УДК 004.021:004.75

ИСПОЛЬЗОВАНИЕ МЕТОДА РАНЖИРОВАНИЯ ПРИ ОПРЕДЕЛЕНИИ СПИСКА ПОТЕНЦИАЛЬНЫХ ЗНАКОМЫХ ПОЛЬЗОВАТЕЛЯ

Е.А. Бугаев
Инженер-программист
ООО «Whitesnake»
buhaev.yauheni@gmail.com

Е.А.Бугаев

Окончил Белорусский государственный технологический университет. Область научных интересов связана с разработкой эффективных и быстрых алгоритмов для обработки больших объемов данных. Интересы включают в себя теоретические и практические аспекты оптимизации, машинного обучения, параллельных и распределенных вычислений, анализа данных и искусственного интеллекта.

Аннотация. Выполнен анализ подхода к планированию системы ранжирования при определении списка потенциальных знакомых. Описан основной алгоритм решения данной проблемы.

Разработана принципиальная схема работы сервиса ранжирования, показаны возможные варианты улучшения системы. Проведена оптимизация входных и выходных данных системы. Проведена оценка и выполнена оптимизация.

Ключевые слова: ранжирование, модель машинного обучения, граф, GNN, ROC-AUC.

Введение. В эпоху цифровых технологий, когда социальные связи играют решающую роль в личной и профессиональной жизни, поиск новых контактов и друзей становится актуальной задачей. Интернет-платформы и социальные сети предлагают множество возможностей для знакомства, однако определить наиболее подходящих людей может быть непросто.

Методы ранжирования социальных контактов предоставляют возможность упорядочивать и оценивать потенциальных знакомых с учетом различных критериев. Эти методы используют сложные алгоритмы для создания списка наилучших кандидатов, что значительно упрощает поиск и выбор новых социальных связей.

Методы машинного обучения. Поточечное ранжирование — один из методов машинного обучения, который используется для оптимизации алгоритмов рекомендаций. Этот метод основывается на бинарной классификации. Он анализирует каждый отдельный объект и определяет, относится ли он к той или иной категории на основе заданных критериев. Этот процесс помогает системе рекомендаций выделить наиболее подходящих потенциальных знакомых пользователю, увеличивая вероятность успешного совпадения интересов и предпочтений. Принципиальная схема работы классификатора представлена на рисунке 1.

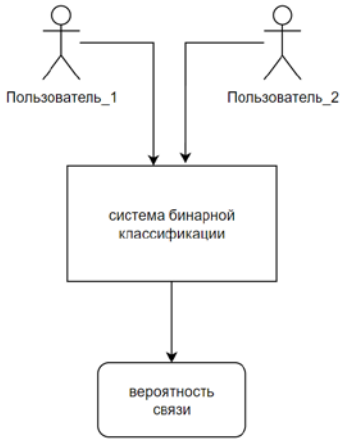

Рисунок 1. Бинарная классификация двух пользователей

Проблемы интеграции социального контекста в рамках бинарной классификации.

В связи с тем, что при определении потенциальных контактов, социальный контекст играет важную роль, его использование необходимо, что заставляет нас рассматривать при классификации не только непосредственно пользователей, но и их окружение. На рисунке 2 представлен простой пример социального контекста двух людей в двух коленах.

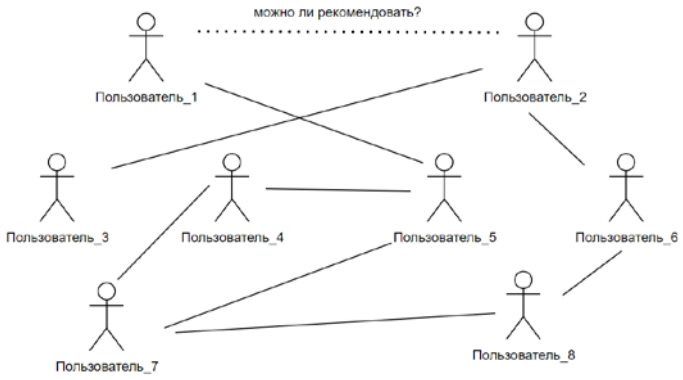

Рисунок 2. Пример социального контекста

Интеграция социального контекста в процессы бинарной классификации сталкивается с рядом проблем, среди которых можно выделить следующие:

