Применение методов машинного обучения для оценки количества биомассы по данным спутниковых снимков Sentinel-1 и Sentinel-2

Автор: Петренко Д.
Руководитель: Озерин А.

Оглавление

Постановка и описание задачи.

Развитие космонавтики привело к удешевлению спутников и снижению стоимости вывода их на орбиту. Что в свою очередь породило поток разнообразных регулярных спутников земной поверхности. Существует ряд инициатив в Европе, США и Китае, которые обеспечивают открытый доступ к гиперспектральным снимкам с высоким разрешением (порядка 20 метров на пиксель). Например, Европейское Космическое Агенство спонсирует миссию Copernicus https://www.copernicus.eu/. Полученные спимки можно посмотреть на http://apps.sentinel-hub.com/. Множество энтузиастов, ученых, институтов и компаний исследуют эти снимки на предмет всевозможных применений. Сейчас они активно используются для наблюдений за ледниками и токсичными водорослями, мониторинга состояния полей, предсказания погоды, отслеживание выбросов метана и многих других приложений.

В данной работе предлагается использовать данные соревнования “The BioMassters”. Цель соревнования заключается в разработке решения для прогнозирования ежегодной надземной биомассы (Above ground biomass - AGB) для финских лесов с использованием спутниковых снимков. Подробную информацию можно найти на странице соревнования: The BioMassters - Competition. Надземная биомасса определяется как стоячая надземная масса живого или мертвого вещества древесных или кустарниковых растений выраженная в виде массы на единицу площади. [1]
В нашем распоряжении два набора данных:

- Спутниковые снимки, полученные Европейским космическим агенством из взаимодополняющих спутниковых программ Sentinel-1 и Sentinel-2, разработанных для получения широкого массива данных о земной поверхности
- Данные измерений ABG собранные при помощи LiDAR в сочетании с измерениями на месте.

Применение LiDAR позволяет получить высококачественную оценку биомассы, однако требует существенно больше времени и усилий, чем спутниковые снимки. Авторы задачи оценили объем биомассы с

помощью LiDAR и предлагают сделать модель предсказывающую AGBM по гораздо более дешевым и доступным спутниковым снимкам. Задача кажется с одной стороны достаточно полезной, а с другой - комплексной: из данных убрана гео-специфичная информация, а по смыслу это задача регрессии на мультимодальных данных с некоторой спецификой временных рядов.

Современные работы проводят исследования методов определения количества биомассы как в городских условиях, например для районов Лондона [1], так и в масштабах страны для оценки интегральных значений биомассы [2]

Ожидаемые результаты

По результатам выполнения данной работы ожидается:

- реализация паплайна для обучения моделей. Это инженерная задача обусловленная с одной стороны большим объемом данных(порядка 200 гигибайт), а с другой стороны нетривиальными данными: есть временная компонента, есть пропуски, данные мультимодальны.
- разработка baseline решения на модели машинного обучения
- разработка решения на базе нейронной сети

Обзор существующих подходов

В настоящее время для оценки надземного уровня биомассы широко применяются различные методы машинного обучения и глубинного обучения.

В частности в работе Somayeh Talebiesfandarani, Ali Shamsoddini [3 показано, что классическием методы машинного обучения могут быть применены для оценки биомассы. Авторы описывают в своей работе классический пайплайн машинного обучения включающий в себя подбор параметров модели, подбор признаков и непосредственно саму модель. Авторы рассматривали

в качестве моделей машинного обучения метод опорных векторов (Support Vector Machine-SVM) и случайный лес (Random Forest - RF), а в качестве модели глубинного обучения - сверточную нейронную сеть (Convolution Neural Network - CNN)

Также в работе Wu [4] было проведено исследование различных методов машинного обучения для оценки биомассы в провинции Чжэцзян Китая по спутниковым снимкам Landsat. Авторы провели анализ качества линейной регрессии, к-ближайших соседей (k-nearest neighbors - KNN), SVM, RF, стохастического градиентного бустинга(stochastic gradient boosting - SGB). В своей работе авторы приходят к выводу, что для имеющихся у них данных наиболее качественные результаты оценки биомассы дает RF.

Детальное исследование качества различных моделей машинного обучения для построения оценок биомассы было также проведено Yuzhen Zhang [5]. В этой работе было показано, что применение современных методов построения ансамблей деревьев, например в реализации CatBoost, могут значимо повысить качество оценки AGB.

Описание данных

Данные - набор изображений для для примерно 13.000 участков. Изображения Sentinel-1 - это радарные данные в четырех каналах, изображения Sentinel-2 - это мультиспектральные изображения (10 каналов и данные облачности).

