Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230643

A study of data heterogeneity in federated learning

ShiKang Wang

International Engineering College, Xi'an University of Technology, Xi'an, 710048, China

3212241036@stu.xaut.edu.cn

Abstract. Data-driven artificial intelligence algorithms cannot do without large amounts of training data. However, the characteristics of data privacy and decentralization make constructing large-scale training data costly, which restricts the further application of artificial intelligence algorithms in different downstream fields. To address the above problems, federated learning has gradually attracted more and more research interests in recent years, which aims to utilize decentralized private data for model training while preserving privacy. However, the non-independent and homogeneous distribution of data across devices causes federated learning to face problems such as data imbalance and label bias, which in turn affects the generalization performance of the model. The problem of data heterogeneity has become a major key challenge in federated learning, and this paper aims to explore the impact of data heterogeneity on federated learning and to synthesize recent research results in this area. By analyzing different solution approaches from the aspects of adaptive data distribution, adding regularization terms, contrastive Learning, and multi-task learning, a comprehensive overview is provided for researchers. This paper further summarizes the existing challenges of data heterogeneity in the research field of federated learning and discusses its potential development directions.

Keywords: Federated Learning, Data Heterogeneity, Artificial Intelligence.

1. Introduction

Machine learning has risen rapidly in recent years, largely due to advances in algorithms, increased computational power, and an influx of available data. With the digitization of all areas of society, data has become a key component driving the success of machine learning. However, machine learning requires more data in a wide range of scenarios to enable more accurate predictions and more robust models. As the demand for data continues to grow, data privacy issues have become more prominent. Especially when data is distributed between different devices, sensitive information involving privacy is more likely to be compromised, such as the regulations implemented in the European Union stemming from the GDPR [1] published in 2018, which aims to protect the security of users' personal information. This regulation emphasizes the need for operators to clearly communicate their commitment to abide by user agreements and strictly prohibits deceptive or leading access to users' private information. In addition, the regulation imposes restrictions on operators, requiring them to refrain from extracting personal data from their training models without obtaining users' permission. The proposed regulation further increases the difficulty and complexity of traditional machine learning training.

In this context, federated learning [2] is proposed as a novel distributed training method. Multiple users can jointly train their respective distributed data under the guidance without need transfer raw data to a centralized server. The private data of each user is retained locally for training, a process coordinated by a centralized server that sends the latest global model to participating training round. The core advantage of federated learning is that it avoids aggregating all data on a single device, thus overcoming the problems associated with privacy breaches and data transfer. But with the continuous expansion of practical applications, the problem of data heterogeneity [3-5] becomes more and more prominent.

Much effort has been devoted to mitigating this issue. Li et al. [6] studied the problem of federated optimization in heterogeneous networks. The problem of slowing down the convergence on the global model when the data from different devices is not evenly distributed was found. Li et al. [7] pointed out that the non-independent homogeneous distribution of data across clients may lead to the degradation of the model's generalization performance and emphasize the urgency of addressing data heterogeneity. Yu et al. [8] explored the trade-off between balancing model robustness and accuracy in distributed learning. They emphasized that data heterogeneity may cause globally shared models to perform poorly on certain devices, thus reducing model robustness and accuracy. This research highlights the important impact of data heterogeneity on striking a balance in federated learning. Hanzely et al [9] questioned the utility of global models, arguing that they are far from typical user usage. In numerous applications, the data distributions of different clients are highly non-independently homogeneous. This statistical heterogeneity makes it difficult to train a single model applicable to all clients. Some algorithms for data heterogeneity emerge in response to the problems that arise.

The challenges posed by the non-independent homogeneous distribution of client-side data in federated learning environments will be explored in depth and summarized, as well as the innovative approaches that have been proposed to overcome these challenges. By examining the merits of different research directions and approaches, we aim to provide valuable insights into solving the data heterogeneity problem and promote the effective application and development of federated learning in facing the challenges of data heterogeneity. In this paper, we will also discuss in detail the proposed solution strategies for the data heterogeneity problem arising in federated learning from the four mainstream ideas: data adaptive strategy, incorporation regularization, comparison learning, and multi-task learning.

