Towards Data Pooling in AI Projects: Opportunities and Challenges

Leonhard Czarnetzki[1], Zsombor Iszak[1], Daniel Bachlechner[1], Ruben Hetfleisch[1], Sisay Adugna Chala[2] and Christian Beecks[3,2]

[1]Fraunhofer Austria Research GmbH, Austria
[2]Fraunhofer FIT, Germany
[3]FernUniversität in Hagen, Germany

Abstract

The pervasive trend of digitization is impacting all sectors, particularly in the realm of small and medium-sized enterprises (SMEs). SMEs are still in the early stages of digitization, with only 60% applying descriptive analytics. As the demand for artificial intelligence (AI) grows, intelligent solutions become crucial, but many SMEs lack the necessary data and competencies, leading to potential failures and risks in AI projects. Data pooling, the practice of combining datasets from multiple sources into a single, larger dataset, enables SMEs to gather, utilize, and share data resources to overcome common data limitations. In this paper, we aim to provide an application-oriented perspective on data pooling by illustrating the concept along different application scenarios and elucidating major opportunities and challenges. Our findings are useful for both scientists and practitioners in the field of AI.

Keywords

Data Pooling, Artificial Intelligence, Machine Learning, AI Projects

1. Introduction

The advent of the digital information age has led to numerous opportunities and challenges in the realm of small and medium-sized enterprises (SMEs) [1], making artificial intelligence (AI) technology prominent in a wide range of applications [2]. Utilizing AI technology to facilitate human-AI teaming [3] and for solving complex business tasks requires a smooth technology integration into existing business processes. Incorporating AI effectively in an SME's overall strategy [4], however, is a challenging endeavor that requires numerous considerations and decisions. Enabling a smooth operation of AI technology requires an appropriate computational infrastructure, knowledge and skills regarding machine learning methods as well as efficient access to valuable data sources. In particular the latter is crucial for business growth [5], its absence can pose major challenges [6]. Lack of data availability and quality, which is particularly prevalent among SMEs, prevents the realisation of many of the potentials that AI holds [7].

One approach to overcome limited data availability and quality is by means of the concept of *data pooling*. Data pooling refers to the practice of combining datasets from multiple sources into a single, larger dataset. Data pooling enables SMEs to gather, utilize, and share data resources within a single company or across multiple companies. In this sense, data pooling helps to acquire and combine data sources with similar or complementary characteristics from different companies to build suitable, task-oriented datasets for the analytical purpose at hand [8]. From a rather non-technical perspective, data pooling is a cooperation among different information holders for the purpose of aggregating data [9]. As such, data pooling enables SMEs to unleash the potential of data, foster cross-company collaboration and drive innovation in a data-driven economy.

In this paper, we propose an application-oriented perspective on data pooling by illustrating its concept with respect to various state-of-the-art AI technologies (Section 2) and by elucidating major

LWDA 2024: Business Intelligence and Analytics (WSBIA 2024)
leonhard.czarnetzki@fraunhofer.at (L. Czarnetzki); zsombor.iszak@fraunhofer.at (Z. Iszak); daniel.bachlechner@fraunhofer.at (D. Bachlechner); ruben.hetfleisch@fraunhofer.at (R. Hetfleisch); sisay.adugna.chala@fit.fraunhofer.de (S. A. Chala); christian.beecks@fernuni-hagen.de (C. Beecks)

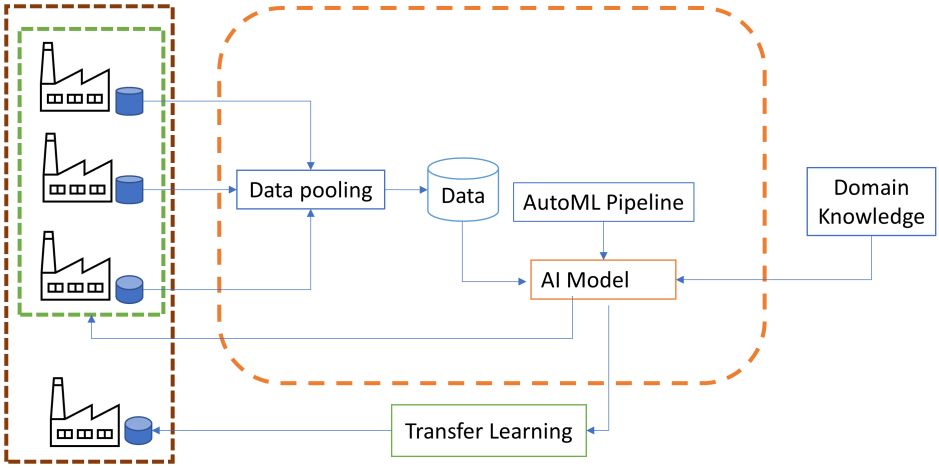



Figure 1: Data pooling application within a machine learning pipeline.

opportunities and challenges (Section 3) along different application scenarios (Section 4). We believe that our insights are useful for both scientists in the field of AI focusing on different research directions (Section 5) as well as practitioners aiming at leveraging this concept in their individual company context.

