ПРОГНОЗИРОВАНИЕ ЗАГРЯЗНЕНИЯ ТЯЖЕЛЫМИ МЕТАЛЛАМИ ИСПОЛЬЗУЯ ДАННЫЕ С КОСМОСНИМКОВ И МАШИННОЕ ОБУЧЕНИЕ

Борисов Максим Сергеевич[1], Ужинский Александр Владимирович[2]

[1] *Студент;*
Государственный университет «Дубна»;
Международная школа по информационным технологиям
«Аналитика больших данных»;
Направление обучения по основной образовательной программе:
Программная инженерия, группа 4252;
e-mail: maksppap@mail.ru.

[2] *Ведущий программист;*
Лаборатория информационных технологий им. М.Г. Мещерякова;
Объединенный институт ядерных исследований.

Ключевые слова: прогнозирование, космоснимки, тяжелые металлы, классификация, градиентный бустинг, нейронная сеть.

Введение

Проблемы загрязнения окружающей среды и экологической безопасности в последние годы крайне актуальны. Загрязнение воздуха оказывает значительное негативное влияние на различные компоненты экосистем, здоровье человека, вымирание животных. Более 90% людей проживают в местах, где загрязнение воздуха превышает безопасные пределы, согласно исследованиям Всемирной организации здравоохранения (ВОЗ) [1]. Промышленные выбросы и отходы — это существенный фактор негативного воздействия на окружающую среду. Критически важно иметь механизмы контроля загрязнений окружающей среды.

Наиболее распространенным методом оценки состояния окружающей среды является определение количества мелкодисперсных частиц и некоторых химических соединений, например, $CO2$. Для получения подробной информации о составе загрязнения используются методы мониторинга, основанные на сборе проб.

Идея использования мхов для измерения атмосферных выпадений тяжелых металлов основана на том факте, что мхи, получают большую часть микроэлементов и питательных веществ непосредственно из атмосферы, при этом поглощение металлов из почвы невелико за счет поверхностного расположения корневой системы [2].

В рамках проекта комиссии Организации Объединенных Наций (ООН) дальнему трансграничному переносу воздушных загрязнений (*ICP Vegetation*) участники собирают образцы мха и используют различные техники, например нейтронно-активационный анализ, чтобы получить данные по содержанию тяжелых металлов, соединений азота, стойких органических соединений и радионуклидов.

Поскольку стандартные методы мониторинга основаны на отборе проб и дальнейшем их анализе, они требуют больших временных и трудовых затрат для прогнозирования в глобальных масштабах, также, при таком подходе сложно получать образцы в труднодоступных зонах. Поэтому хорошим дополнением может стать прогнозирование, оно позволит частично автоматизировать процесс контроля окружающей среды. Общая идея состоит в использовании данных, которые можно получать со спутниковых изображений вместе с данными, полученными после анализа собранных образцов, чтобы обучить модель, а затем использовать только данные со спутниковых изображений для дальнейшего анализа концентрации тяжелых металлов.

1. Исходные данные

Предоставленные данные содержат 5188 точек отбора проб по 14 металлам и 1 неметаллу: *As* (мышьяк), *Cd* (кадмий), *Cr* (хром), *Fe* (железо), *Mg* (магний), *Ni* (никель), *Pb* (свинец), *V* (ванадий), *Zn* (цинк), *Al* (алюминий), *Sb* (сурьма), *N* (азот), *Sr* (стронций), *Mn* (марганец). Они представлены следующим образом: *Longitude* (долгота), *Latitude* (широта), концентрации металлов. Образцы были собраны участниками программы «*ICP Vegetation*» в период 2015–2016 года.

2. Получение данных с космоснимков

Google Earth Engine (*GGE*) обеспечивает веб-доступ к обширному каталогу спутниковых изображений и других геопространственных данных в формате, готовом для анализа. Сервис предоставляет программные интерфейсы (*API*), которые позволяют беспрепятственно реализовывать работу с геопространственными данными.

Наборы данных в *GEE* представлены в виде коллекций (*Collection*) – структура данных, содержащая космоснимки, а также метаданные о них. Набор растровых данных содержит один или несколько слоев, называемых каналами (*bands*). Например, цветное изображение имеет три канала (красный, зеленый и синий), в то время как цифровая модель рельефа имеет один канал (содержащий значения высоты), а мультиспектральное изображение может иметь множество каналов. Каждый растровый канал (*band*) содержит фактические значения ячеек (см. рис. 1), а также некоторые ключевые свойства, такие как: название, разрешение, длина волны, вид спектра, цветовая карта и описание. *Reducer* – это способ агрегирования данных во времени, пространстве, диапазонах, массивах и других структурах данных в *GEE*. Он определяет способ агрегирования данных. В этом классе можно указывать простую статистику для использования агрегирования (например, минимум, максимум, среднее значение, медиана, стандартное отклонение) или более сложную сводку входных данных (гистограмму, линейную регрессию). Визуализация работы *Reducer* приведена на рисунке 1. Индекс включается в себя следующие параметры: космическая программа, исследуемая область, временной промежуток и *Reducer*. Формируется *API*-запрос для получения индекса, он рассчитывается путем произведения вычислительных операций над переданными параметрами и длинной волны в различных спектрах. Именно на основе индексов проводится прогнозирование.

