Dean&Francis
ISSN 2959-6157

FedDyn Combined with Dynamic Federated Distillation Approach on Recommender System

Yu Sun[1, *]

[1]Department of Communications Engineering, Xidian University, Xi'an, China

*Corresponding author: Sunyu000908@gmail.com

Abstract:

This article introduces the background of federated learning, highlighting its emergence due to the challenges posed by data distribution and privacy requirements, which limit traditional methods. However, federated learning also faces issues such as bandwidth constraints and communication power consumption, with existing optimization algorithms having their own shortcomings. The main focus of this paper is to study federated learning-related algorithms and optimize federated recommendation algorithms to address data privacy and communication overhead issues. Two reference algorithms are highlighted: the Federated Dynamic Regularizer, which adjusts the local loss function dynamically to facilitate the convergence of local models toward the global optimum, and the Dynamic Federated Distillation, which compresses models to minimize communication costs while improving the reliability of knowledge and safeguarding privacy. By orthogonally combining the codes of both algorithms, the paper can effectively leverage their advantages. Experiments conducted on three datasets—MovieLens-1M, MovieLens-100K, and Pinterest—compare the optimized algorithm FedDyn-DF with baseline methods such as FedAvg and FedProx, as well as the original algorithms. The results show that FedDyn-DF outperforms these methods, converging faster while also protecting privacy. Finally, the paper discusses the limitations of the experiments and future research directions.

Keywords: FedDyn Algorithm; Federated Learning; Recommender System; Knowledge Distillation.

1. Introduction

In McMahan et al. [1], the authors proposed a concept that utilizes data propagated across multiple devices for distributed learning classification tasks without resorting to data sharing which called Fed-

ISSN 2959-6157

erated Learning (FL). Due to the fact that data is often dispersed among different institutions, enterprises, or individuals in reality, it is difficult to directly integrate and share these data for reasons such as commercial competition, privacy protection, and legal regulations. However, the amount of data from a single party is not sufficient to train high-quality models. And the snowball of privacy protection is gradually increasing, posing challenges to traditional centralized machine learning methods. Thus, federated learning emerged [2].

However, during the federated learning process, many mobile and IoT devices face bandwidth constraints, and the energy required for wireless transmission and reception is greater than that for computation [3]. To tackle these challenges, Federated Distillation [4] has been introduced as a federated adaptation of knowledge distillation, aimed at lowering the communication overhead associated with federated learning. Additionally, data-free knowledge distillation methods have been developed to enhance federated learning. Nonetheless, federated distillation still faces several issues, particularly in Non-IID scenarios where the results may be suboptimal.

Dynamic regularization has been Introduced to address the difference between global and local optimal solutions [5]. This method adds a penalty term sent by the server to the learning objective of each device during each training round, in order to make the model of each device converge towards the global optimum. However, the FedDyn algorithm requires frequent information exchange between the client and server due to the need to construct dynamic proxy datasets and extract knowledge. When the model is large, it still incurs significant communication costs. However, Feddyn and many FL optimization algorithms FedAvg [6], FedProx [7], SCAFFOLD [8] have orthogonality [9], and Feddyn is no exception. Combining the two algorithms can further improve algorithm efficiency.

A method for distilling federated knowledge based on the average of server-side logits has been modified to employ focused distillation, enhancing the reliability of knowledge. Additionally, a local differential privacy technique is utilized to safeguard this knowledge on the client side [10]. Due to the orthogonality between FD and algorithms such as Feddyn, this paper combines the improved distillation method with FedDyn algorithm, called FedDyn DF, to solve the non IID and communication overhead problems that originally had biased knowledge in the local model. Experiments indicate that it achieves a quicker convergence rate and reduced communication resource demand compared to the baseline across three datasets: MovieLens-100K, MovieLens-1M, and Pinterest.

2. Research methods

2.1 Federated Dynamic Regularizer

This section aims to address the issue of:

$$argmin_{\theta ? \mathbb{R}^d} \left[\ell(\theta) \triangleq \frac{1}{m} \sum_{k ? [m]} L_k(\theta) \right] \quad (1)$$

Where m represents the number of client devices, and each N_k training instance is contained devices k, the data is independently and identically extracted from the joint distribution $(x, y)P_k$ of device indices, The experience loss of the k-th device is $L_k(\theta) = \mathbb{E}_{(x,y) D_k} \left[\ell_k(\theta; (x, y)) \right]$, and the neural network's parameter is θ.

The following is the algorithm flow of Federated Dynamic Regularizer:

Algorithm 1: Federated Dynamic Regularizer

Input: T, θ^0, $\alpha > 0$, $\nabla L_k\left(\theta_k^0\right) = 0$.

