A better way to format your document for CEUR-WS

Darren Rawlings[1], Tim Chopard

[1]University of Groningen, Broerstraat 5, 9712 CP Groningen, Netherlands

Abstract

A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the “ceurart” document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

Keywords

Multi-Label classification, Principal component analysis, ResNet, Vision Transformers, XGBoost, GeoLifeCLEF 2024, CEUR-WS

1. Introduction

In this research study, we aim to develop a model for predicting plant species in a specific location and time using various environmental factors as predictors. These predictors include satellite images, climatic time series, and other rasterized environmental data such as land cover, human footprint, bioclimatic variables, and soil characteristics. Our motivation behind this challenge is the potential usefulness of accurate plant species prediction in various scenarios related to biodiversity management and conservation, species identification and inventory tools, and education.

We utilized a large-scale training dataset of approximately 5 million plant occurrences in Europe, as well as validation and test sets with over 5,000 and 20,000 plots, respectively. The predicted output will be multi-label, presence-absence data for all present species at each plot. The data covered over 10,000 different plant species, which created significant challendges associated with this task, including learning from single positive labels, dealing with strong class imbalance, multi-modal learning, and handling large-scale datasets.

The potential applications of accurate plant species prediction are numerous. High-resolution maps of species composition and related biodiversity indicators can be created to aid in scientific ecology studies and conservation efforts. The accuracy of species identification tools can be improved by reducing the list of candidate species observable at a given site. Additionally, location-based recommendation services and educational applications with features such as quests or contextualized educational pathways can be developed to facilitate biodiversity inventories and promote environmental education. We believe that our research will contribute to the advancement of plant species prediction and its practical applications in various fields.

The research was conducted as part of the GeoLifeCLEF 2024 competition on Kaggle [**?**], which is a part of the LifeCLEF initiative. The competition aims to develop models for predicting

Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY

[†] These authors contributed equally.

plant species in a specific location and time using various environmental factors as predictors.

2. Background

The GeoLifeCLEF challenge has been running for a number of years. Each year, participants are tasked with predicting species distribution, but the challenge has evolved over time, with new datasets, evaluation metrics, and research questions introduced each year. Here, we provide an overview of the some of the recent submissions to GeoLifeCLEF challenge and summarize the key contributions.

In 2021, the GeoLifeCLEF challenge focused on fine-grained visual categorization using remote sensing data. The winning submission by [?] leveraged contrastive learning to improve species distribution modeling (SDM) from remote sensing imagery. The authors explored the effectiveness of using only RGB imagery and the impact of adding altitude imagery to the model's performance. They introduced a new consistency-based model selection metric to enhance the model's generalization capabilities. The paper outlined potential areas for further research, including the impact of transformations and the utility of the consistency metric.

In 2022, the GeoLifeCLEF challenge shifted its focus to predicting species distribution across the U.S. and France using remote sensing data and other covariates. The second-place submission by [?] proposed a classification approach with a spatial block-label swap regularization during training and an ensemble of deep learning models. Their method achieved a top-30 accuracy of 31.22% on the private test set, securing second place in the competition. The authors reflected on the results and suggested potential improvements and the importance of species distribution modeling for ecological research.

In 2023, the GeoLifeCLEF challenge introduced a new dataset with single positive labels for each location, making multi-label prediction challenging. The winning submission by [?] proposed a three-step training strategy to leverage the single positive labels effectively. The authors introduced several CNN-based models and demonstrated their effectiveness compared to a simple baseline. The paper discussed the challenges of the new dataset and the proposed models' performance, providing detailed results and comparisons.

3. Data

The training data comprises species observations and environmental data. Below, we explain the data in detail.

3.1. Observations data

The species related training data comprises:

Presence-Absence (PA) surveys: including around 90 thousand surveys with roughly 10,000 species of the European flora. The presence-absence data (PA) is provided to compensate for the problem of false-absences of PO data and calibrate models to avoid associated biases.

Presence-Only (PO) occurrences: combines around five million observations from numerous datasets gathered from the Global Biodiversity Information Facility (GBIF, www.gbif.org). This data constitutes the larger piece of the training data and covers all countries of our study area, but it has been sampled opportunistically (without standardized sampling protocol), leading to various sampling biases. The local absence of a species among PO data doesn't mean it is truly absent. An observer might not have reported it because it was difficult to "see" it at this time of the year, to identify it as not a monitoring target, or just unattractive.

