Методы оценки уровня разнородности данных в федеративном обучении

Е. С. Новикова
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук
novikova@comsec.spb.ru

Я. Чен
Китайский горно-технологический университет
fedora.cy@gmail.com

А. В. Мелешко
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук
meleshko.a@iias.spb.su

Аннотация. **Федеративное обучение позволяет решить задачу обработки данных с ограниченным доступом и, как следствие, дает возможность естественным образом расширять множество данных, на которых обучаются алгоритмы машинного обучения, включая в него конфиденциальные наборы данных. Однако применение федеративного обучения на практике связано с решением множества открытых задач, к таким задачам относится обработка неоднородных данных, которыми владеют разные участники процесса. В данном докладе рассматриваются различные уровни неоднородности, которые возникают при использовании федеративного обучения, и анализируются методы, позволяющие количественно оценить разнородность данных.**

Ключевые слова: федеративное обучение, горизонтальное распределение данных, неидентично распределенные данные, уровень неоднородности, оценка неоднородности данных

I. Введение

В последнее время парадигма федеративного обучения (ФО) получила пристальное внимание со стороны исследователей, поскольку она предлагает практическое решение для построения распределенных аналитических систем, сохраняющих конфиденциальность [1]. Его применение может принести пользу многим предметным областям, таким как цифровое здравоохранение [2], умные города [3] и кибербезопасность [4].

Однако применение ФО по-прежнему сталкивается со многими открытыми исследовательскими проблемами, которые включают в себя необходимость анализа данных разных форматов с разными атрибутами; учета доступности устройств во время обучения модели и неоднородности распределения данных. В последнее время было разработано ряд решений к обработке неоднородности данных в условиях ФО [5]–[7]. Эффективность предлагаемых методов сильно зависит от того, каким образом моделируется неоднородные данные. Решение последней задачи связано с решением другой задачи, обусловленной отсутствием реалистичных наборов данных с различными типами неоднородности данных, которые можно было бы использовать для моделирования взаимодействия клиентов, с разными наборами данных. В [8] авторы показывают, что традиционно используемые в этих целях наборы данных FEMNIST, Shakespeare, Sent140, CelebA и Reddit [4] в основном решают задачу неоднородности предметной области, а другие сценарии не идентично и зависимо распределённых (не-iid, not independent and identically distributed) данных не учитываются.

В настоящее время не существует единого подхода как к моделированию, так и оценки уровня разнородности данных в ФО. В результате сравнение наборов и оценка эффективности различных подходов к формированию глобальной модели в ФО, устойчивых к не-iid данным значительно затруднено. В данной статье рассматриваются различные подходы, используемые для моделирования разделения данных среди клиентов ФО, и исследуются различные методы оценки уровня неоднородности данных между клиентами ФО. Анализируемые метрики тестируются в сценарии ФО,в котором клиенты представлены четырьмя различными наборами данных – UNSW-NB15, BoT-IoT, ToN-IoT и CSE-CIC-IDS2018.

Таким образом, новизна и вклад статьи заключаются в систематизации подходов, используемых для моделирования и оценки уровня неоднородности данных среди клиентов ФО. Авторы также показывают, что наиболее распространенный подход к моделированию гетерогенных данных не решает наиболее важную практическую проблему – ковариантый сдвиг в данных.

Статья построена следующим образом. В разделе 2 рассматриваются и систематизируются различные подходы к моделированию неоднородного разделения данных клиентов. В разделе 3 приводится описание экспериментов и обсуждаются полученные результаты. Завершается статья выводами и уточнением дальнейших направлений работы.

II. Подходы к моделированию распределения данных в федеративном обучении

Прежде чем рассмотреть различные подходы к моделированию разнородных данных, необходимо определить термин разнородных не-iid данных.

