dr hab. Michał Okoniewski
Scientific IT Services
ETH Zurich
Binzmühlestrasse 130
8092 Zürich, Switzerland
michal.okoniewski@id.ethz.ch

Zürich/Altdorf, August 21, 2023

# Review of the PhD thesis

Author: **Mikołaj Markiewicz**
Author's affiliation: **Institute of Computer Science, Faculty of Electronics and Information Technology, Warsaw University of Technology**
Title: **Evaluation of data partitioning strategies for distributed clustering and classification algorithms**
Supervisors: **Prof. Piotr Gawrysiak, Dr Jakub Koperwas**

The PhD thesis of Mr. Mikołaj Markiewicz attempts to explain and solve a particular issue in the scientific area of machine learning from distributed data sources. This discipline is called also federated learning. The task is quite ambitious, as even an implementation of federated querying or searching is often not trivial, due to heterogeneities of data and platforms or network performance issues. Still, the author decided to focus the research on federated learning. This choice is inevitably associated with several compromises.

A first tradeoff to solve is positioning the thesis between the two waves of modern machine learning science. Concepts such as clustering, classification, performance measures of them, have been popular in early 2000's under the umbrella name "data mining". Then the datasets have becoming even bigger, and the second wave of popularity of machine learning, after 2010 came with the concepts of big data, data science and new tools such as deep learning toolboxes. The author chooses to work on the main concepts (clustering, classification) from the well-established old-school data mining, but applied to distributed data sources, which requires more "bringing software to the data" in the style of more modern data science.

It seems that the author feels safer within the earlier wave of the technology. Some of the examples and cases are safe classic ones from the initial time of data science e.g., the use of classic Iris dataset in an example, which is elegant and charming in its simplicity, but it is a really a small dataset. The same with implementations and tools. The authors knows that initially the area used e.g., Java toolboxes, such as Weka or for some years the R machine learning packages have been popular. Also, the first implementations of Apache Spark used Scala, which is JVM-based. The author knows and discusses openly that the more modern implementations are mostly python-based as it is the language in which the younger generation of data scientists "speaks" and implements their new tools almost exclusively. The same with Spark and containerization, represented here by Docker. The author does use these technologies, but in a practical way, making use of them where it makes sense. On the other hand, sticking to older data science practices lets the author use established concepts, especially those related to data modeling in federated database systems. Finding the right balance between

the old and new ways of data science was tricky, but the result is solid. It's done in a way that should make sense to data scientists from different generations.

The main scope and thesis of the work is an attempt to create, use and derive scientific results from a novel test platform that enables quality and performance measurements of machine learning algorithms operating on distributed datasets. This is the goal on the IT engineering side. On the theoretical side of computer science, the author aims into detailed research on the properties of machine learning algorithms in distributed and federated environments, in particular in the case of unusual distributions of the data between the data nodes. This is a challenge often avoided by machine learning researchers, who tend to test their algorithms first on simple, single datasets, or in the case of distributed learning on easy testbeds which are independent and identically distributed data (IID). The author goes against the tide, aiming in processing and analyzing the more difficult to handle, but often closer to real-world scenarios non-IID data. Within the manuscript, the author presents also his favorite and well performing on the non-IID, novel combination of popular algorithms (k-means + OPTICS) in the distributed form.

The thesis is constructed in a logical and almost typical way for PhD theses in IT-related engineering science. The introduction in Chapter 1 includes the motivation part which is a statement of his understanding of the area of data science related to federated learning and explains the room for improvement and problems to solve in it. Then, there is quite clear statement of goals and formulation of the points of the research thesis. The author sets up initial boundaries and expresses the awareness of limitations. This is important, as a PhD thesis is expected usually to have a very deep insight of a well-defined, relatively small but potentially important area of science. Setting up the limitations on the scope of the thesis makes also the whole scientific story clearer.

The Chapter 2, "Foundations" gives a short, focused introduction to the base aspects of machine learning, but also database theory and software techniques relevant for the work presented in the thesis manuscript. First part explains the data partitioning and problems with the data structures in federated data systems, in particular the difference between independent and identically distributed and non-IID.