1 Сложность измерения и оценки: социальный контекст включает в себя множество разнообразных и часто субъективных факторов, таких как интересы, предпочтения, стиль общения и отношения;

2 Многомерность данных: социальный контекст является многомерным и многообразным, что затрудняет его обработку и интеграцию в алгоритмы бинарной классификации;

3 Изменчивость контекста: социальный контекст не является статичным и может меняться со временем. Интересы, предпочтения и социальные связи пользователей могут изменяться, что требует постоянной адаптации и обновления алгоритмов классификации.

4 Конфиденциальность данных: использование социального контекста требует обработки большого объема личных данных пользователей. Это вызывает вопросы конфиденциальности и защиты данных, что требует внедрения строгих мер безопасности и соблюдения нормативных требований.

5 Интерпретируемость моделей: сложные модели машинного обучения, используемые для анализа социального контекста, могут быть трудными для интерпретации и объяснения. **Предсказание связей.** Для решения проблем, описанных выше, модель получает на вход граф, включающий в себя весь социальный контекст, что дает возможность использовать дополнительные данные. На рисунке 3 представлена схема передачи контекста, где пользователи – узлы, а их социальные контакты – ребра.

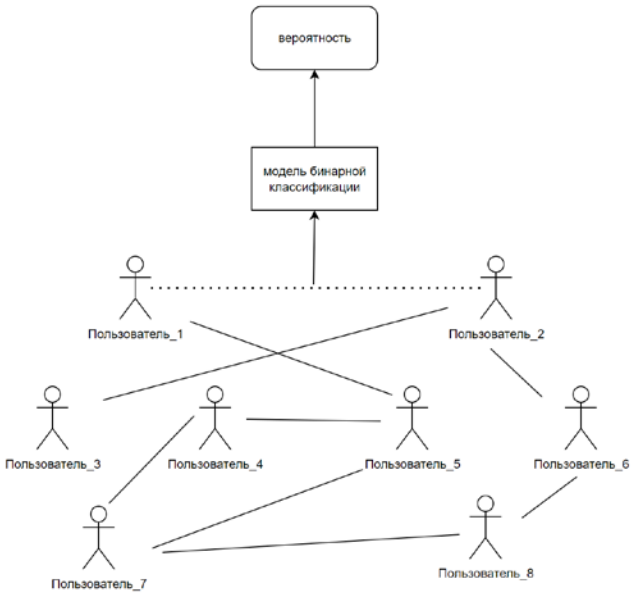

Рисунок 3. Передача социального вектора модели

Любая система машинного обучения состоит из подсистем, выполняющих конкретные функции. Можно выделить такие система, как:

– система данных;
– модель (система обучения);
– система оценки работы модели.

Система данных и конструирование признаков. Данные играют основную роль при разработке и обучению моделей машинного обучения. При работе с системой ранжирования мы работаем, в базовом виде, с тремя абстрактными типами данных, а именно:

– данные о пользователе;
– связи пользователей с другими пользователями;
– взаимодействия пользователей друг с другом.

	affiliate_channel	affiliate_provider	age	country_destination	date_account_created	date_first_booking	first_affiliate_tracked	first_brow
0	direct	direct	NaN	NDF	2010-06-28	NaN	untracked	Chrome
1	seo	google	38	NDF	2011-05-25	NaN	untracked	Chrome
2	direct	direct	56	US	2010-09-28	2010-08-02	untracked	IE
3	direct	direct	42	other	2011-12-05	2012-09-08	untracked	Firefox
4	direct	direct	41	US	2010-09-14	2010-02-18	untracked	Chrome

Рисунок 4. Пример данных о пользователе

На рисунке 4 представлен пример данных о пользователе. В большинстве своем они хранят данные, относящиеся непосредственно к конкретному пользователю.

