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 Käpylä
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** SCI
Supervisor professor Lauri Malmi	
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

Recently, Machine Learning (ML) has been applied extensively to many problems, including customer journey optimization in marketing [7], particle identification in Physics [2], and mental health problems prediction in healthcare [6]. Then, with the increasing number of ML applications and their proven performance, several organizations are seeking third-party ML service providers to improve their operations, creating a new business model — ML as a Service (MLaaS). In MLaaS, there are two major business engagement models. The simpler version is the two-stakeholder engagement model, where the service provider utilizes the data from their customer to provide the ML service, including creating and running that model in production [8]. With the complexity of the problem, we only focus on the two-stakeholder model in this research.

As with other kinds of service, MLaaS requires a contract, 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 job because ML-specific attributes, such as data quality, inference accuracy, and explainability, have created new challenges that demand new solutions. For instance, the popular relationship between data quality and inference accuracy indicates that if the customer submits data unsuitable for the defined purpose, ML models can produce false predictions, causing ML providers to be penalized based on the SLA. Another issue is that in two-stakeholder MLaaS, the customer does not cooperate with the ML provider in the prediction task. Thus, from their perspective, ML models are considered like a black box, which is insufficient in human life-related applications, such as e-health and autonomous vehicles.

One approach to resolve these two challenges is increasing model explainability by accurately explaining the inference result to the customer XAI, which would require tracing down the decreased performance's root cause, i.e., data quality. To obtain that information, ML providers first need to identify data processing from end to end manually, and in the current situation, it would demand considerable work. This is because ML providers do not only utilize one model but an ensemble of many models. Although the ensemble model increases the prediction accuracy and the robustness, it increases the pipeline complexity, e.g., the source data is processed differently for each base model and at different nodes. Thus, this research enhances the current QoA4ML framework with data lineage monitoring as one approach to improve explainability in MLaaS.

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 data quality in the edge and discusses the data lineage definition, characteristic and its relationship with data quality. We then identify the challenges of monitoring data lineage for ML serving.

2.1 Data quality and Data dimensions

In [4], Liu et al. defines data quality as the suitability of the data for the application and data dimension as metrics that describe the level of data quality. As the definitions are pretty general, data dimensions depend highly on the application, and different organizations value data dimensions differently [1]. For example, although completeness is valuable when working with tabular data, it is not as relevant in computer vision. Instead, attributes such as sharpness, noise, or dynamic range, are more significant and can affect the inference accuracy heavily. In edge computing and IoT applications, data characteristics are uncertain, erroneous, and noisy [4], so it demands precise monitoring of the whole pipeline for fault tolerance and problem investigation.

2.2 Data lineage

Data lineage or provenance describes the origin of data, its derivation and changes over time [3]. Based on which questions it can answer, Ikeda and Widom [3] classify data lineage into two types: *where*-lineage and *how*-lineage. Then for each class, there are two *granularities*:

- Schema-level (coarse-grained) answers data lineage questions at the general level. It can be which datasets our current data originate from for where-lineage and which transformations have been applied for how-lineage

- Instance-level (fine-grained), on the other hand, clarifies the origin of a specific data point and how different data points are combined to achieve such results.

The answers that data lineage can provide will shine in edge computing because the complicated data pipeline makes it very time-consuming and costly for ML providers to trace down the root cause of the problems manually. However, with data lineage, they can immediately identify the data input's origin and derivation and then confirm the data quality based on them, helping reduce the required work [5]. Although data lineage is precious in MLaaS, there are some challenges requires considering in this research:

- Storing and querying lineage can be expensive when the number of IoT devices and the complexity of the pipeline increase[3]. Additionally, a resource-constrained environment like the edge demands an even more efficient processes.

- Current ML solution adopts many open source application and library that can be considered as a black box, so how can we accurately capture data provenance for black box operations and many others [5]

In progress

3. Data lineage for ensemble ML serving

4. Experiment and Results

5. Conclusion

Bibliography

[1] Corinna Cichy and Stefan Rass. An Overview of Data Quality Frameworks. *IEEE Access*, 7:24634–24648, 2019.

[2] Denis Derkach, Mikhail Hushchyn, Tatiana Likhomanenko, Alex Rogozhnikov, Nikita Kazeev, Victoria Chekalina, Radoslav Neychev, Stanislav Kirillov, Fedor Ratnikov, and on behalf of the LHCb collaboration. Machine-learning-based global particle-identification algorithms at the lhcb experiment. *Journal of Physics: Conference Series*, 1085(4):042038, sep 2018.

[3] Robert Ikeda and Jennifer Widom. Data lineage: A survey. Technical report, Stanford InfoLab, 2009.

[4] Aimad Karkouch, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73:57–81, September 2016.

[5] Mingjie Tang, Saisai Shao, Weiqing Yang, Yanbo Liang, Yongyang Yu, Bikas Saha, and Dongjoon Hyun. SAC: A System for Big Data Lineage Tracking. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1964–1967, April 2019. ISSN: 2375-026X.

[6] Ashley E. Tate, Ryan C. McCabe, Henrik Larsson, Sebastian Lundström, Paul Lichtenstein, and Ralf Kuja-Halkola. Predicting mental health problems in adolescence using machine learning techniques. *PLOS ONE*, 15(4):1–13, 04 2020.

[7] Alessandro Terragni and Marwan Hassani. Optimizing customer journey using process mining and sequence-aware recommendation. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 57–65, New York, NY, USA, 2019. Association for Computing Machinery.

[8] Linh Truong and Tri Nguyen. QoA4ML – A Framework for Supporting Contracts in Machine Learning Services. 2021 IEEE International Conference on Web Services (ICWS);, page 11, 2021.