Пусть каждый клиент имеет собственный набор данных $D_i = \{(x_i, y_i)\}_{i=0}^{m}$, где $x \in X$ – вектор признаков выборки, а $y \in Y$ является соответствующей меткой, каждая запись данных (x, y) имеет вероятность $P(x, y)$ с кумулятивной функцией распределения $F_{X,Y}(x, y)$. Пусть $D = \cup\{D_i\}_{i=1..N} = \cup\{(X_i, Y_i)\}_{i=1}^{N}$ – множество наборов данных, распределенных по N клиентам, то D считается независимым и одинаково распределенным, если

$$P\big((x,y),(x',y')\big) = P(x,y)\cdot P(x',y') \qquad (1)$$
$$F_{X,Y}(x,y)_{(x,y)\in D_i} = F_{X,Y}(x,y)_{(x,y)\in D_j}$$

где $(x,y) \in D_i, (x',y') \in D_j$ для всех $i,j \in N$

Если какое-либо из этих двух условий не выполняется, данные являются не независимыми и не идентично распределенными данными. В условиях ФО неоднородность в данных может быть определена на уровне меток и на уровне признакового пространства, а также на уровне отношений между признаками и метками [10]. Перекос меток понимается как различия в распределении меток между разными клиентами. Неравномерность распределения (перекос) признаков понимается как различия в распределении признаков, которые выражаются в различном распределении признаков для одного и того же класса на разных клиентах. Под неоднородностью количества данных понимается разница в количестве обучающих данных для разных клиентов.

A. Моделирование и оценка перекоса меток

Перекос меток является одним из наиболее широко используемых и изученных случаев неоднородности данных в ФО [5], [6], [11]–[13]. Обычно этот сценарий данных основан на предположении, что метки между клиентами имеют разное распределение, но признаки в одном классе меток имеют идентичное распределение. Для моделирования такого сценария один набор данных разделяется между клиентами, и существует два способа его разделения: 1) сортировка и разделение, и 2) разделение на основе распределения Дирихле.

Подход сортировки и разделения заключается в сортировке данных по меткам классов, а затем назначении данных, принадлежащих одному классу, одному клиенту. В результате каждый клиент владеет непересекающимся набором классов.

Другой широко используемый подход основан на применении вероятностного распределения Дирихле. Предполагается, что распределение N меток классов между клиентами определяется случайной величиной $\Theta \in R^N$, определенной на симплексе ($\theta_i \geq 0$, $\sum_{i+1}^{N} \theta_i = 1$) и описываемой распределением Дирихле $Dir(\Theta|\alpha)$. Для создания наборов данных, отличных с неидентичным распределением, вектор $\mathbf{q}$ случайным образом выбирается из распределения $Dir(\alpha\mathbf{p})$, где $\mathbf{p}$ определяет априорное распределение классов по N классам, а параметр $\alpha > 0$ характеризует уровень неоднородности. Когда $\alpha \longrightarrow \infty$ все клиенты имеют распределения, идентичные исходному; когда $\alpha \longrightarrow 0$ уровень неоднородности в данных самый высокий, и каждый клиент содержит один случайно выбранный класс. Обычно параметр α имеет значение 0,5.

Параметр α характеризует уровень неоднородности, который используется для создания разнородных наборов данных, и хотя он часто используется в экспериментах для отражения уровня неоднородности данных между клиентами, он не является мерой разнородности данных.

Можно выделить следующие типы показателей, используемых для оценки степени перекоса меток.

- на основе оценки соотношения классов, принадлежащих разным клиентам;
- на основе расчета расстояния между распределением меток
- на основе статистических тестов равенства распределения.

Первый тип мер представлен индексом гетерогенности (HI), предложенным в [12] и χ^2 – расстоянием. Индекс гетерогенности (HI) рассчитывается следующим образом:

$$HI(c) = 1 - \frac{1}{C_{max} - 1}\cdot(c-1) \qquad (1)$$

где c обозначает максимальное количество классов на одного клиента, а C_{max} относится к максимальному количеству классов в наборе данных.

Вторая группа подходов представлена набором метрик в пространстве вероятностных распределений. Наиболее распространенными метриками являются расстояние Хеллингера и дивергенция Дженсена–Шеннона. Они оба ограничены, симметричны и удовлетворяют треугольному неравенству.

Расстояние Хеллингера принимает значения в интервале $[0,1]$, где $H(P,Q) = 0$ относится к идентичным распределениям данных, а $H(P,Q) = 1$ указывает на то, что два распределения далеки друг от друга.

Другим часто используемым методом является метрика Вассерштейна ($W_p(P,Q)$ [14].

