10-2024

Tabular Data-centric AI: Challenges, Techniques and Future Perspectives

Yanjie Fu
Arizona State University

Dongjie Wang
University of Kansas

Hui Xiong
The Hong Kong University of Science and Technology (Guangzhou)

Kunpeng Liu
Portland State University, kunpeng@pdx.edu

Citation Details

Fu, Y., Wang, D., Xiong, H., & Liu, K. (2024, October). Tabular Data-centric AI: Challenges, Techniques and Future Perspectives. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 5522-5525).

Tabular Data-centric AI: Challenges, Techniques and Future Perspectives

Yanjie Fu
Arizona State University
Tempe, Arizona, USA
yanjie.fu@asu.edu

Dongjie Wang
University of Kansas
Lawrence, Kansas, USA
wangdongjie@ku.edu

Hui Xiong
Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
xionghui@ust.hk

Kunpeng Liu
Portland State University
Portland, Oregon, USA
kunpeng@pdx.edu

Abstract

Tabular data are the most widely used data formats in almost every application domain, such as, biology, ecology, and material science. The purpose of tabular data-centric AI is to use AI to augment the predictive power of tabular data to get better AI. Tabular data-centric AI is essential because it can reconstruct distance measures, reshape discriminative patterns, and improve data AI readiness (structural, predictive, interaction, and expression levels), which is significant in industries and real-world deployments. Therefore, our tutorial is designed to capture the interest of professionals with expertise in artificial intelligence, machine learning, and data mining, as well as researchers engaged in specific application areas and interdisciplinary studies. Examples of such applications include quality control, predictive maintenance, supply chain optimization, process efficiency improvements, biomarker identification, material performance screening. In this tutorial, we will explore the emerging field of Tabular Data-Centric AI. Our discussion will provide a comprehensive overview of this domain: (1) We will demonstrate the different settings within this research domain based on distinct application scenarios. (2) We will identify and explain the significant challenges encountered in tabular data-centric AI. (3) We will highlight existing methods and benchmarks. (4) We will discuss future potential directions for this domain and examine its interconnections with other research areas. To enhance the learning experience, this tutorial will include a hands-on section designed to teach participants the fundamental aspects of developing, evaluating and visualizing techniques in tabular data-centric AI. After this tutorial, attendees will have a deep understanding of tabular data-centric AI research, including its key challenges, seminal techniques, and insights into integrating tabular data-centric AI into their own research.

CCS Concepts

• **Computing methodologies → Feature selection**.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679102

Keywords

Machine Learning, Data-Centric AI, Tabular Data Refinement

ACM Reference Format:
Yanjie Fu, Dongjie Wang, Hui Xiong, and Kunpeng Liu. 2024. Tabular Data-centric AI: Challenges, Techniques and Future Perspectives. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3627673.3679102

1 Tutorial outline

Numerous fields, including healthcare, biology, and finance, have collected and accumulated extensive tabular data sets. The refinement of tabular data spaces to enhance decision-making analyses has emerged as an important research area. Existing work, however, necessitates substantial domain expertise and is often ungeneralizable and labor intensive. Therefore, the exploration of methods for automatically operating the rows and columns of tabular data—*i.e.* tabular data-centric AI—to make it more distinctive and informative has gained significant interest. This research domain can be easily incorporated with other domains to conduct many interdisciplinary studies such as bio-marker identification, health care feature analysis, material generation, and etc. Challenges confronting tabular data-centric AI includes processing heterogeneous data types, tackling high dimensionality, ensuring scalability, achieving data space interpretability, and mitigating the effects of distribution shifts, among others. This tutorial aims to discuss the emerging research field of Tabular Data-Centric AI. We will start by outlining various research settings within this domain, tailored to specific application contexts. Following this, we will demonstrate the principal challenges faced, underscore current methodologies and benchmarks, and illustrate prospective directions for future investigation.

2 Our Uniqueness and Comparison with Relevant Tutorials

This tutorial is related to some earlier tutorials:

- NIPS 2024 Data-Centric AI for reliable and responsible AI tutorial [1]
- KDD 2023 Data-Cenetric AI tutorial [2].