Данные о связях позволяют понять отношения одного пользователя к другому. Взаимодействия, в свою очередь описывают непосредственно действия между, как самим пользователями, так и их активностью. Прекрасными примерами взаимодействия в разрезе социальных сетей могут быть:

- реакция на пост;
- заявка в друзья;
- просмотр профиля;
- комментарий.

На основании данных необходимо определить признаки, на базе которых будет производиться обучение модели.

Признаки можно разделить на сотни, если не тысячи различных типов, но в общем стоит рассматривать:

1 индивидуальные признаки пользователя: пол, рост, возраст, страна;

2 общие признаки пользователей: места, хобби, учебные заведения;

3 социальные признаки пользователей: посещаемость страницы профиля, общие друзья.

Корректность выделенных признаков индивидуальна для каждой, отдельно взятой, модели.

Разработка модели машинного обучения. Графовые нейронные сети (GNN) [1] — это тип нейронных сетей, предназначенных для обработки данных, представленных в виде графов. Графы состоят из узлов (вершин) и рёбер (связей), что позволяет GNN эффективно моделировать и анализировать структуру данных, содержащих сложные взаимосвязи между объектами.

Ключевые особенности GNN.

1 Адаптивность к структурированным данным: GNN способны обрабатывать данные, представленные в виде графов, такие как социальные сети, молекулярные структуры, транспортные сети и т. д.

2 Передача и агрегация информации: узлы в графе могут обмениваться информацией с соседними узлами, что позволяет сети "обучаться" и выявлять скрытые зависимости.

3 Обобщающая способность: GNN могут обобщать знания, извлечённые из обучающих данных, на новые, невиданные структуры, что делает их мощным инструментом для множества задач машинного обучения.

Принцип работы GNN. Основой работы GNN являются две основные операции: передача сообщений и агрегация. На каждом этапе (итерации) узлы графа передают и принимают сообщения от своих соседей, обновляя свои представления (векторные состояния) на основе полученной информации. Этот процесс повторяется до достижения сетью стабильного состояния.

Этапы обучения GNN:

1 Подготовка данных: сбор и предварительная обработка данных, которые будут представлены в виде графа. Узлы и рёбра графа должны быть определены и снабжены соответствующими признаками.

2 Инициализация параметров: установка начальных значений параметров модели, которые будут оптимизироваться в процессе обучения.

3 Форвардный проход: на каждом этапе обучения узлы графа передают и получают сообщения от соседей, обновляя свои векторные представления на основе полученной информации.

4 Вычисление функции потерь: оценка текущей производительности модели с использованием метрики, подходящей для решаемой задачи (например, кросс-энтропия для классификации).

5 Обратное распространение ошибки: вычисление градиентов функции потерь по параметрам модели для их последующего обновления.

6 Обновление параметров: корректировка параметров модели с использованием алгоритма оптимизации (например, стохастического градиентного спуска).

7 Повторение процесса: повторение этапов форвардного прохода, вычисления функции потерь, обратного распространения ошибки и обновления параметров до тех пор, пока модель не достигнет требуемого уровня производительности или не исчерпает заданное количество эпох.

Этот процесс позволяет графовым нейронным сетям адаптироваться к структуре и признакам графа, обучаясь на основе доступных данных и улучшая свою способность обобщать знания для новых, невиданных структур.

Этапы работы GNN.

1 Инициализация: каждый узел графа получает начальное представление, основанное на его начальных признаках.

2 Передача сообщений: узлы обмениваются информацией с соседями через рёбра графа.

3 Агрегация: каждый узел обновляет своё представление, агрегируя полученные от соседей сообщения.

4 Финальное представление: процесс передачи и агрегации повторяется несколько раз, после чего сеть выдаёт конечные представления узлов для окончательной оценки.

Поскольку модель GNN предсказывает наличие рёбер, её можно рассматривать как бинарный классификатор. Для оценки её качества будем использовать метрику ROC-AUC.

