A concept for fitness-for-use evaluation in Machine Learning pipelines

David Jonietz
HERE Technologies
Zurich, Switzerland
david.jonietz@here.com

Abstract

Managing data quality is a central task for Machine Learning (ML) applications but often done in an ad hoc manner. Particular challenges for data quality evaluation include ensuring the validity and reproducibility of its results, especially in the context of versioned, multi-step data processing pipelines and increased reuse of data sets for a range of different ML tasks. In this paper, we explore a concept of fitness-for-use which aims to standardize the evaluation process, ensure the validity of the results, and improve its reproducibility by closely intertwining all components of the full ML pipeline in the process. On this basis, we further discuss how relevant historical fitness-for-use scores could be identified to inform the fitness-for-use assessment process for a new ML task.

1 Introduction

Undoubtedly, data plays a central role in modern Machine Learning (ML) research and applications. However, research of ML systems is often driven by widely-used benchmark data sets with sometimes poorly understood quality, while the resulting models are at the same time increasingly applied in real-world applications with considerable impact on people's lives [Paullada et al., 2020]. For this reason, understanding, evaluating, monitoring and ensuring the quality of data should play a central role for ML both in academic and industry settings. Potential quality-related issues such as label noise, class imbalance, data heterogeneity or data incompleteness [Jain et al., 2020] can reduce the efficiency of the training process, lower the bound on achievable model accuracy [Cortes et al., 1995, Gupta et al., 2021], or complicate reproducibility of experimental results [Paullada et al., 2020].

Managing data quality for ML systems, however, is non-trivial and - among others - involves addressing the following challenges:

- How can we ensure that data quality metrics are appropriate and valid with regards to the intended ML task?
- How can data quality metrics be designed which are appropriate and valid across different ML tasks?
- How can efficient and reproducible data quality assessment be managed in the context of versioned data pipelines with multiple processing steps (data provenance/lineage)?

In order to ensure that a measure of data quality is valid for a specific task, task-dependent fitness-for-use (FFU), i.e., the evaluation of the usefulness of data with regards to a particular application [Ziegel, 1990], can be assessed instead of general data quality. In principle, this allows to focus the quality evaluation on a sub-set of data set characteristics of relevance to the downstream ML task, while at the same time acknowledging the fact that insufficient quality of a data set for one ML task does not determine its uselessness for other tasks. For instance, a highly downsampled version of the

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

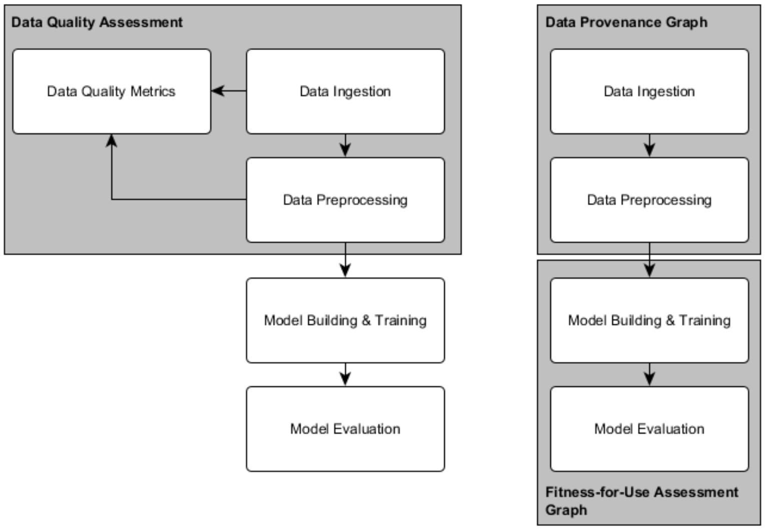

Figure 1: Data quality assessment (left) vs. the proposed FFU evaluation (right) in a ML pipeline

cats and dogs image data set [Parkhi et al., 2012] might no longer allow distinguishing cats from dogs but still be highly useful for the task of segmenting the pet's body from the image background. A practical disadvantage of FFU compared to general data quality, however, is its restricted validity to a specific task. Furthermore, the design of useful metrics requires a high level of domain knowledge and is often performed in an exploratory manner rather than grounded by theory. Finally, the practical problem of how to manage reproducible data quality evaluation in typical multi-step data processing pipelines - where data quality issues might be introduced at each processing step - and possibly across different versions of such pipelines is not addressed.

