International Journal of Data Science and Analytics (2023) 16:403–406
https://doi.org/10.1007/s41060-023-00465-x

EDITORIAL

Theoretical and practical data science and analytics: challenges and solutions

Carson K. Leung[1] · Gabriella Pasi[2] · Li Wang[3]

Accepted: 25 September 2023 / Published online: 6 October 2023

Abstract
Big data have become a core technology for providing innovative solutions in numerical applications and services in many fields. Embedded in these big data is valuable information and knowledge. This calls for data science and analytics, which has emerged as an important paradigm for driving the new economy and domains (e.g., Internet of Things, social and mobile networks, cloud computing), reforming classic disciplines (e.g., telecommunications, biology, health and social science), as well as upgrading core business and economic activity. In this article, we focus on both theoretical and practical data science and analytics. We summarize and highlight some of its challenges and solutions, which are covered in the eight articles in the current Special Issue on "theoretical and practical data science and analytics."

Keywords Data science · Data analytics · Advanced analytics · Theory · Practice

1 Introduction

Nowadays, very large volumes of valuable data can be generated or collected at a rapid speed from a wide variety of rich data sources in numerous real-life applications. Embedded in these big data is valuable information and knowledge. Hence, big data can be considered as new oil. This calls for *data science and analytics* to (a) analyze and mine big data and (b) discover the valuable information and knowledge from the big data.

In general, *data science* [1–4] can be considered as a concept that historically proposed within the statistics and mathematics community with focus on data analysis. Nowadays, it has evolved beyond statistic and mathematics and has included areas like data mining and machine learning. It can now be considered as a new interdisciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge—including statistics, mathematics, informatics, computing, communication, data management, sociology, to study domain data following "data science thinking." Sometimes, data science can be considered as a encompassing of principles, problem definitions, algorithms, and processes for non-trivial extraction and discovery of useful patterns from large data sets. It involves the process of collecting, preparing, managing, analyzing, explaining, and disseminating the data and analysis results.

The current Special Issue discusses some challenges on both theoretical and practical aspects of data science and analytics, as well as solutions to tackle these challenges. On the one hand, solutions for the theoretical data science and analytics focus on foundations and theoretical developments of data science and analytics. These include: advanced analytics and knowledge discovery methods; computer vision and pattern recognition; data science foundations and theories; large-scale databases, big data processing, distributed processing, and ethical analytics; machine, deep and/or statistical learning-based algorithms; mathematics and statistics for data science and analytics; model explainability and provenance; optimization theories and methods; survey and review; theories and methods for evaluation, explanation, visualization, and presentation; as well as understanding data characteristics and complexities. On the other hand, solutions for the practical data science and analytics focus on (a) applications and best practices of data science and analytics across various disciplines and domains (e.g., business, government, health and medical science, natural sciences and engineering, social sciences and humanities); (b) examinations of inspiring results to policy-makers, end-users, or practitioners of

✉ Carson K. Leung
Carson.Leung@UManitoba.ca

[1] University of Manitoba, Winnipeg, MB, Canada

[2] University of Milano-Bicocca, Milan, Italy

[3] Taiyuan University of Technology, Taiyuan, Shanxi, China

data science and analytics; as well as (c) practical solutions to new research challenges motivated by the specific needs and characteristics of application areas. These include: business, economic, environmental, social impact modeling; cloud, crowd, online, mobile and distributed data analytics; deployment, management and policy-making; domain-specific data science and analytics practice (e.g., business, enterprise, environmental, financial, government, health, and/or medical analytics); ethics, social issues, privacy, trust, fairness and bias; operationalizable infrastructures, platforms, and tools; real-world applications, case studies and demonstrations; as well as reflections and lessons for better practice.

2 Challenges and solutions for theoretical and practical data science and analytics

Nowadays, very large volumes of valuable data can be generated or collected at a rapid speed from a wide variety of rich data sources in numerous real-life applications. Embedded in these big data is valuable information and knowledge. Hence, big data can be considered as new oil. This calls for *data science and analytics* to (a) analyze and mine big data and (b) discover the valuable information and knowledge from the big data.

