# Studies in Computational Intelligence

Volume 605

**Series editor**

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

## *About this Series*

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/7092

Stan Matwin · Jan Mielniczuk
Editors

# Challenges in Computational Statistics and Data Mining

Springer

*Editors*
Stan Matwin
Faculty of Computer Science
Dalhousie University
Halifax, NS
Canada

Jan Mielniczuk
Institute of Computer Science
Polish Academy of Sciences
Warsaw
Poland

and

Warsaw University of Technology
Warsaw
Poland

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-319-18780-8 ISBN 978-3-319-18781-5 (eBook)
DOI 10.1007/978-3-319-18781-5

Library of Congress Control Number: 2015940970

Springer Cham Heidelberg New York Dordrecht London


Printed on acid-free paper

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

# Preface

This volume contains 19 research papers belonging, roughly speaking, to the areas of computational statistics, data mining, and their applications. Those papers, all written specifically for this volume, are their authors' contributions to honour and celebrate Professor Jacek Koronacki on the occcasion of his 70th birthday. The volume is the brain-child of Janusz Kacprzyk, who has managed to convey his enthusiasm for the idea of producing this book to us, its editors. Books related and often interconnected topics, represent in a way Jacek Koronacki's research interests and their evolution. They also clearly indicate how close the areas of computational statistics and data mining are.

Mohammad Reza Bonyadi and Zbigniew Michalewicz in their article "Evolutionary Computation for Real-world Problems" describe their experience in applying Evolutionary Algorithms tools to real-life optimization problems. In particular, they discuss the issues of the so-called multi-component problems, the investigation of the feasible and the infeasible parts of the search space, and the search bottlenecks.

Susanne Bornelöv and Jan Komorowski "Selection of Significant Features Using Monte Carlo Feature Selection" address the issue of significant features detection in Monte Carlo Feature Selection method. They propose an alternative way of identifying relevant features based on approximation of permutation p-values by normal p-values and they compare its performance with the performance of built-in selection method.

In his contribution, Łukasz Dębowski "Estimation of Entropy from Subword Complexity" explores possibilities of estimating block entropy of stationary ergodic process by means of word complexity i.e. approximating function $f(k|w)$ which for a given string $w$ yields the number of distinct substrings of length $k$. He constructs two estimates and shows that the first one works well only for iid processes with uniform marginals and the second one is applicable for much broader class of so-called properly skewed processes. The second estimator is used to corroborate Hilberg's hypothesis for block length no larger than 10.

Maik Döring, László Györfi and Harro Walk "Exact Rate of Convergence of Kernel-Based Classification Rule" study a problem in nonparametric classification

concerning excess error probability for kernel classifier and introduce its decomposition into estimation error and approximation error. The general formula is provided for the approximation and, under a weak margin condition, its tight version.

Michał Dramiński in his exposition "ADX Algorithm for Supervised Classification" discusses a final version of rule-based classifier ADX. It summarizes several years of the author's research. It is shown in experiments that inductive methods may work better or on par with popular classifiers such as Random Forests or Support Vector Machines.

Olgierd Hryniewicz "Process Inspection by Attributes Using Predicted Data" studies an interesting model of quality control when instead of observing quality of inspected items directly one predicts it using values of predictors which are easily measured. Popular data mining tools such as linear classifiers and decision trees are employed in this context to decide whether and when to stop the production process.

Szymon Jaroszewicz and Łukasz Zaniewicz "Székely Regularization for Uplift Modeling" study a variant of uplift modeling method which is an approach to assess the causal effect of an applied treatment. The considered modification consists in incorporating Székely regularization into SVM criterion function with the aim to reduce bias introduced by biased treatment assignment. They demonstrate experimentally that indeed such regularization decreases the bias.

Janusz Kacprzyk and Sławomir Zadrożny devote their paper "Compound Bipolar Queries: A Step Towards an Enhanced Human Consistency and Human Friendliness" to the problem of querying of databases in natural language. The authors propose to handle the inherent imprecision of natural language using a specific fuzzy set approach, known as compound bipolar queries, to express imprecise linguistic quantifiers. Such queries combine negative and positive information, representing required and desired conditions of the query.

Miłosz Kadziński, Roman Słowiński, and Marcin Szeląg in their paper "Dominance-Based Rough Set Approach to Multiple Criteria Ranking with Sorting-Specific Preference Information" present an algorithm that learns ranking of a set of instances from a set of pairs that represent user's preferences of one instance over another. Unlike most learning-to-rank algorithms, the proposed approach is highly interactive, and the user has the opportunity to observe the effect of their preferences on the final ranking. The algorithm is extended to become a multiple criteria decision aiding method which incorporates the ordinal intensity of preference, using a rough-set approach.

Marek Kimmel "On Things Not Seen" argues in his contribution that frequently in biological modeling some statistical observations are indicative of phenomena which logically should exist but for which the evidence is thought missing. The claim is supported by insightful discussion of three examples concerning evolution, genetics, and cancer.

Mieczysław Kłopotek, Sławomir Wierzchoń, Robert Kłopotek and Elżbieta Kłopotek in "Network Capacity Bound for Personalized Bipartite PageRank" start from a simplification of a theorem for personalized random walk in an unimodal graph which is fundamental to clustering of its nodes. Then they introduce a novel

notion of Bipartite PageRank and generalize the theorem for unimodal graphs to this setting.

