Uma abordagem simples e rápida para estimar parâmetros do solo a partir de imagens hiperespectrais

Matheus Yasuo Ribeiro Utino

Universidade de São Paulo, São Carlos, Brasil

matheusutino@gmail.com

Abstract - Método utilizando técnicas de aprendizado de máquina para estimar a quantidade de minerais e o pH do solo através de imagens hiperespectrais. Nessa abordagem será realizado o pré-processamento dos dados e introduzido um método semissupervionado para extrair informações inclusive dos dados não rotulados. Além disso, os algoritmos de aprendizado de máquina foram otimizados através de uma abordagem bayesiana visando obter boas soluções em um tempo hábil. Por fim, após essas etapas foi possível estimar os parâmetros desejados bom relativa precisão.

Keywords - Machine Learning; método semissupervisionado; imagens hiperespectrais

1 Introdução

Esse paper refere-se ao desafio proposto pela International Telecommunication Union (ITU), ititulado "GeoAI Challenge Estimating Soil Parameters from Hyperspectral Images by ITU" [1]. O objetivo da competição é estimar a quantidade de potássio (K), pentóxido de fósforo (P_2O_5), magnésio (Mg) e pH presente no solo a partir de imagens hiperespectrais capturadas sob áreas agrícolas presente na Polônia.

Essa área de estudo é vital para diversas finalidades distintas. Entre elas pode-se destacar a preservação do meio ambiente, pois seria possível detectar inconsistências e valores atípicos. Além disso, também é ideal para encontrar solos que melhor se adéquam ao plantio. Dessa forma, isso representaria uma redução no tempo tanto para procurar áreas férteis ideias e também para ajudar na preservação do meio ambiente.

Para realizar a estimação dos parâmetros do projeto serão utilizar técnicas de aprendizado de máquina para automatizar esse processo.

2 Conjunto de dados

O conjunto de dados é composto por 1732 casos de treinamento, em que os rótulos estão presentes, e 1154 elementos para o conjunto de teste. Na figura 1, encontra-se uma visualização das imagens hiperespectrais para um patch.

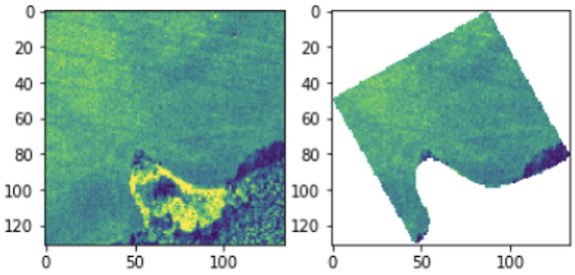

Figure 1. Imagem hiperespectral [1].

3 Metodologia

A metodologia utilizada é bem simples e pode ser ilustrada pela figura 2.

Nota-se que o método utilizado representa uma abordagem semissupervisionada. Nessa metodologia um modelo de aprendizado de máquina é inicialmente treinado com os dados rotulados conhecidos. Após isso, o modelo então prevê o label para dados não rotulados, nesse caso foi utilizado o conjunto de teste disponível. Por fim, esses novos dados rotulados pelo modelo são então incorporados aos dados de treinamento, aumentando a quantidade de dados disponível para o modelo aprender.

4 Pré-processamento

O único pré-processamento utilizado no conjunto de dados foi a normalização dos dados no intervalo entre 0 e 1. Isso foi feito para que os pesos entre as features sejam equivalentes.

5 Regressor

O modelo de aprendizado de máquina utilizado para a tarefa de regressão foi o clássico e bem consolidado K-Nearest Neighbors (KNN).

O KNN é um algoritmo simples e de fácil compreensão. Seu funcionamento se baseia na distância entre as features, em que casos com dados similares tem a tendência de estarem próximos.