This Special Issue also serves as a journal track of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), which provides an international forum for the discussion of up-to-date high-quality research results in all areas related to theoretical and practical data science and analytics. The conference features its strong interdisciplinary synergy between statistics, computing and information/intelligence sciences, and cross-domain interactions between academia and business for data science and analytics. Authors of the accepted submissions for this Special Issue have opportunities to present their work at the journal track of the conference. As such, only those articles (including research articles, survey, position and/or vision articles) that (a) are appropriate for giving conference presentations while (b) meeting the typical journal paper quality are included. Consequently, after rigorous reviews by independent experienced international reviewers, only eight refereed articles were selected for inclusion in this Special Issue. They cover different viewpoints of the challenges and solutions for theoretical or practical data science and analytics. Here, we give an overview of these challenges and solutions.

2.1 Theoretical data science and analytics

Event data—which record when events happen—are essential in various real-world scenarios. One common approach for modeling such events is using a temporal point process, which is a mathematical framework for estimating when the next event will occur based on the history of past events. Dubey et al. [5] observed that, although there have been advances in using neural networks to model these point processes (which surpassing traditional statistical models like the Hawkes process and achieving state-of-the-art results in predicting event times), neural point process models have a limitation. For example, they struggle to provide a measure of uncertainty in their predictions. Having a way to quantify uncertainty in event modeling is crucial for making informed decisions in many practical and critical applications. Hence, Dubey et al. introduced a lightweight point process model called the Bayesian Neural Hawkes Process (BNHP), which combines the uncertainty modeling capabilities of Bayesian methods with the generalization capabilities of neural networks to predict event occurrence times. Moreover, they enhanced BNHP with the ability to handle both the time and location aspects of events, incorporating spatio-temporal modeling. The resulting model can predict not only when an event will happen but also where it will occur with an associated level of uncertainty. Experimental results on various simulated and real-world datasets demonstrated that BNHP significantly improves both prediction accuracy and the quantification of uncertainty when it comes to event prediction.

In functional data analysis (FDA), each observation is a function that constitutes an infinite dimensional object. Hernández et al. [6] noticed that analyzing functional outliers (e.g., magnitude outliers, shape outliers) has become critical. Hence, they introduced a strategy for tackling the issue of identifying unusual patterns in functional data. Their strategy relies on creating a compact and consistent representation of functions using Reproducing Kernel Hilbert Spaces (RKHS). They established a depth metric based on density kernels that possess favorable characteristics. In addition, they also tackled the difficulties linked to estimating the depth of the density kernel. To evaluate the effectiveness of their functional depth measure in identifying outliers, they conducted a Monte Carlo simulation across various scenarios. The results demonstrated the practicality of their method and showcased how it can effectively pinpoint outliers in the analysis of mortality rate curves.

The integration of machine learning (ML) into organizations often involves employing various ML software components. When data scientists construct ML systems from these components, they may encounter practical requirements that extend beyond the well-known challenges of ML (e.g., data engineering and parameter optimization). As such, they may be expected to swiftly identify ML system configurations that strike a suitable balance across multiple performance criteria while also making these options understandable for non-technical users. This poses a challenge for those data scientists who have limited ML experience, necessitating a concept to assist them in finding suitable combinations of ML software. Observing that existing approaches (e.g., AutoML systems) either lack responsiveness or struggle to optimize across diverse performance criteria, Villanueva Zacarias et al. [7] introduced a concept for recommending ML solutions called AssistML (which include software systems equipped with ML models) as an alternative for predictive use cases. Their AssistML collects and pre-processes metadata from existing ML solutions to rapidly identify reusable ML solutions for new use cases. Evaluation results on two illustrative use cases demonstrated that AssistML could recommend ML solutions aligned with user performance preferences within seconds. It provides data scientists with simpler and intuitively explained ML solutions in significantly less time than the existing AutoML.

