Анализ алгоритмов федеративной кластеризации

Д. А. Забалуев, М. А. Колпащиков, Е. С. Новикова

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

dazabaluev@stud.eltech.ru, makolpashikov@gmail.com, esnovikova@etu.ru

***Аннотация*. Федеративное обучение активно исследуется в контексте задач обучения с учителем; однако проблема его адаптации для задач обучения без учителя, в частности, кластеризацию, остается недостаточно исследованным. Применение классических алгоритмов кластеризации в федеративной парадигме является нетривиальной задачей, поскольку они часто требуют обмена конфиденциальной информацией, представленной описанием кластеров. В этой статье представлен систематический обзор существующих алгоритмов федеративной кластеризации с учетом реализованных алгоритмов кластеризации, стратегией формирования глобальных кластеров, и методологии оценки эффективности предложенных решений. Обзор выявляет несколько нерешенных задач, таких как воспроизводимость экспериментов и необходимость разработки надежных метрик качества, адаптированных к федеративному обучению и не предполагающих централизованного доступа к данным.**

Ключевые слова: федеративное обучение, федеративная кластеризация, обучение без учителя, машинное обучение с сохранением конфиденциальности, кластеризация на основе графов, кластеризация на основе разбиения, плотностная кластеризация, иерархическая кластеризация

I. Введение

Федеративное обучение (ФО) — это распределённая парадигма машинного обучения, в которой несколько клиентов совместно обучают общую модель, не передавая исходные данные в централизованное хранилище. Вместо этого локальные обновления модели вычисляются на частных локальных наборах данных и затем агрегируются для построения глобальной модели. Такой подход стал перспективным решением для data-driven приложений, в которых требования к конфиденциальности и приватности данных, а также институциональные ограничения, препятствуют прямому обмену данными. Эти свойства делают FL одним из перспективных направлений исследований в сфере приватно-ориентированного машинного обучения и распределённого анализа данных.

На данный момент основное внимание в исследованиях ФО сосредоточено на задачах обучения с учителем, в частности на построении прогнозных и классификационных моделей. Значительная часть научных публикаций посвящена разработке федеративных процедур оптимизации, коммуникационно-эффективных протоколов обучения, стратегий персонализации, а также механизмов, обеспечивающих сохранение приватности, для задач классификации и регрессии. Например, в [1] представлена систематизация типов неоднородности данных, и анализируются различные методы агрегации, предназначенные для преодоления трудностей, связанных с обработкой неоднородных (non-iid) данных в задачах классификации. В [2] подробно исследуется проблема выбора клиентов, а в работе [3] авторы сосредотачиваются на процессе оценки эффективности ФО, изучают возможные метрики и подчёркивают необходимость создания стандартизированной и комплексной оценочной рамки для федеративных алгоритмов.

Таким образом, эти исследования показывают, что федеративное обучение с учителем превратилось в быстрорастущую область с чётко сформулированными открытыми проблемами.

Параллельно с этим наблюдается рост интереса к распространению ФО на другие задачи машинного обучения, выходящие за рамки предсказания и классификации. В частности, растёт интерес к применению FL в условиях обучения без учителя и для исследовательского анализа, включая кластеризацию, обнаружение аномалий, обучение представлений и анализу процессов.

Кластеризация — это фундаментальная задача обучения без учителя, целью которой является разбиение данных на группы похожих объектов без доступа к «истинным» меткам. Её применение в распределённой среде также может существенно поддерживать исследовательский анализ данных, позволяя выявлять схожие сущности и закономерности. Однако её адаптация к федеративному сценарию не является тривиальной. Хотя ряд недавних работ предложили федеративные версии классических алгоритмов кластеризации и представление-ориентированных подходов, эта область по-прежнему испытывает недостаток всестороннего обзора существующих методов и их ограничений.

Таким образом, данная работа направлена на систематизацию существующих подходов и алгоритмов федеративной кластеризации, а также на выявление ключевых проблем, открытых вопросов и ограничений, связанных с этим направлением исследований. Обобщая текущий массив знаний и выделяя нерешённые методологические вопросы, исследование стремится заложить структурированную основу для будущих работ по обучению без учителя в федеративных средах.

