Bachelor's Programme in Science and Technology

Data lineage and observability for ensemble Machine Learning serving in the edge

Dung Nguyen Anh

Aalto University

BACHELOR'S THESIS

Aalto University
BACHELOR'S THESIS 2023

Data lineage and observability for ensemble Machine Learning serving in the edge

Dung Nguyen Anh

Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Technology.
Otaniemi, 29 Jan 2023

Supervisor: professor Maarit Korpi-Lagg
Advisor: Minh-Tri Nguyen

Aalto University
School of Science
Bachelor's Programme in Science and Technology

Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author

Dung Nguyen Anh

Title

Data lineage and observability for ensemble Machine Learning serving in the edge

School School of Science			
Degree programme Bachelor's Programme in Science and Technology			
Major Data Science			**Code** SCI3095
Supervisor professor Maarit Korpi-Lagg			
Advisor Minh-Tri Nguyen			
Level Bachelor's thesis	**Date** 27 Nov 2017	**Pages** 70	**Language** English

Abstract

Abtract

Keywords keyword

urn https://aaltodoc.aalto.fi

Contents

1. Introduction

In recent years, Machine Learning (ML) has been applied extensively to many problems, including customer journey optimization in marketing [1], particle identification in Physics [2], and mental health problems prediction in healthcare [3]. With their proven performance and increasing number of ML applications , several organizations are seeking third-party ML service providers to improve their operations, thus leading to the creation of a new business model — ML as a Service (MLaaS). As with other kinds of service, MLaaS requires a contract which is usually called a service level agreement (SLA) that sets the expectations and describes the delivered service to preserve the benefits of both stakeholders. However, the current SLA has yet to fulfill its preservation job because ML-specific attributes, such as data quality, inference accuracy, and explainability, have created new challenges. For instance, in Internet of Things (IoT) applications, many issues of the ML models have their root causes in the quality of data, which is impacted by many factors such as erroneous measurement, environmental noise, and discrete observations [4]. Moreover, because of the relationship between data quality and inference accuracy, ML models can produce false predictions when data quality problems exist, causing ML providers to be penalized based on the SLA. Another issue is that in MLaaS, the customers only submit the data and does not cooperate with the ML provider in the prediction task. Thus, from their perspective, ML models are considered to be a black box, which is insufficient in human life-related applications, such as e-health and autonomous vehicles, where uncertain decisions cannot be tolerated.

One approach to resolve these two challenges is increasing model explainability by interpreting the inference result to the customer [5]. However, explainability aspects of ML models have been researched mostly in the training task [4], making it unclear about the appropriate utilization and implementation of ML-specific attributes and constraints. As an approach to support ML-specific service contracts, Linh et al.[6] proposed the QoA4ML framework which outlines the essential components of such contracts, the definition of ML attributes and constraints, and the guidance on the process to monitor and assess these elements. Although QoA4ML has created a foundation for robust ML monitoring, current implementation only focuses on general ML metrics such as the final confidence of the ensemble model. Consequently, because base models are challenging to monitor and explain with the current state of QoA4ML, they remain like a black box to both stakeholders. Such a level of explainability is not sufficient for some sophisticated applications, such as digital assistants, where a complex chain of models is employed. Additionally, with the increasing amount of research in autoscaler for the inference serving system [7] and automated ensemble [8], it demands a new technique to capture and visualize underlying inference graph to ensure dynamic and robust inference capabilities of ML solutions.

Recognizing above problem can be formalized as capturing data lineage, this thesis performs an analysis of data lineage monitoring and its implementation in edge environment. From that, the research proposes an approach to improve QoA4ML framework with data lineage, which can be combined with other techniques to improve the current MLaaS. This paper provides a prototype with QoA4ML as the foundation so that ML developer and provider can integrate it with their deployed service.

The rest of this paper is organized as follows. Section 2 discusses the background of the research, and section 3 explains data lineage implementation in the QoA4ML framework. Section 4 describes the experiment and its result, while section 5 concludes the research.

2. Background

This section presents the background on provenance for AI applications, ensemble ML, and its explainability challenge. Lastly, Section 2.3 discusses one standard for data model of provenance which is utilized in section 3.