2. Data Pooling – Application Perspective

As indicated in the previous section, data pooling can be thought of as (big) data integration [10] that consolidates data from various internal and external sources into a unified view that enables more comprehensive and accurate data analytics and thereupon decision making [11]. Frequently but not necessarily, data pooling is integrated into complex machine learning pipelines and systems, as illustrated in Figure 1. In this particular scenario, data is pooled from several databases, systems and applications of different companies and integrated into a common storage system. This storage system is then used within an automated machine learning pipeline [12] to deliver a suitable AI model that can be enriched by domain knowledge [13] or other informed machine learning approaches [14]. Finally, transfer learning [15] is applied in this scenario to customize existing AI models for novel objectives and further applications.

In fact, the requirements for a successful application of data pooling vary among SMEs, depending on their specific demands and business areas. The first decision that has to be taken into account is whether data pooling is performed within an organization or across organizational boundaries. While the organization-internal data integration has less technical, organizational, security-relevant, and regulatory requirements [16], external data integration comes with the additional challenges of accessing and trusting data sharing platforms such as International Data Spaces[1] or Gaia-X[2].

While both *internal* as well as *external data pooling* aims to bring together disparate data sources, the context, scope, governance and requirements vary significantly depending on the specific application scenario. Requirements such as *data consistency* in terms of data types, formats and representations, *data quality* in terms of errors, inconsistencies, missing values, or outliers, and *data compatibility* in terms of inherent structure and semantics, as well as *data topicality* with regard to the timeliness of updates to the data sources are the major complexity dimensions for SMEs when making use of data pooling for improving the robustness and performance of AI models.

[1] https://internationaldataspaces.org
[2] https://gaia-x.eu

3. Opportunities and Challenges

Data pooling can be a valuable strategy for organizations with limited data in general and SMEs in particular to improve the outcome of AI projects. Depending on the application scenario, data pooling offers a number of opportunities.

From a **technical perspective**, data pooling enables an *increased overall sample size* for the model training phase and helps to mitigate overfitting issues and to improve generalization ability of AI models. In addition, combining data from different sources can lead to *higher diversity*, which helps to capture a broader range of patterns and variations and to address biases, leading to more robust and adaptable models. Furthermore, additional data sources can contain *new features and variables* which enhance the richness of the model and improve its performance. Finally, data pooling can leverage *transfer learning techniques* to fine-tune or adapt existing models.

From a **business perspective**, data pooling enables organizations to gain a *more comprehensive and holistic understanding* of the issue they are investigating and to *collaborate with external partners* in order to leverage complementary business expertise. Moreover, data pooling opens up opportunities for organizations to *monetize their data assets* and to *reduce efforts and costs* associated with data collection, storage and maintenance.

In addition to the aforementioned opportunities, data pooling also entails a number of challenges. In this regard, one has to differentiate between challenges that appear by aggregating data and those that appear by contributing data for pooling purposes. A major challenge in the scope of data pooling is concerned with the questions of *data ownership* and *data privacy*. Determining ownership and usage rights as well as control over pooled data can lead to disputes and conflicts, where missing anonymization can lead to privacy concerns. In a similar manner, *data security, confidentiality and trust* epitomize further challenges when sharing sensitive data, as well as *data compliance and governance*. But also technical challenges such as *data interoperability and compatibility*, which arise among different IT systems, platforms and data formats, can complicate the utilization of data pooling. On a data-centric level, the challenges are *data quality* and *data transferability*, when it comes to matching the statistical properties of pooled data resources.

