Columbia Univ., USA
Arun Sharma	Queensland Univ. of Technology, Brisbane, Australia
Ayumi Shinohara	Kyushu Univ., Japan
Eiji Takimoto	Tohoku Univ., Japan
Sandra Zilles	Univ. Kaiserslautern, Germany

Local Arrangements

Massimo Melucci	University of Padova, Italy

Subreferees

Kazuyuki Amano
Pedro Felzenswalb
Kouichi Hirata
Klaus P. Jantke
Adam Klivans
Stephen Kwek
Martin Memmel
Jochen Nessel
Ryan O'Donnell
Daniel Reidenbach
Yoshifumi Sakai
Gilles Stoltz
Rolf Wiehagen

Sponsoring Institutions

Institute for Theoretical Computer Science, University at Lübeck
Division of Computer Science, Hokkaido University

此为试读,需要完整PDF请访问: www.ertongbook.com

Remembering Carl Smith, 1950–2004

Sadly, Carl Smith passed away 10:30PM, July 21, 2004. He had had a 1.5 year battle with an aggressive brain tumor. He fought this battle with calm optimism, dignity, and grace. He is survived by his wife, Patricia, his son, Austin, and his sister, Karen Martin.

Carl was very active in the algorithmic or computational learning communities, especially in the inductive inference subarea which applies recursive function theory techniques.

I first met Carl when I interviewed for my faculty position at SUNY/Buffalo in the Spring of 1973. He was then a graduate student there and told me he was interested in recursive function theory. After I joined there, he naturally became my Ph.D. student, and that's when we both began working on inductive inference. We spent a lot of time together, pleasantly blurring the distinction between the relationships of friendship and advisor-student.

After Buffalo, Carl had faculty positions at Purdue and, then, the University of Maryland.

Carl had a very productive career. He was a master collaborator working with many teams around the world. Of course he also produced a number of papers about inductive inference by teams — as well as papers about anomalies, queries, memory limitation, procrastination, and measuring mind changes by counting down from notations for ordinals. I had the reaction to some of his papers of wishing I'd thought of the idea. This especially struck me with his 1989 *TCS* paper (with Angluin and Gasarch) in which it is elegantly shown that the learning of some classes of tasks can be done only sequentially after or in parallel with other classes.

Carl played a significant leadership role in theoretical computer science. In 1981, with the help of Paul Young, Carl organized the Workshop on Recursion Theoretic Aspects of Computer Science. This became the well known, continuing series of Computational Complexity conferences. Carl provided an improvement in general theoretical computer science funding level during his year as Theory Program Director at NSF. He was involved, in many cases from the beginning, in the COLT, AII, ALT, EuroCOLT, and DS conferences, as a presenter of papers, as a member of many of their program committees and, in some cases, steering committees. He spearheaded the development of COLT's Mark Fulk Award for best student papers and managed the finances.

Carl was very likable. He had a knack for finding funding to make good things happen. He was a good friend and colleague. He is missed.

August 2004

John Case

Table of Contents

Statistical Supervised Learning

Statistical Analysis of Unlabeled Data

Online Sequence Prediction

Approximate Optimization Algorithms

Logic Based Learning

Query and Reinforcement Learning

Tutorial Papers

String Pattern Discovery

Ayumi Shinohara

Department of Informatics, Kyushu University 33, Fukuoka 812-8581, JAPAN
PRESTO, Japan Science and Technology Agency
ayumi@i.kyushu-u.ac.jp

Abstract. Finding a good pattern which discriminates one set of strings from the other set is a critical task in knowledge discovery. In this paper, we review a series of our works concerning with the string pattern discovery. It includes theoretical analyses of learnabilities of some pattern classes, as well as development of practical data structures which support efficient string processing.

1 Introduction

A huge amount of text data or sequential data are accessible in these days. Especially, the growing popularity of Internet have caused an enormous increase of text data in the last decade. Moreover, a lot of biological sequences are also available due to various genome sequencing projects. Many of these data are stored as raw strings, or in semi-structured form such as HTML and XML, which are essentially strings. *String pattern discovery*, where one is interested in extracting patterns which characterizes a set of strings or sequential data, has attracted widespread attentions [1,36,13,24,12,3,4,30]. Discovering a *good rule* to separate two given sets, often referred as *positive examples* and *negative examples*, is a critical task in Machine Learning and Knowledge Discovery. In this paper, we review a series of our works for finding best string patterns efficiently, together with their theoretical background.

Our motivations originated in the development of a machine discovery system BONSAI [31], that produces a decision tree over regular patterns with alphabet indexing, from given positive set and negative set of strings. The core part of the system is to generate a decision tree which classifies positive examples and negative examples as correctly as possible. For that purpose, we have to find a *pattern* that maximizes the goodness according to the entropy information gain measure, recursively at each node of trees. In the initial implementation, a pattern associated with each node is restricted to a *substring pattern*, due to the limit of computation time. In order to allow more expressive patterns while keeping the computation in reasonable time, we have introduced various techniques gradually [15,17,16,7,19,8,21,32,6,18]. Essentially, they are combinations of pruning heuristics in the huge search space without sacrificing the optimality of the solution, and efficient data structures which support various string processing.

S. Ben-David, J. Case, A. Maruoka (Eds.): ALT 2004, LNAI 3244, pp. 1–13, 2004.
© Springer-Verlag Berlin Heidelberg 2004

此为试读,需要完整PDF请访问: www.ertongbook.com