2.1 Provenance and Data Quality

The terms lineage, provenance, and traceability can be used interchangeably to refer to the process of constructing a final product, whether it is a digital file or a physical item [9]. Regarding lineage for the digital file, Wang et al [10] was the first paper to discuss those issues, which helped formally define provenance as "data source tagging" and "intermediate source tagging" problems [11, p. 5]. From its original interpretation, provenance has been applied in many domains such as scientific databases [12] [13] [14] [15], data warehouses [16] [17], and recently big data platforms [18] [19], and IoT [20]. In different applications, the types of lineages, their generating techniques, and main issue that they can answer are diverse [9]. For instance, there are four main provenance classes, i.e., data provenance, information systems provenance, provenance meta-data, and workflow provenance.

With the development of many AI applications, researchers have tried to incorporate provenance into the AI systems [21] by linking the input and output of the model, which can be a valuable source in interpreting the inference result. However, in this domain, provenance is commonly utilized in a rather general way, where entire algorithms or data transformations are merely represented by semantic relationships [22]. Consequently, while entire pipelines can be documented with provenance, the specific inner workings of individual models remain opaque.

Additionally, provenance is often considered when assessing data quality tasks, e.g., evaluating integrity, trust, and accuracy. By analyzing provenance, the AI systems can detect errors in data generation and processing, which is valuable for IoT applications where data is uncertain, erroneous, and noisy [23].

2.2 Ensemble ML and its Explainability Challenges

Ensemble ML is a conventional technique that involves combining multiple models, which can reduce the variance of prediction, to improve the accuracy and robustness of predictions. This is achieved by utilzing multiple base models with different algorithms, hyperparameters, or subsets of features on the same input dataset, and then combining their results by applying various aggregation methods such as averaging, voting, or stacking to produce the final decision.

Such a technique has been shown to be highly effective in many applications including financial forecasting [24], image recognition [25], and natural language processing [26]. Then, as the demand for more precise forecasts and advanced applications, such as AI assistants or autonomous vehicles, increases, it requires a complex dataflow graph, usually represented by a direct acyclic graph (DAG), comprised of multiple base models that are interconnected and orchestrated to handle the input data sequentially or in parallel [7]. For each prediction request, the whole DAGs or only a subset of it can be involved, requiring ML providers to record the data flow of the inference task for further analysis and explanation. Additionally, in resource-constrained environments like the edge, ensemble selection rules can be applied so that only models that provide the best inference performance are utilized [8]. Although this approach can help resolve the challenges of high computation complexity, high resource occupation, and the moderate inference time of ensemble ML, it introduces uncertainty to the inference flow, making the data pipeline less transparent and explainable. Consequently, it is very challenging for current interpreting techniques to produce an explainable and transparent ML solutions without the data lineage of the ensemble model.

2.3 PROV Family

Researchers have proposed various models, languages, and tools to facilitate the documentation of provenance, including those tailored for AI/ML models [28]. However, it is still crucial to have a standard reusable ontology to capture provenance in heterogeneous environments, such as the web, for further development of provenance documentation. As an approach to resolve this problem, the PROV

PROV Concepts	Classes	Name
Entity Activity Agent	PROV-DM types	Entity Activity Agent
Generation Usage Communication Derivation Attribution Association Delegation	PROV-DM relations	WasGeneratedBy Used WasInformedBy WasDerivedFrom WasAttributedTo WasAssociatedWith ActedOnBehalfOf

Table 2.1. PROV-DM core concepts and types [27]

family of documents (PROV-OVERALL) was introduced by The World Wide Web Consortium (W3C) in 2013 [27]. In those documents, the most significant one is the PROV data model (PROV-DM) which provides a universal data model for provenance that can be applied to translate domain-specific and application-specific representations of provenance into a standardized format that can be shared between different systems. PROV-DM has ten core concepts which can be divided into types and relations like in Table 2.1. As illustrated in Figure 2.1, PROV-DM outlines the employment and production of *entities* by *activities*, which may be influenced in several manner by *agent*. In the following paragraphs, the proper definition and application of those data models will be discussed

Firstly, entities are the fundamental building blocks of the PROV-DM data model. They represent physical or digital objects involved in a process, such as a document, dataset, or prediction result. Entities can have fixed attributes that describe their characteristics, such as their size, quality, or creation time. Then, when the value of an attribute is adjusted, a new entity is produced and related to the previous version ("wasDerivedFrom").

Secondly, activities are actions or processes that transform or manipulate entities. For instance, activities include data processing, predicting, and request serving. Activities can have attributes describing their properties, such as the start and end times of the task or the ML models employed to generate the prediction. Then, this data model can be "used" by entities as their inputs, and it can output a new entity which "wasGeneratedBy" the activities. Moreover, an activity can be informed by other activities with "wasInformedBy" relations.

