Оценка временной сложности для задачи поиска идентичных товаров для электронной торговой площадки на основании композиции моделей машинного обучения

Ф.В. Краснов

***Аннотация*—В данном исследовании рассматриваются решение задачи поиска идентичных товаров для электронной торговой площадки. Данная задача является одной из составляющих рекомендательной системы электронной торговой площадки и относится к классу рекомендательных систем item2item — товар-товарные рекомендации. В рамках этой задачи необходимо найти все идентичные товары с различной ценой и временем доставки от разных поставщиков из каталога товаров, который может содержать сотни миллионов товаров. Математическая постановка для данной задачи относится к классу NP-полных задач. Для сокращения сложности автор применил композицию моделей машинного обучения для решения этой задачи – более быстрые и универсальные модели производят отбор пар-кандидатов идентичных товаров, а затем более ресурсоёмкие и точные модели производят скоринг идентичности пар-кандидатов товаров. Завершающая модель определяет порядок в группе идентичных товаров на основании модели ранжирования. В результате получается список групп идентичных товаров отсортированных по степени идентичности внутри группы. Метрика временной сложности для композиции моделей составила O(N*N*LOG(P)/M), где N — общее количество товаров в каталоге, P — количество модальностей товара, а M — количество категорий товаров.**

***Ключевые слова*— RecSys, Approximate Nearest Neighbors, Price Matcher, Time complexity.**

I. Введение

Идентичные товары по различным оценкам составляют в отдельных категориях каталога товаров электронных площадок до 40 %. Бизнес значение от владения информацией об идентичности товаров позволяет получать выгоду для электронной площадки, покупателя и продавца. Площадка получает выгоду от интенсификации товарооборота в товарных группах за счет объективизации процесса ценовой конкуренции. Покупатель получает возможность выбора оптимального товара по цене, видит равнозначные замены при отсутствии на складе выбранного ранее в корзине товара, экономит время при поиске за счет группировки информации об идентичных товарах в поисковой выдаче. Продавец получает доступ к информации о конкурентных товарах и возможность более интеллектуального и сфокусированного ценообразования.

Задача поиска идентичных товаров относится к классу «задач перебора». Постановка задачи нахождения идентичных товаров сводится к перебору и сравнению свойств товаров «каждый с каждым». Так как изначально любой товар из каталога электронной площадки может быть идентичен по своим свойствам любому товару из каталога.

II. Методика

Как и все задачи связанные с поиском информации мы можем оценивать качество результата с помощью метрик *полноты (recall) и точности (precision)* (или их производных). Переход от полного перебора к выборочному сравнению товаров значительно ускорит решение, но будет ухудшать метрики качества. *Точность* решения задачи поиска идентичных товаров зависит только от качества сравнения свойств одного товара с другим. В то время как, *полнота* может значительно упасть при переходе к выборочному сравнению товаров. С другой стороны сравнение товаров может быть ускоренно за счет неполного перебора свойств при сравнении, что уменьшает *точность*. Ограничение по времени решения задачи поиска идентичных товаров не является абстрактным в случае электронной площадки. Каталог товаров ежедневно обновляется примерно на 1 %. Информация об идентичности товаров устаревает и перестает наносить пользу всем участникам процесса электронной коммерции.

Таким образом, речь идет об оптимальном решении задачи задачи поиска идентичных товаров для электронной торговой площадки с точки зрения времени решения и метрик качества. Для оценки метрик качества обычно используют размеченный набор данных. С помощью асессоров для каждой выбранной пары товаров кандидатов на идентичность проставляется бинарная метка класса «товары идентичны» (Да/Нет). Отдельная разметка данных

Статья получена 29 ноября 2022.
Ф.В.Краснов, Исследовательский центр ООО "ВБ СК" на базе Инновационного Центра Сколково e-mail: krasnov.fedor2@wb.ru

производится для определения порядка идентичных товаров группе.