В качестве целевой переменной используются данные LiDAR измерений для каждого пикселя изображения участка. Для каждого участка предоставлены ежемесячные спутниковые изображения за год. Для каждого участка должны быть доступны 24 изображения, но из-за перебоев с передачей данных некоторые из них могут быть потеряны. Также мультиспектральные изображения могут отсутствовать за какой-то период из-за высокой облачности - спутник технически не смог получить какое-либо изображение поверхности. У каждого канала многоканального изображения исходно было разное разрешение. Для упрощения работы авторы соревнования пересэмплирование снимки чтобы привести их

разрешению 10 метров на пиксель. Каждое изображение представляет из себя агрегированные за месяц данные в формате GeoTIFF без геолокации.

Метаданные

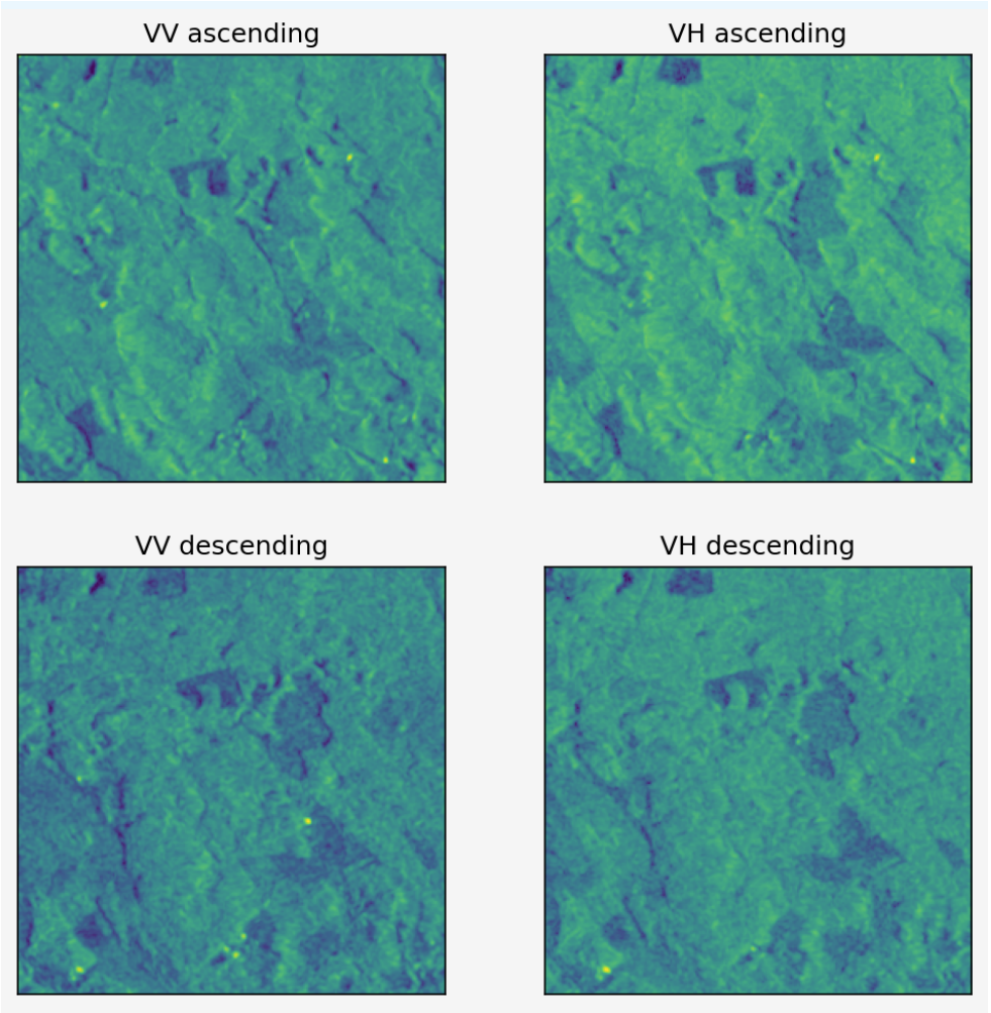
Для спутниковых изображений доступны сопутствующие метаданные. Наибольший интерес представляют следующие поля:

- chip_id - уникальный идентификатор для одного участка леса
- filename - имя файла состоящее из идентификатора участка, названия спутника и месяца съемки
- satellite - спутник, который сделал этот снимок
- month - размер файла в байтах
- corresponding_agbm - имя файла с данными измерений AGBM