Lastly, agents are entities that are responsible for carrying out activities ("wasAssociatedWith"), and when one activity is finished, an entity will be produced which is attributed back to the agent ("wasAttributedTo"). Agents can be people, software tools, ML models, or other entities that can initiate or control activities. Additionally, an agent can "actedOnBehalfOf" other agents, describing the delegation relationship among them.

Section 3 will discuss the implementation of PROV-DM to capture data lineage of ensemble ML models.

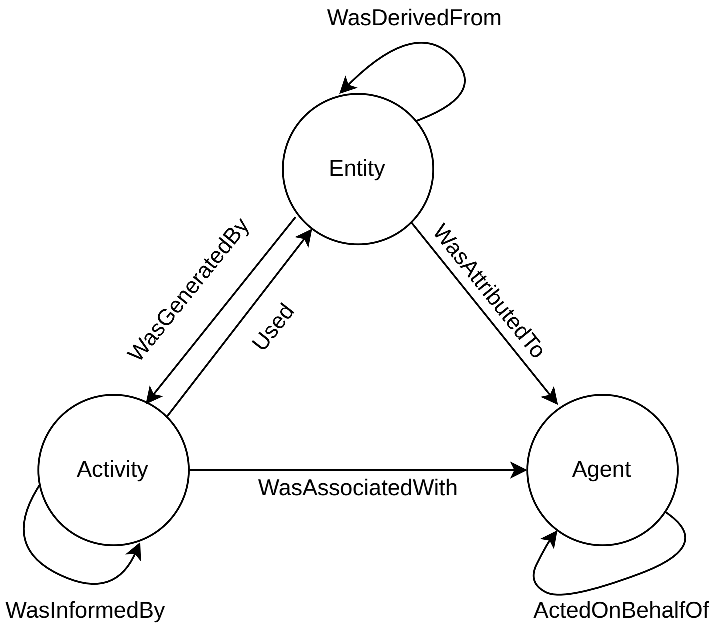

Figure 2.1. W3C PROV-DM structure

2.4 Related Work

Data lineage/provenance in AI systems: Data provenance is getting more attention when the demand for Explainable AI (XAI) increases. For instance, some researchers have proposed how provenance should be implemented: [29] introduced 'Six Ws' framework for provenance graph-based XAI; [30] combined abstraction and reasoning support offered by models with provenance graph which allows tracing the process of producing the current state.

3. Data Lineage for Ensemble ML Serving

3.1 PROV-DM Comparasion with Other Approaches

Before the introduction of PROV data model in 2013, the first version of Open Provenance Model (OPM) [31] has been released in 2007 as the result of the Provenance Challenges series in 2006. This model represents the provenance graph with a DAG which contains three types of vertices: artifact, process, and agent. While *artifact* is the 'immutable piece of state' that portray the focused entities of provenance, the *process* illustrates the course of actions which is enabled, facilitated, controlled, and affected by the *agent* [31, p. 3]. In the provenance graph of OPM, the edges describe the dependencies and causal relationship between the entities, such as the generated by, triggered by, or used relation. From the descriptions, we can recognize that OPM created a foundation for the later development of PROV-DM, and most of their entities and relationship can be directly mapped to the other according to Table 3.1. However, with less relationship, OPM is a more lightweight model that can be employed to present provenance information. However, there are some aspects of PROV-DM that are more beneficial when applying to ML applications than OPM, so the following paragraph will discuss these problems.

Firstly, while OPM was originally developed for scientific workflows, PROV-DM is a domain-agnostics data model which can be applied and extended to describe provenance data in various fields. For its extensibility, PROV-DM is more suitable to capture data lineage of ML as there are numerous applications of ML. For example, Souza et al. [32] included new *referred* and *hadStore* relationship to represent data references in heterogeneous database or Pina et al. [33] proposed a new domain-specific data model based on PROV-DM to represent training-specific data from deep learning experiments called DNNProv-Df.

OPM	PROV
Artifact	Entity
Process	Activity
Agent	Agent
Used	Used
WasGeneratedBy	WasGeneratedBy
WasTriggeredBy	WasInformedBy
WasDerivedFrom	WasDerivedFrom
WasControlledBy	WasAssociatedWith
	ActedOnBehalfOf
	WasAttributedTo

Table 3.1. Comparision between OPM and PROV-DM

Secondly, as OPM is a generic provenance model, its granularity level is lower than PROV-DM. Specifically, while OPM can only describe causal relationships, PROV-DM can represent attribution and delegation in its core and many others in its extended model. In the context of XAI, such a low level of granularity is not sufficient for explaining ensemble ML which requires a decent amount of data about all the base models.

