**Gradient Matching for Categorical Data Distillation in CTR Prediction**

Cheng Wang, Jiacheng Sun, Zhenhua Dong,

Ruixuan Li, Rui Zhang

Дата публикации: сентябрь 2023

RecSys '23: 17th ACM Conference on Recommender Systems

<https://dl.acm.org/doi/10.1145/3604915.3608769>

[Ссылка на статью](https://www.ruizhang.info/publications/RecSys23_Gradient%20Matching%20for%20Categorical%20Data%20Distillation%20in%20CTR%20Prediction.pdf) (если вдруг удалили, то [вот в моём гугл-диске](https://drive.google.com/file/d/1orEWUsz5KnbWjZn2ngqRO2KrxZ2Dwkq2/view?usp=sharing))

[Репозиторий университета](https://github.com/huawei-noah), который представляют многие авторы статьи (но кода по именно этой статье там пока что нет)

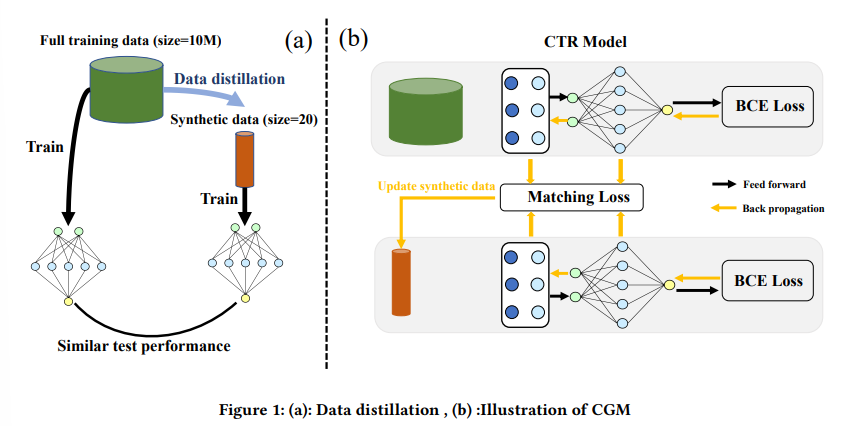
# **Краткое содержание**

Авторы статьи предлагают новый метод обработки данных для предсказания метрики click-through rate (CTR) в рекомендательных системах, названный ими категориальная дистилляция данных с применением градиентного мэтчинга (Categorical data distillation with Gradient Matching, CGM). В рамках данного подхода большой датасет разреженных данных (разреженных ввиду большого содержания категориальных фичей в нём) преобразовывается со значительным сокращением размерности в виде синтетических фичей. Такой подход в целом напоминает понижение размерности через метод главных компонент (PCA). Под градиентным матчингом авторы подразумевают обучение такой предсказательной модели, которая на исходных и синтетических данных давала бы сопоставимое мат ожидание функции потерь. Показанный метод, по словам авторов, превосходит прочие методы обработки данных для задачи предсказания CTR за счёт лучшего сохранения информативности в полученном синтетическом датасете.

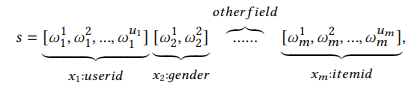
# **Ключевые идеи и гипотезы**

1. Улучшение существующих методов сокращения размерности разреженных данных за счёт снижения затрат на обучение модели, генерирующей синтетический датасет, с минимальным понижением качества данных
2. Решение проблем дистилляции данных для рекомендательных систем
   1. Большая размерность и множество категориальных данных -> преобразование в непрерывные переменные
   2. Отказ от предположения об одинаковом распределении данных в рамках одной метки целевой переменной (что совершенно точно не так для RecSys) -> рандомное сплитование данных
   3. Отказ от [bi-level optimization](https://en.wikipedia.org/wiki/Bilevel_optimization) -> проведение градиентного мэтчинга в один шаг

# **Краткое описание модели**



1. Преобразование исходных категориальных признаков в сконкатенированные по признакам тензоры той же размерности, что и one-hot encoded исходные признаки.



При этом w сэмплируется из равномерного распределения так, чтобы сумма по исходному признаку сходилась к 1.

1. Обозначение эмбеддинга размерности u\_i x m (количество объектов, по которым берётся пространство фичей х количество категориальных фичей в пространстве), представляющего собой смаппленные векторы эмбеддингов на векторы сгенерированных w по каждому категориальному признаку
2. Сокращение размерности эмбеддинга до 1 x m за счёт их агрегации
3. Обучение преобразованного признакового пространства таким же образом, как и исходного. Но при этом в рамках преобразованного признакового пространства обучаем и эмбеддинги, и сэмплы категориальных данных.
4. Минимизация разницы лоссов (относительно целевой метки решаемой задачи) между исходным и синтетическим признаковым пространством без двойного цикла прохода, когда в исходной постановке сначала во внутреннем цикле оптимзируются параметры эмбеддинга, а затем во внешнем цикле оптимизируются значения синтетических данных. Вместо этого предлагается считать полный лосс только на первом проходе данных, понять направление изменения синтетических данных по полученным параметрам модели и больше не менять их, сосредоточившись лишь на оптимизации значений синтетических данных.