2. Definition of Statistical Heterogeneity

Few studies have fully addressed the data heterogeneity problem. Therefore, to further improve the effectiveness, it becomes especially necessary to address the data heterogeneity problem This section will focus on the classification of data heterogeneity. Data heterogeneity refers to the differences in statistical properties, feature distribution, label distribution, etc. of data held by different devices in a federated learning environment. This variability may stem from a variety of factors such as the geographical location of the device, user characteristics, and data collection methods.

(1) Quantitative skew. When constructing an ITS model, the traffic flow data of developed cities may be much more than that of ordinary cities, and this difference in data volume will lead to a quantitative imbalance in traffic data between different cities. In this case, the federated learning model may be more influenced by the data of developed cities in the global model training process, while the data of ordinary cities may have less influence on the training of the global model. This may lead to poor prediction performance of the global model on ordinary city-data [10].

(2) Skewed distribution of characteristics. The transaction behaviors and risk characteristics of different users in the financial domain may differ, resulting in a non-independent homogeneous distribution of the transaction data held by them. The financial transaction data of individual users and corporate users may differ in terms of transaction amount, transaction type, risk preference, etc., and this data heterogeneity makes it difficult to generalize the federated learning model to all users on a global scale [11].

(3) Skewed distribution of labels. Medical record data from different hospitals show different distributions due to the type of disease, age of the patient, and other factors. Certain hospitals may be more focused on cardiac patient data, while others are more involved in neurological disease data. This

data heterogeneity may lead to difficulties in finding a uniform model for all hospital data in the global context of federated learning models [12].

(4) Same features but different labels. In the field of automated driving, although the traffic environments in different cities are similar, due to geographic differences and differences in laws and regulations, there may be different classification labels for the same features (e.g., vehicle speed or road type), resulting in the same features but unbalanced labels [13].

(5) Different users may express similar sentiments for the same social media posts, but use different expressions and vocabulary due to factors such as culture, personal preferences, etc., resulting in the same labels (similar sentiments) but different characteristics [14].

3. Method

This section focuses on some of the mainstream ideas and their algorithms that currently address the heterogeneity of federated learning data.

3.1. Adaptive data distribution

Data distribution self-adaptation [15] refers to the ability of algorithms to automatically adjust to the distributional characteristics of different data in machine learning tasks, especially in distributed learning frameworks such as federated learning, to better adapt to the characteristics and distribution of different data. McMahan et al. (2017) first introduced the concept of federated learning and proposed the corresponding Fedavg [2] algorithm, but Zhao (2018) [3] et al. found that non-independent identically-distributed data (non-IID) led to reduced accuracy in the traditional federated learning algorithm FedAvg. To solve this problem, an improved algorithm based on globally shared data is introduced, which focuses on achieving a balance between the globally shared dataset and the local data by creating a small subset of data that is globally shared among all edge devices and pre-training the model in the initial phase of the FedAvg algorithm [2] to randomly select a portion of the globally shared data for distribution to each client. The client uses this shared data and local data to train the model, and the cloud aggregates the local model to update the global model. This approach reduces the weighting differences in the global dataset with a small amount of shared data, allowing each device to have some degree of information from the global shared dataset. This approach is like letting each device "adaptively" utilize the information from the global dataset, thus mitigating the impact of data heterogeneity on model performance.

3.2. Adding regularization terms

Regularization [16] is a technique used in statistical modeling to control the complexity of the model and prevent overfitting. When training a model, adding regularization restricts the range of values of the model's parameters or reduces the absolute values of the parameters. This can help improve the generalization ability of the model and make it perform better on new data.

Li et al [7] proposed a novel framework called FedProx to solve the data heterogeneity problem in federated learning. The FedProx framework solves the data heterogeneity problem by introducing a new regularization term, i.e., Proximal Term. The role of the Proximal Term is to impose a constraint on each device's local model when the global model is updated, making each device's local model closer to the global model. constraints that make the local model of each device closer to the global model when the global model is updated. The introduction of this regularization term can effectively reduce the differences between devices, thus improving the convergence speed and accuracy of the global model. FedProx introduces the "Proximal Weighting" weight allocation strategy to solve the problem of device heterogeneity in federated learning. Different devices have different computing power, storage capacity, and network bandwidth, so the traditional method may lead to the over-contribution of some devices, which may affect the overall model effect. "Proximal Weighting" dynamically adjusts the weight of each device by calculating the similarity between devices to achieve a more balanced contribution. Compared with the current approach FedAvg [2], the FedProx framework performs more robustly in highly heterogeneous federated networks.

