Model Assisted Statistics and Applications 12 (2017) 193–194
DOI 10.3233/MAS-170404
IOS Press

Editorial

Machine Learning in Applied Statistics

Stan Lipovetsky[a] and Jong-Min Kim[b,*]
[a]*MASA Co-Editor-in-Chief, Minneapolis, Minnesota, MN, USA*
[b]*Morris, Minnesota, MN, USA*

This special issue of Model Assisted Statistics and Applications (MASA) focused on knowing how current machine learning methods can be applied to diverse statistics areas. We have ten papers about the recent machine learning developments and applications, including survey sampling, biostatistics, bioinformatics, genetics, time series analysis, and technology forecasting.

The issue starts with a research work "Variance Estimation by Multivariate Imputation Methods in Complex Survey Designs" by Profs. Jong-Min Kim, Kee-Jae Lee, and Wonkuk Kim describing application of modern machine learning methods to survey sampling imputation.

Drs. Insuk Sohn, Sujong Kim, Profs. Jae Won Lee, Ja-Yong Koo, and Dr. Junsu Ko present a paper "Identifying Novel NF-kB-Regulated Immune Genes in the Human Genome using a Discrete Kernel Structured Support Vector Machine". This paper develops a Discrete Kernel Structured Support Vector Machine (DSSVM) and applied this method to the promoters of 58 known NF-kB-target genes to find characteristic patterns of transcription factor binding sites (TFBSs) in their promoters.

Profs. Jooyong Shim and Changha Hwang present a paper "Kernel-Based Orthogonal Quantile Regression Model". This paper proposes a kernel-based quantile regression model that effectively considers errors on both input and output variables.

Profs. Hye-Seung Lee and Jeffrey P. Krischer present a paper "A New Framework for Prediction and Variable Selection for Uncommon Events in a Large Prospective Cohort Study". This paper describes a framework illustrated with an application of featuring high-dimensional variable selection in a large prospective cohort study.

Profs. Jong-Min Kim, Jea-Bok Ryu, Seung-Joo Lee, and Sunghae Jun present a paper "Penalized Regression Models for Patent Keyword Analysis". The authors analyze the patent keywords extracted from the patent documents using ridge regression, least absolute shrinkage and selection operator, elastic net, and random forest. In addition, to show how the research could be applied to real problem efficiently, the authors carry out a case study of Apple technology.

Miss Soyeon Park and Prof. Wonkuk Kim present a paper "Multifactor Dimensionality Reduction Method Based on Area under Receiver Operating Characteristic Curve". The authors explain multifactor dimensionality reduction (MDR) method which is a machine learning algorithm to detect nonlinear interactions, and compare performance of the standard with the modified multifactor dimensionality reduction method in which the best model is selected by the area under receiver operating characteristic curve (ROC) and cross-validation consistency of the area under ROC curve.

Drs. Dipankar Mitra and Ranjit Kumar Paul present a paper "Hybrid Time-Series Models for Forecasting Agricultural Commodity Prices". The authors apply the hybrid methodology namely ARIMA-GARCH and ARIMA-ANN for modelling and forecasting of wholesale potato price in Agra market of India. The comparison of forecast performance among the ARIMA, GARCH, ARIMA-GARCH and ARIMA-ANN hybrid models shows that the hybrid models perform better with respect to minimum of MAPE and RMSE values.

*Guest Editor for this special issue.

ISSN 1574-1699/17/$35.00

Drs. Yoonha Choi, Joshua Babiarz, Ed Tom, Giulia C. Kennedy, and Jing Huang present a paper "Repurposing Kinship Coefficients as a Sample Integrity Method for Next Generation Sequencing Data in a Clinical Setting". The authors first describe the general concept of kinship coefficients and focus on the novel adaptations on feature (i.e. variants and/or SNPs) selection utilizing expressed variants to make it suitable for clinical setting.

Drs. Kath Bogie, Yifan Xu, Junheng Ma, Adah Zhang, Yuanyuan Wang, Kristine Zanotti, and Prof. Jiayang Sun present a paper "Associations between Diagnostic Patternsand Stages in Ovarian Cancer". The authors identify diagnostic patterns from a complex multivariate data source and investigate their association with ovarian cancer stages, given disease symptoms, severity, and patients' self-advocacy. The analytic approach uses tree-based models providing a versatile robust approach without requiring restrictive parametric assumptions about the model that may not fit the data well or is subject to challenges in dealing with missing categorical values.

Dr. Haoda Fu and Prof. Jin Zhou present a paper "A Unified Approach for Subgroup Identification and Individualized Treatment Recommendation with Applications to Randomized Control Trials and Observational Studies". The authors explain that the precision medicine is important in the new era of medical product development. The authors describe the limitations of traditional subgroup identification methods and propose a general framework which connects the subgroup identification methods and individualized treatment recommendation rules. The proposed framework permits to handle two or more treatments from both randomized control trials and observation studies.

The problems discussed in the current issue enrich both statistical and biostatistics research theoretical modeling and their practical applications in a wide variety of problems of modern science, business, and medicine.

We appreciate Dr. Ying Lu, Professor of Biomedical Data Science from Stanford University, for reviewing the last three wonderful research papers in this issue which were presented at the conference of DahShu 2017: Data Science and Computational Health in Feb 20–22, 2017 held in San Francisco, California (http://dahshu.org/events/cph2017/).