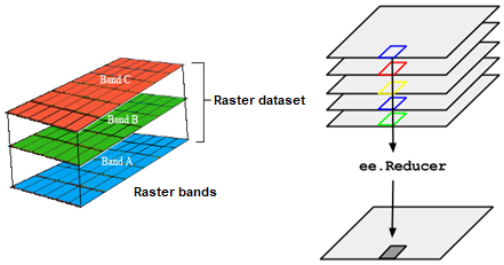

Рис. 1. Формирование растрового канала и визуализация работы Reducer

Для каждой точки отбора производится получение индексов (из различных космических программ). Но существует более 100 космических программ, со множеством каналов, при этом не все из них имеют значимую корреляцию между концентрацией металла и полученным индексом, поэтому каналы необходимо отбирать. Для отбора индексов применяется ранговая корреляция Спирмена (r) и *phik*-корреляция (между концентрацией металла и полученными индексами). Отбираются индексы с лучшей ($|r| > 0.45$) корреляцией, отбрасывая коллинеарные (оказывают негативное влияние на качество обучаемых моделей).

Выявление статистических зависимостей, путем нахождения корреляции, служит только для первичного анализа данных. Корреляция не предоставляет точной информации о взаимосвязи признаков.

3. Прогнозирование

Задачу прогнозирования (с учителем) можно решать 2 способами: регрессия и классификация.

3.1. Регрессия

Прогнозное регрессионное моделирование — это задача аппроксимации функции отображения (f) от входных переменных (x) до непрерывной выходной переменной (y). Решая задачу регрессии, предсказывается непрерывное значение. В качестве модели регрессии выбрана библиотека *XGBoost* основанная на алгоритме дерева решений и реализующая методы градиентного бустинга. Идея градиентного бустинга заключается в построении последовательно уточняющих друг друга элементарных моделей, последняя из которых обучается на ошибках предыдущих и ответы моделей суммируются. Также библиотека *XGBoost* имеет встроенные методы регуляризации – что позволяет избежать переобучения моделей.

Модель регрессии *XGBoostRegressor* обучалась с использованием алгоритма подбора гиперпараметров *GridSearchCV*, перебираемые параметры: *n_estimators*, *max_depth*, *booster*, *gamma*, *subsample* и другие. При помощи механизма *feature_importances* (показывает вклад каждого признака в модель), были дополнительно отобраны индексы, что способствовало увеличению точности модели.

Для оценки качества регрессионной модели использовался коэффициент детерминации ($R2$) и среднеквадратическая ошибка (*RMSE*). Результаты лучшей модели (для прогнозирования концентрации алюминия): $R2 = 0.54$, $RMSE = 4.06$, что недостаточно для прогнозирования, поэтому было принято решение использовать классификацию с разбиением концентраций на диапазоны.

3.2. Классификация

Задача алгоритма классификации состоит в том, чтобы найти функцию отображения для сопоставления входа (x) с дискретным выходом (y). Для задачи классификации необходимо разбить показатели концентрации на диапазоны. На примере алюминия (*Al*) классов получилось 8. После разбиения на классы выборка получилась несбалансированная. При обучении моделей классификации необходимо сбалансировать выборку. Выборку можно балансировать разными способами (*undersampling*, *oversampling*, добавление весов классам). В рамках данной задачи оптимальным способом является *oversampling*, реализован с помощью алгоритма *SMOTE*. Суть алгоритма заключается в создании искусственных примеров минорных классов, не используя дублирования. Были протестированы различные алгоритмы классификации: *Decision Tree*, *Random Forest*, *MLP* (*encoder*), *XGBoostClassifier* показал лучший результат ($accuracy = 82\%$). Использовались техники: подбора гиперпараметров, уменьшение скорости обучения, ранний выход из процесса обучения. Визуализация работы модели *XGBoostClassifier* продемонстрирована на рисунке 2.