Jiang et al. [17] also proposed a better model named FedMatch, where the model improves generalization performance by ensuring that the model produces similar outputs on data from different clients through a consistent regularization method. The algorithm is divided into two consistency regularization methods: Inter-client Consistency Loss uses the predictions of the consensus model to improve the generalization performance, and Data-level Consistency Regularization consistently regularizes the unlabeled data from each client, allowing the model to learn from small changes with valuable information. valuable information from small changes. Together, these approaches enable FedMatch to better utilize unlabeled data in a federated learning environment.

3.3. Contrastive Learning

Contrastive Learning (CLE) [18] is a machine learning method which learns more discriminative feature representations by comparing similarities and differences between different samples. The method is typically used in unsupervised or self-supervised learning tasks, where the model is trained such that similar samples are brought closer together in the feature space and dissimilar samples are pushed away. The core idea of contrast learning is to learn more meaningful feature by maximizing the similarity between similar samples and minimizing the similarity between dissimilar samples, thus performing better generalization performance in subsequent tasks. Tailin Zhou et al [19] proposed a framework called FedFA, the core idea of which is to use a feature alignment method to solve the variability of data from different clients. Each client uses local data to train the model and generate local feature mappings. Subsequently, the client compares the local feature mappings with the shared feature anchors to evaluate the differences between them. These differences are then used to fine-tune the local feature mappings to better match the shared feature space. Finally, the local model is updated using the adjusted feature mappings to achieve consistent classifiers and feature mappings. This approach aims to eliminate data discrepancies and generalization of federated learning models.

3.4. Multi-task learning

In federated learning, the multi-task learning [20] is applied to solve data heterogeneity. Different devices may cover different tasks or label distributions, resulting in data that differ in task dimensions. By jointly training multiple tasks, the model can learn shared feature representations from multiple tasks, and thus migrate and share features across different devices, thus improving the performance and generalization ability of the model. Virgin et al [21] proposed a method called Federated Multi-Task Learning (Fed-MTL) and introduced an innovative solution in this method. The core idea of this method is to improve accuracy of the model by mapping the data from different devices into a common feature space, eliminating the effect of data heterogeneity and thus improving accuracy of the model. Specifically, Fed-MTL uses the kernel trick technique to map the data on different devices to a common feature space through kernel functions. In this feature space, the data relationships between devices are more explicit, overcoming the difficulties caused by heterogeneity. During the training process, each task (device) has a local model, and these local models are trained by joint optimization to produce a global model. This can simultaneously process data from each device, thus improving the performance of the overall model.

4. Future work and challenge

In recent years, the problem of data heterogeneity has limited development in areas such as environmental monitoring, social network analysis, and energy management. In the field of environmental monitoring, differences in environmental data in different regions can lead to models that perform well in a specific region but have insufficient generalization ability to adapt to other regions. In social network analysis, data heterogeneity on different platforms makes it difficult to obtain consistent behavioral patterns of users across platforms. In the field of energy management, the heterogeneity of energy data can lead to inaccuracies in forecasting and planning. The related algorithms provided in this paper have the potential to be applied to it. By using these specific algorithms for data heterogeneity, it is possible to overcome data discrepancies and train to obtain specific models for domains such as environmental monitoring, social network analysis, and energy management to empower industries while protecting data privacy. It is worth noting that when applying various types of federated learning algorithms, such as FedFA and FedMTL, even though important progress has been made in solving problems such as data heterogeneity, it is still necessary to scrutinize whether new challenges are likely to emerge in the process of problem-solving. Such comprehensive thinking helps to better cope with the complexity of real-world applications and ensure the robustness of the algorithms. For example, when applying FedFA, the stability of feature alignment and feature mapping errors may be problematic. For this reason, additional discriminator networks can be introduced to reduce the noise introduced by feature alignment with the help of adversarial training techniques to improve the stability of feature alignment. Also, an adaptive feature alignment mechanism can be designed to adjust the feature mapping according to the data distribution to reduce the error. For using FedMTL, the problems of high computational complexity of Kernelized and difficulty in modeling multi-task relationships may be encountered. To overcome these, the introduction of approximate computational methods, such as stochastic feature mapping or approximate kernel tricks, can be considered to reduce the computational complexity. In addition, complex models, such as graph neural networks, can also be tried to improve the model performance by improving the ability to model relationships between multiple tasks. These solutions help to deal with the problems that may arise in practical applications more comprehensively, thus promoting the further development and application of federated learning.