Lastly, the higher adoption of PROV-DM than OPM is beneficial for the ML provider. As all stages of ML lifecyle require provenance for explainability, traceability, and replicability, a holistic data model is advantageous for provenance data analysis throughout the lifecycle.

3.2 Architecture

3.2.1 Overview and Design Principles

As the target of the research is data lineage in the production stage, there are some requirements that must be incorporated in the design principle of the library. Firstly, as ML serving requires aligning with strict SLA about response time, ML provider cannot afford high runtime overhead during prediction. Thus, the proposed library focuses on diminishing runtime provenance capture which will not affect the overall performance of the applications. Secondly, given the numerous numbers of ML applications, the library should be easily integrated into the existing solutions without too much code instrumentation. Moreover, the low instrumentation helps reduce time and resources to incorporate latest updates to the model. Finally, the choice of the database for storing the data lineage needs to be scalable to support large amount of writing request while foster various kinds of provenance and data analysis. With those mentioned requirements, I implemented the library following these main principles:

- Lightweight:
- Scalability
- Asynchronicity

3.2.2 Library

4. Experiment and Results

4.1 Supported Analysis with Data Lineage

4.2 Pricing Model based on Resource Usage and Accuracy

5. Conclusion

Bibliography

[1] A. Terragni and M. Hassani, "Optimizing customer journey using process mining and sequence-aware recommendation," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, (New York, NY, USA), p. 57–65, Association for Computing Machinery, 2019.

[2] D. Derkach, M. Hushchyn, T. Likhomanenko, A. Rogozhnikov, N. Kazeev, V. Chekalina, R. Neychev, S. Kirillov, F. Ratnikov, and on behalf of the LHCb collaboration, "Machine-learning-based global particle-identification algorithms at the lhcb experiment," *Journal of Physics: Conference Series*, vol. 1085, p. 042038, sep 2018.

[3] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, "Predicting mental health problems in adolescence using machine learning techniques," *PLOS ONE*, vol. 15, pp. 1–13, 04 2020.

[4] M.-L. Nguyen, T. Phung, D.-H. Ly, and H.-L. Truong, "Holistic Explainability Requirements for End-to-End Machine Learning in IoT Cloud Systems," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, (Notre Dame, IN, USA), pp. 188–194, IEEE, Sept. 2021.

[5] S. K. Jagatheesaperumal, Q.-V. Pham, R. Ruby, Z. Yang, C. Xu, and Z. Zhang, "Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 2106–2136, 2022. Conference Name: IEEE Open Journal of the Communications Society.

[6] L. Truong and T. Nguyen, "QoA4ML – A Framework for Supporting Contracts in Machine Learning Services," 2021 IEEE International Conference on Web Services (ICWS);, p. 11, 2021.

[7] K. Razavi, M. Luthra, B. Koldehofe, M. Mühlhäuser, and L. Wang, "FA2: Fast, Accurate Autoscaling for Serving Deep Learning Inference with SLA Guarantees," in *2022 IEEE 28th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 146–159, May 2022. ISSN: 2642-7346.

[8] Y. Bai, L. Chen, M. Abdel-Mottaleb, and J. Xu, "Automated Ensemble for Deep Learning Inference on Edge Computing Platforms," *IEEE Internet of Things Journal*, vol. 9, pp. 4202–4213, Mar. 2022.

[9] M. Herschel, R. Diestelkämper, and H. Ben Lahmar, "A survey on provenance: What for? What form? What from?," *The VLDB Journal*, vol. 26, pp. 881–906, Dec. 2017.

[10] Y. R. Wang and S. E. Madnick, "A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective,"

[11] M. Schmitz, "Survey on Data Quality and Provenance,"

[12] A. Woodruff and M. Stonebraker, "Supporting fine-grained data lineage in a database visualization environment," in *Proceedings 13th International Conference on Data Engineering*, pp. 91–102, 1997.

[13] P. Buneman, A. Chapman, and J. Cheney, "Provenance management in curated databases," in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, (New York, NY, USA), p. 539–550, Association for Computing Machinery, 2006.

[14] C. Ré and D. Suciu, "Approximate lineage for probabilistic databases," *Proc. VLDB Endow.*, vol. 1, p. 797–808, aug 2008.

