
Multi-Center Federated Learning to Cluster Clients with non-IID data

*A thesis submitted in fulfilment of the requirements
for the degree of*

Doctor of Philosophy

in
Analytics

by

Ming Xie

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

August 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Ming Xie* declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science, Faculty of Engg. & IT* at the University of Technology Sydney, Australia.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note:
- Signature removed prior to publication.
[Ming Xie]

DATE: 30th August, 2022

PLACE: Sydney, Australia

ABSTRACT

Federated learning (FL) is a new machine learning paradigm to collaboratively learn an intelligent model across many clients without uploading local data to the server. Non-IID data across clients is a significant challenge for the FL system because its inherited distributed machine learning framework is designed for the scenario of IID data across clients. Clustered FL is a type of FL method to solve non-IID challenges using a client clustering method in the FL context. However, even adopts a client clustering FL method still facing minor problems such as unstable against client-wise outliers and the drop of model performance with model poisoning attack.

To face the aforementioned challenges, the main research objective of the thesis is to study that how to make FL effectively, seamlessly solved non-IID data across clients in horizontal clients partition settings.

The main research objective has been studied from four coherently linked perspectives: (i) how to make FL to address the non-IID distribution of data across different clients in a effective and scalable manner so that they can be applied to real world cases which consists of thousands of client and varies type of devices, (ii) how to make cluster FL methods more robust to client-wise outliers, (iii) how to make better balance between the performance of global models and the extent of personalisation of local models, (iv) how to make FL training more robust to model poisoning attack by density methods.

This thesis proposes a novel FL framework with robust clustering algorithm and secure the models to tackle client-wise outliers as well as model poisoning in the FL system. Specifically, we will develop a robust federated aggregation operator using a bootstrap median-of-means mechanism that can produce a higher breakdown point to tolerate a larger proportion of outliers. All work experiments on three benchmark datasets have demonstrated the effectiveness of the proposed method that outperforms other baseline methods in terms of evaluation criteria.

In short, we develop a original, effective clustered FL baseline algorithm which can improve FL performance in horizontal clients partition settings. We compared proposed work against several state-of-the-art FL algorithms using both synthetic and real-world data.

DEDICATION

dedicate my thesis to my mother and my wife for their love

ACKNOWLEDGMENTS

I, Ming Xie, acknowledge Dr. Chandranath Adak for providing this amazing thesis template. I am truly grateful to my supervisor, Prof. Guodong Long, for his great guidance through every facet of the research world. In the past few years, not only I acquired critical technical knowledge in Federated Learning research topic, but also how to clearly write and communicate those my ideas formally and accurately. In addition, I realize the importance of preparing presentations and engaging audience plays a vital role in research. Likewise, interacting with other academics and building research collaborations. All of this was crucial to my future work and life.

A very special gratitude goes to A/Prof. Lu Qin, who has provided many life-changing ideas and advice not only on academic but also on personal life. Words cannot express my emotions in this regard.

Thanks to all those who shared endless discussion and conversation in UTS building 2. It was great spending time in the building and in the laboratory with all of you.

I am also grateful to all the following people, who have helped and supported me along the way: Prof. Chengqi Zhang, Dr. Jing Jiang, Dr. Xueping Peng, Mr. Wensi Tang, Mr. Jie ma, Dr. Shen Tao, Dr. Xubo Wang, Dr. Wentao Li, Mr. Yang Wang.

I am indebted to Faculty of Engineering & IT of the University of Technology Sydney for the wonderful, inclusive and productive work environment, to LINKAGE scholarship for initially funding my PhD. Thank you all the staff, admins and proof readers of my Faculty, they are very professional. I am extremely grateful for living in Sydney, even though I moved out from time to time, it is a city full of culture diversity, beautiful beach and I love to work here.

Finally, I would like to thank my family, for all the obstacles they quietly removed so I could realize my dreams.

LIST OF PUBLICATIONS

Related to the Thesis :

1. Ming Xie, Jie Ma, Guodong Long, “Personalized Federated Learning with Robust Clustering against Model Poisoning”, Advanced Data Mining and Applications, 2022 [CORE rank B] (Accepted by 10 Aug 2022)
2. Ming Xie, Jie Ma, Guodong Long, Chengqi Zhang, “Robust Clustered Federated Learning”, The Asia Pacific Web (APWeb), 1-15, 2022 [CORE rank B] (Accepted by 7 June 2022)
3. Ming Xie, Jing Jiang, Tao Shen, Yang Wang, Leah Gerrard, Allison Clarke, “A Green Pipeline for Out-of-Domain Public Sentiment Analysis”, International Conference on Advanced Data Mining and Applications, 190-202, 2021, DOI: https://doi.org/10.1007/978-3-030-95405-5_14, [CORE rank B] (**Best Student Paper Award**)
4. Guodong Long*, Ming Xie*, Tao Shen, Tianyi Zhou, Xianzhi Wang, Jing Jiang, “Multi-center Federated Learning: Client Clustering for Better Personalization”, World Wide Web Journal, June 2022, DOI: <https://doi.org/10.1007/s11280-022-01046-x> [**CORE rank A**]

Notes: * indicates equal contributions.

