

Lecture Notes in Artificial Intelligence

9047

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

More information about this series at <http://www.springer.com/series/1244>

Alexander Gammerman · Vladimir Vovk
Harris Papadopoulos (Eds.)

Statistical Learning and Data Sciences

Third International Symposium, SLDS 2015
Egham, UK, April 20–23, 2015
Proceedings

Editors

Alexander Gammerman
University of London
Egham, Surrey
UK

Harris Papadopoulos
Frederick University
Nicosia
Cyprus

Vladimir Vovk
University of London
Egham, Surrey
UK

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Artificial Intelligence

ISBN 978-3-319-17090-9

ISBN 978-3-319-17091-6 (eBook)

DOI 10.1007/978-3-319-17091-6

Library of Congress Control Number: 2015935220

LNCS Sublibrary: SL7 – Artificial Intelligence

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

In memory of Alexey Chervonenkis

Preface

This volume contains the Proceedings of the Third Symposium on Statistical Learning and Data Sciences, which was held at Royal Holloway, University of London, UK, during April 20–23, 2015. The original idea of the Symposium on Statistical Learning and Data Sciences is due to two French academics – Professors Mireille Gettler Summa and Myriam Touati – from Paris Dauphine University. Back in 2009 they thought that a “bridge” was required between various academic groups that were involved in research on Machine Learning, Statistical Inference, Pattern Recognition, Data Mining, Data Analysis, and so on; a sort of multilayer bridge to connect those fields. This is reflected in the symposium logo with the Passerelle Simone-de-Beauvoir bridge. The idea was implemented and the First Symposium on Statistical Learning and Data Sciences was held in Paris in 2009. The event was indeed a great “bridge” between various communities with interesting talks by J.-P. Benzecri, V. Vapnik, A. Chervonenkis, D. Hand, L. Bottou, and many others. Papers based on those talks were later presented in a volume of the *Modulad* journal and separately in a post-symposium book entitled *Statistical Learning and Data Sciences*, published by Chapman & Hall, CRC Press. The second symposium, which was equally successful, was held in Florence, Italy, in 2012.

Over the last 6 years since the first symposium, the progress in the theory and applications of learning and data mining has been very impressive. In particular, the arrival of technologies for collecting huge amounts of data has raised many new questions about how to store it and what type of analytics are able to handle it – what is now known as Big Data. Indeed, the sheer scale of the data is very impressive – for example, the Large Hadron Collider computers have to store 15 petabytes a year (1 petabyte = 10^{15} bytes). Obviously, handling this requires the usage of distributed clusters of computers, streaming, parallel processing, and other technologies. This volume is concerned with various modern techniques, some of which could be very useful for handling Big Data.

The volume is divided into five parts. The first part is devoted to two invited papers by Vladimir Vapnik. The first paper, “Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer,” is a further development of his research on learning with privileged information, with a special attention to the knowledge representation problem. The second, “Statistical Inference Problems and their Rigorous Solutions,” suggests a novel approach to pattern recognition and regression estimation. Both papers promise to become milestones in the developing field of statistical learning.

The second part consists of 16 papers that were accepted for presentation at the main event, while the other three parts reflect new research in important areas of statistical learning to which the symposium devoted special sessions. Specifically the special sessions included in the symposium’s program were:

- Special Session on Conformal Prediction and its Applications (CoPA 2015), organized by Harris Papadopoulos (Frederick University, Cyprus), Alexander Gammernan (Royal Holloway, University of London, UK), and Vladimir Vovk (Royal Holloway, University of London, UK).

- Special Session on New Frontiers in Data Analysis for Nuclear Fusion, organized by Jesus Vega (Asociacion EURATOM/CIEMAT para Fusion, Spain).
- Special Session on Geometric Data Analysis, organized by Fionn Murtagh (Goldsmith College London, UK).

Overall, 36 papers were accepted for presentation at the symposium after being reviewed by at least two independent academic referees. The authors of these papers come from 17 different countries, namely: Brazil, Canada, Chile, China, Cyprus, Finland, France, Germany, Greece, Hungary, India, Italy, Russia, Spain, Sweden, UK, and USA.