III.Обзор литературы

Подбор товаров является центральной задачей в приложениях электронной коммерции [1], таких как порталы сравнения цен и электронные торговые площадки. Современные методы подбора товаров позволяют получить значения F1-score выше 0.9, используя методы глубокого обучения в сочетании с огромными объемами обучающих данных . Что достаточно дорого и трудоемко. Для поиска идентичных товаров используют различные модальности — Название, Описание, Изображения. Авторы исследований [2] и [3] предложили использовать поисковые системы в Интернете с целью обогащения названий товаров несколькими важными пропущенными словами и изображениями для поиска идентичных товаров в получившихся кластерах. Тем не менее, в случае больших наборов данных их подход довольно неэффективен или даже невозможен, поскольку выполнение тысяч запросов к коммерческой поисковой системе является i) непомерно дорогим и ii) запрещенным условиями использования этих систем. С другой стороны, в работе [4] предложено использовать морфологический анализ названий товаров в качестве дополнительной информации для поиска идентичных товаров. А в работе [5] исследованы методы поиска идентичных товаров в семантическом многомерном представлении товаров.

Для повышения вычислительной эффективности во многих исследованиях [6,7] используют приближенные методы поиска в векторном пространстве товаров. Алгоритмы поиска «ближайших соседей» не смотря на свою простоту становятся все более важными, особенно в области электронной коммерции и голосовых помощников с десятками публикаций научных статей за последние годы. Во многом этот энтузиазм обусловлен успешным новыми приближенными подходами на основе метрических деревьев и в частности на локально-чувствительном хешировании (LSH). В этой статье мы рассматриваем специальный случай большого количества возможных соседей, при котором использование приближенных методов поиска ближайших соседей драматически влияет на метрику *полнота*.

IV.Эксперимент

Ансамбли моделей хорошо зарекомендовали себя как относительно простой и результативный подход к решению задач машинного обучения с помощью усреднения результата работы нескольких моделей [8,9,10,11,12]. Композиция моделей дает более широкие возможности выстраивания моделей в виде сети (графа), поиск гиперпараметров одних моделей с помощью других моделей, сбалансированное обучение по метрикам и другое. Но это не глубокое машинное обучение, так как не обязательно присутствует и не всегда возможно сквозное обучение. В качестве примера и подтверждения актуальности подхода на основе композиций моделей приведем ссылку на соревнование на платформе Kaggle[1] 2022 года для поиска идентичных товаров на электронной площадке.

Так как мы рассматриваем задачу поиска, то у нее есть онлайн и оффлайн составляющие. К онлайн составляющей относится компоненты портала электронной площадки и веб-сервис, который обрабатывает запрос по id товара и предоставляет группу, состоящую из id идентичных товаров. К оффлайн составляющей относятся процесс подготовки данных для веб-сервиса и процесс обучения модели. Для оффлайн подготовки данных для веб-сервиса автор разработал следующую композицию моделей,