TABLE OF CONTENTS

List of Publications	ix
List of Figures	xv
List of Tables	xvii
1 Introduction	1
1.1 Motivation	1
1.1.1 Background	1
1.1.2 Federated Learning	2
1.1.3 Research Problems	5
1.2 Structure of the Thesis	9
1.3 Claims of the Thesis	10
2 Related Work	13
2.1 Federated Learning Preliminaries	13
2.2 Federated Learning Challenges	15
2.3 Federated Learning with Non-IID Data	18
2.4 Clustering Methods for Federated Learning	24
2.5 Robust Methods for Federated Learning	26
3 Multi-center Federated Learning	37
3.1 Related work	39
3.2 Background	39
3.2.1 Problem Setting	39
3.2.2 Motivation	41
3.3 Methodology	42
3.3.1 Multi-Center Model Aggregation	42
3.3.2 Problem Formulation	43

TABLE OF CONTENTS

3.3.3	Optimization Algorithm	44
3.4	Some Possible Extensions	46
3.4.1	Model Aggregation with Neuron Matching	46
3.4.2	Selection of K	47
3.5	Experiments	47
3.6	Training Setups	48
3.6.1	Experimental Study	49
3.7	Conclusion and Remarks	51
4	Robust Clustering for Mutli-center Federated Learning	53
4.1	Introduction	53
4.2	Methodology	55
4.2.1	Problem Definition	55
4.2.2	Robust clustered FL with bMOM	57
4.2.3	Algorithm	58
4.3	Experiment	58
4.3.1	Training Setups	58
4.3.2	Experiment Analysis	62
4.4	Conclusion and Remarks	63
5	Personalized Federated Learning with LOF against Model Poisoning	65
5.1	Introduction	65
5.2	Motivation	66
5.2.1	Model poisoning and anomaly detection	66
5.2.2	Problem	67
5.3	Methodology	68
5.3.1	PFL	68
5.3.2	LOF	69
5.3.3	Proposed method	70
5.4	Algorithm	71
5.5	Experiments	72
5.5.1	Experimental settings	73
5.5.2	Experimental study	75
5.6	Conclusion and Remarks	76
6	Conclusion	79

6.1 GreenSAP in Federated Learning	80
Bibliography	83

LIST OF FIGURES

FIGURE	Page
2.1 Downpour-SGD Algorithm [34]	15
2.2 Horizontal Federated Learning, Vertical Federated Learning, Federated Transfer Learning [42]	16
2.3 Non-IID data learning in Decentralized ML [56]	22
2.4 Overview of IFCA architecture [46]	23
2.5 Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints [90]	24
2.6 A multi-task learning approach included for Federated Learning [97]	27
2.7 Case study of impact of Neuron matching	34
2.8 Neuron Matching Steps	34
3.1 Comparison between single-center aggregation in vanilla FL (left) and multi-center aggregation in the proposed one (right). Each W_i represents the local model's parameters collected from the i -th device, which is denoted as a node in the space. \bar{W} represents the aggregation result of multiple local models.	41
3.2 A high-level view of Federated Learning	43
3.3 Convergence analysis for the proposed FeSEM with different cluster number (in parenthesis) in terms of micro-accuracy.	50
3.4 Clustering analysis for different local models (using PCA) derived from FeSEM(4) using FEMNIST and Celeba data.	51
3.5 Figure shows the clustering effect of FeSEM on dataset FEMNIST by writers, on the left are three writers handwritten digits which are smaller and lighter than on the right ones	51
4.1 Convergence Analysis from Benchmarks with Model Poisoning Attack	61
5.1 Framework of classical FL	69
5.2 Reachability distance of o, p_1 and o, p_2 , respectively, for $n = 5$	70

LIST OF FIGURES

5.3	Framework of proposed method	71
5.4	Convergence analysis for the proposed FedPRC with different cluster number (in parenthesis) in terms of micro-accuracy.	77

LIST OF TABLES

TABLE	Page
1.1 Close Existing Research	4
3.1 Comparison of our proposed FeSEM(K) algorithm with the baselines on FEM-NIST and FedCelebA datasets. Note the number in parenthesis following “FeSEM” denotes the number of clusters, K	46
3.2 Statistics of datasets.	47
4.1 Statistics of datasets. “# of inst. per dev.” represents the average number of instances per device.	60
4.2 FeSEM v.s. FedRoC	62
4.3 Comparison of our proposed FedRoC(K) algorithm with the baselines on FEMNIST and FedCelebA datasets. Note the number in parenthesis denotes the number of clusters, K	63
5.1 Table of Notations	68
5.2 Statistics of datasets. “#” represents the number of instances.	73
5.3 Comparison of our proposed FedPRC(K) algorithm with the baselines on FEMNIST and FedCelebA datasets. Note the number in parenthesis following “FedPRC” denotes the number of clusters, K	76