Последняя группа подходов представлена тестом Колмогорова–Смирнова (KS-тест), который проверяет, имеют ли два набора данных одну и ту же функцию распределения вероятностей. KS-тест основан на расчете KS-статистики ($KS(P,Q)$), которая вычисляется следующим образом:

$$KS(P,Q) = \sup_x |P(x) - Q(x)| \qquad (1)$$

где $P(x)$ и $Q(x)$ – эмпирические функции распределения двух выборок, а sup – супремум-функция. В [11] $KS(P,Q)$ используется в качестве меры сравнения степени подобия данных в различных экспериментальных сценариях.

B. Моделирование и оценка перекоса признаков

Перекос признаков и перекос отношений между признаками и метками являются наиболее сложными случаями как для построения ФО, так и для моделирования распределения данных. Можно выделить два основных подхода к неоднородному распределению данных. Один основан на использовании нескольких наборов данных из одной предметной области.

Второй подход основан на разделении одного набора данных и включает в себя несколько методов, предлагаемых для разных типов наборов данных. Типичный способ создания не-iid распределения признаков, состоит в добавлении шума к исходной выборке данных [17][18]. Примером такого набора данных является FEMNIST [9], который разработан

специально для проверки чувствительности алгоритмов ФО к не-iid данным. Он является расширенной версией набора данных MNIST (EMNIST), который состоит из рукописных символов, созданных разными авторами и искаженных, созданных для моделирования асимметрии признаков.

Другой способ моделирования перекоса признаков – применение техники переворота меток, когда метки одного класса заменяются на другие [6].

При моделировании неоднородного распределения признаков во временных рядов исходный набор данных обычно разбивается на основе временных меток. Например, в [19] для моделирования неоднородных условий в ФО данные SWAT, моделирующие функционирование водоочистного сооружения в течение 14 дней, были разбиты на несколько временных периодов. В [20] авторы используют другой подход к моделированию разнородных настроек ФО: они используют набор данных датчиков коммерческих транспортных средств [21], который представляет собой многомерный временной ряд, описывающий движение транспортных средств – самосвалов двух разных типов, и разделяют его между клиентами на основе источника данных, т. е. тип самосвала. Аналогичный подход использовался авторами в [22], разделение одного набора данных на подмножества было выполнено с учетом типа датчиков, которые генерируют данные.

C. Оценка степени неоднородности с сохранением конфиденциальности

Все упомянутые подходы требуют обмена информацией, касающейся данных, и распространения меток среди клиентов. Эта необходимость не является ограничением при проектировании и оценке алгоритмов агрегации на не-iid данных, однако на практике она неприемлема. Чтобы решить эту проблему, в [23] предлагается новая метрика $DataSkew$. Он оценивает степень неоднородности на основе точности локальной модели МО и точности модели на других наборах данных:

$$DataSkew = n \cdot \frac{\max\left(\Delta Accuracy_{pairwise}\right)}{\sum_{i=0}^{n} Accuracy_{client_i}} \quad (1)$$

где n – количество клиентов.

III. Практический пример: моделирование неоднородных данных в федеративном обучении

Для проверки поведения описанных метрик мы использовали подход, предложенный в [15], поскольку он отражает наиболее реалистичный сценарий взаимодействия субъектов. Четыре выбранных набора данных сильно отличаются друг от друга: они содержат различные типы атак, а соотношение нормального и вредоносного трафика варьируется от набора к набору. Краткие характеристики выбранных наборов данных приведены в табл. 1.

ТАБЛИЦА I. Характеристики выбранных наборов данных

Набор данных	Год и место создания	Способ генерации данных	Число меток
CICIDS2017	2017, Canada CyberSecurity Hub, Канада	Сетевой трафик от физических машин	Норма и 8 различных типов атак: DoS и DDoS атаки, эксплуатирования уязвимостей, взлом паролей, сканирование портов, веб атаки, ботнет
ToN-IoT	2019, Intelligent Security Group, UNSW, Австралия	Не указано	Норма и 7 различных типов атак: DoS и DDoS атаки, эксплуатирования уязвимостей, взлом паролей, сканирование портов, веб атаки, программы-вымогатели
BoT-IoT	2021, Intelligent Security Group, UNSW, Австралия	инструмент Ostinato	Норма и 5 типов атак: DoS и DDoS атаки, эксплуатирования уязвимостей, сканирование портов, клавиатурный шпион,
UNSW-NB15	2015, Intelligent Security Group, UNSW, Австралия	Программно-аппаратный комплекс IXIA PerfectStorm	Норма и 9 типов атак: генераторы случайного трафика (Fuzzers), сканирование портов, DoS –атаки, взломщики шифров, использование аязвимостей, устновка бэкдоров и запуск шеллкодов, черви