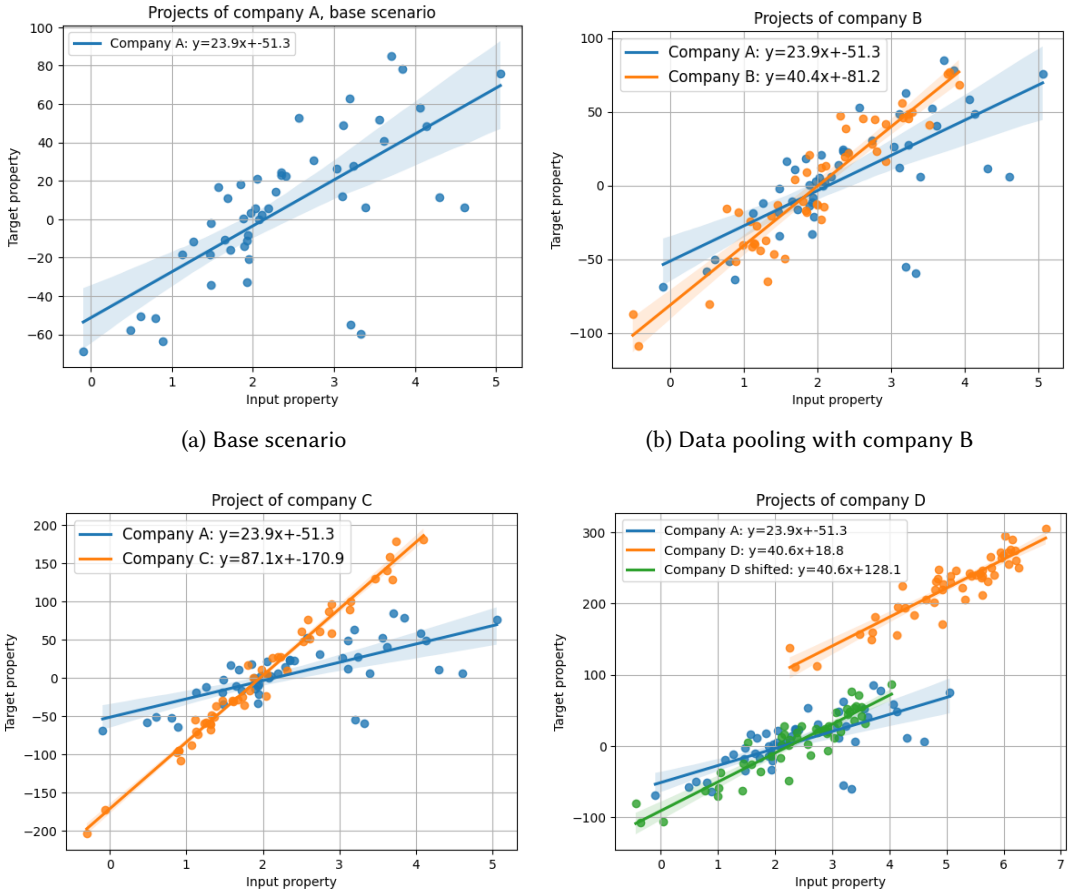
4. Application Scenario

In order to illustrate the key aspects of data pooling, we use an application scenario based on a real industry use case that has already been discussed in [8]. In order to experiment with different dataset characteristics in a controlled manner and to illustrate relevant aspects in a comprehensible way, we decided to synthetically generate data based on the real use case, instead of using the real data.

To this end, we consider a small consulting company A that would like to predict financial success of their prospective consulting projects based on information about projects implemented already. The available dataset comprises a total of 50 projects whose financial output is shown in Figure 2(a) as a function of a single input property of a project, i.e. the number of employees working on that project. Based on this dataset, consulting company A generates a simple linear AI model for predicting future project scuccess.

The different pooling possibilities for company A are shown in Figure 2(b)–(d). While the projects of company B, as shown in Figure 2(b), show a high degree of similarity in terms of input-target relation, the projects of company C, as shown in Figure 2(c), differ to those of company A as they generally produce more financial success with increasing number of employees. Finally, the data available from company D, as illustrated in Figure 2(d), shows a similar tendency at a different scale. The data provided by company D indicates the transferability challenge which can be tackled by pre-processing, i.e. shifting, the data to company A's data as illustrated in Figure 2(d).