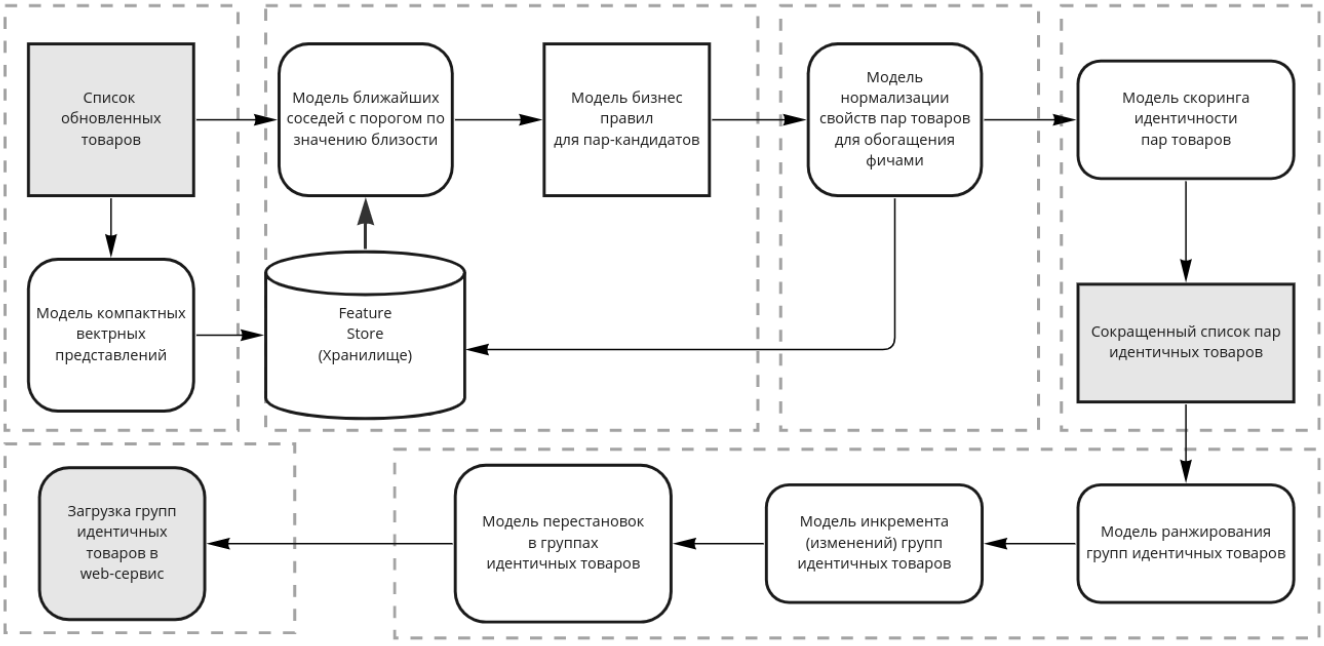

Рис. 1 Композиция моделей для поиска Идентичных товаров для оффлайн обработки новых товаров

[1] https://www.kaggle.com/c/shopee-product-matching

изображенную на рисунке 1.

Остановимся подробнее на наиболее важных из моделей приведенных на рисунке 1. Каталог товаров непрерывно изменяется, измененные карточки товаров поступают на вход Композиции моделей. Содержание карточки товара состоит из различных модальностей – это Название, Описание, Изображения, Свойства товара. Каждая из модальностей содержания трансформируется в компактное векторное представление с помощью дообученной языковой модели в случае текстовой природы модальности [13] и с помощью сверточной нейронной сети в случае изображений [14]. Содержание карточки необходимо проецировать на единое векторное пространство с возможностью определения похожих карточек товаров по содержанию, это обеспечивает Модель ближайших соседей с порогом близости. Для Модели ближайших соседей с порогом близости автор использовал метод соединенных пространств (connected vector spaces, CVecS), поясненный на рисунке 2.

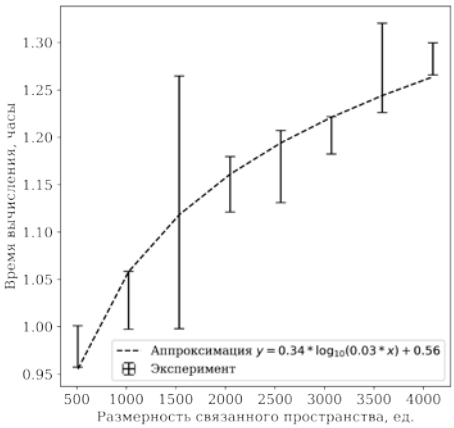

Рис. 3 Аппроксимация экспериментальных данных о временной сложности расчета.