Применение метрики ROC-AUC для GNN. ROC-AUC (Receiver Operating Characteristic - Area Under Curve) [1] — это метрика, широко используемая для оценки качества бинарных классификаторов. Она основывается на построении кривой зависимости истинно положительных срабатываний (True Positive Rate) от ложно положительных (False Positive Rate) при различных порогах классификации. Пример кривой представлен на рисунке 5.

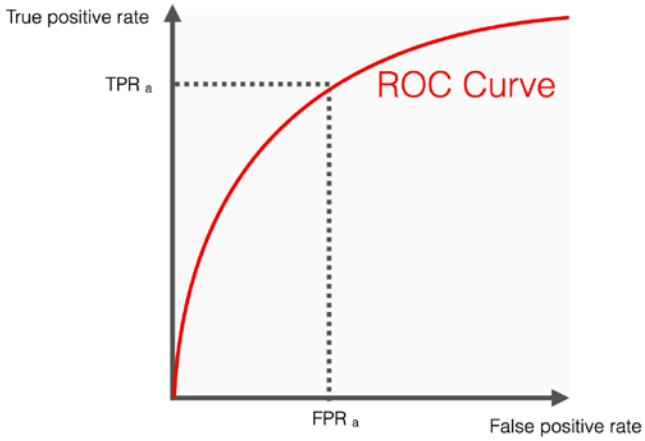

Рисунок 5. Пример ROC-AUC кривой

Применение ROC-AUC при работе с GNN:

1 Классификация узлов: в задачах, где GNN используется для классификации узлов графа (например, для определения категории узла в социальной сети или типологии молекулы), ROC-AUC позволяет оценить, насколько хорошо модель различает положительные и отрицательные классы узлов при различных порогах.

2 Классификация связей: в задачах, связанных с прогнозированием наличия или отсутствия связи между двумя узлами (например, для рекомендаций друзей или предсказания химических взаимодействий), ROC-AUC помогает понять, насколько

эффективно модель идентифицирует истинные положительные связи при минимальном количестве ложных срабатываний.

Использование метрики ROC-AUC для оценки GNN позволяет получить обобщённое представление о качестве модели вне зависимости от выбора конкретного порога классификации, что делает её особенно полезной при работе с несбалансированными данными, где классы могут быть представлены в разных пропорциях.

Схема системы и возможные улучшения. На рисунке 6 представлена базовая версия системы ранжирования.

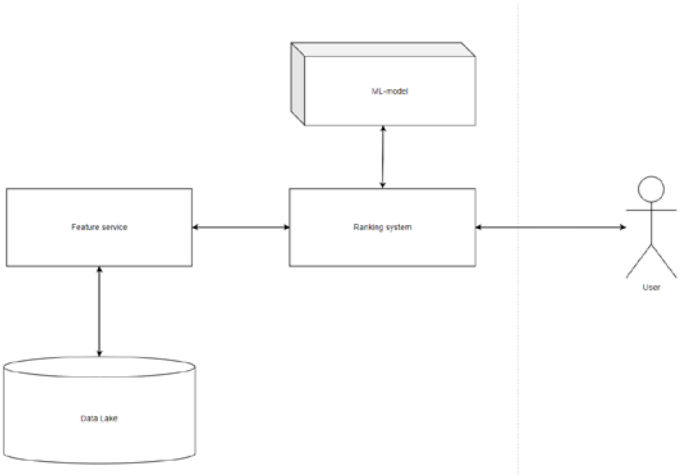

Рисунок 6. Принципиальная схема системы ранжирования

Принцип ее действия прост. Сервис ранжирования получает запрос, и собирает социальный вектор. Вектор отправляется на вход модели машинного обучения, которая, в свою очередь, отдает свои предсказания, которые передаются пользователю в качестве рекомендаций.

Однако подобная система не сможет справиться с большим количеством запросов, а количество узлов социального графа может насчитывать миллионы. На основании данных исследования [2], новые дружеские связи с вероятностью в 92% формируются через «друзей друзей». На основании данного факта можно заключить, что анализ следует проводить только в отношении «друзей друзей», что позволит значительно сократить количество узлов.

На рисунке 7 представлена базовая схема с добавлением сервиса определения «друзей друзей».

