статистических критериев. // Вестник пермского университета. Серия: Математика. Механика. Информатика. 2020 № 1 (48), с.26-32

2. Иванов А.И. Искусственные математические молекулы: повышение точности статистических оценок на малых выборках (программы на языке MathCAD): препринт // Пенза, из-во «Пензенского государственного университета», 2020 г., 36 с. ISBN 978-5-907262-42-3.

Сардор Каримов Илхом угли

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ С ИСПОЛЬЗОВАНИЕМ СПУТНИКОВЫХ ИЗОБРАЖЕНИЙ SENTINEL-2

(Санкт-Петербургский государственный университет)

Аннотация. Прогнозирование изменчивости урожая в пределах поля может помочь фермерам принимать правильные решения в различных ситуациях. Текущие достижения в области дистанционного зондирования и доступность изображений высокого разрешения, высокой частоты и бесплатных изображений Sentinel-2 улучшают внедрение точного земледелия для более широкого круга фермеров.

Ключевые слова. Deep learning, XGBoost, LightGBM, зондирования, Sentinel

Введение

Цель этой статьи — создать модель, способную оценить урожайность полей в Восточной Азии. Дан временной ряд изображений Sentinel 2 и климатических переменных. Модель сможет оценить пространственную изменчивость урожайности зерна кукурузы в тоннах на акр.

Данные Sentinel-2 открывают новые возможности для регионального, а также глобального сельскохозяйственного мониторинга, позволяя просматривать Землю в 12 спектральных диапазонах с пространственным разрешением 10–20 м, с глобальным охватом и 5-дневной периодичностью повторных посещений и совместимыми с текущие и исторические миссии Landsat. Мониторинг свойств почвы и состояния посевов, наряду с картированием обработки почвы, помогает исследователям и фермерам оценивать землепользование, прогнозировать урожай, отслеживать сезонные изменения и помогать в реализации политики устойчивого развития. С ростом числа доступных источников спутниковых данных, многие из которых можно использовать бесплатно, потенциал огромен (Sentinel-hub, 2021) [1].

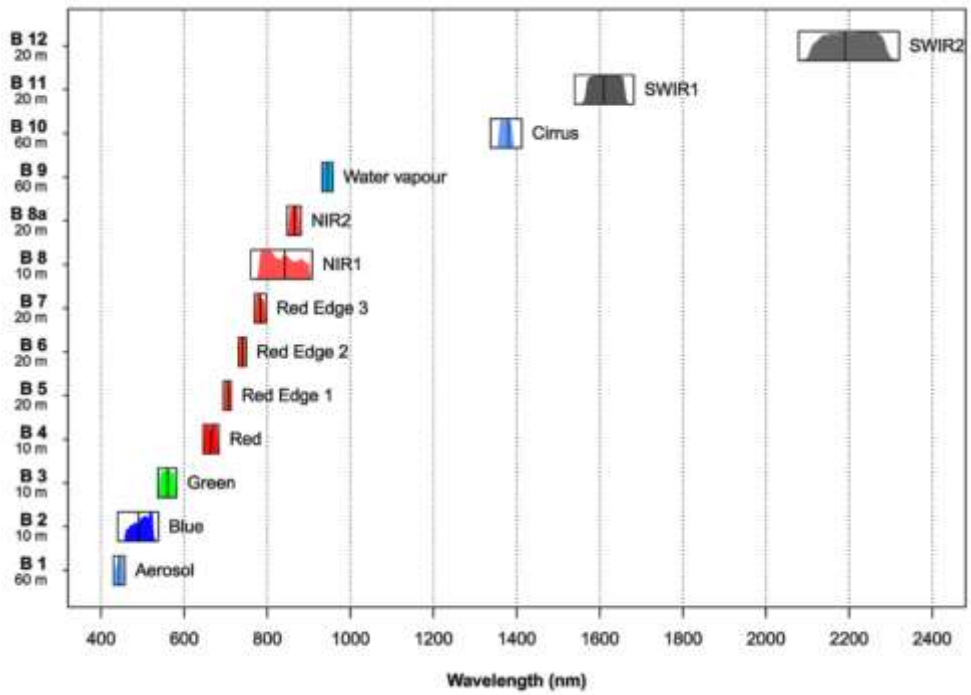

Рис. 1. Важные полосы Sentinel-2 для прогнозирования урожайности кукурузы

После того, как каждая модель была обучена и использована для прогнозирования, была рассчитана важность всех функций, используемых для прогнозирования. Было обнаружено, что значимые полосы Sentinel-2, которые в большей степени способствовали точности модели, были в основном полосами с красным краем, особенно с красным краем 3 (полоса 7). Кроме того, коротковолновый инфракрасный диапазон (диапазон 10), также известный как Cirrus, и узкий ближний инфракрасный диапазон (диапазон 8A) были полезны для модели прогнозирования[3].