A special session at the symposium was devoted to the life and work of Alexey Chervonenkis, who tragically died in September 2014. He was one of the founders of modern Machine Learning, a beloved colleague and friend. All his life he was connected with the Institute of Control Problems in Moscow, over the last 15 years he worked at Royal Holloway, University of London, while over the last 7 years he also worked for the Yandex Internet company in Moscow. This special session included talks in memory of Alexey by Vladimir Vapnik – his long standing colleague and friend – and by Alexey’s former students and colleagues.

We are very grateful to the Program and Organizing Committees, the success of the symposium would have been impossible without their hard work. We are indebted to the sponsors: the Royal Statistical Society, the British Computer Society, the British Classification Society, Royal Holloway, University of London, and Paris Dauphine University. Our special thanks to Yandex for their help and support in organizing the symposium and the special session in memory of Alexey Chervonenkis. This volume of the proceedings of the symposium is also dedicated to his memory. Rest in peace, dear friend.

February 2015

Alexander Gammerman
Vladimir Vovk
Harris Papadopoulos



Organization

General Chairs

Alexander Gammerman, UK
Vladimir Vovk, UK

Organizing Committee

Zhiyuan Luo, UK
Mireille Summa, France
Yuri Kalnishkan, UK

Myriam Touati, France
Janet Hales, UK

Program Committee Chairs

Harris Papadopoulos, Cyprus
Xiaohui Liu, UK
Fionn Murtagh, UK

Program Committee Members

Vineeth Balasubramanian, India
Giacomo Boracchi, Italy
Paula Brito, Portugal
Léon Bottou, USA
Lars Carlsson, Sweden
Jane Chang, UK
Frank Coolen, UK
Gert de Cooman, Belgium
Jesus Manuel de la Cruz, Spain
Jose-Carlos Gonzalez-Cristobal, Spain
Anna Fukshansky, Germany
Barbara Hammer, Germany

Shenshyang Ho, Singapore
Carlo Lauro, Italy
Guang Li, China
David Lindsay, UK
Henrik Linusson, Sweden
Hans-J. Lenz, Germany
Ilia Nouretdinov, UK
Matilde Santos, Spain
Victor Solovyev, Saudi Arabia
Jesus Vega, Spain
Rosanna Verde, Italy

Contents

Invited Papers

Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer.	3
<i>Vladimir Vapnik and Rauf Izmailov</i>	
Statistical Inference Problems and Their Rigorous Solutions	33
<i>Vladimir Vapnik and Rauf Izmailov</i>	

Statistical Learning and its Applications

Feature Mapping Through Maximization of the Atomic Interclass Distances.	75
<i>Savvas Karatsiolis and Christos N. Schizas</i>	
Adaptive Design of Experiments for Sobol Indices Estimation Based on Quadratic Metamodel	86
<i>Evgeny Burnaev and Ivan Panin</i>	
GoldenEye++: A Closer Look into the Black Box	96
<i>Andreas Henelius, Kai Puolamäki, Isak Karlsson, Jing Zhao, Lars Asker, Henrik Boström, and Panagiotis Papapetrou</i>	
Gaussian Process Regression for Structured Data Sets.	106
<i>Mikhail Belyaev, Evgeny Burnaev, and Yermek Kapushev</i>	
Adaptive Design of Experiments Based on Gaussian Processes	116
<i>Evgeny Burnaev and Maxim Panov</i>	
Forests of Randomized Shapelet Trees	126
<i>Isak Karlsson, Panagotis Papapetrou, and Henrik Boström</i>	
Aggregation of Adaptive Forecasting Algorithms Under Asymmetric Loss Function.	137
<i>Alexey Romanenko</i>	
Visualization and Analysis of Multiple Time Series by Beanplot PCA	147
<i>Carlo Drago, Carlo Natale Lauro, and Germana Scepi</i>	
Recursive SVM Based on TEDA	156
<i>Dmitry Kangin and Plamen Angelov</i>	