Мы измеряли неоднородность данных на уровне меток и на уровне признаков. Индекс гетерогенности сначала рассчитывался для каждого набора данных и усреднялся для получения интегрального значения. Аналогичным образом рассчитывалась KS-статистика, сначала попарно, а затем усреднялась для всех случаев. Для оценки метрики *DataSkew* мы использовали простую двухслойную сверточную нейронную сеть. Сначала она была обучена на одном наборе данных, а затем ее точность была вычислена на трех других наборах данных, эта процедура была повторена для всех четырех наборов данных. Результаты приведены в табл. 2.

Интересно, что согласно расстоянию Хеллингера степень неоднородности выше на уровне меток, и она выше, чем индекс гетерогенности, поскольку учитывает не только соотношение классов, но и их частотное распределение внутри клиента. KS-статистика не очень показательна, прежде всего потому, что не ограничена сверху и может использоваться только для сравнения наборов данных. Эксперименты показали, что наборы BoT-IoT и Ton-IoT имеют наиболее схожее распределение меток, а попарное сравнение распределения меток с помощью расстояния Хеллингера показало, что наиболее схожи наборы CICIDS2017 и Ton-IoT, также оно показало, что эти два набора имеют наиболее схожее распределение признаков. Расстояние Вассерштейна оказалось крайне малым, а поскольку оно не ограничено сверху, трудно понять, какое значение соответствует крайне неоднородной степени.

Метрика DataSkew указывает на чрезвычайно высокую степень неоднородности данных. В результате ФО на этих наборах данных с использованием функции агрегирования FedAvg [1], которая не подходит для не идентично распределенных данных, точность ФО для клиентов оказалась следующей: 0.32 (CICIDS2017), 0.67 (Ton-IoT), 0.95 (BoT-IoT) и 0.49 (UNSW-NB15).

ТАБЛИЦА II. ОЦЕНКИ НЕОДНОРОДНОСТИ, ВЫЧИСЛЕННЫЕ ДЛЯ НАБОРОВ ДАННЫХ CICIDS2017, TON-IOT, BOT-IOT И UNSW-NB15

Название показателя	Значение
Перекос на уровне меток	
Индекс гетерогенности	0.43
Растояния Хелингера	0.69
KS-статистика	0.42
Перекос на уровне признаков	
Растояния Хелинжера	0.53
Метрика Вассерштейна	0.04
Метрики, сохраняющие конфиденциальность	
Dataskew	8.01

IV. ЗАКЛЮЧЕНИЕ

Наиболее серьезным препятствием для применения FL на практике является неидентичное распределение данных по клиентам, что является наиболее реалистичным сценарием работы с данными. Несмотря на то что было проведено множество исследований, направленных на разработку решений, устойчивых к неидентичным условиям применения FL, все еще существует необходимость в унификации методов тестирования таких подходов, поскольку не существует стандартной методологии для моделирования и оценки неоднородности степени данных.

В данной статье авторы рассмотрели различные методы, которые используются как для моделирования, так и для оценки уровня неоднородности данных у клиентов. Эксперименты показали, что методы, основанные на оценке дивергенции распределений вероятности, являются наиболее подходящим решением при разработке новых алгоритмов обработки неидентичных данных. Однако их применение требует знания распределения признаков на стороне клиента, что делает их неприменимыми в реальных сценариях. Метрика DataSkew предлагает решение, сохраняющее конфиденциальность данных, однако ее применение требует обучения модели на каждом наборе данных с последующим тестированием на других, поэтому ее расчет требует дополнительных вычислительных ресурсов, что может стать проблемой в распределенной среде с ограниченными ресурсами.

Обозначенные проблемы определяют дальнейшее направление работ, заключающееся в разработке единого подхода к измерению и диагностике неоднородностей в данных с сохранением их конфиденциальности.

СПИСОК ЛИТЕРАТУРЫ

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Int. Conf. on Artificial Intelligence and Statistics, 2016.

[2] N. Rieke, J. Hancox, W. Li, F. Milletar et al., "The future of digital health with federated learning," npj Digit. Med, vol. 3, no. 119, 2020.