[1] https://www.vanderschaar-lab.com/neurips-2023-data-centric-ai-tutorial/
[2] Project page: https://dcaitutorial.github.io

However, our tutorial has the following uniquenesses:

- From the research problem perspective, this tutorial will focus on tabular data-centric AI, particularly on augmenting the AI readiness of tabular data for diverse tabular data modeling enabled applications. Among the two previous tuotorials, one is focusing on address data erros, missing, shifts, drifts, bias, and quality issues for reliable and responsible AI; another one is focusing on covering how to improve AI via the machine learning life cycle: training data, inference data, and data maintenance.
- From the underlying AI Tasks perspective, this tutorial will focus on introducing the feature-based DCAI, instance-based DCAI, and joint feature-instance based DCAI under the contexts of tabular data. Among the two previous DCAI tutorials,the KDD23 tutorial more focuses techniques about collection, labeling, preparation, cleaning, in or out of distributions, storage, retrieval for general data types; the NIPS23 tutorial focues on data characterization, generating synthetic data, data privacy, as well as practices with existing tools.
- From the technical framework perspective, this tutorial will focus on advanced automated intelligent tabular DCAI techniques, including reinforcement intelligence, generative intelligence, neuro-symbolic methods to enable auto data augmentation. The two previous tutorials focus on comprehensively reviewing classic methods.

As a result, our tutorial will provide a deeper focused tutorial on intorducing advnaced reinforcement, generative AI for auto improvement of tabular data AI readiness from three focused angles: feature-based, instance-based, and joint feature-instance based. After this tutorial, participants will gain a comprehensive understanding of tabular DCAI and its developmental trajectory. Furthermore, they will obtain in-depth knowledge on the formulation and implementation of strategies for this domain, as well as insights into future research directions and potential advancements.

3 Prerequisite knowledge

Our tutorial targets the audience with a basic knowledge of databases, data mining, machine learning such as reinforcement learning, deep learning, generative learning. The knowledge of data-centric AI is not required.

4 Content

Tabular datasets ubiquitously exist in many scenarios, including healthcare, biology, and finance, have collected and accumulated extensive. Tabular data space refinement has emerged as a hot research topic in recent days. However, existing works in this domain require substantial domain knowledge and are often ungeneralizable and labor-intensive. Therefore, the exploration of methods for automatically operating the feature and instance of tabular data—*i.e.* tabular data-centric AI—to make the data space more distinctive and informative has gained significant interest [2, 6, 8, 16, 18, 22, 29, 30].

Owing to the diversity in data types and the distinct distributions observed across various application contexts, coupled with the imperative for interpretability, tabular data-centric AI fundamentally different from and exhibit greater complexity than conventional data-centric AI methodologies.

To begin with, there are extensive techniques to refine the pattern and quality of tabular data space. In this tutorial, we will focus on making improvements by managing the features or instances of the data space. We can divide the corresponding tabular data-centric AI tasks into three categories:

1. *Exclusive Feature Manipulation* refers to the refinement of features within tabular data, focusing on enhancing the discriminative patterns of the data space.
2. *Exclusive Instance Manipulation* focuses on handling of individual instances within tabular data, aimed at identifying valuable data samples for model learning.
3. *Joint Feature and Instance Manipulation* refers to the simultaneous optimization of both features and instances within tabular data, aiming to broadly improve the configuration and utility of the data space.

Based on these aforementioned settings, we will introduce the key challenges of the tabular data-centric AI tasks. The first major challenge of tabular data-centric AI is to efficiently handle large search space when confronted with large-scale tabular data space(*i.e.* large feature dimension or large instance number) [1, 9, 21, 23, 24]. The second challenges is that substantial domain knowledge is required to get refine the tabular data space for better quality [10, 17, 31]. The third challenge is that the learned data-oriented knowledge relevant to data-centric AI is hard to generalize to different domains, distributions and scenarios [15, 19, 20]. There are also other challenges like the robustness and interpretability of tabular data-centric AI techniques, which will be detailed in our tutorial.