Marzena Kryszkiewicz devotes her article "Dependence Factor as a Rule Evaluation Measure" to the presentation and discussion of a new evaluation measure for evaluation of associations rules. In particular, she shows how the dependence factor realizes the requirements for interestingness measures postulated by Piatetsky-Shapiro, and how it addresses some of the shortcomings of the classical certainty factor measure.

Adam Krzyżak "Recent Results on Nonparametric Quantile Estimation in a Simulation Model" considers a problem of quantile estimation of the random variable $m(X)$ where $X$ has a given density by means of importance sampling using a regression estimate of $m$. It is shown that such yields a quantile estimator with a better asymptotic properties than the classical one. Similar results are valid when recursive Robbins-Monro importance sampling is employed.

The contribution of Błażej Miasojedov, Wojciech Niemiro, Jan Palczewski, and Wojciech Rejchel in "Adaptive Monte Carlo Maximum Likelihood" deal with approximation to the maximum likelihood estimator in models with intractable constants by adaptive Monte Carlo method. Adaptive importance sampling and a new algorithm which uses resampling and MCMC is investigated. Among others, asymptotic results, such that consistency and asymptotic law of the approximative ML estimators of the parameter are proved.

Jan Mielniczuk and Paweł Teisseyre in "What do We Choose When We Err? Model Selection and Testing for Misspecified Logistic Regression Revisited" consider common modeling situation of fitting logistic model when the actual response function is different from logistic one and provide conditions under which Generalized Information Criterion is consistent for set $t^*$ of the predictors pertaining to the Kullback-Leibler projection of true model $t$. The interplay between $t$ and $t^*$ is also discussed.

Mirosław Pawlak in his contribution "Semiparametric Inference in Identification of Block-Oriented Systems" gives a broad overview of semiparametric statistical methods used for identification in a subclass of nonlinear-dynamic systems called block oriented systems. They are jointly parametrized by finite-dimensional parameters and an infinite-dimensional set of nonlinear functional characteristics. He shows that using semiparametric approach classical nonparametric estimates are amenable to the incorporation of constraints and avoid high-dimensionality/high-complexity problems.

Marina Sokolova and Stan Matwin in their article "Personal Privacy Protection in Time of Big Data" look at some aspects of data privacy in the context of big data analytics. They categorize different sources of personal health information and emphasize the potential of Big Data techniques for linking of these various sources. Among others, the authors discuss the timely topic of inadvertent disclosure of personal health information by people participating in social networks discussions.

Jerzy Stefanowski in his article "Dealing with Data Difficulty Factors while Learning from Imbalanced Data" provides a thorough review of the approaches to learning classifiers in the situation when one of the classes is severely

underrepresented, resulting in a skewed, or imbalanced distribution. The article presents all the existing methods and discusses their advantages and shortcomings, and recommends their applicability depending on the specific characteristics of the imbalanced learning task.

In his article James Thompson "Data Based Modeling" builds a strong case for a data-based modeling using two examples: one concerning portfolio management and second being the analysis of hugely inadequate action of American health service to stop AIDS epidemic. The main tool in the analysis of the first example is an algorithm called MaxMedian Rule developed by the author and L. Baggett.

We are very happy that we were able to collect in this volume so many contributions intimately intertwined with Jacek's research and his scientific interests. Indeed, he is one of the authors of Monte Carlo Feature Selection system which is discussed here and widely contributed to nonparametric curve estimation and classification (subject of Döring et al. and Krzyżak's paper). He started his career with research in optimization and stochastic approximation—the themes being addressed in Bonyadi and Michalewicz as well as in Miasojedow et al. papers. He held long-lasting interests in Statistical Process Control discussed by Hryniewicz. He also has, as the contributors to this volume and his colleagues from Rice University, Thompson and Kimmel, keen interests in methodology of science and stochastic modeling.

Jacek Koronacki has been not only very active in research but also has generously contributed his time to the Polish and international research communities. He has been active in the International Organization of Standardization and in the European Regional Committee of the Bernoulli Society. He has been and is a longtime director of Institute of Computer Science of Polish Academy of Sciences in Warsaw. Administrative work has not prevented him from being an active researcher, which he continues up to now. He holds unabated interests in new developments of computational statistics and data mining (one of the editors vividly recalls learning about Székely distance, also appearing in one of the contributed papers here, from him). He has co-authored (with Jan Ćwik) the first Polish textbook in statistical Machine Learning. He exerts profound influence on the Polish data mining community by his research, teaching, sharing of his knowledge, refereeing, editorial work, and by exercising his very high professional standards. His friendliness and sense of humour are appreciated by all his colleagues and collaborators. In recognition of all his achievements and contributions, we join the authors of all the articles in this volume in dedicating to him this book as an expression of our gratitude. Thank you, Jacku; dziękujemy.

We would like to thank all the authors who contributed to this endeavor, and the Springer editorial team for perfect editing of the volume.

Ottawa, Warsaw, March 2015 Stan Matwin
Jan Mielniczuk

# Contents