5. Conclusion

Federated learning has been a popular research topic in the machine learning community in recent years, and it aims to address the data privacy issues that are becoming prominent in the context of the growing demand for data. Federated learning utilizes decentralized private data for model training, however, the homogeneous distribution of data across devices leads federated learning to face problems such as data imbalance and label bias, which in turn affects the generalization performance of the model. Aiming at the above data heterogeneity problem, this paper explores the data heterogeneity on federated learning in detail by analyzing the existing research and introduces the latest research results in this field. Specifically, this paper analyzes different solutions to data heterogeneity in terms of adaptive data distribution, addition of regularization terms, comparative learning, and multi-task learning, etc. This paper further summarizes the existing data heterogeneity challenges and potential development directions in the field of joint learning research.

References

[1] P. Voigt and A. Von dem Bussche, “The eu general data protection regulation (gdpr),” A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, et al., “Communication-effificient learning of deep networks from decentralized data,” arXiv preprint arXiv:1602.05629, 2016.

[3] Zhao, Yue, et al. "Federated learning with non-iid data." arXiv preprint arXiv:1806.00582 (2018).

[4] Li, Qinbin, et al. "Federated learning on non-iid data silos: An experimental study." 2022 IEEE 38th International Conference on Data Engineering (ICDE). IEEE, 2022.

[5] Yang, Liuyan, et al. "Federated Learning for Medical Imaging Segmentation via Dynamic Aggregation on Non-IID Data Silos." Electronics 12.7 (2023): 1687.

[6] Li, T., Konečný, J., Recht, B., & Srebro, N. (2018). On the convergence of federated optimization in heterogeneous networks.

[7] Li, M., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions.

[8] Yu, H. F., Mohamed, A. R., & Smith, M. R. (2020). Understanding and mitigating the tradeoff between robustness and accuracy.

[9] F. Hanzely and P. Richt´arik, “Federated learning of a mixture of global and local models,” arXiv preprint arXiv:2002.05516, 2020.

[10] Kashyap, Anirudh Ameya, et al. "Traffic flow prediction models–A review of deep learning techniques." Cogent Engineering 9.1 (2022): 2010510.

[11] Wen, Fenghua, Zhifang He, and Xiaohong Chen. "Investors' risk preference characteristics and conditional skewness." Mathematical Problems in Engineering 2014 (2014).

[12] Malehi, Amal Saki, Fatemeh Pourmotahari, and Kambiz Ahmadi Angali. "Statistical models for the analysis of skewed healthcare cost data: a simulation study." Health economics review 5 (2015): 1-16.

[13] Lin, Jiaxin, et al. "Road traffic law adaptive decision-making for self-driving vehicles." 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022.

[14] Nam, Benjamin H., Yicheng Yang, and Richard Draeger Jr. "Intercultural communication between Chinese college students and foreign teachers through the English corner at an elite language university in Shanghai." International Journal of Intercultural Relations 93 (2023): 101776.

[15] Fan, Yang, et al. "Learning what data to learn." arXiv preprint arXiv:1702.08635 (2017).

[16] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the lasso and generalizations.CRC press,2015.

[17] Chen, Jiangui, et al. "FedMatch: federated learning over heterogeneous question answering data." Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021.

[18] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

[19] Zhou, Tailin, Jun Zhang, and Danny Tsang. "FedFA: federated learning with feature anchors to align feature and classifier for heterogeneous data." arXiv preprint arXiv:2211.09299 (2022).

[20] Caruana, Rich. "Multitask learning." Machine learning 28 (1997): 41-75.

[21] Smith, Virginia, et al. "Federated multi-task learning." Advances in neural information processing systems 30 (2017).