Based on the given information, consulting company A has to make a decision, which data to pool. The impact of the individual datasets on the prediction performance in terms of the coefficient of determination (R^2) is shown in Figure 3(a). As can be seen, by pooling the dataset of company B, the



(a) Base scenario

(b) Data pooling with company B

(c) Data pooling with company C

(d) Data pooling with company D

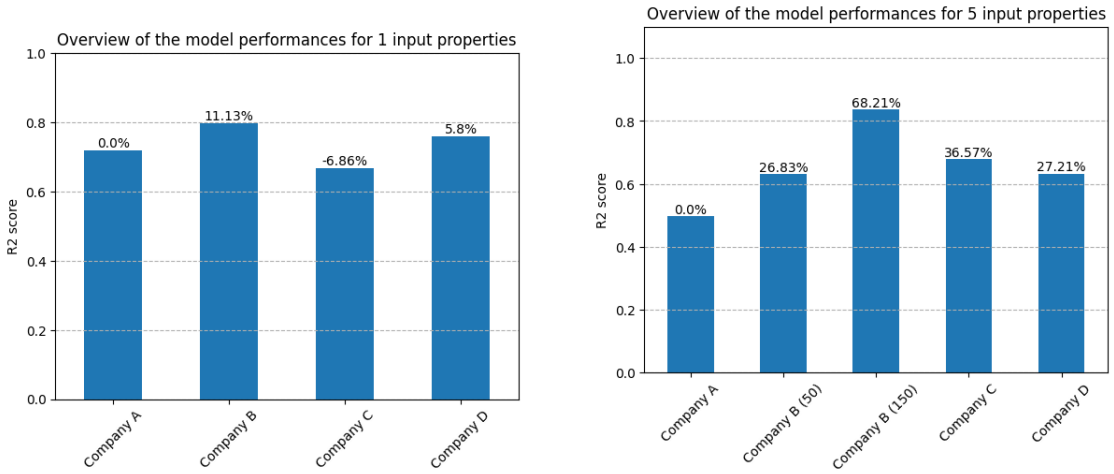
Figure 2: Various pooling possibilities for company A.

value of R^2 increases by 11.13%, whereas pooling the dataset provided by company C leads to a decrease of more than 6.8%. Pre-processing the data of company D allows us to overcome the transferability challenge and to achieve a performance gain of 5.8%. To conclude, even in case of a simple linear prediction model, data pooling enables more accurate prediction results.

With increasing complexity of this scenario, e.g. by taking into account more project properties, by dropping the linear dependency assumption, and by including outliers, a more complex prediction model is required. While higher complexity leads to a performance drop in the base scenario, it increases the potential of data pooling. For this purpose, let us consider that the financial success of each consulting project is no longer solely determined by the number of employees, but that additional factors such as project duration and project performance indicators describing complexity or experience, are also taken into account, leading to a total of five project input properties.

The prediction performance of the different data pooling strategies for this scenario is shown in Figure 3(b). As before, the available dataset of consulting company A comprises information about 50 projects, but with five input properties each. As can be seen in the figure, the increased scenario complexity results in a significant drop of prediction performance for company A. The potential of data pooling, however, becomes obvious when increasing the pooling size with data from company B from 50 to 150 samples. In this case, the prediction performance in terms of R^2 value almost doubles and increases by 26.83% and 68.21% for 50 and 150 samples, respectively. Figure 3(b) also shows that pooling data from companies C and D can be beneficial in this complex setting and can lead to an increase of prediction performance by 36.57% and 27.21%, respectively.

To sum up, we have presented an application scenario for data pooling based on a real industry use case. The examples shown above indicate the significant potential of data pooling. With growing



(a) Results including a single project input property

(b) Results including five project input properties

Figure 3: Prediction performance for different data pooling strategies.

complexity of the underlying data, the advantages of data pooling become more evident, as long as the challenges regarding data quality, compatibility and transferability are adequately resolved.

5. Conclusions and Future Work

Data pooling can be a valuable strategy for organizations with limited data to improve the outcome of AI projects. However, to benefit from the opportunities, challenges must be overcome. Overcoming these challenges causes costs that are not always easy to cope with, especially for SMEs. It is therefore important that both the benefits and the costs of applying a certain data pooling strategy can be well assessed before making a significant investment. Moreover, it is crucial to find the best data pooling strategy for a given application scenario.

General guidelines with regard to the pros and cons of certain strategies under certain circumstances in combination with experiments based on synthetic data can provide a good basis for assessing the benefits and costs of data pooling. An automated machine learning pipeline can be valuable for carrying out experiments efficiently. Experiments can not only provide important insights for concrete data pooling decisions, but can also help to identify further guidelines that may be useful in the future independently of experiments. We can already make initial, still fairly general recommendations, but plan to further investigate data pooling based on automated experiments and develop more specific guidelines.

