Intelligent Systems Reference Library, Volume 21

Editors-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Prof. Lakhmi C. Jain
University of South Australia
Adelaide
Mawson Lakes Campus
South Australia 5095
Australia
E-mail: Lakhmi.jain@unisa.edu.au

Further volumes of this series can be found on our homepage: springer.com

Vol. 5. George A. Anastassiou
Intelligent Mathematics: Computational Analysis, 2010
ISBN 978-3-642-17097-3

Vol. 6. Ludmila Dymowa
Soft Computing in Economics and Finance, 2011
ISBN 978-3-642-17718-7

Vol. 7. Gerasimos G. Rigatos
Modelling and Control for Intelligent Industrial Systems, 2011
ISBN 978-3-642-17874-0

Vol. 8. Edward H.Y. Lim, James N.K. Liu, and
Raymond S.T. Lee
Knowledge Seeker – Ontology Modelling for Information
Search and Management, 2011
ISBN 978-3-642-17915-0

Vol. 9. Menahem Friedman and Abraham Kandel
Calculus Light, 2011
ISBN 978-3-642-17847-4

Vol. 10. Andreas Tolk and Lakhmi C. Jain
Intelligence-Based Systems Engineering, 2011
ISBN 978-3-642-17930-3

Vol. 11. Samuli Niiranen and Andre Ribeiro (Eds.)
Information Processing and Biological Systems, 2011
ISBN 978-3-642-19620-1

Vol. 12. Florin Gorunescu
Data Mining, 2011
ISBN 978-3-642-19720-8

Vol. 13. Witold Pedrycz and Shyi-Ming Chen (Eds.)
Granular Computing and Intelligent Systems, 2011
ISBN 978-3-642-19819-9

Vol. 14. George A. Anastassiou and Oktay Duman
Towards Intelligent Modeling: Statistical Approximation Theory, 2011
ISBN 978-3-642-19825-0

Vol. 15. Antonino Freno and Edmondo Trentin
Hybrid Random Fields, 2011
ISBN 978-3-642-20307-7

Vol. 16. Alexiei Dingli
Knowledge Annotation: Making Implicit Knowledge Explicit, 2011
ISBN 978-3-642-20322-0

Vol. 17. Crina Grosan and Ajith Abraham
Intelligent Systems, 2011
ISBN 978-3-642-21003-7

Vol. 18. Achim Zielesny
From Curve Fitting to Machine Learning, 2011
ISBN 978-3-642-21279-6

Vol. 19. George A. Anastassiou
Intelligent Systems: Approximation by Artificial Neural Networks, 2011
ISBN 978-3-642-21430-1

Vol. 20. Lech Polkowski
Approximate Reasoning by Parts, 2011
ISBN 978-3-642-22278-8

Vol. 21. Igor Chikalov
Average Time Complexity of Decision Trees, 2011
ISBN 978-3-642-22660-1

Igor Chikalov

Average Time Complexity of Decision Trees

Springer

Dr. Igor Chikalov
Mathematical and Computer Sciences
and Engineering Division
4700 King Abdullah University of Science
and Technology
Thuwal 23955-6900
Kingdom of Saudi Arabia
E-mail: igor.chikalov@kaust.edu.sa

ISBN 978-3-642-22660-1 e-ISBN 978-3-642-22661-8

DOI 10.1007/978-3-642-22661-8

Intelligent Systems Reference Library ISSN 1868-4394

Library of Congress Control Number: 2011932437

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To my wife Julia and our daughter Svetlana, for always believing in me and loving me, no matter what.

Foreword

It is our great pleasure to welcome a new book "Average Time Complexity of Decision Trees" by Igor Chikalov. This book is devoted to the study of average time complexity (average depth and weighted average depth) of decision trees over finite and infinite sets of attributes. It contains exact and approximate algorithms for decision tree optimization, and bounds on minimum average time complexity of decision trees. The average time complexity measures can be used in searching for the minimum description length of induced data models. Hence, there exist relationships of the presented results with the minimum description length principle (MDL).

The considered applications include the study of average depth of decision trees for Boolean functions from closed classes, the comparison of results of the performance of greedy heuristics for average depth minimization with optimal decision trees constructed by dynamic programming algorithm, and optimization of decision trees for the corner point recognition problem from computer vision.

The book can be interesting for researchers working on time complexity of algorithms and specialists in machine learning.

The author, Igor Chikalov, received his PhD degree in 2002 from Nizhny Novgorod State University, Russia. During nine years he was working for Intel Corp. as a senior software engineer/research scientist in machine learning applications to the control and diagnostic problems of semiconductor manufacturing. Since 2009 he is a senior research scientist in King Abdullah University of Science and Technology, Saudi Arabia. His current research interests include supervised machine learning and extensions of dynamic programming to the optimization of decision trees and decision rules.

The author deserves the highest appreciation for his outstanding work.

Mikhail Moshkov
May 2011 Andrzej Skowron

Preface

The monograph is devoted to theoretical and experimental study of decision trees with a focus on minimizing the average time complexity. The study resulted in upper and lower bounds on the minimum average time complexity of decision trees for identification problems. Previously known bounds from information theory are extended to the case of identification problem with an arbitrary set of attributes. Some examples of identification problems are presented giving an evidence that the obtained bounds are close to unimprovable. In addition to universal bounds, we study effectiveness of representing several types of discrete functions in a form of decision trees. In particular, for each closed class of Boolean functions we obtained upper bounds on the average depth of decision trees implementing functions from this class.

The monograph also studies the problem of algorithm design for optimal decision tree construction. An algorithm based on dynamic programming is proposed that describes a set of optimal trees and allows for subsequent optimization on other criteria. Experimental results show applicability of the algorithm to real-life applications that are represented by decision tables containing dozens of attributes and several thousands of objects.

Beside individual identification problems, infinite classes of problems are considered. It describes necessary conditions on such classes in order to have polynomial complexity algorithms for optimal decision tree construction.

The presented results can be of interest for researchers in test theory, rough set theory and machine learning. Some results may be considered for including in graduate courses on discrete mathematics and computer science. The monograph can be used as a reference to prior results in the area.

Some results were obtained in collaboration with Dr. Mikhail Moshkov and published in joint papers [51, 52, 53, 54, 56]. I am heartily thankful to Dr. Moshkov for help in preparing this book.

I would like to acknowledge and extend my gratitude to Victor Eruhimov for fruitful discussions about applications of decision trees and Dr. Andrzej Skowron for constructive criticism and suggestions for improvement of the book.

Thuwal, Saudi Arabia,
April 2011 Igor Chikalov

Contents