Behaviormetrika (2024) 51:287–289
https://doi.org/10.1007/s41237-023-00222-1

PREFACE

Issues in behavioral data science

Alfonso Iodice D'Enza[1] · Angelos Markos[2] · Koji Kurihara[3]

Published online: 9 January 2024

Analyzing data about different aspects of human behavior is a valuable and challenging task and it represents the general aim of behavioral data science. Both behavioral science and data science are interdisciplinary fields. In particular, behavioral science ranges from economics and finance to psychology and sociology, up to health-related behaviors; similarly, under the umbrella of data science are statistics, machine learning, computer science, to name a few.

Therefore, behavioral data science requires a faceted approach, where specific knowledge domains must drive the choice and the development of methodological tools: while this is an important driver in any scientific field, in behavioral data science it becomes mandatory. When the data analysis goal is to understand human behavior, obtaining explanations is key, no black boxes can be used or trusted. Domain specific knowledge is of course essential throughout the learning pipeline, from pre-processing and feature engineering to the interpretation of the results. However, it is at the same time important to develop tools that support domain experts and ease their interpretation of the results: visualization tools can be crucial in this respect.

This special issue features a range of papers that fit the above description: the contributions explore complex behaviors from multiple angles and show how behavioral data science blends elements from finance, education, healthcare, sociology, and text analysis, all through the lens of sophisticated data science methods, with a special focus on the interpretability of the results. A rough taxonomy of the articles in the special issue is: (i) contributions that tailor methods to specific application fields; (ii) contributions that present methodological enhancements to extend applicability and explainability of specific data science methods.

Papers of the former type are by Richert and Buch (2023), Palazzo et al. (2023), Ingusci et al. (2023) and Krazinger et al. (2023).

✉ Alfonso Iodice D'Enza
iodicede@unina.it

[1] University of Naples Federico II, Naples, Italy

[2] Democritus University of Thrace, Komotini, Greece

[3] Kyoto Women's University, Kyoto, Japan

The article by Richert and Buch (2023) addresses the challenge of missing data in market-implied volatility of European swaptions. Due to the illiquidity of various swaption instruments, significant gaps in data often occur, making it difficult to calibrate standard models accurately. The authors propose a novel approach using variational autoencoders to infer the full structure of the implied volatility cube from limited data. This method, tested for robustness on both synthetic and real market data, shows promising results in accurately calibrating stochastic volatility models and is also effective for practical financial applications like setting up delta-neutral portfolios for hedging.

The article by Palazzo et al. (2023) emphasizes the importance of understanding financial knowledge as a latent variable and uses statistical models to profile individuals' awareness of financial market risks. The proposed approach combines Item Response Theory (IRT) models with Archetypal Analysis (AA) to find homogeneous groups based on knowledge levels, assisting policymakers and decision-makers.

The study by Kranzinger et al. (2023), standing at the intersection of health sciences and data science, assesses the applicability of machine learning models for classifying sleep stages. It not only validates these models across diverse populations but also paves the way for future improvements, showcasing the evolving landscape of machine learning in healthcare.

Ingusci et al. (2023) follow a component-based approach to validate the Italian version of the Life Crafting Scale. The findings offer substantial insights into the construct of life crafting, particularly within educational contexts.

The second part of the special issue consists of proposals that aim at enhancing complex methods by increasing interpretability and the visualization of the results.

The article by de Rooij (2023) tackles the complexity of handling multiple binary response variables by introducing a novel algorithm for logistic reduced rank regression. The integration of triplots for visualization clarifies complex statistical models, merging sophisticated analysis with user-friendly interpretation.

Focusing on the interpretability of random forest models, Szepannek and von Holt (2023) delve into concepts like representative and surrogate trees. The study offers an insightful look into striking a balance between the complexity of models and their comprehensibility, an essential endeavor in an age dominated by 'black box' models.

The article by Babaeic and Giudici (2023) presents InstanceSHAP, an innovative method refining the SHapley Additive exPlanations approach. By offering more focused and interpretable Shapley values, it becomes especially relevant in the context of risk assessment in financial models.

The contribution by (Giordani and Kiers 2023) introduces a technique for archetypal analysis on datasets with missing entries, using a weighting system for null values and weighted least squares. This method shows effectiveness in handling incomplete data, compared to existing approaches.

Reflecting the dynamic nature of text data, Preis and Schwaar (2023) introduce a multinomial change-point model to identify structural changes in text, thereby enhancing document classification systems. This approach underlines the need for adaptive methodologies in contemporary text analysis.

Funding The authors have not disclosed any funding.

Declarations

Conflict of interest The authors have not disclosed any competing interests.

References

Babaeic G, Giudici P (2023) InstanceSHAP: an instance-based estimation approach for Shapley values. Behaviormetrika. https://doi.org/10.1007/s41237-023-00208-z

de Rooij M (2023) A new algorithm and a discussion about visualization for logistic reduced rank regression. Behaviormetrika. https://doi.org/10.1007/s41237-023-00204-3

Giordani P, Kiers HAL (2023) Weighted least squares for archetypal analysis with missing data. Behaviormetrika. https://doi.org/10.1007/s41237-023-00220-3

Ingusci E, Angelelli M, Sternativo GA et al (2023) A higher-order life crafting scale validation using PLS-CCA: the Italian version. Behaviormetrika. https://doi.org/10.1007/s41237-023-00209-y

Kranzinger S, Baron S, Kranzinger C et al (2023) Generalisability of sleep stage classification based on interbeat intervals: validating three machine learning approaches on self-recorded test data. Behaviormetrika. https://doi.org/10.1007/s41237-023-00199-x

Palazzo L, Iannario M, Palumbo F (2023) Integrated assessment of financial knowledge through a latent profile analysis. Behaviormetrika. https://doi.org/10.1007/s41237-023-00217-y

Preis A, Schwaar S (2023) Change point detection in text data. Behaviormetrika. https://doi.org/10.1007/s41237-023-00207-0

Richert I, Buch R (2023) Interpolation of missing swaption volatility data using variational autoencoders. Behaviormetrika. https://doi.org/10.1007/s41237-023-00213-2

Szepannek G, von Holt BH (2023) Can't see the forest for the trees. Behaviormetrika. https://doi.org/10.1007/s41237-023-00205-2