Значимые индексы вегетации для прогнозирования урожайности кукурузы

Я использовал различные спектральные индексы растительности (VI), которые обнаружил в репозитории пользовательских скриптов Sentinel. Однако из 40 различных VI я обнаружил, что 25 были значимы только в отношении прогнозирования урожайности кукурузы. В таблице 1 ниже представлен кандидат VI, используемый при оценке урожайности. Значение этих вегетационных индексов было выявлено в недавних исследованиях, посвященных спутниковой оценке урожайности и урожайности сельскохозяйственных культур. Однако мой обзор показал, что вегетационные индексы, тесно связанные с урожайностью, часто включали длины волн с красным краем или были разработаны так, чтобы быть чувствительными к содержанию хлорофилла в растительном покрове. Заметив это, я решил использовать функцию в скрипте, написанном победителями конкурса Crop Yield Prediction Challenge. Функция получает важные характеристики из трех красных полос изображений Sentinel-2[4].

Таблица 1 - Значимые индексы растительности для прогнозирования урожайности кукурузы

INDICES	FORMULA
NDVI	(NIR-Red)/(NIR+Red)
SAVI	((NIR - RED)/ (NIR + RED + L))*(1 + L)
ARVI	(NIR – (2 * Red) + Blue) / (NIR + (2 * Red) + Blue)
GCI	(NIR) / (Green) – 1
NDWI	(NIR-SWIR)/(NIR+SWIR)
RVI	(NIR/RED)
OSAVI	((1.0 + 0.16) * (NIR - Red) / (NIR + Red+ 0.16))
SLAVI	NIR / (Red+ SWIR2)))
NPCRI	(Red-Blue)/(Red+ Blue)
SCI	(SWIR1- NIR) / (SWIR1+ NIR)
SARVI	(1.0 + L) * (NIR - (Rr - y * (RB - Rr))) / (NIR + -(Rr - y * (RB - Rr)) + L)
NDMI	(NIR - SWIR1) / (NIR + SWIR1)
EVI_2	2.5 (NIR - RED)/(NIR + C2 * Blue + l)
GNDVI	(NIR - Green)/(NIR + Green)
NDRE1	(Red_Edge2-Red_Edge1)/(Red_Edge2 +Red_Edge1)
SeLI	(Near_infrared_narrow-Red_Edge1)/(Near_infrared_narrow+ Red_Edge1)
SAVI2	NIR / (Red+ b / a)
AFRI2100	NIR - 0.5 * SWIR2/ (NIR + 0.56 * SWIR2)
BSI	((SWIR1+ Red) - (NIR + Blue)) / ((SWIR1+ Red) + (NIR + Blue))
EVI	2.5 (NIR - RED)/(NIR + C1 * RED - C2 * Blue + l)
Red Edge NDVI 705	(NIR - Red_Edge1)/(NIR + Red_Edge1)
MTCI	(Red_Edge2-Red_Edge1)/(Red_Edge1-Red)
MSI	(SWIR)/(NIR)
BWDRVI	(0.1 * NIR - Blue) / (0.1 * NIR + Blue)
CCCI	((NIR- NIR - Red_Edge1) / (NIR+ Edge1)) / ((NIR- Red) / (NIR+ Red))

В предоставленном наборе данных климатические переменные были получены из Terraclimate. TerraClimate — это набор данных о ежемесячном климате и климатическом водном балансе земной поверхности мира за период с 1958 по 2019 год. Данные TerraClimate имеют месячное временное разрешение и пространственное разрешение ~4 км (1/24 градуса).

Данные о почве для статьи были получены из ISRIC World Soil Information. Каждый актив данных сетки почвы представляет собой изображение с 6 каналами, по каналу для каждой глубины (0–5 см, 5–15 см, 15–30 см, 30–60 см, 60–100 см, 100–200 см). Однако данные о почве для соревнований содержали только полосы глубиной от 5 до 15 см.