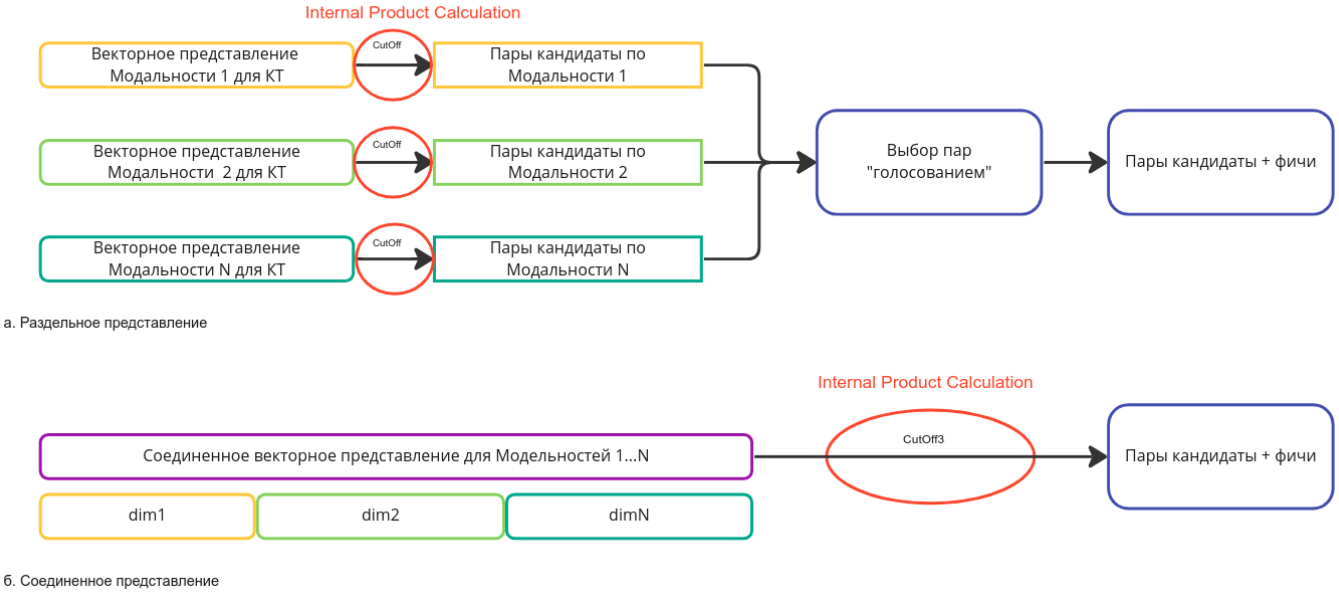

Рис. 2 Соединенное пространство CVecS. а. Раздельное представление каждой модальности карточки товара (КТ), б. Соединенное пространство модальностей карточек товаров.

Основная идея Метода соединенных пространств состоит в том, чтобы вместо раздельных векторов для каждой модальности использовать один «длинный» вектор. Полученный составлением из векторов для каждой модальности. Важным отличием является то, что получившийся «длинный» вектор не обладает собственной нормировкой, не «единичный». Это позволяет, с одной стороны, сохранить поведение в каждой модальности при операциях сравнения. А с другой, рассматривать «длинный» вектор как единое представление товара в векторном пространстве.

На рис. 3 приведен результат эксперимента по измерению Временной сложности алгоритма Метода соединенных пространств.

Вторым механизмом для уменьшения временной сложности является следующая за Моделью ближайших соседей с порогом близости Модель бизнес правил для пар-кандидатов идентичных товаров. К наиболее важным бизнес правилам относятся следующие:

- БП1. Оба товара должны быть в наличии на складе,
- БП2. Продавцы товаров должен отличаться,
- БП3. Оба товара должны принадлежать одной торговой категории каталога.

Покажем степень влияния бизнес правил на временную сложность для БП3. В каталоге N товаров и M категорий, так что M << N. Тогда N = $\sum_i^M M_i$, где M_i — это количество товаров в *i*-й категории. А

математическое ожидание $E[M_i] = N/M$. Отсюда делаем оценку для изменения временной сложности:

$$\sum_i^M M_i * M_i \sim {}^{N^2}/_{M} \quad (1)$$

то есть в M раз меньше. В случае с крупной электронной торговой площадкой при $N \sim 10^8$ и $M \sim 10^4$ выигрыш по временной сложности от применения БПЗ получается значительный.

Следующей в композиции является Модель нормализации свойств пар товаров для обогащения фичами. Всременная сложность данной модель оценивается как количество фич (n) умноженное на количество пар-кандидатов на идентичность товаров. Здесь следует пояснить, почему методы приближенного поиска *K*-ближайших соседей драматически влияют на *полноту,* получающихся пар-кандидатов. Так как, среди пар-кандидатов может быть любое неизвестное зараниее количество потенциально идентичных товаров, невозможно заранее выбрать *K* так чтобы быть уверенным, что часть идентичных товаров не осталась отброшенными.

Скоринг пар-кандидатов использует CatBoosting эстиматор [15], обученный на размеченных ассесорами парах идентичных товаров и имеет временную сложность:

$$O(N_{predict}) = O(s*n), \quad (2)$$

где s — количество примеров, а n — количество фич. Вклад в общую сложность Композиции моделей является незначитеьльным. Так же как и вклады от Модедей ранжирования и перестановок в группах.