II. Методология поиска и анализа релевантных работ

Поиск и анализ научных статей, посвящённых федеративной кластеризации, выполнялся в соответствии с рекомендациями по систематическому анализу и обзору научной литературы [4]. Этот процесс

Работа выполнена при поддержке гранта Российского научного фонда №25-11-20020, https://rscf.ru/project/25-11-20020/ и Санкт-Петербургского научного фонда.

включает: 1) формулировку исследовательских вопросов, 2) разработку стратегий поиска и отбора научных публикаций и 3) определение критериев включения и исключения работ в обзор. Были сформулированы следующие исследовательские вопросы:

Какие типы алгоритмов кластеризации реализованы в федеративном режиме?

Какие стратегии агрегации используются для построения глобальных кластеров?

Какие сценарии применяются для оценки эффективности предложенных федеративных алгоритмов кластеризации?

В поиске учитывались как англоязычные библиографические системы, так и русскоязычная научная электронная библиотека eLibrary. В качестве ключевых слов были определены следующие термины: федеративная кластеризация, федеративный dbscan, федеративный k-means, федеративные графовые алгоритмы.

Были определены следующие критерии включения (КВ) и исключения (КИ):

КВ1. Статья явно описывает подход к федеративной кластеризации, представляет экспериментальный сценарий, используемые наборы данных и псевдокод алгоритма.

КВ2. Публикация имеет чёткую структуру. Методы изложены ясно и наглядно. Статья написана на английском/русском языке в соответствии с принятыми стилистическими и грамматическими нормами.

КИ1. Статья представляет собой обзор работ или сравнение методов.

КИ2. Предложенный подход описан недостаточно подробно, а публикация не имеет чёткой структуры и/или написана на ненаучном языке.

КИ3. Статья предлагает алгоритм кластеризации для предварительной обработки non-iid данных с целью использования в федеративном обучении.

В ходе исследования удалось отобрать в общей сложности 12 статей, соответствующих определённым критериям включения. Следует отметить, что на момент проведения поиска публикаций на русском языке по данной тематике выявлено не было. Это позволяет предположить, что рассматриваемая тема пока недостаточно исследована российскими учёными.

III. Анализ федеративных алгоритмов кластеризации

Анализ отобранных статей показал, что федеративно адаптированные алгоритмы кластеризации можно разбить на четыре основные группы:

- алгоритмы на основе разбиения (partition-based);
- плотностные алгоритмы (density-based);
- иерархические алгоритмы, а также
- граф-ориентированные алгоритмы.

Рис. 1 иллюстрирует распределение публикаций за последние 6 лет по типам предложенных федеративных алгоритмов кластеризации. Ниже рассматриваются особенности стратегий агрегации, предложенных для каждого из этих типов.