Экспериментальная часть

Моделирование

```
# import libraries
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler
import lightgbm as lgb
import sklearn
from lightgbm import LGBMRegressor
from sklearn.ensemble import RandomForestRegressor
import random
RANDOM_STATE = 42
import pprint
import seaborn as sns
pd.set_option('display.max_columns', None)
```

Рис. 2. Импорт библиотек

Функция сокращает использование памяти, преобразовывая тип данных каждого столбца к минимуму — например, с плавающей запятой 64 в число с плавающей запятой 16. Это связано с тем, что более высокий тип данных потребляет больше памяти по сравнению с более низкими типами данных, такими как float16.

```
link = 'C:/Crop Yield Prediction/'
train_df = pd.read_csv(link + 'Train.csv') # training dataframe
test_field_ids_years_df = pd.read_csv(link + 'test_field_ids_with_year.csv') # years for the test fields
add_info_df = pd.read_csv(link + 'fields_w_additional_info.csv') # additional soil and climate information
sample_sub_df = pd.read_csv(link + 'SampleSubmission.csv')
```

Рис. 3. Загрузка наборов данных

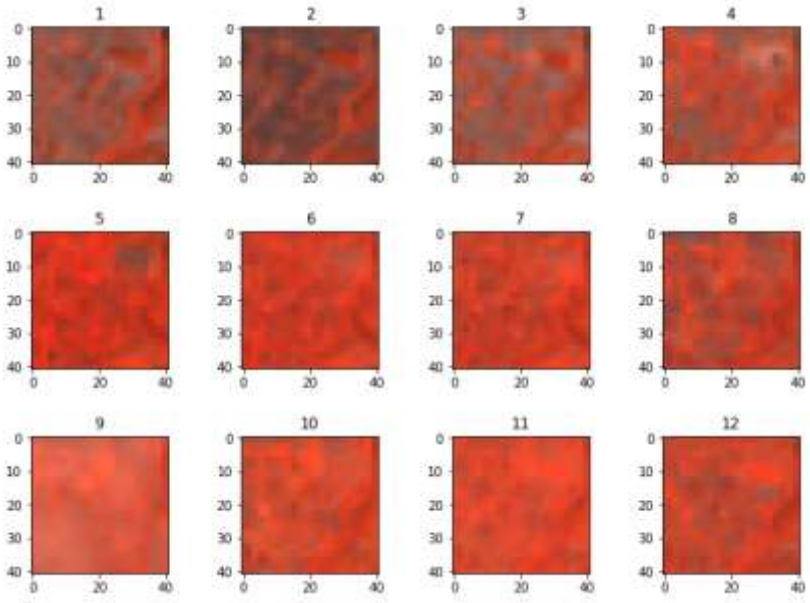

Рис. 4. Просмотр изображений в искусственных цветах в течение 12 месяцев

Предложение по разработке функций:

1. Указать полосы интереса и средние центральные точки
2. Средние климатические значения за 4 года по месяцам сезона кукурузы для некоторых климатических переменных
3. Создать статистику для индексов растительности, охватывающую весь снимок Sentinel-2.
4. Использование важных индексов растительности
5. Создать статистики из индексов растительности (макс., мин., медиана)
6. Создайте медианные функции из красных полос

Рис. 5. Разработка признаков с использованием индексов

Обработка и подготовка — данные и характеристики:

1. Применение функции «Указать полосы интереса» и «Средние центральные точки».
2. Применение функцию уменьшения использования памяти
3. Применение функцию «Средние климатические значения за 4 года».
4. Применение статистику для индексов, охватывающих всю функцию изображения Sentinel-2.
5. Применение функцию индексов растительности
6. Применение функцию «Статистика из индексов растительности».
7. Применение функцию Срединные характеристики красных полос.

Из-за математических операций, выполняемых с данными, могут возникать пропущенные значения. Чтобы обнаружить, присутствуют ли пропущенные значения в данных — пропущенные значения были проверены на наличие [5].

