Хакатон “Нефтекод”

Алиев Т.А.[1], Медведев М.Г.[1], Муравьев А.А.[1], Сушков О.А.[2], Кожевников Д.А.[3], Петров В.А.[3], Скорб Е.В.[1]

[1] НОЦ Инфохимии, Университет ИТМО, ул. Ломоносова 9, Санкт-Петербург, 191002, Россия, skorb@itmo.ru

[2] ООО Газпромнефть – Цифровые решения, Киевская,5 к 4, Санкт-Петербург, Россия, 190013, Россия

[3] ООО Газпромнефть – смазочные материалы, Ленинградский проспект, д. 37А, корп. 4, БЦ «Arcus III», Москва, 125167, Россия

Оглавление

Введение

Хакатон разработан для реализации алгоритма предсказания параметров многокомпонентных рецептур масел с использованием данных об отдельных компонентах рецептуры. Компоненты необходимо представлять с использованием дескрипторов для отражения их химической природы, структуры (SMILES), геометрии (координаты атомов) и т.д. в машиночитаемом виде. Основной проблемой в создании алгоритма является разное число и содержание компонентов в рецептурах масел, что требует использование современных алгоритмов, таких как трансформеры, LSTM и LLM. Кроме того, при использовании квантово-механических молекулярных дескрипторов потребуется отбор полезных признаков. Для работы с формами представления молекул могут оказаться полезными графовые нейронные сети. Работа с различным набором входных данных, в соответствии с количеством компонентов в масле, а также разнообразие входных данных может потребовать разработку универсального конвейера. Создание подобного универсального алгоритма может быть использовано для предсказания других свойств различных многокомпонентных смесей.

Экспертиза представленных решений будет осуществляться на наборе рецептур, не представленных в датасете для участников, на платформе с лидербородом.

Данные

Для работы участникам предоставлены 2 вида данных.

Открытые данные

Молекулы компонентов масла в формате SMILES и необходимый для предсказания параметр. Участники могут извлекать любые признаки из SMILES (дескрипторы, графовую репрезентацию и т.д.). Помимо этого, к каждому SMILES предоставлены результаты квантово-механических расчетов методом DFT, как дополнительный дескриптор.

Приватные обезличенные данные

Структура данных

Данные будут представлены в виде таблицы (рисунок 1).

	oil_type	oil_property_param_title	oil_property_param_value	component_name	component_type_title	component_property_param_title	component_property_param_value
0	1ca56158-6fe1-47fb-b19f-61cf5a26e2da	6a84c27d-cfa4-4c39-8c98-c7ca82c999a4	NaN	84d7e484-f2d2-44e0-969c-ef78bea6720f	59eeacb3-1764-4088-9e90-8f459b8630c5	ff4a1b84-d577-44a3-b369-ab1a0fc2c4fc	NaN
1	1ca56158-6fe1-47fb-b19f-61cf5a26e2da	6a84c27d-cfa4-4c39-8c98-c7ca82c999a4	NaN	84d7e484-f2d2-44e0-969c-ef78bea6720f	59eeacb3-1764-4088-9e90-8f459b8630c5	aae8479c-7373-463f-95e7-c3909db72f26	NaN
2	1ca56158-6fe1-47fb-b19f-61cf5a26e2da	6a84c27d-cfa4-4c39-8c98-c7ca82c999a4	NaN	84d7e484-f2d2-44e0-969c-ef78bea6720f	59eeacb3-1764-4088-9e90-8f459b8630c5	28d3183f-4a1b-4d41-a7b5-7e38cbf79faf	0.00014
3	1ca56158-6fe1-47fb-b19f-61cf5a26e2da	6a84c27d-cfa4-4c39-8c98-c7ca82c999a4	NaN	84d7e484-f2d2-44e0-969c-ef78bea6720f	59eeacb3-1764-4088-9e90-8f459b8630c5	cdf8b7fc-d205-42ba-a33e-c7cde16c2ca2	0.00110
4	1ca56158-6fe1-47fb-b19f-61cf5a26e2da	6a84c27d-cfa4-4c39-8c98-c7ca82c999a4	NaN	84d7e484-f2d2-44e0-969c-ef78bea6720f	59eeacb3-1764-4088-9e90-8f459b8630c5	8aaadd15-56c7-4dc9-b6e0-67b0736b55be	0.00011

Рисунок 1. Приватные обезличенные данные. Под UUID скрыты типы масел, свойства масел, свойства каждого компонента. Каждому UUID соответствует определенное название, поэтому с помощью них можно выделить каждое масло/компонент или их свойство.