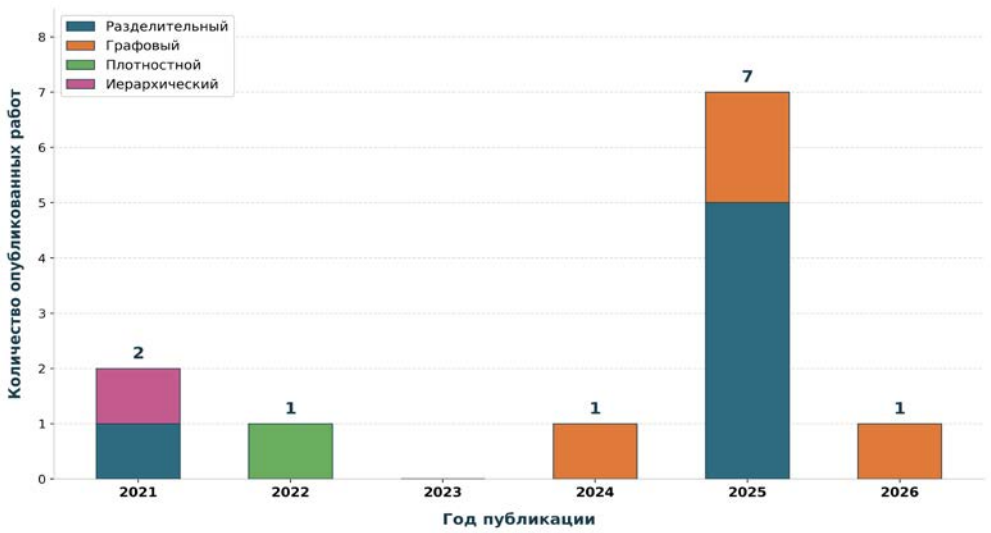

Рис. 1. Распределение работ по годам, выполнена группировка по типу алгоритма кластеризации.

Алгоритмы на основе разбиения. Эти алгоритмы являются наиболее часто реализуемыми в федеративном режиме. Мы полагаем, что это связано с тем, что методы разбиения являются самыми популярными и простыми по своей природе алгоритмами кластеризации. Алгоритмы FKM [5] и FKmeansCB [6] весьма схожи по своей архитектуре. Сначала локальные кластеры вычисляются на каждом клиенте, затем центроиды отправляются на глобальный сервер, который рассматривает центроиды как точки данных и кластеризует их. Такой процесс повторяется итеративно. Алгоритм FKM учитывает количество точек данных в локальных кластерах и использует их в качестве весов при определении глобальных кластерных центроидов. Для сохранения приватности передаваемых центров кластеров в [6, 7] предлагается добавлять к центроидам шум Лапласа. В [6] для решения проблемы избыточной коммуникационной нагрузки авторы предлагают процедуру однократной (one-shot) агрегации, основанную на оценке плотности локальных кластеров.

Плотность центроида определяется числом ближайших соседей и структурной схожестью их окрестности; центроиды с наибольшей локальной плотностью выбираются в качестве финальных кластерных центроидов.

Алгоритм Fed-MVKM [8] представляет федеративную версию метода multi-view кластеризации, который ищет кластеры в нескольких различных представлениях одних и тех же данных. Аналогично FKM и FKmeansCB, каждый клиент вычисляет локальные multi-view центроиды, которые получаются усреднением по каждому представлению.

В работе [10] задача федеративной кластеризации рассматривается как распределённая задача оптимизации на основе двойственной декомпозиции, и именно двойственность используется для определения нижних оценок целевой функции глобальной задачи кластеризации, тем самым направляя кластеризацию между клиентами.

Плотностные алгоритмы. Ключевая идея федеративного DBSCAN-алгоритма, предложенного в [11], заключается в отображении исходного пространства данных в двумерную сетку (data grid). Клиенты локально определяют ячейки, содержащие точки данных, и передают информацию о таких ячейках на сервер. Сервер выделяет ячейки с высокой плотностью и при необходимости расширяет области вокруг этих плотных ячеек. Затем он возвращает клиентам информацию о принадлежности каждой плотной ячейки кластеру. Клиенты завершают процедуру кластеризации, относя точки к ближайшей плотной ячейке и выявляя выбросы. Авторы также показывают, что предложенная идея может быть распространена и на кластеризацию вертикально разбиенных данных.

Алгоритмы на основе графов. Рост интереса к федеративной граф-ориентированной кластеризации связан с эффективностью графовых нейросетей (GNN) в задачах обучения без учителя. GNN эффективно кодируют структурную информацию о признаках и, следовательно, могут существенно повысить качество кластеризации.

В [12] авторы предлагают преобразовывать локальные данные в графы, а затем выполнять кластеризацию по граф-ориентированным признакам вместо передачи сырых данных или эмбеддингов. Поскольку в ходе коммуникации обмениваются только графы схожести и их признаки, приватность данных клиентов сохраняется. Клиенты на основе таких структурных графов вычисляют прототипы кластеров и отправляют на сервер как прототипы, так и графы. Сервер, в свою очередь, агрегирует эту информацию в одну глобальную структурную матрицу **E**, которая инкапсулирует как локальную структурную информацию, так и схожесть между клиентами. Применяя к глобальной матрице **E** лагранжевую функцию и теорему Кая Фана, извлекается точная информация о кластерах.

