Math. Appl. **7** (2018), 1–2

THE MULTIFACETED IMPACT OF STATISTICAL METHODOLOGY AND THEORY IN DATA SCIENCE

ERNEST FOKOUÉ AND BORIS BRIMKOV

The vast amount of recorded data and the exponential growth of computational power in the last two decades have enabled the extraction and processing of information in unprecedented ways. This has led to the emergence of data science as a bridge between data mining, algorithm design, modeling, machine learning, visualization, and artificial intelligence, in an effort to understand and gain deeper insights from data in various forms. In turn, this has caused a resurgence of statistical science, due to the prominent role of statistical methodology and statistical theory in data science.

This special issue of Mathematics for Applications features a compilation of contributions dealing with the multifaceted impact of statistical methodology and theory in data science. The works featured in this special issue have initially been reported at the conferences UP-STAT 2016 "Data Science, Statistical Practice, and Education" and UP-STAT 2017 "Data Science, Statistics, and the Environment." After a careful selection and a rigorous review process, seven substantially extended papers out of over 100 works presented at both conferences have been selected for publication in this issue. These works contribute to the theoretical foundations of data science and explore mathematical models and methods for solving problems in applied fields.

One of the most prominent ways in which statistical science has enriched data science is through the introduction of regularization methods, such as ridge regression and LASSO, which are used to handle ultra-high dimensional or multicollinear datasets. In the first paper of this special issue, the authors Y. Zhang, J. Thakar, D.J. Topham, A.R. Falsey, D. Zeng and X. Qiu construct two useful equivalence relationships for regularized regression – one for efficiently fitting the concurrent functional regression model, and the second for efficiently solving weighted principal component regression.

The development of ensemble learning methods has complemented traditional model selection and regularization approaches, giving experimenters and end-users a rich arsenal of powerful statistical learning methods. Spearheaded by techniques such as random forest, bagging, adaptive boosting, gradient boosting, and random subspace learning, ensemble learning methods have proven valuable in the study and analysis of increasingly larger and more complex data sets. In the second paper, M. Elshrif and E. Fokoué present a novel adaptation of the random subspace learning approach to regression analysis and classification of high-dimension, low-sample-size data.

The ubiquitous presence of statistical ideas and methods in artificial intelligence has made statistical machine learning a field of central importance in the unfolding of the data science era. In particular, kernel methods, as encountered in support vector machines, Gaussian process learning machines, kernel k-means, and kernel principal component analysis, have given a new life to reproducing kernel Hilbert spaces, allowing unprecedented extensions and applications to existing statistical machine learning methods. The authors of the third paper, G. Olinto and E. Fokoué, propose a new approach for kernelized cost-sensitive listwise ranking, give the theoretical framework for the algorithm, and compare it with other methods through computational experiments. The fourth paper by L. Khinkis, M. Crotzer, and A. Oprisan addresses the case in nonlinear regression analysis when the residual sum of squares has multiple local minima, which complicates finding the global minimum and adversely affects the reliability of the relevant statistical methods. The authors propose a solution through the use of an equidistant function.

The Bayesian statistical learning paradigm, spearheaded by the ubiquity of cheap and powerful computing resources, has been a great asset to data science. The regularization framework enjoys a natural relationship with the Bayesian school of thought, and Bayesian model averaging enjoys an optimality property that is valuable in the theory of linear regression modeling. The modeling flexibility inherent in the Bayesian paradigm has been harnessed in fields like biology, biostatistics, medicine, image processing, computer vision, and text mining. The authors of the fifth paper, A. LaLonde and T. Love, use a Dirichlet process mixture model implemented through a Bayesian Markov chain Monte Carlo approach to improve estimation of the overall effect of mercury on child neurodevelopment.

The role of statistical methodology in education, often aimed at understanding various aspects of classroom learning, has steadily increased and led to the birth of educational data mining. In the sixth paper, S. E. Mason and E. M. Reid examine the relationship between anxiety and performance in a statistics class, and find that self-reported anxiety levels are negatively correlated with both grade expectations and final course grades, while several other factors lead to decreased student anxiety.

Traditional statistical methods along with statistical machine learning methods have helped establish sports analytics as both a household and corporate approach to various aspects of sports. The authors of the last paper, G. C. Phelan and J. T. Whelan, describe a hierarchical Bayesian version of the Bradley–Terry model suitable for use in ranking and prediction problems, such as major league baseball.

It is our hope that the wide variety of themes and topics addressed in this volume will satisfy our readers' interest and prompt further research. We would like to thank the Editor-in-Chief of Mathematics for Applications, Prof. Josef Šlapal, for the publication of this special issue. We are also grateful to the reviewers for their helpful comments, and to all authors who submitted articles to this special issue.

Ernest Fokoué
School of Mathematical Sciences
Rochester Institute of Technology
NY, USA

Boris Brimkov
Department of Computational
and Applied Mathematics
Rice University
TX, USA