V.Заключение

В настоящем исследовании представлены результаты применения новых подходов к задаче нахождения идентичных товаров на электронной площадке. Применение композиции моделей машинного обучения позволило автору построить инженерно-техническое решение в условиях ограниченных временных и вычислительных ресурсов не уменьшая метрики качества. Основной метрикой качества для данной задачи является степень покрытия рекомендациями идентичных товаров в товарных категориях — Fraction Cover, составляющая от 5% до 20%. В частности для таких категорий, как Смартфоны с помощью Композиции Моделей получено значение Fraction Cover более 40%, что соответствует индустриальным бенчмаркам. А в таких категориях товаров как Рукоделие и Авторские картины Cover Fraction составляет менее 1%.

Подход, предложенный и исследованный автором не уменьшает значение метрики Fraction Cover. Но подход на основе Композиции Моделей позволил автору сократить время вычисления за счет применения нового метода Соединенных представлений для модальностей карточек товаров и уменьшить вычислительную сложность с O(N*N*P) до O(N*N*LOG(P)/M) для модели формирования пар-кандидатов карточек товаров, где N — общее количество товаров в каталоге, P — количество модальностей товара, а M — количество категорий товаров.

Библиография

[1] Peeters, Ralph, et al. "Using schema. org annotations for training and maintaining product matchers." Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics. 2020.

[2] V. Gopalakrishnan, S. P. Iyengar, A. Madaan, R. Rastogi, and S. Sengamedu, “Matching Product Titles using Web-based Enrichment,” in Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 605–614.

[3] N. Londhe, V. Gopalakrishnan, A. Zhang, H. Q. Ngo, and R. Srihari, “Matching Titles with Cross Title Web-search Enrichment and Community Detection,” Proceedings of the VLDB Endowment, vol. 7, no. 12, pp. 1167–1178, 2014.

[4] Akritidis, Leonidas, and Panayiotis Bozanis. "Effective unsupervised matching of product titles with k-combinations and permutations." 2018 Innovations in Intelligent Systems and Applications (INISTA). IEEE, 2018.

[5] Bhutani P, Baranwal SK, Jain S (2021) Semantic framework for facilitating product discovery. In: 2021 Advances in Computational Intelligence, its Concepts & Applications (ACI 2021), vol. CEUR-WS, pp 30–36

[6] Arya, Sunil, et al. "An optimal algorithm for approximate nearest neighbor searching fixed dimensions." Journal of the ACM (JACM) 45.6 (1998): 891-923.

[7] Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with gpus." IEEE Transactions on Big Data 7.3 (2019): 535-547.

[8] Nan Wu, Yuan Xie, A Survey of Machine Learning for Computer Architecture and Systems, ACM Computing Surveys, 10.1145/3494523, 55, 3, (1-39), (2023).

[9] Nikolay O. Nikitin, Pavel Vychuzhanin, Mikhail Sarafanov, Iana S. Polonskaia, Ilia Revin, Irina V. Barabanova, Gleb Maximov, Anna V. Kalyuzhnaya, Alexander Boukhanovsky, Automated evolutionary approach for the design of composite machine learning pipelines, Future Generation Computer Systems, 10.1016/j.future.2021.08.022, 127, (109-125), (2022).

[10] S. B. Goyal, Pradeep Bedi, Anand Singh Rajawat, Rabindra Nath Shaw, Ankush Ghosh, Multi-objective Fuzzy-Swarm Optimizer for Data Partitioning, Advanced Computing and Intelligent Technologies, 10.1007/978-981-16-2164-2_25, (307-318), (2022).

[11] Ya-Lin Zhang, Qitao Shi, Meng Li, Xinxing Yang, Longfei Li, Jun Zhou, A Classification Based Ensemble Pruning Framework with Multi-metric Consideration, Intelligent Systems and Applications, 10.1007/978-3-030-82193-7_44, (650-667), (2022).

[12] Jianrong Yao, Zhongyi Wang, Lu Wang, Zhebin Zhang, Hui Jiang, Surong Yan, A hybrid model with novel feature selection method and enhanced voting method for credit scoring, Journal of Intelligent & Fuzzy Systems, 10.3233/JIFS-211828, 42, 3, (2565-2579), (2022).

[13] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[14] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.