Addressing the aforementioned challenges, we introduce existing methodologies and demonstrate how these approaches mitigate challenges across various tabular data-centric tasks. Initially, we delve into feature selection and generation within the task of Exclusive Feature Manipulation, drawing on insights from the realms of reinforcement learning and generative AI [12, 13, 25–27]. Subsequently, we discuss the application of coreset selection and sample reweighting in the context of Exclusive Instance Manipulation [3, 7, 11]. Furthermore, we demonstrate the contributions in the area of Joint Feature and Instance Manipulation, providing a comprehensive overview of the integrated strategies employed [5, 14]. Following these techniques, we will introduce relevant benchmark works [4, 15, 28], which set up standard experimental settings and provide a fair platform for developing new tabular data-centric AI techniques.

After introducing the current progress of tabular data-centric AI, as a newly emerging and fast-growing area, there are still many challenges to tackle and various promising directions to explore. Tabular data-centric AI is actually highly relevant to many research areas, and could potentially benefit a vast majority of the research community, such as finance, health care. We will first discuss some challenges within the tabular data-centric AI field awaiting to be tackled, then summarize the topics that transcend the boundary of current tabular data-centric AI research and intersect with other areas.

5 Supplementary materials

1. The first DCAI workshop https://data-centric-ai-dev.github.io/ICDM/

(2) The second DCAI workshop https://data-centric-ai-dev.github.io/BigData/
(3) The third DCAI workshop https://data-centric-ai-dev.github.io/BigData2024/

6 Short biographies of presenters:

There are three presenters on tabular data-centric AI.

Dr. Yanjie Fu is an associate professor in the School of Computing and AI at Arizona State University. He received his Ph.D. degree from Rutgers, the State University of New Jersey in 2016, the B.E. degree from the University of Science and Technology of China in 2008, and the M.E. degree from the Chinese Academy of Sciences in 2011. He has research experience in industry research labs, such as Microsoft Research Asia and IBM Thomas J. Watson Research Center. He has published prolifically in refereed journals and conference proceedings, such as IEEE TKDE, IEEE TMC, ACM TKDD, ACM SIGKDD, AAAI, IJCAI, VLDB, WWW, ACM SIGIR. His research has been recognized by: 1) two federal junior faculty awards: US NSF CAREER and NSF CRII awards; 2) five best paper (runner-up, finalist) awards, including ACM KDD18 Best Student Paper Finalist, IEEE ICDM14, 21, 22 Best Paper Finalist, ACM SIGSpatial20 Best Paper Runner-up; 3) three industrial awards: 2016 Microsoft Azure Research Award, 2022 Baidu Scholar global top Chinese young scholars in AI, 2021 Aminer.org AI 2000 Most Influential Scholar Award Honorable Mention in Data Mining; 4) several other university-level awards: Reach the Stars Award, University System Research Board Award and University Interdisciplinary Research Award. He was chosen for the nation's early career engineers by the National Academy of Engineering 2023 Grainger Foundation Frontiers of Engineering Symposium. He is committed to data science education. His graduated Ph.D. students have joined academia as tenure-track faculty members.

Dr. Dongjie Wang is an assistant professor in the Department of Computer Science at the University of Kansas. His academic focus revolves around data-centric AI, with particular interests in data-centric AI, causal graph learning, root cause analysis, outlier detection, spatial-temporal data mining, user profiling, and graph mining. He has gained precious industry experience through internships at prestigious institutions such as Nokia Bell Labs and NEC Labs America. He has published 25+ leading journals (e.g. TKDE, KAIS) and conferences (e.g. NeurIPS, KDD, AAAI, WWW). Notably, three of his papers (SIGSPATIAL2020, ICDM2021, ICDM2021) were recognized as best paper runner-ups. In addition to the research contributions, he also actively contributes to the academic community by serving as a PC member for certain prestigious conferences and journals, including KDD, IJCAI, AAAI, WSDM, CIKM, TNNLS, KBS, and TBD.