Observing that prior research on Fourier transform requires calculating closed-form solutions for continuous random variables, Sorvisto [8] established a discrete version of the characteristic function designed for discrete random variables and created computational techniques for determining this discrete characteristic function for various widely recognized discrete random variables. Moreover, a precise definition of the Fourier transform for a probability mass function is provided, which outlines the process rigorously. Furthermore, Sorvisto elucidated how to reverse this process, allowing the retrieval of the probability mass function of a discrete random variable, given a method for calculating the characteristic function.

2.2 Practical data science and analytics

Tree-structured techniques are widely recognized as potent tools for tasks involving classification and regression. In the present era, there is a growing interest in employing machine learning for functional data across various domains. However, when it comes to tree-based models tailored for functional variables, Hael [9] observed that the options are somewhat limited. Consequently, there is a demand to introduce an effective classifier within this dynamic field. In response, Hael expanded the traditional binary tree approach into a functional context, giving rise to the functional classification tree (FCT) approach. It focuses on the supervised classification of functional data, taking into account functional covariates and scenarios with multiple response classes. To process the data, two data-driven methods—namely, Fourier transformation and functional derivation—are employed. Hael also utilized conditional recursive partitioning techniques and functional permutation tests to uncover the underlying independence structure among the functional inputs. Moreover, the Bonferroni P-value adjustment and certain hyperparameters are incorporated to make informed decisions about splitting and controlling the growth of the tree. Simulation results demonstrated that the proposed FCT method outperforms existing approaches in terms of both computational efficiency and classification accuracy. Application of this method to an electrocardiogram dataset showcased its utility and advantages: It enhances the comprehensibility and informativeness of medical data classification tasks.

In recent years, significant advancements have been made in generative modeling for time series data, with generative adversarial networks (GANs) based on deep recurrent or convolutional neural networks being the predominant models. Many existing GANs were designed for generating time series data primarily focus on preserving correlations over time. Ahmed and Schmidt-Thieme [10] observed that, although these models can capture long-term dependencies to some extent, their ability to allocate varying levels of attention across different time steps can be limited. Hence, they introduced an approach called SparseGAN, which is based on sparse self-attention mechanisms within GANs. SparseGAN enables attention-driven, long-memory modeling for the generation of both regular and irregular time series data through a learned embedding space. Moreover, SparseGAN leads to a more informative representation for generating time series while utilizing the original data for guidance and supervision. Experimental results on both synthetic and real-world datasets demonstrated that forecasting models trained on data generated by SparseGAN perform similarly to models trained on real data for both regularly and irregularly sampled time series. The results also demonstrated that SparseGAN outperforms the current state-of-the-art models when dealing with limited data resources. Furthermore, SparseGAN introduces a method to generate realistic synthetic time series data by leveraging both long-term structural and temporal information.

Machine learning—particularly with the advancements in deep learning—has displayed significant potential in the analysis of time series data. However, in many scenarios, there exists additional information that has the potential to enhance predictions. This can be crucial for datasets originating from sources (e.g., sensor networks, which include valuable information about sensor locations). In these cases, spatial information can be effectively leveraged by representing it through graph structures, alongside the sequential time series data. Bloemheuvel et al. [11] observed that, although recent progress has been made in adapting deep learning techniques to graphs for various tasks, there has been relatively limited adaptation of these methods for time series tasks. Most of the efforts in this area have primarily centered on time series forecasting with short sequence lengths. These architectures may be less suitable for regression or classification tasks where the prediction depends not only on the most recent values but also on the entire history of the time series. In response to this gap, they introduced TISER-GCN as a graph convolutional network (GCN) architecture designed specifically for time series extrinsic regression (TSER) of long multivariate time series. Evaluation results on two seismic datasets containing earthquake waveforms showed a reduction in mean squared error (MSE) when compared to the best-performing baseline models. Moreover, the results also showed that TISER-GCN achieves similar performance while requiring only half the input size of the baseline models.