RDE with Forgetting: An Approximate Solution for Large Values of k with an Application to Fault Detection Problems	169
<i>Clauber Gomes Bezerra, Bruno Sielly Jales Costa, Luiz Affonso Guedes, and Plamen Parvanov Angelov</i>	
Sit-to-Stand Movement Recognition Using Kinect	179
<i>Erik Acorn, Nikos Dipsis, Tamar Pincus, and Kostas Stathis</i>	
Additive Regularization of Topic Models for Topic Selection and Sparse Factorization	193
<i>Konstantin Vorontsov, Anna Potapenko, and Alexander Plavin</i>	
Social Web-Based Anxiety Index's Predictive Information on S&P 500 Revisited	203
<i>Rapheal Olaniyan, Daniel Stamate, and Doina Logofatu</i>	
Exploring the Link Between Gene Expression and Protein Binding by Integrating mRNA Microarray and ChIP-Seq Data	214
<i>Mohsina Mahmuda Ferdous, Veronica Vinciotti, Xiaohui Liu, and Paul Wilson</i>	
Evolving Smart URL Filter in a Zone-Based Policy Firewall for Detecting Algorithmically Generated Malicious Domains	223
<i>Konstantinos Demertzis and Lazaros Iliadis</i>	
Lattice-Theoretic Approach to Version Spaces in Qualitative Decision Making	234
<i>Miguel Couceiro and Tamás Waldhauser</i>	
Conformal Prediction and its Applications	
A Comparison of Three Implementations of Multi-Label Conformal Prediction	241
<i>Huazhen Wang, Xin Liu, Ilia Nouruddinov, and Zhiyuan Luo</i>	
Modifications to p-Values of Conformal Predictors	251
<i>Lars Carlsson, Ernst Ahlberg, Henrik Boström, Ulf Johansson, and Henrik Linusson</i>	
Cross-Conformal Prediction with Ridge Regression	260
<i>Harris Papadopoulos</i>	
Handling Small Calibration Sets in Mondrian Inductive Conformal Regressors	271
<i>Ulf Johansson, Ernst Ahlberg, Henrik Boström, Lars Carlsson, Henrik Linusson, and Cecilia Sönströd</i>	

Conformal Anomaly Detection of Trajectories with a Multi-class Hierarchy	281
<i>James Smith, Ilia Nouretdinov, Rachel Craddock, Charles Offer, and Alexander Gammerman</i>	
Model Selection Using Efficiency of Conformal Predictors	291
<i>Ritvik Jaiswal and Vineeth N. Balasubramanian</i>	
Confidence Sets for Classification.	301
<i>Christophe Denis and Mohamed Hebiri</i>	
Conformal Clustering and Its Application to Botnet Traffic	313
<i>Giovanni Cherubin, Ilia Nouretdinov, Alexander Gammerman, Roberto Jordaney, Zhi Wang, Davide Papini, and Lorenzo Cavallaro</i>	
Interpretation of Conformal Prediction Classification Models.	323
<i>Ernst Ahlberg, Ola Spjuth, Catrin Hasselgren, and Lars Carlsson</i>	
New Frontiers in Data Analysis for Nuclear Fusion	
Confinement Regime Identification Using Artificial Intelligence Methods . . .	337
<i>G.A. Rattá and Jesús Vega</i>	
How to Handle Error Bars in Symbolic Regression for Data Mining in Scientific Applications.	347
<i>A. Murari, E. Peluso, M. Gelfusa, M. Lungaroni, and P. Gaudio</i>	
Applying Forecasting to Fusion Databases	356
<i>Gonzalo Farias, Sebastián Dormido-Canto, Jesús Vega, and Norman Díaz</i>	
Computationally Efficient Five-Class Image Classifier Based on Venn Predictors	366
<i>Jesús Vega, Sebastián Dormido-Canto, F. Martínez, I. Pastor, and M.C. Rodríguez</i>	
SOM and Feature Weights Based Method for Dimensionality Reduction in Large Gauss Linear Models	376
<i>Fernando Pavón, Jesús Vega, and Sebastián Dormido Canto</i>	
Geometric Data Analysis	
Assigning Objects to Classes of a Euclidean Ascending Hierarchical Clustering	389
<i>Brigitte Le Roux and Frédérik Cassor</i>	
The Structure of Argument: Semantic Mapping of US Supreme Court Cases	397
<i>Fionn Murtagh and Mohsen Farid</i>	

Supporting Data Analytics for Smart Cities: An Overview of Data Models and Topology	406
<i>Patrick E. Bradley</i>	
Manifold Learning in Regression Tasks.	414
<i>Alexander Bernstein, Alexander Kuleshov, and Yury Yanovich</i>	
Random Projection Towards the Baire Metric for High Dimensional Clustering	424
<i>Fionn Murtagh and Pedro Contreras</i>	
Optimal Coding for Discrete Random Vector.	432
<i>Bernard Colin, François Dubeau, and Jules de Tibeiro</i>	
Author Index	443