3.2. Environmental data

Besides species data, we provide spatialized geographic and environmental data as additional input variables (see Figure 1). More precisely, For each species observation location, we provide:

3.2.1. Satellite image patches

Satellite image patches: 3-band (RGB) and 1-band (NIR) 128x128 JPEG images, a color JPEG file for RGB data and a grayscale one for Near-Infrared images at 10m resolution. The source for these images is Sentinel2 remote sensing data pre-processed by the Ecodatacube platform.

3.2.2. Satellite time series

Satellite time series: Up to 20 years of values for six satellite bands (R, G, B, NIR, SWIR1, and SWIR2). Each observation is associated with the time series of the satellite median point values over each season since the winter of 1999 for six satellite bands (R, G, B, NIR, SWIR1, and SWIR2). This data carries a high-resolution local signature of the past 20 years' succession of seasonal vegetation changes, potential extreme natural events (fires), or land use changes. The original satellite data has a resolution of 30m per pixel. The source for this is the Landsat remote sensing data pre-processed by the Ecodatacube platform

3.2.3. Environmental rasters

Environmental rasters Various climatic, pedologic, land use, and human footprint variables at the European scale. We provide scalar values, time-series, and original rasters from which you may extract local 2D images.

Four climatic variables computed monthly (mean, minimum and maximum temperature, and total precipitation) from January 2000 to December 2019, yielding 960 low-resolution (30 arcsec 1 kilometer) rasters covering Europe. The source for these rasters is the CHELSA climate dataset.

Environmental rasters, for each observation, we were provided additional environmental data such as GeoTIFF rasters and scalar values already extracted from the rasters. We provide CSV files, one per band raster type, i.e., Climate, Elevation, Human Footprint, LandCover, and SoilGrids.

1. Bioclimatic rasters: 19 low-resolution rasters covering Europe; commonly used in species distribution modeling. Provided in longitude/latitude coordinates (WGS84). These were provided as GeoTIFF files with compression and CSV file with extracted values, with a resolution of 30 arcsec (1 kilometer). The source for these rasters is the CHELSA climate dataset.

2. Soil rasters: Nine pedologic low-resolution rasters covering Europe. Provided variables describe the soil properties from 5 to 15cm depth and are determinant of plant species distributions. Check the definition.txt file about the provided variables (e.g., pH, clay, organic carbon and nitrogen contents, etc.). The format is GeoTIFF files with compression and CSV file with extracted values, with a resolution of 1 kilometer. The source for these rasters is Soilgrids.
3. Elevation: High-resolution raster covering Europe. Provided as a GeoTIFF file and CSV file with extracted values, with a resolution of 1 arc second (30 meters). The source for this raster is the ASTER Global Digital Elevation Model V3.
4. Land Cover: A medium-resolution multi-band land cover raster covering Europe. Each band describes either the land cover class prediction or its confidence under various classifications. We recommend the use of IGBP (17 classes) or LCCS (43 classes) layers, often used in species distribution modeling. The format is GeoTIFF file with compression and CSV file with extracted values, with a resolution of 500 meters. The source for this raster is MODIS Terra+Aqua 500m.
5. Human footprint: Several low-resolution rasters describing human footprint, encapsulating seven pressures on the environment (e.g., nighlight level, population density) induced by human presence and activity, are provided for two time periods, the early 90's (1993) and late 2000' (2009). We provide two summary rasters combining all human pressures and two detailed rasters per pressure, which avoid an arbitrary degradation of the original data. The format is GeoTIFF files with compression and CSV file with extracted values, with a resolution of 1 kilometer. The source for these rasters is [?].

4. Method

In order to manage the multilabel classification required for this research we implemented an ensemble approach. This section details the individual architectures and the methods used to combine them.

4.1. Architectures

The core architectures used in this project were 18 layer ResNet, XGBoost and [TRANSFORMER]. The outputs of these were then weighted and combined before the maximum arguments were selected.