Реализация федеративной multi-view кластеризации, опирающейся на графическое кодирование, представлена в [13]. Авторы рассматривают проблему, когда признаки в разных представлениях различаются. На практике эта проблема соответствует отсутствию данных (missing data). В предложенном алгоритме клиенты сначала извлекают признаки с помощью GNN, а именно графового автоэнкодера, а затем выполняют мягкие (soft) локальные присвоения кластеров на основе локальной матрицы принадлежности кластерам. Авторы также вводят глобальные псевдометки, которые используются для повышения эффективности локальной извлечения признаков и сохранения глобальной согласованности между клиентами. После получения признаков Z_i и кластерных центров U_i сервер выполняет неоднородную взвешенную агрегацию совпадающих признаков в глобальные векторные представления Z и обновляет псевдометки, после чего возвращает их клиентам. В [14] вместо стандартной взвешенной функции агрегации авторы используют схему оптимизации консенсуса прототипов, основанную на модифицированной гауссовской оценке, адаптированной к нестрогим гауссовским распределениям и выбросам.

Чжоу и др. [15] исследуют эволюцию графовых структур во времени с целью выявления скрытых структур в динамических графовых данных.

Иерархическая кластеризация. Федеративный алгоритм кластеризации, представленный в [16], строит иерархические кластерные деревья (HCT) на графах в федеративной настройке с локальной дифференциальной приватностью. Ключевая трудность в такой постановке – сохранение приватности локального списка смежности пользователей. Для преодоления этого препятствия авторы вводят векторы степени как грубые структурные дескрипторы, в которых каждый элемент указывает количество соседей в случайно выбранном наборе пользователей, искажённое шумом Лапласа. Основное предположение подхода состоит в том, что, если две вершины принадлежат «близким» сообществам, их паттерны окрестностей будут похожи по разным ячейкам. Федеративная иерархическая кластеризация выполняется в два этапа: во-первых, приватный алгоритм PrivaCT строит матрицу несходства для локально зашумленных векторов степени, во-вторых, сервер запускает неприватный иерархический кластерный алгоритм GenTree, оптимизирующий целевую функцию, определённую поверх матрицы несходства.

Построение сценариев тестирования. Для оценки качества кластеризации исследователи обычно используют аннотированные наборы данных. Наличие аннотаций позволяет применять метрики, основанные на сравнении истинных меток (ground truth) и прогнозов алгоритма. Одними из наиболее распространённых метрик являются нормированная взаимная информация (NMI) и скорректированный индекс Ранда (ARI).

При использовании синтетических наборов данных типичная процедура генерации заключается в выборе заданного числа центров кластеров, равноудалённых друг от друга, а все остальные точки данных сэмплируются вокруг каждого центра с помощью нормального распределения (с дисперсией 1). Использование реальных данных позволяет получать кластеры более сложной формы в многомерном пространстве. Проблема моделирования федеративной среды раскрывается недостаточно полно, и не всегда ясно, как именно в экспериментах выполняется разбиение данных. Анализ также показал, что не все исследовательские публикации предоставляют исходный

код, позволяющий воспроизвести эксперименты на различных наборах данных.

IV. Заключение

Кластеризация в федеративной среде может существенно помочь в создании распределённых аннотированных наборов данных. Отсутствие меток для данных клиентов делает федеративное обучение с учителем неэффективным. Однако проведённое исследование показало, что в литературе по-прежнему мало работ, посвящённых федеративной кластеризации. Классические алгоритмы кластеризации, такие как разбиение-ориентированные или плотностные, нельзя напрямую реализовать в федеративной настройке, поскольку они требуют передачи чувствительной информации, представленной центроидами кластеров. Наиболее распространённым решением для сохранения приватности данных является механизм дифференциальной приватности, который добавляет случайный шум к чувствительной информации.

