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.

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

[†]These authors contributed equally.

abc@def.ghi (D. Rawlings); timchopard@pm.me (T. Chopard)

https://startung.github.io/ (D. Rawlings); http://cloudberries.io (T. Chopard)

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

2. 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.

2.1. 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.

2.1.1. Principal Component Analysis

Principal Component Analysis (PCA) is a method for reducing dimensionality [1]. It functions by projecting the high dimensional data onto the direction of maximum variance, thus retaining key features and reducing noise.

2.1.2. XGBoost

XGBoost is an open source gradient tree boosting package [2]. 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 [3].

2.1.3. ResNet

ResNet (Residual Network) is an architecture for deep neural networks that speeds up the training process through the use of residual connections [4]. 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.

2.1.4. [Transformer]

2.2. Process

2.2.1. Preprocessing

2.2.2. 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 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 1 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.

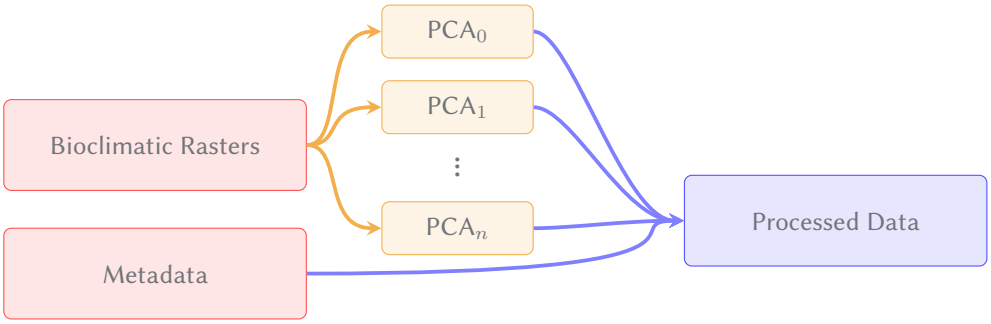

Figure 1: Preprocessing the data for use in the count prediction and XGBoost pseudo-probability models

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

References

[1] L. KPFRS, On lines and planes of closest fit to systems of points in space, in: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (SIGMOD), 1901, p. 19.

[2] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794. URL: http://doi.acm.org/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.

[3] L. Ferreira, A. Pilastri, C. M. Martins, P. M. Pires, P. Cortez, A comparison of automl tools for machine learning, deep learning and xgboost, in: 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1–8. doi:10.1109/IJCNN52387.2021.9534091.

[4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385. arXiv:1512.03385.