In this paper, we argue that some of the challenges mentioned previously can be addressed by closely tying the FFU evaluation process to the entire ML pipeline. Concretely, our approach:

- assigns a FFU score not to a data set, but to a specific version of the entire data pipeline from the import of the (raw) data up until its ingestion by the ML model (data provenance/lineage) to standardize the FFU evaluation process and ensure reproducibility
- infers FFU directly from the evaluation results of a ML model instead of designing a hand-crafted quality measure to ensure its task-specific validity
- proposes a possible direction towards evaluating the potential validity of FFU scores across different ML tasks

2 Evaluating fitness-for-use in ML pipelines

As figure 1 illustrates (on the left side), a data quality evaluation procedure is typically decoupled from the downstream ML task, and can occur at any of the stages of the data processing pipeline, e.g., immediately after ingestion of the raw data or further downstream after several preprocessing steps. In contrast, in this paper we propose to intertwine the full ML pipeline into the FFU evaluation procedure, as illustrated in figure 1 (on the right side). For this, we conceptually divide the computational graph which represents the entire ML pipeline on an abstract level into two sub-graphs with different purposes:

- Data Provenance Graph (DPG_ϵ): This sub-graph represents the set of computational operations which implement the data processing pipeline from the import of a (raw) data set to its preprocessing steps such as filtering or feature engineering up until its ingestion by a ML model. The abstract sub-graph is made concrete by parameters ϵ specifying the actual data set and the set of concrete computational operations. In our concept, FFU scores will be calculated for and assigned to DPG_ϵ objects.

- Fitness-for-use Assessment Graph (FUG_θ): This sub-graph represents the set of computational operations which implement the training and evaluation of a ML model on the preprocessed data (i.e., the output of DPG_ϵ), as well as the translation of the task-specific evaluation results to a universal FFU score which is agnostic to the exact evaluation metric used and therefore comparable across different FUG objects. The abstract sub-graph is made concrete by parameters θ specifying the concrete set of computational operations implementing e.g., the ML model architecture, training and evaluation process and FFU calculation procedure.

The FFU Assessment Graph can therefore also be re-interpreted as a function

$$FUG_\theta : DPG_\epsilon \longrightarrow k_{\theta,\epsilon} \tag{1}$$

which describes how a certain version of a data processing pipeline (DPG_ϵ) - i.e., the raw data plus the processing operations - is mapped to a FFU score $k_{\theta,\epsilon}$ by using its outputs to train and evaluate a ML model, and translate the evaluation results to the final FFU score (FUG_θ). The final translation step from the evaluation metric to a FFU score can be implemented via any suitable function f which translates a vector (or scalar if only a single metric is used) of evaluation metrics $\mathbf{m}_{\theta,\epsilon}$ to a scalar FFU score $k_{\theta,\epsilon} \in \mathbb{R}$:

$$f_\theta : \mathbf{m}_{\theta,\epsilon} \longrightarrow k_{\theta,\epsilon} \tag{2}$$

A concrete, simple example for an implementation of f could be a conversion into a binary measure which indicates whether the ML task was solved with sufficient accuracy by the ML pipeline (e.g., based on a user-defined threshold value for the evaluation metrics). More complex functions could include e.g., a relative measure setting the evaluation results of the ML model in relation to human-level performance or another suitable baseline. The resulting FFU score $k_{\theta,\epsilon}$ links DPG_ϵ and FUG_θ objects, and can be stored as part of the meta-data of DPG_ϵ.

3 Towards evaluating the validity of fitness-for-use metrics across ML tasks

In general a FFU score is intrinsically valid exclusively for the specific task it was designed for. Thus, estimating the potential usefulness of a given data set for a certain ML task would normally require designing and computing a new FFU score specifically for this task. In many cases, however, data sets are reused multiple times for varying purposes [Koesten et al., 2020], using different preprocessing steps, ML model architectures and evaluation metrics. In practice, how useful a data set proved to be for other, similar ML tasks in earlier experiments often provides a first indication of its expected usefulness for a new ML task - even before designing, implementing and evaluating a novel, task-specific FFU score. As an example, the results of a cat vs. dog classification experiment might to a degree indicate (but not determine) the potential usefulness of the used data set for a new breed prediction task, while the FFU observed on a very different task such as segmenting the pet from the image background might be of lower validity here. The concept presented in this paper can support such inference processes, in particular the identification and retrieval of historical FFU scores of potential validity and relevance for a new ML task. In particular, the validity of the FFU score for the original ML task is ensured since it is derived from the task-dependent model evaluation results directly while - due to the translation to a universal score - it is also meaningful with regards to the new ML task. Further, the similarity of ML tasks can partly be assessed based on their respective Fitness-for-use Assessment Graphs FUG_θ. Finally, since FFU scores are not assigned to data sets, but rather to entire data processing pipelines (Data Provenance Graphs DPG_ϵ), the process can inform about potentially suitable data processing workflows for a new ML task.