Примеры данных: Sentinel-1

Спутники поколения Sentinel-1 используют для формирования изображений метод радиолокационного синтезирования апертуры в C-диапазоне (C-band Synthetic Aperture Radar - SAR [https://www.earthdata.nasa.gov/learn/backgrounders/what-is-sar]), что позволяет получать изображения вне зависимости от метеоусловий на земле. Спутники поколения Sentinel-1 работают на полярных орбитах. Так движение спутника из южного полушария в северное называется восходящим, а наоборот нисходящим.
Предоставленные данные имеют два диапазона: с горизонтальной и вертикальной поляризацией. Таким образом для данных Sentinel-1 мы имеем 4 различных варианта снимков:

- VV ascending
- VH ascending
- VV descending
- VH descending



Поскольку спутник Sentinel-1 такую орбиту, что спутник возвращается на одно и то же место раз в 6 дней, то один район фотографируется примерно пять раз в месяц. Мы будем использовать изображение полученное усреднением всех снимков сделанных в течение месяца.

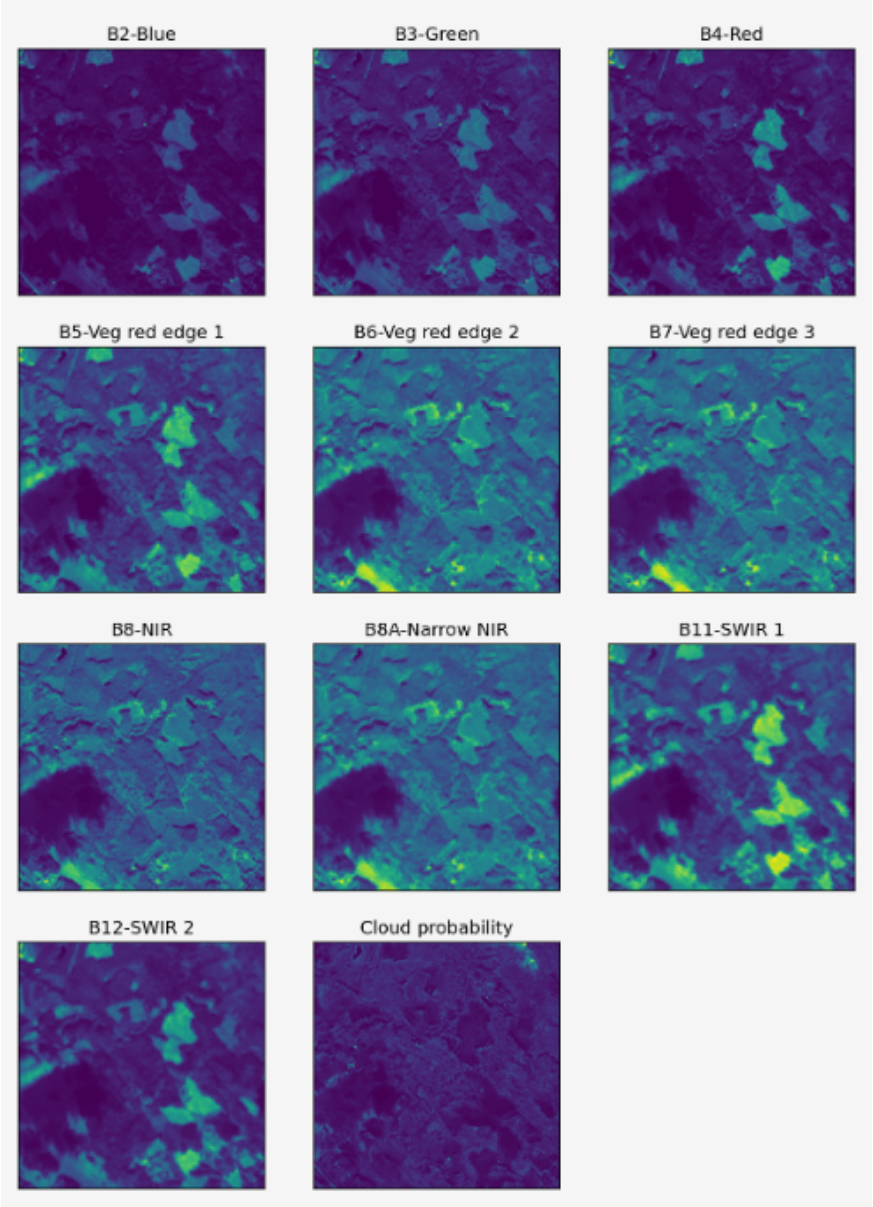
Примеры данных: Sentinel-2

Sentinel-2 - это космический аппарат для получения изображений в высоком качестве для мониторинга растительности, почвы, водного покрова и прибрежных районов. Sentinel-2 использует мультиспектральное оборудование, которое собирает данные в видимом, инфракрасном, коротковолновом и электромагнитном спектрах. Поскольку качество данных этого спутника зависит от погодных условий во время съемки мы будем

использовать лучшее из имеющихся изображений полученных в каждом месяце.