Обучение и тестирование

Был реализован алгоритм XGBoost. XGBoost — это алгоритм, который в последнее время доминирует в прикладном машинном обучении для структу-

рированных или табличных данных. XGBoost — это реализация деревьев решений с градиентным усилением, разработанных для обеспечения скорости и производительности. Говоря о скорости выполнения, XGBoost быстр. Действительно быстро по сравнению с другими реализациями повышения градиента. XGBoost доминирует над структурированными или табличными наборами данных в задачах моделирования классификации и регрессионного прогнозирования.

```
from sklearn.model_selection import KFold
import xgboost as xgb
from sklearn.metrics import mean_squared_error


X, y  = train_sampled.drop(['Field_ID'],axis=1), train_df['Yield']


kf = KFold(n_splits =5,shuffle=True,random_state=160)
feats = pd.DataFrame({'features': X.columns})
gbm_predictions = []
cv_score_ = 0
oof_preds = np.zeros((train_df.shape[0],))


for i,(tr_index,test_index) in enumerate(kf.split(X,y)):
    print()
    print(f'######### FOLD {i+1} / {kf.n_splits} ')


X_train,y_train = X.iloc[tr_index,:],y[tr_index]
    X_test,y_test = X.iloc[test_index,:],y[test_index]

    gbm = xgb.XGBRegressor(eval_metric = 'rmse',n_estimators =
2000,learning_rate = 0.001,seed=162,random_state =
162,colsample_bytree=0.65)


gbm.fit(X_train,y_train,eval_set = [(X_test,
y_test)],early_stopping_rounds = 200,verbose=100)
```

Рис. 6. Алгоритм XGBoost

Могут быть реализованы другие деревья решений с градиентным усилением, такие как LightGBM. Однако для дальнейшего повышения точности модели прогнозирования урожайности необходимо использовать дополнительные приемы для создания более важных функций, а настройка гиперпараметров модели также добавит возможности прогнозирования модели.

```
model = lgb.LGBMRegressor(n_estimators=1000)
model.fit(X_train, y_train,
        eval_set=[(X_test, y_test)],
          early_stopping_rounds=10)
# Score with RMSE
print(" ")
print('Score:', mean_squared_error(y_test, model.predict(X_test), squared=False))
```

Рис. 7. Алгоритм LightGBM

Заключение

Мы предлагаем прозрачную и переводимую систему прогнозирования урожайности на основе спутниковых изображений Sentinel-2, чтобы удовлетворить насущную потребность адаптироваться к растущей нестабильности в глобальном поставках продовольствия. Мы создаем собственную систему маркировки для интерпретируемых результатов классификации и проводим первоначальную пробную фазу классификации системы. Наша методология основана на предыдущей работе по классификации изображений и прогнозированию урожайности путем классификации нескольких глобальных основных сельскохозяйственных культур, а также включения слоя актуальных атмосферных данных, необходимых для точного прогнозирования урожайности. Первоначальные результаты классификации, которые мы получаем, сопоставимы с различными методами ансамбля, проверенными на спутниковых снимках Sentinel-2, а наши результаты перекрестной проверки показывают еще большую точность в четырех из пяти проверенных мест, а также указывают на потенциальные улучшения сбора данных. Что наиболее важно, процедура классификации и уровень дополнительных данных могут быть применены к практическим вариантам использования.

Литература

1. ESA Earth Observation Portal. Available at: https://directory.eoportal.org/web/eoportal/ satellite-missions/c-missions/copernicus-sentinel-2/ (accessed 30. 06. 2021)

2. Copernicus Open Access Hub. Available at: https://scihub.copernicus.eu/dhus/ (accessed 02. 07. 2021)

3. В.М.Гришкин, С.И.Каримов. / Сравнение данных мультиресурсного дистанционного зондирования для вегетационных индексов / Advanced Information Technologies and Scientific Computing. (PIT 2021). 14-с.

4. В.М.Гришкин, С.И.Каримов. / Models and methods of data processing remote sensing / The American journal of Engineering and technology. ISSN 2689-0984. Volume 3. 2021.

5. В.М.Гришкин, С.И.Каримов. / Общее описание приема и изучения данных, поступающих через спутник / Труды III Международной научно-практической конференции, посвященной 90-летию Брянского государственного инженерно-технологического университета «ЦИФРОВОЙ РЕГИОН: ОПЫТ, КОМПЕТЕНЦИИ, ПРОЕКТЫ» Брянск. 2020. 1044-с.

6. Grishkin V.M., Karimov S.I. / Use of satellite imagery and index control to monitor and analyze the agricultural lands of Bukhara region, which is a world historical heritage / 1st International Conference on Problems and Perspectives of Modern Science(ICPPMS-2021).