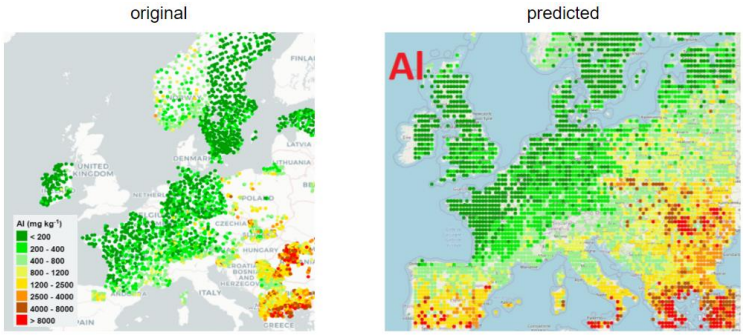

Рис. 2. Визуализация работы модели XGBoostClassifier

Также для решения задачи классификации была протестирована сиамская архитектура (*SNN*). Такая архитектура содержит две или более идентичных подсетей (с одинаковыми весами), используемых для генерации векторов признаков. Входы *SNN* состоят из 3 признаков: якорь (*anchor*) – произвольный класс, *positive* – класс, совпадающий с *anchor*, *negative* – класс отличающийся от *anchor*. В сети используется триплетная функция потерь (*triplet loss*), она высчитывает расстояние между 3 выходными векторами признаков. Минимизируя данную функцию, негативный класс отдаляется от якоря, тем самым упрощая классификацию векторов признаков. После обучения положительные примеры будут ближе к якорю, а отрицательные — дальше от него. В ходе работы не удалось достичь корректной работы сети. На рисунке 3 изображен эффект минимизации триплетной функции потерь.

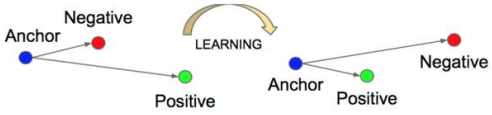

Рис.3. Визуализация работы минимизации триплетной функции ошибки

4. Веб-приложение

Для визуализации и анализа результатов было реализовано веб-приложение. В качестве программной платформы для создания интернет-приложений на языке программирования *Python* выбран *Dash*, разработан с использованием библиотек *Flask, React.JS, Plotly*. Он хорошо оптимизирован для создания и развертывания приложений, которые работают с данными и настраиваемыми пользовательскими интерфейсами. Все данные поступают из *NoSQL* базы данных *MongoDB*. Реализован функционал: выбора металла для отображения, выбор диапазона концентраций, визуализация гистограммы распределения классов, загрузка выгрузка результатов, объединение точек в области (гексагоны), в них отображается средняя концентрация. Интерфейс приложения изображен на рисунке 4.

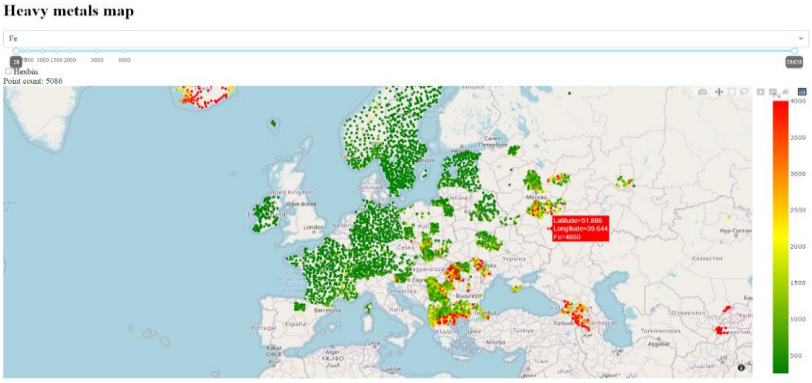

Рис. 4. Интерфейс веб-приложения

Заключение

В ходе исследовательской были протестированы 2-метода прогнозирования: регрессия и классификация. Модели регрессии не показали приемлемой точности, максимальный коэффициент детерминации равен $R2 \approx 0.54$, что недостаточно для прогнозирования. Лучшей моделью прогнозирования тяжелых металлов в задаче классификации проявил себя *XGBoostClassifier*, точность (*accuracy*) для 8 классов при прогнозировании алюминия (*Al*) составила 82%, при этом, получаемые ошибки находятся в соседних классовых диапазонах. Таких показателей будет достаточно, чтобы производить мониторинг. Так же были протестированы нейросетевые классификаторы, в том числе архитектуры сиамских сетей с триплетной функцией ошибок, но с их помощью не удалось добиться точности, сопоставимой с алгоритмами градиентного бустинга.

Дальнейшие исследования должны быть направлены на воспроизведение результатов в глобальном масштабе. Планируется протестировать *state of the art* (*SOTA*) решения в области *metric learning*, для задач прогнозирования. Апробировать архитектуры глубокого обучения для работы с табличными данными.

Список литературы

1. Исследования ВОЗ. — [Электронный ресурс]. URL: https://www.who.int/ru/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action.
2. Marina Frontasyeva, Harry Harmens, Alexander Uzhinskiy, Omar Chaligav // Mosses as biomonitors of air pollution: 2015/2016 survey on heavy metals, nitrogen and POPs in Europe and beyond, 2020. P.1-136.