Xiao et al. [12] observed that perplexity serves as a critical parameter in the dimensionality reduction algorithm known as t-distributed stochastic neighbor embedding (t-SNE). Hence, they investigated the connection between t-SNE perplexity and various graph layout evaluation metrics—including graph stress—preserved neighborhood information and visual inspection. Their investigation revealed that a low perplexity is associated with a relatively higher normalized stress, indicating better preservation of neighborhood information with higher precision but at the cost of reduced global structural information. To address this trade-off, Xiao et al. proposed a method to estimate an appropriate perplexity value. This estimation can be based on either a modified standard t-SNE approach or the sklearn Barnes-Hut t-SNE method. Experimental results on a collection of benchmark datasets demonstrated the effectiveness and user-friendliness of their method.

3 Conclusion

Nowadays, in the era of big data, very large volumes of valuable data can be generated or collected at a rapid speed from a wide variety of rich data sources in numerous real-life applications. Embedded in these big data are valuable information and knowledge that can be discovered by data science and analytics. In this article, we focused on both theoretical and practical data science and analytics. We summarized and highlighted some of its challenges and solutions, which are covered in the eight articles in the current Special Issue on "theoretical and practical data science and analytics."

Acknowledgments This Special Issue would not have been possible without the help and effort of many people and organizations. We thank the editors (both the Founding Editor-in-Chief Longbing Cao and the current Editor-in-Chief João Gama) for the International Journal of Data Science and Analytics, as well as staff at Springer Nature (especially, Subbulakshmi Devanathan, Danielle Lupo, Priya Verma) for their support. We also express our thanks to all authors who have contributed to this Special Issue, and all reviewers who have provided constructive comments and suggestions to these authors.

Declarations

Conflict of interest The authors declare they have no financial interests.

References

1. Cao, L.: Data science and analytics: a new era. Int. J. Data Sci. Anal. **1**, 1–2 (2016)
2. Özsu, M.T.: Data science - a systematic treatment. Commun. ACM **66**(7), 106–116 (2023)
3. Cao, L.: Data science: challenges and directions. Commun. ACM **60**(8), 59–68 (2017)
4. Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. **50**(3), 43 (2017)
5. Dubey, M., Palakkadavath, R., Srijith, P.K.: Bayesian neural Hawkes process for event uncertainty prediction. Int. J. Data Sci. Anal. (2023). https://doi.org/10.1007/s41060-023-00443-3
6. Hernández, N., Muñoz, A., Martos, G.: Density kernel depth for outlier detection in functional data. Int. J. Data Sci. Anal. (2023). https://doi.org/10.1007/s41060-023-00420-w
7. Villanueva Zacarias, A.G., Reimann, P., Weber, C., Mitschang, B.: AssistML: an approach to manage, recommend and reuse ML solutions. Int. J. Data Sci. Anal. (2023). https://doi.org/10.1007/s41060-023-00417-5
8. Sorvisto, D.: Applications of the discrete-time Fourier transform to data analysis. Int. J. Data Sci. Anal. (2023). https://doi.org/10.1007/s41060-023-00409-5
9. Hael, M.A.: Unbiased recursive decision tree for supervised functional data classification with applying on electrocardiogram signals. Int. J. Data Sci. Anal. (2023). https://doi.org/10.1007/s41060-023-00410-y
10. Ahmed, N., Schmidt-Thieme, L.: Sparse self-attention guided generative adversarial networks for time-series generation. Int. J. Data Sci. Anal. (2023). https://doi.org/10.1007/s41060-023-00416-6
11. Bloemheuvel, S., van den Hoogen, J., Jozinović, D., Michelini, A., Atzmueller, M.: Graph neural networks for multivariate time series regression with application to seismic data. Int. J. Data Sci. Anal. **16**, 317–332 (2023)
12. Xiao, C., Hong, S., Huang, W.: Optimizing graph layout by t-SNE perplexity estimation. Int. J. Data Sci. Anal. **15**, 159–171 (2023)