[15] Anna Veronika Dorogush, Vasily Ershov, Andrey Gulin "CatBoost: gradient boosting with categorical features support". Workshop on ML Systems at NIPS 2017

Estimation of time complexity for the task of retrieval for identical products for an electronic trading platform based on the decomposition of machine learning models

F.V. Krasnov

Abstract—This paper scrutinize the solution of the problem of matching identical products for an electronic trading platform. This task is one of the components of the recommendation system of the electronic trading platform and belongs to the class of item2item recommendation systems— product-product recommendations. As part of this task, it is important to find all identical products with different prices and delivery conditions from different suppliers from a product catalog that contains hundreds of millions of products. The mathematical formulation for this problem belongs to the class of NP-complete problems. The author applied a decomposition of machine learning models to solve this problem – faster and more versatile models select candidate pairs of identical products, and then more CPU-intensive and accurate models score the identity of candidate pairs of products. The final model determines the order in a group of identical products based on the ranking model. The result is a list of groups of identical products sorted by the rank of identity within the group. The time complexity metric for the model composition was O(N*N*LOG(P)/M), where N is the total number of products in the catalog, P is the number of product modalities, and M is the number of product categories.

Keywords— RecSys, Approximate Nearest Neighbors, Price Matcher, Time complexity.

REFERENCES

[1] Peeters, Ralph, et al. "Using schema. org annotations for training and maintaining product matchers." Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics. 2020.

[2] V. Gopalakrishnan, S. P. Iyengar, A. Madaan, R. Rastogi, and S. Sengamedu, “Matching Product Titles using Web-based Enrichment,” in Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 605–614.

[3] N. Londhe, V. Gopalakrishnan, A. Zhang, H. Q. Ngo, and R. Srihari, “Matching Titles with Cross Title Web-search Enrichment and Community Detection,” Proceedings of the VLDB Endowment, vol. 7, no. 12, pp. 1167–1178, 2014.

[4] Akritidis, Leonidas, and Panayiotis Bozanis. "Effective unsupervised matching of product titles with k-combinations and permutations." 2018 Innovations in Intelligent Systems and Applications (INISTA). IEEE, 2018.

[5] Bhutani P, Baranwal SK, Jain S (2021) Semantic framework for facilitating product discovery. In: 2021 Advances in Computational Intelligence, its Concepts & Applications (ACI 2021), vol. CEUR-WS, pp 30–36

[6] Arya, Sunil, et al. "An optimal algorithm for approximate nearest neighbor searching fixed dimensions." Journal of the ACM (JACM) 45.6 (1998): 891-923.

[7] Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with gpus." IEEE Transactions on Big Data 7.3 (2019): 535-547.

[8] Nan Wu, Yuan Xie, A Survey of Machine Learning for Computer Architecture and Systems, ACM Computing Surveys, 10.1145/3494523, 55, 3, (1-39), (2023).

[9] Nikolay O. Nikitin, Pavel Vychuzhanin, Mikhail Sarafanov, Iana S. Polonskaia, Ilia Revin, Irina V. Barabanova, Gleb Maximov, Anna V. Kalyuzhnaya, Alexander Boukhanovsky, Automated evolutionary approach for the design of composite machine learning pipelines, Future Generation Computer Systems, 10.1016/j.future.2021.08.022, 127, (109-125), (2022).

[10] S. B. Goyal, Pradeep Bedi, Anand Singh Rajawat, Rabindra Nath Shaw, Ankush Ghosh, Multi-objective Fuzzy-Swarm Optimizer for Data Partitioning, Advanced Computing and Intelligent Technologies, 10.1007/978-981-16-2164-2_25, (307-318), (2022).

[11] Ya-Lin Zhang, Qitao Shi, Meng Li, Xinxing Yang, Longfei Li, Jun Zhou, A Classification Based Ensemble Pruning Framework with Multi-metric Consideration, Intelligent Systems and Applications, 10.1007/978-3-030-82193-7_44, (650-667), (2022).

[12] Jianrong Yao, Zhongyi Wang, Lu Wang, Zhebin Zhang, Hui Jiang, Surong Yan, A hybrid model with novel feature selection method and enhanced voting method for credit scoring, Journal of Intelligent & Fuzzy Systems, 10.3233/JIFS-211828, 42, 3, (2565-2579), (2022).

[13] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[14] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.

[15] Anna Veronika Dorogush, Vasily Ershov, Andrey Gulin "CatBoost: gradient boosting with categorical features support". Workshop on ML Systems at NIPS 2017