4.1.1. Principal Component Analysis

Principal Component Analysis (PCA) is a method for reducing dimensionality [?]. It functions by projecting the high dimensional data onto the direction of maximum variance, thus retaining key features and reducing noise.

4.1.2. XGBoost

XGBoost is an open source gradient tree boosting package [?]. For this research we used the xgbregression model It has shown broad success accross a range of tasks, performing on par with or better than most equivalent and Automated Machine learning approaches [?].

4.1.3. ResNet

ResNet (Residual Network) is an architecture for deep neural networks that speeds up the training process through the use of residual connections [?]. For the purposes of this research we focused on ResNet18, the 18 layer deep variation with some modifications to accomodate the shape of the input data and the required output.

4.1.4. Transformer

Vision transformers (ViTs) have emerged as a novel approach to image classification, often outperforming traditional convolutional neural networks (CNNs) by leveraging self-attention mechanisms originally developed for machine translation [?]. Transformer now underpin some of the most powerful Large Language models [?]. By tokenizing images into fixed-size patches and encoding them into embeddings for transformer layers to process, ViTs excel at capturing long-range dependencies within an image, enabling a more holistic interpretation of visual context compared to CNN's local focus [?]. This methodological shift allows the models to effectively decipher complex scenes and interactions that are typically challenging for traditional architectures.

Often ViT-based models are often developed for large high resolution imagery. This was not the case for this application where the satellite images were a modest 128x128 pixels. We believed that the ability of transformer based models to capture to intricate regional relationships. Would be beneficial to the task of species classification.

4.2. Process

4.2.1. Preprocessing

The raw data included some missing, or infinite values that needed to be processed prior to model fitting. When a column was both deemed important to model fitting and contained such values, these values were replaced by the mean so as to avoid excessive influence from outliers without overly effecting the shape of the data [?].

TODO: Further explanation when models finalized

The Metadata and Bioclimatic Rasters were preprocessed for use in the count and XGBoost models in order to both reduce dimensionality and remove potential noise.

In addition, the number of viable output species was reduced from 11255 (the total species present accross all data) down to 1141 (the total species with more than 100 occurrences in the presence/absence data). This reduction served to remove several edge cases and focus the models more on likelier species.

4.2.2. Main Process

There were two key steps to the process of species selection. First the core models which generated pseudo-probabilities and were then weighted and combined. Second separeate models

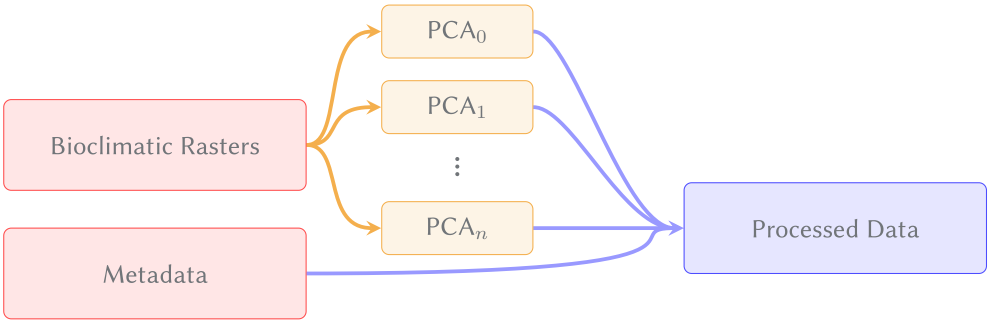

Figure 1: Preprocessing the data for use in the count prediction and XGBoost pseudo-probability models

were used to determine the expected number (N_{species}) of species per survey. These steps came together with the expected counts being used to select the top N_{species} psuedo-probabilities per survey.

The main scoring metric used was the micro-averaged F1 score as shown in equation **??** which is calculated from the precision $\frac{\text{FP}}{\text{FP+FP}}$ and the recall $\frac{\text{FP}}{\text{FP+FN}}$ for each individual class i. Where FP is the true positives, FP is the false positives and FN is the false negatives.

$$\text{F1}_{micro} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \cdot \text{FP}_i}{2 \cdot \text{FP}_i + \text{FP}_i + \text{FN}_i} \tag{1}$$