Перспективным направлением исследований является представление-ориентированная кластеризация, например кластеризация на графах. Такие подходы сохраняют приватность данных, преобразуя исходное пространство данных в пространство эмбеддингов. Однако основным препятствием здесь является создание единого пространства, общего для всех клиентов.

Ещё одной проблемой является отсутствие унифицированного сценария оценки. Оценочная процедура должна включать шаги, связанные с настройкой неоднородной (non-iid) федеративной среды. Более того, оценка качества кластеризации в условиях федеративного обучения остаётся трудной задачей, поскольку традиционные внутренние и внешние метрики валидации, как правило, предполагают доступ к централизованным данным или аннотациям.

Список литературы

[1] B. S. Guendouzi, S. Ouchani, H. EL Assaad, and M. EL Zaher, “A systematic review of federated learning: Challenges, aggregation methods, and development tools,” Journal of Network and Computer Applications, vol. 220, p. 103714, 2023. [Online]. Available https://www.sciencedirect.com/science/article/pii/S1084804523001339

[2] J. Li, T. Chen, and S. Teng, “A comprehensive survey on client selection strategies in federated learning,” Computer Networks, vol. 251, p. 110663, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138912862400495X

[3] D. Chai, L. Wang, L. Yang, J. Zhang, K. Chen, and Q. Yang, “A survey for federated learning evaluations: Goals and measures,” 2024. [Online]. Available: https://arxiv.org/abs/2308.11841

[4] B. A. Kitchenham, “Procedures for performing systematic reviews,” 2004.

[5] S. Garst and M. Reinders, “Federated k-means clustering,” in International Conference on Pattern Recognition. Springer, 2024, pp. 107–122.

[6] Z. Deng, Y. Wang, and M. M. Alobaedy, “Federated k-means based on clusters backbone,” Plos one, vol. 20, no. 6, p. e0326145, 2025.

[7] Y. Wang, W. Pang, D. Wang, and W. Pedrycz, “One-shot secure federated k-means clustering based on density cores,” IEEE Transactions on Neural Networks and Learning Systems, 2025.

[8] M.-S. Yang and K. P. Sinaga, “Federated multi-view k-means clustering,” IEEE transactions on pattern analysis and machine intelligence, vol. 47, no. 4, pp. 2446–2459, 2024.

[9] B. Xie, X. Dong, and C. Wang, “An improved k-means clustering intrusion detection algorithm for wireless networks based on federated learning,” Wireless Communications and Mobile Computing, vol. 2021, no. 1, p. 9322368, 2021.

[10] V. Yfantis, A. Wagner, and M. Ruskowski, “Federated k-means clustering via dual decomposition-based distributed optimization,” Franklin Open, vol. 10, p. 100204, 2025.

[11] G. Marino, “Federated dbscan,” Bachelor’s thesis, Dipartimento di Ingegneria dell’Informazione, Università di Pisa, 2022.

[12] G. He, Z. Wang, J. Wang, L. Tang, R. Wang, and F. Nie, “Towards federated clustering: A client-wise private graph aggregation framework,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, no. 26, 2026, pp. 21 619–21 627.

[13] X. Yan, Z. Wang, and Y. Jin, “Federated incomplete multi-view clustering with heterogeneous graph neural networks,” in International Workshop on Trustworthy Federated Learning. Springer, 2024, pp. 61–76.

[14] J. Liu, J. Cheng, R. Han, W. Tu, J. Wang, and X. Peng, “Federated graph-level clustering network,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 18, 2025, pp. 18 870–18 878.

[15] Z. Zhou, Y. Liu, X. Xu, and Q. Li, “Federated temporal graph clustering,” arXiv preprint arXiv:2410.12343, 2024.

[16] A. Kolluri, T. Baluta, and P. Saxena, “Private hierarchical clustering in federated networks,” in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 2342–2360.