Chapter 3, "Related work and methods" gives more overview on the related research. Terminology in this area is heterogeneous and confusing at times, which is well pointed out in the beginning (p 29). The chapter also draws attention to the separate area of data simulations, which is non-trivial, especially in the context of the federated learning and important, as non-IID data partitioning simulation is a part of the solution presented in the manuscript. Simulations of useful and meaningful datasets are not trivial also in various specialized data science areas, e.g., in RNA-seq bioinformatics (see Soneson and DeLorenzi, 2013, Bastide, Soneson et al, 2022, compcodeR package of Bioconductor). So, it is convincing that in a more general case it requires a special work, attention and tuning. The section on benchmarking and evaluation platforms also mentions the recent change of focus in data science, which uses more frequently python implementations. Perhaps the whole chapter could get integrated with Chapter 2. It is also hard in this area to find balance between the unprecise story-telling and precise, crisp applications to specific data models. But the author is trying, and he emphasizes rightly the importance of that part of theoretical foundations of the thesis.

In the Chapter 4, there is the main contribution of the thesis described. It is done in mostly systematic way, describing partitioning solution for finding parametrized splits of the data,

splitting based on the distance and splitting based on real densities algorithms. Those two are in the form of pseudocode formalism, while Algorithm 3 is presented descriptively, probably due to potential difficulties in formalization. Then the author introduces the proposed combination of distributed clustering algorithms. Finally presents the architecture of the evaluation platform, comparing it with other two platforms in the Table 4.1 and briefly mentioning its limitations. The author implemented the execution pipeline in an own architecture, which likely gave him the flexibility of working with self-created software. Still, would be interesting to consider here e.g., the use of out-of-the box workflow management systems. But I can imagine that a proper solution with workflows could be laborious, perhaps a topic for a good part of yet another PhD.

Chapter 5 is the description of the experiments performed with the selected evaluation method and data distributions. Tables 5.2 and 5.3 clearly describe algorithms and datasets used for this evaluation. The author discusses and presents numerical results of non-distributed and distributed data execution in a detailed way (partly as the supplementary data in Chapter 7). The main technological and scientific finding of the experimental part is presented convincingly in the Table 5.9.

Chapter 6 includes summarized conclusions with an outline of potential future directions of research, while Chapter 7 is a collection of additional technical details of the evaluation system and supplementary figures.

The overall message of the thesis is clear and convincing. It surely proves the author's capacity to perform independent scientific research in the complex areas of modern computer science. The research area is rather challenging and spans many theoretical and technological aspects and techniques. The author has an extensive knowledge in a large part of it and shows a lot of capabilities of extending the research tasks if needed also into related and supplementary areas such as map-reduce systems, containerization, python implementations in modern data science. The thesis could perhaps benefit from additional discussions and implementations, such as using workflow systems, providing a reproducible code and data with public repositories or containers, or discussing non-numerical datasets applications. But within the scope defined by the author, the research goals have been defined well and solved in a satisfactory way. The findings thesis may be thus used as research-based set of guidelines for the designers and power users of systems utilizing federated machine learning techniques.

In a description of such ambitious topics some mistakes and presentation issues are practically unavoidable. Below there is a listing of various editorial and methodological problems and inconsistencies noticed in the thesis. All the remarks are minor compared with the scientific and educational value of the whole thesis:

Abstract:

- "this paper" is not appropriate in a PhD thesis. In the Polish version is translated well: "ta praca".

Streszczenie:

- "... przez ukazanie jego niedoskonałości" not clear if it is about the algorithm or data distribution. In the English version there is no problem in this sentence.
- "zaadresowany" is a wrongly used anglicism.

Chapter 2. "Foundations":