[15] T. Heinis and G. Alonso, "Efficient lineage tracking for scientific workflows," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, (New York, NY, USA), p. 1007–1018, Association for Computing Machinery, 2008.

[16] Y. Cui, J. Widom, and J. L. Wiener, "Tracing the lineage of view data in a warehousing environment," *ACM Trans. Database Syst.*, vol. 25, p. 179–227, jun 2000.

[17] Y. Cui and J. Widom, "Lineage tracing for general data warehouse transformations," *The VLDB Journal — The International Journal on Very Large Data Bases*, vol. 12, pp. 41–58, 09 2001.

[18] D. Logothetis, S. De, and K. Yocum, "Scalable lineage capture for debugging disc analytics," in *Proceedings of the 4th Annual Symposium on Cloud Computing*, SOCC '13, (New York, NY, USA), Association for Computing Machinery, 2013.

[19] B. Glavic, *Big Data Provenance: Challenges and Implications for Benchmarking*, vol. 8163, pp. 72–80. 01 2014.

[20] X. Rong-na, L. Hui, S. Guo-zhen, G. Yun-chuan, N. Ben, and S. Mang, "Provenance-based data flow control mechanism for internet of things," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 5, p. e3934, 2021.

[21] L. G. Azevedo, R. Souza, R. M. Thiago, E. Soares, and M. Moreno, "Experiencing ProvLake to Manage the Data Lineage of AI Workflows," in *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação (Anais Estendidos do SBSI 2020)*, (Brasil), pp. 206–209, Sociedade Brasileira de Computação (SBC), Nov. 2020.

[22] F. T. Jaigirdar, C. Rudolph, G. Oliver, D. Watts, and C. Bain, "What Information is Required for Explainable AI? : A Provenance-based Research Agenda and Future Challenges," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, (Atlanta, GA, USA), pp. 177–183, IEEE, Dec. 2020.

[23] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, "Data quality in internet of things: A state-of-the-art survey," *Journal of Network and Computer Applications*, vol. 73, pp. 57–81, Sept. 2016.

[24] S. Sun, S. Wang, and Y. Wei, "A new ensemble deep learning approach for exchange rates forecasting and trading," *Advanced Engineering Informatics*, vol. 46, p. 101160, 2020.

[25] Y. Chen, Y. Wang, Y. Gu, X. He, P. Ghamisi, and X. Jia, "Deep learning ensemble for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 6, pp. 1882–1897, 2019.

[26] S. Das, M. Ashrafuzzaman, F. T. Sheldon, and S. Shiva, "Network intrusion detection using natural language processing and ensemble machine learning," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 829–835, 2020.

[27] P. M. Luc Moreau, "Prov-dm: The prov data model,"

[28] A. Kale, T. Nguyen, F. C. Harris, C. Li, J. Zhang, and X. Ma, "Provenance documentation to enable explainable and trustworthy AI: A literature review," *Data Intelligence*, vol. 5, pp. 139–162, Mar. 2023.

[29] F. T. Jaigirdar, C. Rudolph, G. Oliver, D. Watts, and C. Bain, "What Information is Required for Explainable AI? : A Provenance-based Research Agenda and Future Challenges," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, (Atlanta, GA, USA), pp. 177–183, IEEE, Dec. 2020.

[30] O. Reynolds, A. García-Domínguez, and N. Bencomo, "Automated provenance graphs for models@run.time," in *Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, (Virtual Event Canada), pp. 1–10, ACM, Oct. 2020.

[31] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson, "The Open Provenance Model: An Overview," in *Provenance and Annotation of Data and Processes* (J. Freire, D. Koop, and L. Moreau, eds.), vol. 5272, pp. 323–326, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. Series Title: Lecture Notes in Computer Science.

[32] R. Souza, L. Azevedo, R. Thiago, E. Soares, M. Nery, M. A. S. Netto, E. Vital, R. Cerqueira, P. Valduriez, and M. Mattoso, "Efficient Runtime Capture of Multiworkflow Data Using Provenance," in *2019 15th International Conference on eScience (eScience)*, (San Diego, CA, USA), pp. 359–368, IEEE, Sept. 2019.

[33] D. Pina, L. Kunstmann, F. Bevilaqua, I. Siqueira, A. Lyra, D. De Oliveira, and M. Mattoso, "Capturing Provenance from Deep Learning Applications Using Keras-Prov and Colab: a Practical Approach," *Journal of Information and Data Management*, vol. 13, Dec. 2022.