Dr. Hui Xiong received his Ph.D. in Computer Science from the University of Minnesota - Twin Cities, USA, in 2005, the B.E. degree in Automation from the University of Science and Technology of China (USTC), Hefei, China, and the M.S. degree in Computer Science from the National University of Singapore (NUS), Singapore. He is currently a Distinguished Professor at Rutgers, the State University of New Jersey, where he received the 2018 Ram Charan Management Practice Award as the Grand Prix winner from the Harvard Business Review, RBS Dean's Research Professorship (2016), two-year early promotion/tenure (2009), the Rutgers University Board of Trustees Research Fellowship for Scholarly Excellence (2009), the ICDM-2011 Best Research Paper Award (2011), the Junior Faculty Teaching Excellence Award (2007), Dean's Award for Meritorious Research (2010, 2011, 2013, 2015) at Rutgers Business School, the 2017 IEEE ICDM Outstanding Service Award (2017), and the AAAI-2021 Best Paper Award (2021). Dr. Xiong is also a Distinguished Guest Professor (Grand Master Chair Professor) at the University of Science and Technology of China (USTC). For his outstanding contributions to data mining and mobile computing, he was elected an ACM Distinguished Scientist in 2014, an IEEE Fellow, and an AAAS Fellow in 2020.

Dr. Kunpeng Liu joined the CS Department of Portland State University as a tenure-track assistant professor in 2022 Fall. Prior to that, he received the Ph.D. degree from the CS Department of the University of Central Florida. He received both his M.E. degree and B.E. degree from the Department of Automation, University of Science and Technology of China (USTC). He has rich research experience in industry research labs, such as Geisinger Medical Research, Microsoft Research Asia, and Nokia Bell Labs. He has published in refereed journals and conference proceedings, such as IEEE TKDE, ACM SIGKDD, IEEE ICDM, AAAI, and IJCAI. He has received a Best Paper Runner-up Award from IEEE ICDM 2021.

7 Acknowledgement

This research was partially supported by the National Science Foundation (NSF) via the grant numbers: 2421864, 2421803, 2421865, and National academy of engineering Grainger Foundation Frontiers of Engineering Grant.

References

[1] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. 2016. Feature selection for high-dimensional data. *Progress in Artificial Intelligence* 5 (2016), 65–75.

[2] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[3] Chengliang Chai, Jiayi Wang, Nan Tang, Ye Yuan, Jiabin Liu, Yuhao Deng, and Guoren Wang. 2023. Efficient coreset selection with cluster-based methods. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 167–178.

[4] Sabri Eyuboglu, Bojan Karlaš, Christopher Ré, Ce Zhang, and James Zou. 2022. dcbench: A benchmark for data-centric ai systems. In *Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning*. 1–4.

[5] Wei Fan, Kunpeng Liu, Hao Liu, Hengshu Zhu, Hui Xiong, and Yanjie Fu. 2022. Feature and instance joint selection: A reinforcement learning perspective. *arXiv preprint arXiv:2205.07867* (2022).

[6] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* 34 (2021), 18932–18943.

[7] Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*. Springer, 181–195.

[8] Lasse Hansen, Nabeel Seedat, Mihaela van der Schaar, and Andrija Petrovic. 2023. Reimagining Synthetic Tabular Data Generation through Data-Centric AI: A Comprehensive Benchmark. *Advances in Neural Information Processing Systems* 36 (2023), 33781–33823.

[9] Jie Hao, Kaiyi Ji, and Mingrui Liu. 2024. Bilevel Coreset Selection in Continual Learning: A New Formulation and Algorithm. *Advances in Neural Information Processing Systems* 36 (2024).

[10] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-centric artificial intelligence. *Business & Information Systems Engineering* (2024), 1–9.

[11] Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. 2021. Retrieve: Coreset selection for efficient and robust semi-supervised learning. *Advances in neural information processing systems* 34 (2021), 14488–14501.