Dessa forma, o KNN compara uma instância, que deseja-se prever, com todos os outros exemplos já rotu-

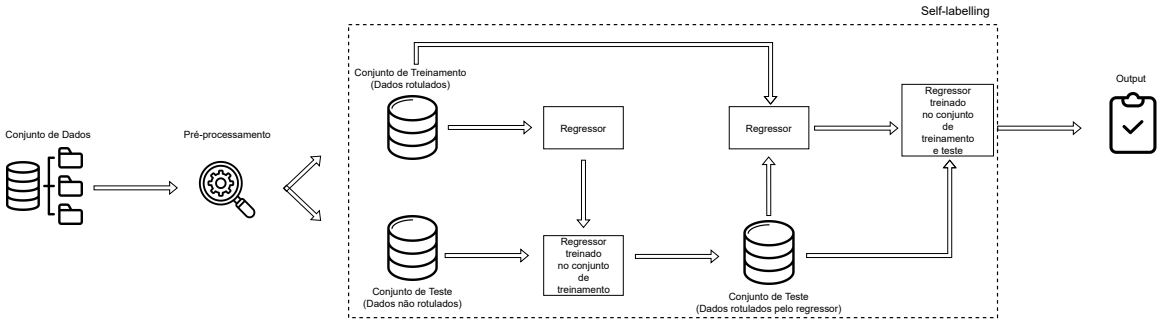

Figure 2. Metodologia.

lados, então ele seleciona os K elementos mais próximos e realiza uma operação para calcular o valor do dado não rotulado, como utilizar a média dos K vizinhos.

6 Otimização

O KNN apresenta primordialmente três parâmetros vitais que devem ser otimizados para obter melhores resultados: o número de vizinhos, o peso utilizado nas predições e a métrica de distância.

Quanto ao peso nas predições do modelo foi utilizado de forma fixa, em que o peso seria o inverso da distância obtida, com isso dados mais próximos tem maior influência em relação a pontos mais distantes.

Já em relação ao número de vizinhos e a métrica de distância utilizada, foi utilizado uma abordagem bayesiana para uma otimização em um número menor de passo, pois existem diversas combinações possíveis e seria inviável testar todas manualmente.

Com isso, foi utilizado a biblioteca optuna, em que o número de vizinhos foi variado entre 3 e 600. E as métricas de distância testadas foram: cosseno, l1, l2 e chebyshev. Além disso, o número de tentativas durante a otimização foi de 400 para cada modelo.

O resultado dos melhores parâmetros obtidos para cada modelo, considerando a técnica de cross validation com 5 folds, estão presentes na tabela 1.

Table 1. Melhores parâmetros obtidos por otimização bayesiana.

Alvo	Número de vizinhos	Métrica de distância
K	224	cosseno
P_2O_5	55	cosseno
Mg	49	cosseno
pH	80	chebyshev

7 Tempo de execução

O código foi executado no Kaggle, em que o hardware disponível é composto por um Intel Xeon 2.2GHz utilizando 2 cores. O tempo de realizar o pré-processamento; treinamento e inferência estão presentes na tabela 2, considerando 100 execuções.

Table 2. Tempo de execução.

Pré-processamento	Treinamento e inferência
3.264 ms	1.239 s

8 Métrica

Como métrica de comparação entre os modelos, principalmente durante a otimização, foi utilizado a Root Mean Squared Error (RMSE) representa pela equação 1.

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

em que $\hat{y}_i$ são os valores preditos pelo modelo de machine learning, y_i são os valores reais e n corresponde ao número de observações.

Valor baixos de RMSE representam melhores resultados, dessa forma o objetivo da competição é reduzir ao máximo essa métrica.

9 Resultados

A técnica de self-labelling obteve um resultado para o conjunto de dados públicos de 0.277308199, enquanto para o conjunto privado de 0.278582187. Esse resultado garantiu que essa abordagem atingisse a 15ª colocação no ranking final da competição.

10 Conclusão

Nota-se que essa abordagem simples usando técnicas de aprendizado de máquina é capaz de trazer resultados satisfatórios, sendo vital para a busca de áreas agrícolas ideias e na preservação do meio ambiente.

Além disso, essa metodologia destaca-se por ser extremamente leve e não necessitar do uso de GPUs para o

treinamento e inferência do modelo, o que torna ele mais acessível para todos os que desejam utilizar essa abordagem.

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, Acessado em: 26/11/2023.