A simple and quick approach to estimate soil parameters from hyperspectral images

Matheus Yasuo Ribeiro Utino

Student at University of São Paulo, São Carlos, Brazil

matheusutino@gmail.com

Abstract - Method using machine learning techniques to estimate the quantity of minerals and soil pH through hyperspectral images. In this approach, data preprocessing will be carried out, and a semi-supervised method will be introduced to extract information, including from unlabeled data. Additionally, machine learning algorithms have been optimized through a Bayesian approach to achieve good solutions within a reasonable time frame. Finally, after these steps, it was possible to estimate the desired parameters with relative accuracy.

Keywords - Machine Learning; semi-supervised method; hyperspectral images

1 Introduction

This paper refers to the challenge proposed by the International Telecommunication Union (ITU), titled "GeoAI Challenge Estimating Soil Parameters from Hyperspectral Images by ITU" [1]. The objective of the competition is to estimate the quantity of potassium (K), phosphorus pentoxide (P_2O_5), magnesium (Mg), and pH present in the soil from hyperspectral images captured in agricultural areas in Poland.

This area of study is vital for various distinct purposes. Among them, one can highlight environmental preservation, as it would be possible to detect inconsistencies and outliers. Additionally, it is also ideal for finding soils that best suit cultivation. Thus, this would represent a reduction in time both to search for ideal fertile areas and to contribute to environmental preservation.

To estimate the project parameters, machine learning techniques will be employed to automate this process.

2 Dataset

The dataset consists of 1732 training cases with available labels and 1154 elements for the test set. Figure 1 provides a visualization of the hyperspectral images for a patch.

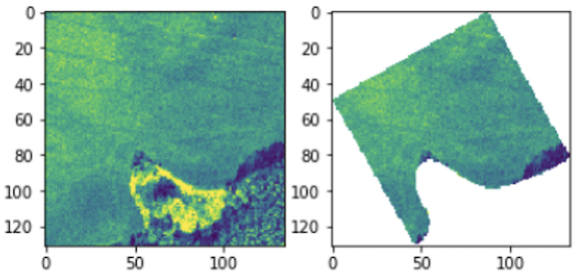

Figure 1. Hyperspectral image [1].

3 Methodology

The methodology used is quite simple and can be illustrated by Figure 2.

It is noteworthy that the method used represents a semi-supervised approach. In this methodology, a machine learning model is initially trained with known labeled data. Afterward, the model predicts labels for unlabeled data, in this case, the available test set. Finally, these newly labeled data from the model are then incorporated into the training data, increasing the amount of data available for the model to learn.

4 Preprocessing

The only preprocessing applied to the dataset was the normalization of data within the range of 0 to 1. This was done to ensure equivalent weights among the features.

5 Regressor

The machine learning model used for the regression task was the classic and well-established K-Nearest Neighbors (KNN).

KNN is a simple and easily understandable algorithm. Its operation is based on the distance between features, where cases with similar data tend to be close to each other.

In this way, KNN compares an instance that needs prediction with all other labeled examples. It then selects the K closest elements and performs an operation to calculate

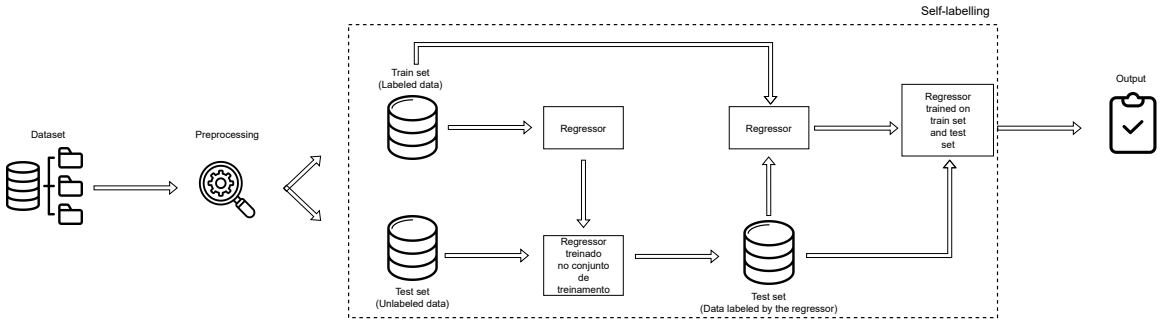

Figure 2. Methodology.

the value of the unlabeled data, such as using the average of the K neighbors.

6 Optimization

The KNN primarily has three vital parameters that must be optimized to achieve better results: the number of neighbors, the weight used in predictions, and the distance metric.

Regarding the weight in the model's predictions, a fixed approach was used, where the weight is the inverse of the obtained distance. Thus, closer data has a greater influence compared to more distant points.

As for the number of neighbors and the distance metric used, a Bayesian approach was employed for optimization in a smaller number of steps, as there are numerous possible combinations, and testing all manually would be impractical.

To achieve this, the optuna library was used, where the number of neighbors was varied between 3 and 600. The tested distance metrics included cosine, l1, l2, and chebyshev. Additionally, the number of attempts during optimization was set to 400 for each model.

The results of the best parameters obtained for each model, considering the 5-fold cross-validation technique, are presented in Table 1.

Table 1. Best parameters obtained by Bayesian optimization.

Target	Number of Neighbors	Distance Metric
K	224	cosseno
P_2O_5	55	cosseno
Mg	49	cosseno
pH	80	chebyshev

7 Execution Time

The code was executed on Kaggle, where the available hardware consists of an Intel Xeon 2.2GHz processor with 2 cores. The times for preprocessing, training and inference are shown in Table 2, considering the mean of 100 executions.

Table 2. Execution Time.

Preprocessing	Training and Inference
3.264 ms	1.239 s

8 Metric

As a metric for comparing models, especially during optimization, the Root Mean Squared Error (RMSE) was used, represented by Equation 1.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

where $\hat{y}_i$ represents the predicted values by the machine learning model, y_i represents the actual values and n corresponds to the number of observations.

Low RMSE values indicate better results; thus, the goal of the competition is to minimize this metric as much as possible.

9 Results

The self-labeling technique achieved a result of 0.277308199 for the public dataset and 0.278582187 for the private dataset. This outcome secured the 15th position in the final competition ranking for this approach.

10 Conclusão

It is evident that this simple approach using machine learning techniques is capable of delivering satisfactory results, proving vital for the search for ideal agricultural areas and environmental preservation.

Furthermore, this methodology stands out for being extremely lightweight and not requiring the use of GPUs for model training and inference, making it more accessible to all who wish to adopt this approach.

References

[1] Zindi. Geoai challenge estimating soil parameters from hyperspectral images by itu. On-line: https://zindi.africa/competitions/geoai-challenge-estimating-soil-parameters-from-hyperspectral-images, Accessed on: 26/11/2023.