[12] Kunpeng Liu, Yanjie Fu, Pengfei Wang, Le Wu, Rui Bo, and Xiaolin Li. 2019. Automating feature subspace exploration via multi-agent reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 207–215.

[13] Kunpeng Liu, Dongjie Wang, Wan Du, Dapeng Oliver Wu, and Yanjie Fu. 2023. Interactive reinforced feature selection with traverse strategy. *Knowledge and Information Systems* 65, 5 (2023), 1935–1962.

[14] Yasitha Warahena Liyanage, Daphney-Stavroula Zois, and Charalampos Chelmis. 2021. Dynamic instance-wise joint feature selection and classification. *IEEE Transactions on Artificial Intelligence* 2, 2 (2021), 169–184.

[15] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, et al. 2024. Dataperf: Benchmarks for data-centric ai development. *Advances in Neural Information Processing Systems* 36 (2024).

[16] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. 2024. When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems* 36 (2024).

[17] Kaushikkumar Patel. 2024. Ethical reflections on data-centric AI: balancing benefits and risks. *International Journal of Artificial Intelligence Research and Development* 2, 1 (2024), 1–17.

[18] Neoklis Polyzotis and Matei Zaharia. 2021. What can data-centric AI learn from data and ML engineering? *arXiv preprint arXiv:2112.06439* (2021).

[19] Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. 2024. Dissecting Sample Hardness: A Fine-Grained Analysis of Hardness Characterization Methods for Data-Centric AI. *arXiv preprint arXiv:2403.04551* (2024).

[20] Assaf B Spanier, Dor Steiner, Navon Sahalo, Yoel Abecassis, Dan Ziv, Ido Hefetz, and Shimon Kimchi. 2024. Enhancing Fingerprint Forensics: A Comprehensive Study of Gender Classification Based on Advanced Data-Centric AI Approaches and Multi-Database Analysis. *Applied Sciences* 14, 1 (2024), 417.

[21] Siva Sankari Subbiah and Jayakumar Chinnappan. 2021. Opportunities and Challenges of Feature Selection Methods for High Dimensional Data: A Review. *Ingénierie des Systèmes d'Information* 26, 1 (2021).

[22] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 645–654.

[23] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. 2024. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems* 36 (2024).

[24] Zhijing Wan, Zhixiang Wang, Yuran Wang, Zheng Wang, Hongyuan Zhu, and Shin'ichi Satoh. 2024. Contributing Dimension Structure of Deep Feature for Coreset Selection. *arXiv preprint arXiv:2401.16193* (2024).

[25] Dongjie Wang, Meng Xiao, Min Wu, Yuanchun Zhou, Yanjie Fu, et al. 2024. Reinforcement-enhanced autoregressive feature transformation: Gradient-steered search in continuous space for postfix expressions. *Advances in Neural Information Processing Systems* 36 (2024).

[26] Meng Xiao, Dongjie Wang, Min Wu, Kunpeng Liu, Hui Xiong, Yuanchun Zhou, and Yanjie Fu. 2024. Traceable group-wise self-optimizing feature transformation learning: A dual optimization perspective. *ACM Transactions on Knowledge Discovery from Data* 18, 4 (2024), 1–22.

[27] Meng Xiao, Dongjie Wang, Min Wu, Pengfei Wang, Yuanchun Zhou, and Yanjie Fu. 2023. Beyond discrete selection: Continuous embedding space optimization for generative feature selection. In *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 688–697.

[28] Liang Xu, Jiacheng Liu, Xiang Pan, Xiaojing Lu, and Xiaofeng Hou. 2021. Dataclue: A benchmark suite for data-centric nlp. *arXiv preprint arXiv:2111.08647* (2021).

[29] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems* 32 (2019).

[30] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158* (2023).

[31] Yunkun Zhang, Jin Gao, Zheling Tan, Lingfeng Zhou, Kexin Ding, Mu Zhou, Shaoting Zhang, and Dequan Wang. 2024. Data-centric foundation models in computational healthcare: A survey. *arXiv preprint arXiv:2401.02458* (2024).