Каждый столбец соответствует:

- oil_type – тип смазочного масла;
- oil_property_param_title – название свойства каждого смазочного масла;
- oil_property_param_value – значение свойства каждого смазочного масла;
- component_name – название компонента смазочного масла;
- component_type_title – тип компонента смазочного масла;
- component_property_param_title – название свойства компонента смазочного масла;
- component_property_param_value – значение свойства компонента смазочного масла.

Таким образом, в таблице представлен набор масел и их свойства. Одно из свойств требуется предсказать. Для каждого масла представлены компоненты, для каждого компонента – свойства. Эти свойства могут быть использованы в составе обучающей выборки. На рисунке 2 дана схема распределения данных. Зеленым отмечены признаки, которые можно использовать для обучения, красным – признак, который нужно предсказать.

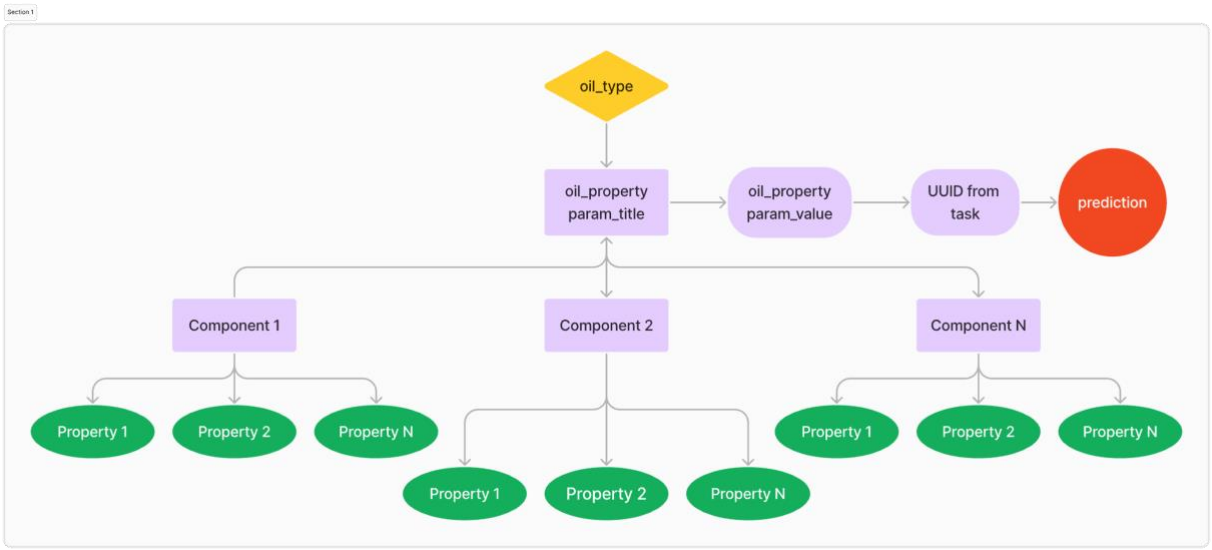

Рисунок 2. Схема распределения данных в таблице.

Наполнение данных

Представленные данные имеют пропуски, от которых можно избавиться любым доступным способом. Однако не рекомендуется использовать исключительно восстановленные данные для обучения, а изначально полные данные для валидации, так как модель может себя неправильно повести на скрытой выборке.

Постановка задачи

Разработать модель, предсказывающую исследуемый параметр масла на основе представленных открытых и приватных обезличенных данных. Модель должна работать как при подаче только нескольких SMILES (как компонентов), так и при подаче параметров, похожих на свойства компонентов, как в приватных обезличенных данных (модель должна сама восполнить пропуски в них, если они будут). Для решения этой задачи рекомендуется использовать следующий план:

1. Проанализировать данные;
2. Проанализировать публикации, за счет них увеличить размер выборки;
3. Рассмотреть способы преобразования SMILES в машиночитаемый вид с помощью трансформеров, графовых нейросетей, квантово-химических дескрипторов и т.д.;
4. Разработать первую часть модели, способную предсказывать необходимый параметр по SMILES. Количество SMILES для одного масла может быть произвольным;
5. Разработать вторую часть модели, предсказывающую нужный параметр по приватным обезличенным данным. При этом модель должна автоматически восполнять пропуски в данных, если таковые имеются;
6. Собрать обе части модели в конвейер, оптимизировать гиперпараметры;
7. Оценить жизнеспособность модели на тестовых данных основными метриками;
8. Подготовить инференс;
9. Подготовить презентацию вашего решения.