Concretely, in case of a data set having been applied to different ML tasks before, there would be a set of pre-computed FFU scores $K = \{k_{\theta,\epsilon}, ..., k_{\delta,\omega}\}$ where k refer to the FFU scores resulting from previous applications of ML pipelines consisting of different Data Provenance Graphs $\{DPG_\epsilon, ..., DPG_\omega\}$ and different Fitness-for-use Assessment Graphs $\{FUG_\theta, ..., FUG_\delta\}$. Please note that in this example, all versions of the data processing pipeline $\{DPG_\epsilon, ..., DPG_\omega\}$ share the same raw data set (e.g., cats and dogs), whereas different preprocessing steps might have been applied for different analytical tasks. The problem of inferring the potential FFU of the data set (plus processing pipeline) $DPG_\kappa \in \{DPG_\epsilon, ..., DPG_\omega\}$ for the new ML task $FUG_\rho \notin \{FUG_\theta, ..., FUG_\delta\}$

can therefore be reinterpreted as extracting from the set K of precomputed FFU scores the subset of those which are potentially valid for the current ML task as well, and can therefore provide a first estimate of the expected $k_{\rho,\kappa}$. Please note that FUG_ρ needs to be known, i.e., the ML model architecture, training and evaluation procedures needed to solve the new ML task should be pre-defined at this stage.

This process involves two separate steps: First, the subset $K' \subseteq K$ needs to be extracted where the sub-graphs $\{FUG_\theta, ..., FUG_\delta\}$ are of sufficient similarity to FUG_ρ, i.e., similar ML model architectures, training regimes and evaluation procedures have been used compared to what is planned with regards to the new ML task. For instance, if both the cat vs. dog classification task and the breed prediction task involve similar shallow Convolutional Neural Network (CNN) architectures, their respective FUG sub-graphs would be considered more similar (at least in terms of the model architecture) than e.g., the image segmentation task which might require a more complex encoder-decoder architecture. Similar computational operations in the FUG sub-graphs indicate similar characteristics of the learned patterns (e.g., locality) required for solving both ML tasks. High FFU scores for ML tasks requiring such types of patterns therefore suggest that quality issues which would generally obscure such patterns (e.g., bad lighting conditions, obstructions or strong distortions in the image data) are neither present in the raw data nor accidentally introduced at some step of the data processing pipeline. This alone, however, does not guarantee the existence of the specific patterns in the input data which are required to successfully solve the new ML task. Ultimately, apart from experimental work this can only be pre-assessed by a domain expert evaluating the similarity or overlap of patterns required to solve both respective tasks.

Of course, a main challenge is how to assess the similarity between sub-graphs of type FUG in an automated manner. In principle, any function $sim()$ could be suitable which assesses the pairwise similarity of the sub-graph FUG_ρ to all elements of the set $\{FUG_\theta, ..., FUG_\delta\}$, or in other words maps (θ, ρ) to a similarity score which can then be compared to a user defined threshold value t:

$$K' = \{k_{\theta,\epsilon} \in K \mid sim(\theta, \rho) > t\} \tag{3}$$

After additional filtering by a domain expert, the sub-graph DPG can be extracted from the subset K' which maximizes k, or in other words, from the sub-set of FFU scores which have been achieved for sufficiently similar ML tasks, the preprocessing pipeline is identified which achieved best results:

$$DPG_\kappa = \underset{k_{\theta,\epsilon} \in K'}{\arg\max}\, F : k_{\theta,\epsilon} \longrightarrow DPG_\epsilon \tag{4}$$

Intuitively, while k provides a first indication into the potential usefulness of a data set for a new ML task, the retrieved DPG_κ can provide orientation to the concrete data processing steps which were applied in similar ML tasks in the past, and might therefore be suitable for the new task as well.

4 Conclusion

The concept presented in this paper strongly couples the FFU assessment to the individual parts of the ML pipeline itself. On the one hand, this is achieved by assessing the FFU of the entire Data Provenance Graph instead of an individual 'snapshot' of the data set either before or after various preprocessing steps. In our view, this approach standardizes the FFU assessment process and increases reproducibility, as well as allows to store the resulting scores in the meta-data of the specific data processing pipeline version (while still maintaining a clear link to the FFU Assessment Graph used). This acknowledges the fact that the final predictive performance is not determined by the ML model alone, but also by the data and its preprocessing workflow. On the other hand, the fact that FFU is not a detached quality but directly inferred from the model evaluation results ensures its validity to the actual use case. As we further discuss, this concept can support the process of identifying and retrieving historical FFU scores of potential validity for a new ML task.

Being a preliminary concept, however, there are multiple questions which are left unanswered and subject to future work. For instance, the question of how to translate model evaluation results to comparable FFU scores has only been superficially explored in this paper. Furthermore, the task of computing useful similarity values of Fitness-for-Use Assessment Graphs is highly challenging.

References

C. Cortes, L. D. Jackel, W.-P. Chiang, et al. Limits on learning machine accuracy imposed by data quality. In *KDD*, volume 95, pages 57–62, 1995.

N. Gupta, S. Mujumdar, H. Patel, S. Masuda, N. Panwar, S. Bandyopadhyay, S. Mehta, S. Guttula, S. Afzal, R. Sharma Mittal, et al. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4040–4041, 2021.

A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, and V. Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3561–3562, 2020.

L. Koesten, P. Vougiouklis, E. Simperl, and P. Groth. Dataset reuse: Toward translating principles to practice. *Patterns*, page 100136, 2020.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020.

E. R. Ziegel. Juran's quality control handbook, 1990.