# **Результаты экспериментов**

Авторы провели эксперименты на 3 DeepCTR моделях (DCN, DeepFM, Wide&Deep) для внедрения в них Categorical data distillation with Gradient Matching (здесь и далее, CGM), оценив их на 3 наборах данных:

1. Avazu
2. Criteo
3. Movielens

Сравнение производилось на основе метрик AUC и Logloss.

В качестве бейзлайна подготовки обучающей выборки были взяты следующие методы:

1. Random (случайные столбцы из исходного датасета)
2. K-center (сэмплирование центральных столбцов из датасета)
3. SVP-CF (сэмплирование для коллаборативной фильтрации)
4. Distill-CF (похожий подход без сэмплирования исходных категориальных фичей с bilevel градиентным мэтчингом)

На всех предложенных наборах данных обе метрики превзошли все бейзлайны, в результате чего авторы приходят к выводу, что их подход с новым методом подсчёта сэмплирования категориальных переменных, одношагового градиентного мэтчинга, стратегии разделения данных на батчи превосходят существующие бейзлайны, то есть лучше сокращает размерность разреженных данных для рекомендательных систем. К тому же, метод может быть обобщён на разные модели для обучения (поскольку был прикручен к трём разным DL-моделям для рекомендательных систем).

В целом, последний вывод авторов вряд ли стоит назвать выдающимся, поскольку все методы преобразования обучающей выборки среди бейзлайнов также обобщаются на почти любые модели (кроме Distill-CF). Более того, метрики на датасетах Avazu & Criteo у лучших моделей CGM не сильно отличаются от бейзлайнового Random+DCN, который, вероятно, мог бы выступить лучше при выборе другой модели глубинного обучения для предсказания CTR. Однако это не отменяет того факта, что предложенный метод сэмплирования действительно достойно проявил себя на рассмотренных наборах данных

# **Критический анализ статьи**

1. Ваш общий критический анализ статьи: сильные и слабые стороны прочитанной статьи

Сильными сторонами стоит назвать то, что авторы разработали и внедрили сразу несколько улучшений существующих методов снижения размерности разреженных данных для предсказания CTR, подробно проанализировав математическую постановку существующих подходов и приведя результаты экспериментов.

Однако у данной статьи выделяется сразу несколько слабых сторон:

1. Следовало бы итеративно показывать вклад их нововведений в существующие методы генерации синтетических данных, чтобы понять, что именно играет существенную роль (какое-то конкретное логическое изменение или их совокупность)
2. Слабая аргументация отказа от bi-level optimization (на первой эпохе наибольший лосс у параметров модели, поэтому можем его зафиксировать и заняться только тюнингом синтетических значений данных)
3. Отсутствие таблиц сравнения по времени генерации датасета, хотя сокращение времени обучения приводится как ключевая фича нового подхода
4. Отсутствие Github с наглядным представлением проделанных экспериментов, что отсылает нас к статье [Are we really making much progress?](https://arxiv.org/pdf/1907.06902.pdf) и заставляет задуматься, не отбросили ли авторы более удачные бейзлайны, чтобы приукрасить картину значимости своего подхода

# **Оценка вклада статьи**

Кажется, что статья вносит довольно значительный вклад в подход к разработке признакового пространства, но на данный момент ограничивается лишь предсказанием CTR в относительно узкой сфере рекомендательных систем (рекламные платформы).

Научная новизна статьи обуславливается инновацией в подходе градиентного матчинга, которая должна ускорить процесс обучения синтетического набора данных.

# **Дальнейшие исследования**

Авторы заинтересованы в исследовании проблемы сохранения приватности данных на основе полученных синтетических данных. Также авторы хотят рассмотреть предложенный метод и данную проблему в контексте графовых данных, анализе речи и сентимента.

На мой взгляд, такие дальнейшие шаги исследования являются несколько скудными и странными, поскольку вообще не вытекают из логики исследования, да и в целом уже существует [ряд более ранних подходов к Data Distillation](https://github.com/Guang000/Awesome-Dataset-Distillation), в том числе и от университета, опубликовавшего данное исследование.

# **Шаги по продолжению исследования**

Заявленный метод было бы важно воспроизвести на других типах данных и типах задач в контексте сохранения качества полученного синтетического датасета. В целом, было бы очень важно также проверить другие бейзлайны и представить корректное сравнение, которое не вызывало бы никакого скепсиса и вопросов по честности экспериментов.

Также было бы любопытно представить динамику изменения качество предсказания CTR в зависимости от количества эпох, на котором обучаются параметры модели генерации синтетических данных в градиентном мэтчинге.