Полученная модель после загрузки в систему будет оцениваться по средней абсолютной ошибке (MAE). В качестве приватного датасета будут использоваться экспериментальные данные, не предоставленные участникам для хакатона. Целевая переменная будет находится в столбце oil_property_param_value с совпаданием UUID в столбце oil_property_param_title - ad7e6027-00b8-4c27-918c-d1561f949ad8.

Образ финального решения

Образ финального решения состоит из модели, способной предсказывать исследуемый параметр смазочного масла с произвольным количеством компонентов, как для компонентов, представленных с помощью SMILES, так и для известных компонентов, представленных экспериментальными параметрами из приватного обезличенного датасета. На рисунке 3 представлен примерный образ финального решения.

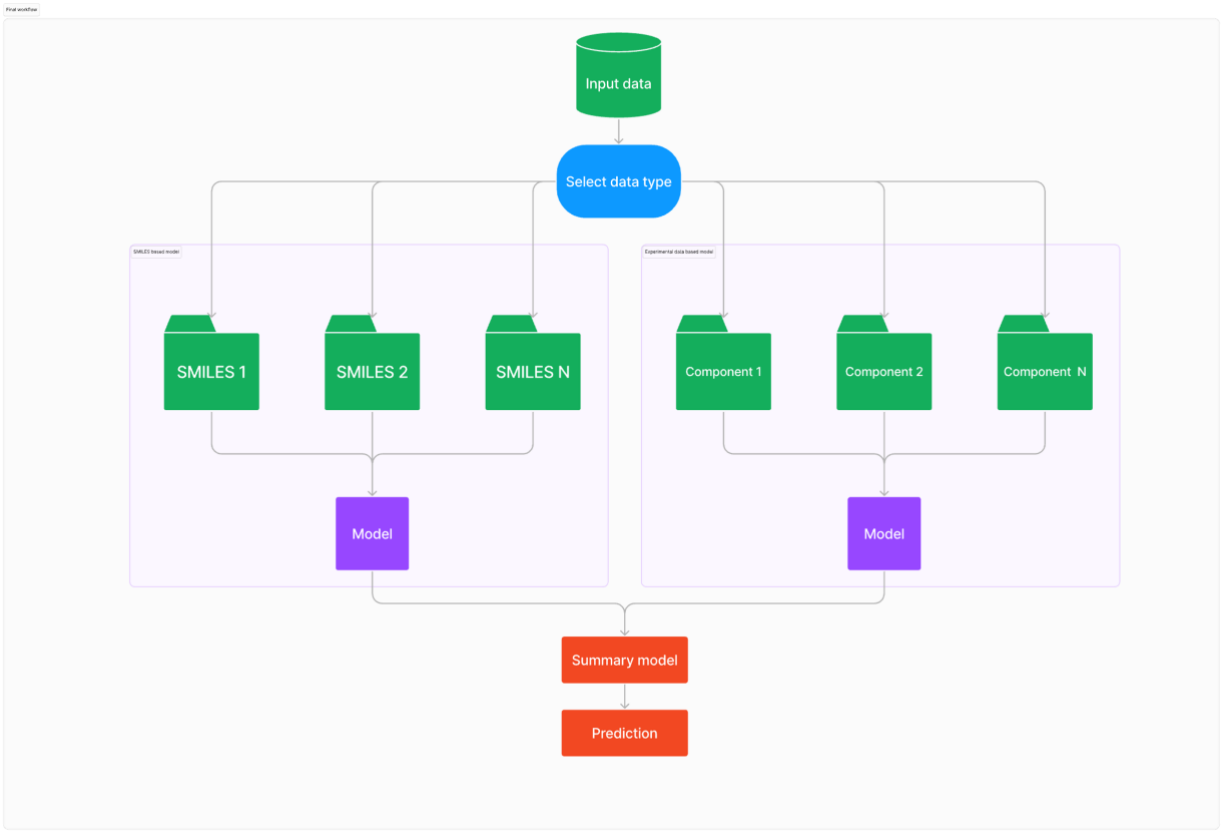

Рисунок 3. Образ финального решения.