Каждый снимок Sentinel-2 имеет 11 каналов: B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12 и слой вероятности облачности (cloud probability - CLP) Более подробную информацию можно найти на странице спутника: Sentinel-2 L2A.

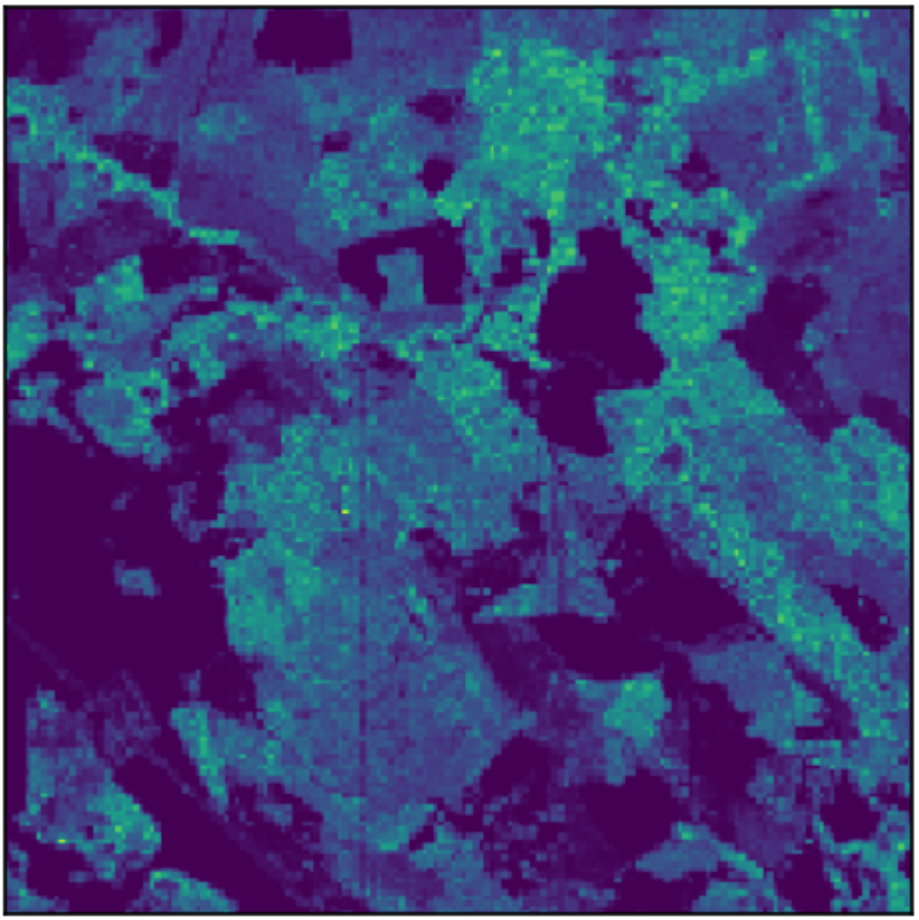
Канал CLP необходим, поскольку Sentinel-2 чувствителен к облачности. CLP указывает вероятность облачности для данного пикселя в диапазоне от 0 до 100. Значение 255 используется для разметки случаев чрезмерного шума



Наземные данные

Целевая переменная - истинные значения AGBM получены в результате применения технологии дистанционного зондирования LiDAR. Метка для каждого участка представляет собой пиковое значение биомассы измеренное в течение лета.

Аналогично спутниковым данным, данные LiDAR представляют из себя изображения областей размером 2560 на 2560 метров с разрешением 10 метров т.е. 256 на 256 пикселей. Значение 0 в пикселе означает, что это район с нулевой AGBM или что для этого района отсутствуют данные



Дальнейшие шаги

- Провести дополнительный анализ данных. В частности:
  - Рассчитать долю участков у которых потеряны изображения.
  - Провести анализ распределения таргета

- Оценить качество baseline решений:
  - Наивный прогноз: взять среднее значение для каждого пикселя и использовать его в качестве предсказания
  - Применить классические методы машинного обучения

- Реализовать модель на базе нейронных сетей.

Список использованной литературы

1. Estimating urban above ground biomass with multi-scale LiDAR | Carbon Balance and Management | Full Text

2. [Can Machine Learning Algorithms Successfully Predict Grassland Aboveground Biomass?]

3. https://doi.org/10.1016/j.rsase.2022.100868

4. (PDF) Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery

5. An Evaluation of Eight Machine Learning Regression Algorithms for Forest Aboveground Biomass Estimation from Multiple Satellite Data Products