First of all, organisations must acknowledge the potential value of data pooling. By combining datasets from different organizations, they can significantly enhance the quality and quantity of their data, leading to more robust AI models. However, it is crucial to identify compatible partners. Organizations should look for partners with similar goals and data privacy standards. In any case, organizations should prioritize privacy and security throughout the data pooling process. Implementing stringent measures helps to safeguard sensitive information and comply with relevant regulations. Additionally, it is reasonable to establish clear data sharing agreements. Specifying the types of data to be shared as well as data usage rights and mechanisms helps resolving disputes or breaches. Data spaces are a promising way to cost-effectively find compatible partners and negotiate important agreements. To increase compatibility and consistency, organizations should standardize data formats, preprocess datasets and employ advanced data integration techniques. Finally, organizations should continuously monitor model performance and evaluate the effectiveness of data pooling strategies. By adhering to these guidelines, organizations can harness the power of data pooling across organizational boundaries to overcome data limitations and unlock the full potential of AI for their business objectives.

Data pooling is a rapidly evolving field and future directions are likely to focus on overcoming current challenges and unlocking new opportunities. While it is expected that data pooling leverages transfer learning techniques to fine-tune or adapt existing models, the extent of improvement is yet to be quantified and established. Moreover studies need to be conducted to understand the relevance of data spaces for data pooling as well as the advantages of data pooling over alternative approaches to data enhancement such as informed machine learning or the use of synthetic data.

Acknowledgments

This research has been supported by the Fraunhofer-Gesellschaft, the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK) and the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the projects MEMENTO and champI4.0ns (891793).

Data and code for generating the plots of the application scenario section can be provided upon request.

References

[1] A. Thrassou, N. Uzunboylu, D. Vrontis, M. Christofi, Digitalization of SMEs: A Review of Opportunities and Challenges, Springer International Publishing, Cham, 2020, pp. 179–200. doi:10.1007/978-3-030-45835-5_9.

[2] S. Borah, C. Kama, S. Rakshit, N. R. Vajjhala, Applications of artificial intelligence in small-and medium-sized enterprises (smes), in: Cognitive Informatics and Soft Computing: Proceeding of CISC 2021, Springer, 2022, pp. 717–726.

[3] R. Zhang, N. J. McNeese, G. Freeman, G. Musick, "an ideal human" expectations of ai teammates in human-ai teaming, Proceedings of the ACM on Human-Computer Interaction 4 (2021) 1–25.

[4] T. H. Davenport, The AI advantage: How to put the artificial intelligence revolution to work, mit Press, 2018.

[5] M. Iqbal, S. H. A. Kazmi, A. Manzoor, A. R. Soomrani, S. H. Butt, K. A. Shaikh, A study of big data for business growth in smes: Opportunities challenges, in: 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018, pp. 1–7. doi:10.1109/ICOMET.2018.8346368.

[6] J. Kádárová, L. Lachvajderová, D. Sukopová, Impact of digitalization on sme performance of the eu27: Panel data analysis, Sustainability 15 (2023) 9973.

[7] A. Bettoni, D. Matteri, E. Montini, B. Gladysz, E. Carpanzano, An ai adoption model for smes: a conceptual framework, IFAC-PapersOnLine 54 (2021) 702–708. doi:10.1016/j.ifacol.2021.08.082.

[8] L. Czarnetzki, F. Kainz, F. Lächler, C. Laflamme, D. Bachlechner, Enabling supervised machine learning through data pooling: A case study with small and medium-sized enterprises in the service industry, in: German Conference on Artificial Intelligence (Künstliche Intelligenz), Springer, 2022, pp. 53–59.

[9] M. Mattioli, The data-pooling problem, Berkeley Technology Law Journal 32 (2017) 179–236.

[10] X. L. Dong, D. Srivastava, Big data integration, in: 2013 IEEE 29th international conference on data engineering (ICDE), IEEE, 2013, pp. 1245–1248.

[11] S. C. Albright, W. L. Winston, C. J. Zappe, M. N. Broadie, Data analysis and decision making, volume 577, Citeseer, 2011.

[12] F. Hutter, L. Kotthoff, J. Vanschoren, Automated machine learning: methods, systems, challenges, Springer Nature, 2019.

[13] H. Zhang, U. Roy, Y.-T. T. Lee, Enriching analytics models with domain knowledge for smart manufacturing data analysis, International journal of production research 58 (2020) 6399–6415.

[14] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, et al., Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems, IEEE Transactions on Knowledge and Data Engineering 35 (2021) 614–633.

[15] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (2020) 43–76.

[16] J. Kutz, V. P. Göbels, D. Brajovic, B. Fresz, N. Renner, S. Omri, J. Neuhüttler, M. Huber, B. Bienzeisler, Ki-zertifizierung und absicherung im kontext des eu ai act, 2023. URL: https://doi.org/10.24406/publica-1875. doi:10.24406/publica-1875.