Noname manuscript No.
(will be inserted by the editor)

Report: GEO-AI Challenge for Cropland Mapping by ITU

Mohsen Ahmadkhani

Date: October 14, 2023

1 Data Preparation

The project addressed challenges associated with handling datasets spanning distinct regions and time frames. While data for Iran and Sudan (SUIR) were available from July 1, 2019, to June 30, 2020, the dataset for Afghanistan (AF) was confined to April 2022, creating complexities in structuring a uniform training dataset. To augment the dataset, an additional 250 data points for each class per country were manually integrated from the geojson.io website, culminating in a training set of 3000 entries with columns for ID, Lat, Lon, and Target.

A DBSCAN classifier was utilized to categorize data points into three classes, with class 0 assigned to AF and 1 or 2 to the other countries. This classification aided in partitioning the train and test datasets into distinct AF and SUIR data frames. Sentinel 2 bands (R, G, B, NIR) were procured and aggregated over 12 months for SUIR using a "sentineller" function, while

Mohsen Ahmadkhani
Geography Environment and Society
University of Minnesota
Minneapolis, MN
E-mail: ahmad178@umn.edu

for AF, a "sentineller_interval" function was employed to capture data in 12 distinct intervals in April 2022 due to the shorter time span. This approach standardized the datasets for SUIR and AF, featuring rows corresponding to data points and columns representing various attributes, including ID, Lat, Lon, Target, Label, and spectral bands.

To enhance machine learning applicability, the datasets were reshaped, featuring initial columns and multiple instances of R, G, B, NIR values, each tagged to different intervals. NDVI columns for these intervals across all countries were then computed, expanding the dataset. Understanding the relational dynamics of Lat and Lon, a "distancer" function was introduced, computing the haversine distance of each point to the dataset's extremities and conducting a PCA on these distances. This method introduced five new columns to the dataset.

The "landuser" function was deployed to acquire land use data from the 'COPERNICUS/Landcover/100m/Proba-V/Global' collection, offering detailed 2014 land composition data. Data on soil organic carbon content from the "OpenLandMap/SOL/SOL_ORGANIC-CARBON_USDA-6A1C_M/v02" dataset and topographical slope data from the 'USGS/SRTMGL1_003' dataset were also integrated, resulting in a comprehensive 3000x73 dataframe across all countries.

1.1 Addressing Data Completeness

A minor setback was encountered with five instances of missing 'soil_data' in the 3000-point training set. This was resolved by employing a "missing_filler" function that inferred these missing values from the nearest neighboring point, based on the autocorrelation assumption inherent in soil features.

2 Model Implementation

Through rigorous testing, the Random Forest model, with its 100 trees, emerged as the most promising in terms of test accuracy. However, an ensemble method was also constructed, incorporating decision-tree algorithms (Random Forest, XGBoost, CatBoost, and Gradient Boosting) alongside parametric algorithms (Ridge Classifier, SVM, and Logistic Regression), with the final classification determined by majority voting. It's crucial to note that data standardization was imperative for enhancing the efficacy of parametric methods.

3 Feature Optimization

Several feature selection strategies were explored, including backward multiple linear regression, Chi-Square Test, and LASSO. However, the most effective was an iterative feature elimination strategy, encapsulated in the "optimal_feature_selection" function. This methodology involved a stepwise removal of features, gauging the impact on Random Forest model accuracy, and reinstating any feature whose absence detrimentally affected performance.