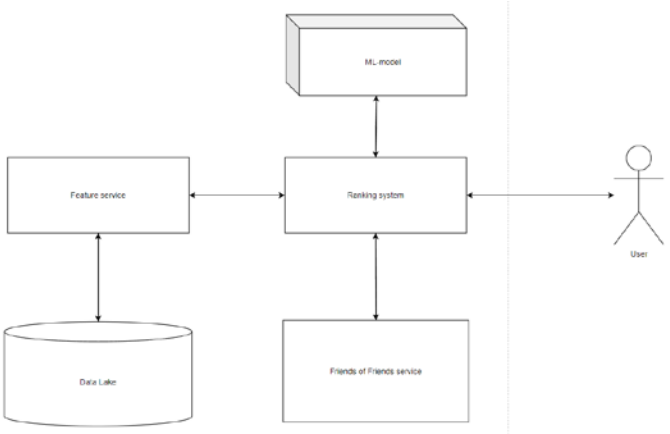

Рисунок 7. Принципиальная схема системы ранжирование с добавочным сервисом определения потенциальных друзей.

Дальнейшая оптимизация может быть направлена на отказ от работы системы по запросу. Для это достаточно разработать сервис предварительных расчетов, который будет обновлять предсказания в зависимости от приходящих изменений и выдавать пользователям заблаговременно подготовленные ответы. Принципиальная схема отображена на рисунке 8.

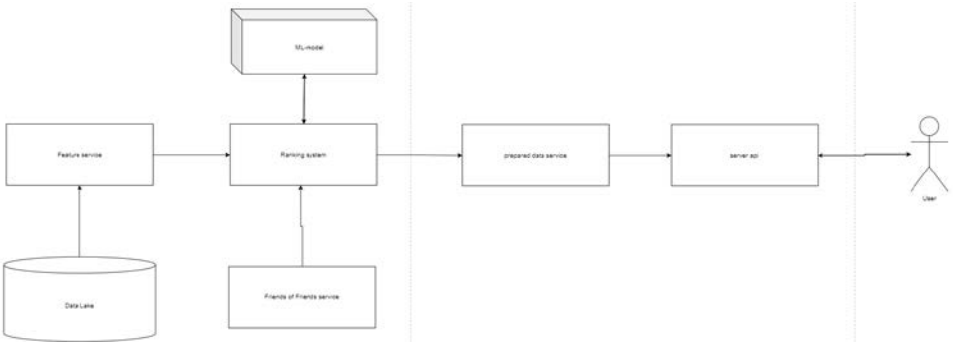

Рисунок 8. Принципиальная схема системы ранжирование с добавочным сервисом определения потенциальных друзей и предварительной генерацией рекомендация.

Заключение. Грамотно реализованная система, корректно подобранные признаки и рационально смоделированная модель машинного обучения позволяет добиться высоких результатов при выполнении предсказаний и предоставлении пользовательских рекомендаций. Уменьшение размерности графов способствует увеличению производительности системы. Использование системы ранжирования позволяет улучшить пользовательский опыт пользователя приложения социальной направленности, что, в свою очередь, благоприятно влияет на популярность системы.

Список литературы

[1] Рашка С., Лю Ю., Мирджалили В. Машинное обучение с PyTorch и Scikit-Learn: учебное пособие. Москва: Издательство, 2023. – 600 с.

[2] интернет ресурс https://youtu.be/Xpx5RYNTQvg?t=1823

Авторский вклад

Бугаев Евгений Анатольевич – анализ и описание системы ранжирования. Разработка схемы системы, практическое описание.

USING THE RANKING METHOD TO DETERMINE A LIST OF POTENTIAL USER ACQUAINTANCES

E. A. Bugaev
Software Engineer
Whitesnake

Annotation. An analysis of the approach to planning the ranking system when determining the list of potential acquaintances is carried out. The main algorithm for solving this problem is described. A basic scheme of the ranking service has been developed, possible options for improving the system are shown. Optimization of the input and output data of the system was carried out. Assessed and optimized**.**

Keywords: ranking, machine learning model, graph, GNN, ROC-AUC**.**