For t = 1,2,...T do

Sample devices $P_t \subseteq$ [m] and transmit θ^{t-1} to each selected device,

For each device $k \in P_t$, and in parallel do

Set $\theta_k^0 = argminL_k(\theta) - ?\nabla L_k\left(\theta_k^{t-1}\right), \theta ?+ \frac{\alpha}{2} \| \theta - \theta^{t-1} \|^2$,

Set $\nabla L_k\left(\theta_k^t\right) = \nabla L_k\left(\theta_k^{t-1}\right) - \alpha(\theta_k^t - \theta^{t-1})$,

Transmit device model θ_k^t to server,

end for

For each device $k \notin P_t$, and in parallel do

Set $\theta_k^t = \theta_k^{t-1}, \nabla L_k\left(\theta_k^t\right) = \nabla L_k\left(\theta_k^{t-1}\right)$,

end for

Set $h^t = h^{t-1} - \alpha \frac{1}{m}\left(\sum_{k \in P_t} \theta_k^t - \theta^{t-1}\right)$,

Set $\theta^t = \left(\frac{1}{|P_t|}\sum_{k \in P_t} \theta_k^t\right) - \frac{1}{\alpha} h^t$

end for

The risk objective outlined in Algorithm 1 modifies the local loss function dynamically, guaranteeing that if the local model achieves a consensus point, it corresponds with the stagnation point of the global loss. When the local device models converge, they move towards the server model, with the convergence point identified as the global loss's stagnation point. By observing the update equation in the algorithm, a series of relationships can be derived when $\theta_k^t \to \theta_k^\infty$, resulting in $\sum_k \nabla L_k\left(\theta_k^t\right) \to \sum_k \nabla L_k\left(\theta^\infty\right) = 0$, which converges to a global risk stagnation point [5]. FedDyn relies on precise minimization, allowing each participating device to dynamically adjust its regularizer in each round. This method guarantees that the optimal model for the regularization loss corresponds with the global empirical loss. Across various federated learning scenarios, it can be effectively trained, exhibiting strong convergence and robustness to device heterogeneity, a large number of devices, partial participation, and imbalanced data. Compared to existing methods such as FedAvg, FedProx, and SCAFFOLD, it has advantages in communication efficiency, convergence speed, and accuracy.

2.2 Dynamic federated distillation

This section is mainly designed based on a proposed dynamic federal distillation method [10]. The algorithm's overall architecture utilizes a client-server model based on federated learning, where each user participating acts as an individual client, uploading only local knowledge (logits) to the server during the training process. The process consists of several key steps: 1. Initialization: All clients, along with the server, initialize their model parameters. 2. Client selection and training involve the server choosing a subset of clients and sending them signals. After receiving these signals, the clients conduct multiple rounds of training with their local data. 3. Knowledge extraction and uploading: Once client training is finished, an ensemble operation is executed to extract and upload local knowledge (including features and logits) to the server. 4. Server update and feedback: The local knowledge is received by the server, which then updates the global model through distillation and returns the updated model to the clients. 5. Steps 2-5 are repeated until the model converges. The process diagram is illustrated in Fig.1.

ISSN 2959-6157

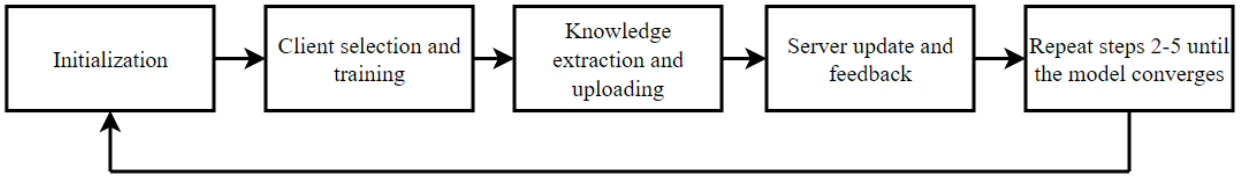

Fig.1 Dynamic federated distillation algorithm process

Section Ensemble is a knowledge extraction method aimed at tackling the problem of randomly constructed data potentially failing to guarantee the system's final convergence, all while improving training efficiency. The operation process can be summarized as the client evaluating all project ID (related to the user ID associated with the client u_k) with a local model and selecting the K projects that have the highest engagement likelihood. Simultaneously, generate fake data using random user IDs and K random item IDs to calculate the features of the chosen user item pairs which are including both synthetic and real data, and upload these features and corresponding logits to the server. Upon receiving the data, the server shuffles it and integrates it to construct the proxy data. The proxy data includes the features d_r and logits l_r which are extracted.

What sets this distillation method apart is that while other federated distillation techniques typically utilize the average of logits, this algorithm employs the logits produced by the client model that correspond to the same user as input. Consequently, this algorithm enhances reliable knowledge and addresses the non-IID issue associated with biased local model knowledge, making server-side optimization more stable and efficient. And the use of methods similar to local differential privacy technology has strengthened privacy protection.