[3] S. Pandya, G. Srivastava, R. Jhaveri, M. R. Babu, et al. "Federated learning for smart cities: A comprehensive survey," Sustainable Energy Technologies and Assessments, vol. 55, p. 102987, 2023.

[4] E. Fedorchenko, E. Novikova, and A. Shulepov, "Comparative review of the intrusion detection systems based on federated learning: Advantages and open challenges," Algorithms, vol. 15, no. 7, 2022.

[5] K. Hsieh, A. Phanishayee, O. Mutlu, and P. B. Gibbons, "The non-iid data quagmire of decentralized machine learning," in Proc. of the 37th Int. Conf. on Machine Learning, ser. ICML'20. JMLR.org, 2020.

[6] X. Ma, J. Zhu, Z. Lin, S. Chen, and Y. Qin, "A state-of-the-art survey on solving non-iid data in federated learning," FGCS, vol. 135, pp. 244–258, 2022.

[7] W. Lu, J. Cheng, X. Li, and J. He, "A review of solving non-iid data in federated learning: Current status and future directions," in Artificial Intelligence and Machine Learning, Singapore: Springer Nature Singapore, 2024, pp. 58–72.

[8] M. F. Criado, F. E. Casado, R. Iglesias, C. V. Regueiro, and S. Barro, "Non-iid data and continual learning processes in federated learning: A long road ahead," Information Fusion, vol. 88, pp. 263–280, 2022.

[9] S. Caldas, P. Wu, T. Li, J. Konecny, et al, "LEAF: A benchmark for federated settings," CoRR, vol. abs/1812.01097, 2018.

[10] P. Kairouz, H. B. McMahan, B. Avent, and et al., "Advances and open problems in federated learning," CoRR, vol. abs/1912.04977, 2019.

[11] L. Qu, Y. Zhou, P. Liang, Y. Xia, et al, "Rethinking architecture design for tackling data heterogeneity in federated learning," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun 2022, pp. 10 051–10 061.

[12] S. Zawad, A. Ali, P. Chen, A. Anwar, et al., "Curse or redemption? how data heterogeneity affects the robustness of federated learning," CoRR, vol. abs/2102.00655, 2021.

[13] M. Yurochkin, M. Agarwal, S. S. Ghosh, K. H. Greenewald, et al., "Bayesian nonparametric federated learning of neural networks," ArXiv, vol. abs/1905.12022, 2019.

[14] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," 2018.

[15] S. I. Popoola, G. Gui, B. Adebisi, M. Hammoudeh, and H. Gacanin, "Federated deep learning for collaborative intrusion detection in heterogeneous networks," in 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), 2021, pp. 1–6.

[16] M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, "Netflow datasets for machine learning-based network intrusion detection systems," in Big Data Technologies and Applications, Cham: Springer International Publishing, 2021, pp. 117–135.

[17] P. Qi, D. Chiaro, A. Guzzo, M. Ianni, G. Fortino, and F. Piccialli, "Model aggregation techniques in federated learning: A comprehensive survey," Future Generation Computer Systems, vol. 150, pp. 272–293, 2024.

[18] A. Kundu, P. Yu, L. Wynter, and S. H. Lim, "Robustness and personalization in federated learning: A unified approach via regularization," in 2022 IEEE Int. Conf. on Edge Computing and Communications (EDGE), 2022, pp. 1–11.

[19] K. Zhang, Y. Jiang, L. Seversky, C. Xu, D. Liu, and H. Song, "Federated variational learning for anomaly detection in multivariate time series," in 2021 IEEE Int. Performance, Computing, and Communications Conf. (IPCCC). Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 1–9

[20] I. Kholod, E. Yanaki, D. Fomichev, E. Shalugin, et al., "Open-source federated learning frameworks for iot: A comparative review and analysis," Sensors, vol. 21, no. 1, 2021.

[21] "Commercial vehicles sensor data set," accessed on 12.04.2024 (https://www.kaggle.com/datasets/smartilizer/commercial-vehicles-sensor-data-set)

[22] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, "D¨iot: A federated self-learning anomaly detection system for iot," 2019 IEEE 39th Int. Conf. on Distributed Computing Systems (ICDCS), pp. 756–767, 2018.

[23] M. Haller, C. Lenz, R. Nachtigall, F. M. Awayshehl, and S. Alawadi, "Handling non-iid data in federated learning: An experimental evaluation towards unified metrics," in 2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Int Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2023, pp. 0762–0770.