Загрузка решений

Загрузка решений будет осуществляться через платформу. Решение необходимо упаковать в zip архив. Zip архив должен содержать код модели и ее инференс (notebook или script), а также ответы к предоставленным тестовым данным в формате файла csv, сохраняя последовательность.

При загрузке архива csv файл с ответами вашего решения будет сверяться с правильными ответами и высчитываться средняя абсолютная ошибка. Полученная ошибка будет использоваться при ранжировании лидерборда.

На весь срок хакатона у каждой команды есть 50 попыток загрузить свое решение в систему (таймаут между загрузками 1 мин).

В день завершения хакатона необходимо прислать ваше лучшее решение на почту itmo.hack@itmo.ru в следующем виде:

- файлы с исходным кодом к обучению модели и инференсу (по возможности использовать Docker-контейнер)
- презентация решения
- веса обученной модели
- обучающая выборка

** в теме письма обязательно указать название команды*

Тестирование моделей

После завершения времени, отведенного на загрузку решений (21.04.2024 23:59), будут выбраны 10 лучших решений в соответствии с лидербордом. Каждое решение будет проверено экспертами вручную в соответствии с критериями оценивания и допущено до финала либо же нет.

Критерии оценивания

Основные критерии оценивания – уникальность решения, подготовка датасета, точность и устойчивость модели, и представление результатов.

Экспертная оценка решений. В период с 22.04 по 23.04 2024 г. экспертами производится ручная проверка предложенных решений по следующим критериям:

1. Увеличение предоставленного датасета. Необоснованное увеличение датасета (например, сбор случайных SMILES из открытых источников) не учитывается;
2. Подготовка модели на SMILES: уникальность подхода, масштабируемость и скорость работы;
3. Подготовка модели на экспериментальных данных, подобных приватным обезличенным данным, либо на SMILES: уникальность подхода, масштабируемость и скорость работы;
4. Подготовка инференса: скорость работы, удобство развертывания;
5. Реализация алгоритма работы с неизвестным количеством компонентов в смазочном масле: уникальность подхода, масштабируемость и скорость работы

Проверка будет осуществляться на операционной системе Ubuntu 22.04. Для избежания проблем с запуском решения следует уделить внимание инференсу модели и, по возможности, использовать Docker-контейнер.

При невозможности запустить решение дальнейшая проверка не осуществляется. Вместо этого решения будет взято следующее в ранжировании решение команды, не попавшей в топ-10 лидерборда.

При обнаружении попытки умышленного подгона результатов под платформу лидерборда с использованием скриптов, случайных или необоснованных данных решение аннулируется. Вместо этого решения будет взято следующее в ранжировании решение команды, не попавшей в топ-10 лидерборда.

24 апреля 2024г. участникам будет оглашен список топ-10 команд, которые будут допущены до финальной защиты решения перед расширенной командой экспертов. Защита пройдет 26.04 на базе Университета ИТМО по адресу г. Санкт-Петербург, Кронверкский пр-кт 49 (возможен гибридный формат защит, если ваша команда находится не в г. Санкт-Петербург).

Рекомендации к применению алгоритмов

Ниже предложены алгоритмы для повышения точности и качества модели:

1. Получение эмбеддингов из SMILES. Можно брать готовые большие лингвистические модели, напр., ChemBERT (Seyone Chithrananda 2020);
2. При реализации алгоритмов работы с репрезентацией молекулы, отличной от SMILES (2D, 3D и т.д.) – графовое представление с использованием фреймворка PyTorch Geometric (Matthias Fey 2019);
3. При реализации алгоритма работы с многокомпонентной смесью – графовые нейросети, а также решения, применяемые в трансформерах (Ashish Vaswani 2017), LSTM (Sepp Hochreiter 1997) и LLM.

Применение этих алгоритмов носит рекомендательный характер.

Список литературы

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. «Attention Is All You Need.» *arXiv.*

Matthias Fey, Jan Eric Lenssen. 2019. «Fast Graph Representation Learning with PyTorch Geometric .» *arXiv.*

Sepp Hochreiter, Jürgen Schmidhuber. 1997. «Long Short-Term Memory.» *Neural Computation* 1735-1780.

Seyone Chithrananda, Gabriel Grand, Bharath Ramsundar. 2020. «ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction .» *arXiv.*