- It is not clear in some cases where the definitions come from, there are no citations or clear suggestion that they were own statements of the author e.g. 2.2.1, 2.4.1, 2.4.2
- It would be good to explain how the criteria 2.1 and 2.2 may relate to classic statistical tests for distribution independence, such as Kolmogorov-Smirnov test.
- In 2.3 (p 16) the statements about "modern techniques" could be supported with a citation. Similarly, e.g., for the listing of types of classification (pp 19-20), as it is not clear if it comes from e.g. a review paper or is an own consideration of the author. Similarly, no citation for Rand Index (p 24). Similarly: no citation for "FL introduced by Google"
- In the subsection 2.6.2 the criticism of Spark sounds a bit harsh and not quite justified (end of p27) and is not clear if it is author's own opinion or experience or comes from a publication.
- "seamlessly" p 27, IT-jargon buzzword that should be avoided in a scientific text.
- In the subsection 2.6.3 the description of Docker is brief and general. The author misses two important points about it. Docker and containerization in general is an important vehicle for reproducible data analysis and reproducible research. Security is mentioned, but in case of Docker one of the main problems, not described by the author, is the fact that it has potential vulnerabilities in the daemon and the container runtime that may give access to the root account of the computing platform. This makes Docker container forbidden, eg on many academic computing cluster platforms. Instead, typically Singularity or Kubernetes containers are used. The description of Docker does not include also the use of the containers in modern workflow systems (eg NextFlow, AirFlow, Snakemake). Not mentioned is also the containerization as a vehicle of bringing the software to the data, one of the dogmas of modern "big data" science. Both aspects could be relevant for the thesis.

Chapter 3 "Related work":

- The quite confusing bit, likely coming from the dividing of the theoretical part between the separate chapters is repeated three times introduction of non-IID data distributions taxonomy of Kairouz et al, 2021. It comes in the thesis in sections 1.1 (p8), subsection 2.2.2 (pp 15-16) in a descriptive way and again, third time in 4.1 (pp 37-38). Also, in 3.1 (p30), there is a comparison of it with the categorization of Zhu et al, 2021, which in fact explains the difference between the Figures 3.1 and 4.1. Perhaps instead of fragmented and repeated introduction would be good to have a separate section just describing the two fundamental papers.
- In the beginning of 3.2.1 it would be enough to start with "There are"

Chapter 4, "Partitioning methods..."

- The chapter is almost clearly the main contribution one, but it still includes elements of theoretical introduction (pp 37-41). It is hard to clearly see the border between external theory and authors own contributions, likely the later starts around the p 42.
- The covariate shift example on Iris dataset is on one hand simple and educative, on the other hand can be mentioned that it is an old-fashioned but very simplistic classic example.
- What is confusing and at times annoying is the frequent use of first person plural, which I think should not be done in a single-author PhD thesis. There is a lot of "we" and "our" in this chapter and further on (pp 37, 38, 39, 51, 56, 57, 61, 66, 75, 78, 99). It seems not to be a self-use of "pluralis majestatis" or a modern fashion of using personalized pronouns. My guess is that the author copied the style of conference and

journal publications, where in multi-authorship mode it is not that problematic. In case of doubt, perhaps more impersonal style can be applied, such one as eg in the English abstract.

- Figures 4.13 and 4.14, and many others in the thesis, could benefit from more extended descriptions.

Chapter 5 "Experiments"

- In the Table 5.1, it is not clear why 3,6,12 partitions have been selected for testing.
- only in Table 5.3 the author admits openly that the focus is mostly on numerical datasets. This can be guessed with the selection of clustering as a main machine learning method tested but would be good to mention among the limitations. Some of the theoretical descriptions on data partitioning in the first chapters sound like general ones, for all types of data, thus are slightly misleading.

Chapter 6 "Conclusions"

- "PoC" in the meaning of prototype is used only in this and next chapter. It is not fully clear if the author means his platform to be prototypical and what will be a proper, production one?
- "is easy to use and extend" p 100, IT-jargon buzzwords should be avoided in a scientific text

Chapter 7 "Attachments"

- Perhaps it should be named "supplement" or similarly.
- The legends in the figures are very small and hard to read.

Overall, I can conclude that Mikołaj Markiewicz has done a considerable amount of work that proves his academic maturity as a data science researcher. The novel findings of the dissertations can be important in future design and improvement of distributed systems with federated machine learning capacities.
The presented dissertation is original, innovative and makes a creative use of modern complex IT technologies. I liked in particular the selection of the ambitious and non-trivial research goal and independence and persistence in pursuing the solution. The value of the work presented in the thesis has been also confirmed by one publication in a journal and two conference ones. Therefore, I can declare that the thesis meets the formal and customary requirements for doctoral dissertations and Mr Mikołaj Markiewicz can be admitted to the further stages of the PhD procedure.

Dr hab. Michał Okoniewski