Due to the orthogonality between Federated Dynamic Regularizer and Conventional Distillation, combining Federated Dynamic Regularizer with Dynamic Federated Distillation can further combine the advantages of both algorithms on the basis of the original algorithm. In comparison to the baseline and the original algorithm, it demonstrates improved convergence and greater robustness to device heterogeneity, partial participation, a substantial number of devices, and imbalanced data.

3. Experimental results

3.1 datasets

This experiment involves using three public datasets: MovieLens-1M, MovieLens-100K, and Pinterest. The two datasets of MovieLens are movie rating data, which are converted into implicit data (indicated by 0 and 1 to show whether users rate items), with each user interacting with at least 20 items. Pinterest is an implicit feedback dataset utilized for assessing content-based image recommendations, specifically filtered to include only users who have had a minimum of 20 interactions.

3.2 Analysis of experimental results

This article selects FedAvg, FedProx, FedDF, and the original algorithm Dynamic federated distortion (FedDyn) [10] as baseline methods for comparison. This optimization algorithm primarily focuses on the global model's accuracy on the test dataset., leading to the omission of some method comparisons that are intended for communication efficiency and robustness.

Table 1. Recommendation Performance comparing

		FedDyn-DF	FedDyn	FedAvg	FedProx	FedDF
ML1M	HR@10	0.564	0.552	0.547	0.556	0.345
	NDCG@10	0.320	0.313	0.297	0.301	0.232
ML100K	HR@10	0.598	0.585	0.544	0.557	0.361
	NDCG@10	0.379	0.367	0.346	0.359	0.231
Pinterest	HR@10	0.356	0.343	0.338	0.330	0.204
	NDCG@10	0.201	0.194	0.179	0.174	0.140

The results of the experiments shown in Table 1 demonstrate that the optimized FedDyn-DF surpasses both the original and baseline algorithms, confirming its effectiveness in training accurate recommendation models while

safeguarding user privacy. It is superior to FedAvg, as FedAvg struggles with non-IID issues in Federated Recommendations (FR). FedDF shows the poorest performance, likely due to the non-IID characteristics of user behavior, which hinder the creation of effective proxy datasets. Although FedProx have the ability to guarantee convergence in non-IID situations, As a trade-off, it requires more training epochs. In contrast, FedDyn converges more quickly and can rapidly approach a level close to the optimal value across all evaluation datasets. FedDyn DF adds regularization optimization to FedDyn, making it have a higher convergence speed.

4. Conclusion

The experiment confirmed that the optimized algorithm delivers Improved recommendation Efficiency than the meta algorithm and outperforms baseline methods such as FedAvg, FedProx, and FedDF. It has been shown to effectively train accurate recommendation models while safeguarding user privacy. Similar to the original algorithm, it exhibits advantages in addressing identically distributed and non-independent(non-IID) data issues. In comparative experiments, the optimized algorithm demonstrates superior convergence speed and effectively reduces communication overhead. The limitation of this experiment is the lack of further comparison of methods for designing the robustness and communication efficiency of the algorithm. No adjustments were made to the characteristics of the two algorithms during the combination process. Subsequent research can improve the original algorithm derived from the algorithm characteristics of Federated Dynamic Regularizer and Dynamic Federated Distillation, so that the combination of the two can have faster convergence speed and stronger robustness.

References

[1] McMahan H B, Moore E B, Ramage D., Hampson S, Arcas B A. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv: Learning, 2016.

[2] Mcmahan H, Moore E, Ramage D, Hampson S. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[3] Halgamuge M N, Zukerman M, Ramamohanarao K, Hai L. AN ESTIMATION OF SENSOR ENERGY CONSUMPTION. Progress in Electromagnetics Research B, 2009.

[4] Zhu Z, Hong J, Zhou J. Data-Free Knowledge Distillation for Heterogeneous Federated Learning. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:12878-12889, 2021.

[5] Acar E, Zhao Y, Navarro R, Mattina M, Whatmough P, Saligrama V. FEDERATED LEARNING BASED ONDYNAMIC REGULARIZATION. ICLR 2021.

[6] Collins L, Hassani H, Mokhtari A, Shakkottai S. FedAvg with Fine Tuning: Local Updates Lead to Representation Learning. Advances in Neural Information Processing Systems, 2022.

[7] Li T, Sahu A, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated Optimization in Heterogeneous Networks. arXiv: Learning, 2018.

[8] Karimireddy S, Kale S, Mohri M, Reddi S, Stich S, Suresh A. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. International Conference on Machine Learning, 2020.

[9] Zhang L, Shen L, Ding L, Tao D, Duan L. Fine-tuning Global Model via Data-Free Knowledge Distillation for Non-IID Federated Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10174-10183, 2022.

[10] Jin C, Chen X, Gu Y, Li Q. FedDyn: A dynamic and efficient federated distillation approach on Recommender System. International Conference on Parallel and Distributed Systems (ICPADS), 2023.