Novo Método de Análise de Dados de Poucas Amostras de Solo Baseado em Deep Learning

Gabriel Tonon Cimatti - Universidade Estadual de Ponta Grossa - UEPG – 22001926@uepg.br
Gabriel Passos de Jesus - Universidade Estadual de Ponta Grossa - UEPG – gabrielpasje@gmail.com
Alaine Margarete Guimarães - Universidade Estadual de Ponta Grossa - UEPG - alainemg@uepg.br

Resumo: Uma parte essencial para produtividade na agricultura de precisão é a saúde do solo. O carbono orgânico total é um grande indicador dessa saúde e desenvolver um método rápido e acessível para estimar esse indicador é de extrema importância. Esse trabalho tem como objetivo desenvolver um método para a predição do carbono orgânico total no solo utilizando dados obtidos através da técnica de espectroscopia e métodos de machine learning, em específico, deep learning.

Palavras-chave: Agricultura de precisão, Deep Learning, aprendizado de máquina, análise de dados agrícolas.

New Method for Data Analysis of Few Soil Samples Based on Deep Learning

Abstract: *An essential part of productivity in precision agriculture is the soil health. The total organic carbon is a great indicator of this health and developing a quick and accessible method to estimate this indicator is of utmost importance. This article aims to develop a method for predicting total organic carbon in the soil using data obtained through spectroscopy techniques and machine learning models, specifically, deep learning.*

1. Introdução

Uma grande base de dados é um dos principais requisitos para a criação de um modelo computacional de predição baseado em abordagens de Inteligência Artificial, mais especificamente, de Aprendizado de Máquina (AM). Porém, nem sempre isso é possível por questões de dificuldades operacionais e de custos na geração dos dados. Isso é o que ocorre na Agricultura em que existe um custo considerável para obtenção de amostras, bem como um esforço operacional alto, como no caso de amostragem de solos.

Portanto, a determinação de modelos computacionais de predição baseados em AM deve considerar a existência de bases de dados consideradas pequenas, com menos de 300 instâncias. Ao mesmo tempo, uma base de dados mesmo com poucas instâncias pode apresentar muitos atributos (variáveis), levando ao problema da alta dimensionalidade nos dados. Assim criar um modelo computacional baseado em AM que envolta tanto poucas instâncias como alta dimensionalidade não é um processo trivial e demanda muita pesquisa para se chegar a resultados realmente satisfatórios.

Um caso, na agricultura, em que essa problemática está inserida (poucas instâncias e alta dimensionalidade) é na análise de amostras do solo por espectroscopia, que consiste em um método em que se busca correlacionar a resposta espectral de amostras do solo com seus componentes físico-químicos.

Por esses fatores citados esse trabalho tem como objetivo a predição do teor de carbono orgânico total (COT) no solo com base nos índices de absorbância, transmitância e refletância obtidos através de amostras do solo expostas a diferentes faixas de radiação infravermelho.

2. Material e Métodos

2.1. Coleta dos Dados

A base de dados foi composta por 252 amostras de solo obtidas de um experimento de longa duração realizado sob Sistema Plantio Direto (SPD). Este experimento teve início em 1989 no Centro de Experimentação para Assistência e Divulgação Técnica

Agropecuária da Fundação ABC, localizado no município de Ponta Grossa, PR (25º 00’ 35"S e 50º 09’ 16" W), a uma altitude de 890 metros.

A região apresenta clima subtropical úmido mesotérmico do tipo Cfb, de acordo com a classificação de Köppen-Geiger, caracterizado pela ausência de uma estação seca, invernos com geadas frequentes e verões de temperaturas amenas, conforme descrito por Marinho et al. (2022). As temperaturas médias variam entre 13°C nos meses mais frios e 24°C nos meses mais quentes. Segundo o IAPAR (2023), a precipitação média anual é de 1554 mm, e a umidade relativa média do ar é de 77%.

O solo da área experimental foi classificado como Latossolo Vermelho mesoférrico com textura argilosa (EMBRAPA, 2013), em um relevo de inclinação suavemente ondulada, com declividade em torno de 3%.

Dessas amostras foram estabelecidos os índices de absorbância, refletância e transmitância após serem expostas a diferentes comprimentos de onda de raios infravermelho. Isso foi feito com o uso de um espectrofotômetro NIRS (Near-Infrared Spectroscopy) para as faixas de infravermelho próximo (de 1100nm a 1648nm) e de um espectrômetro FTIR (Fourier Transform Infrared) para as faixas de infravermelho distante (de 2.500 nm a 25.026 nm), ambos disponibilizados pelo Complexo de Laboratórios Multiusuários (CLABMU) da UEPG.

Tabela 01 – Partição da tabela de refletância NIRS.

Índice	Amostra	Nível	TOC	R1100	R1102
1	23	F	6	0.194	0.194
2	4	D	11	0.192	0.192
3	17	F	5	0.191	0.191
4	26	B	17	0.196	0.196
5	4	C	14	0.185	0.185
6	26	C	10	0.179	0.179
7	4	E	12	0.185	0.185

Elaboração: Os autores

Como mostra na Tabela 01, as amostras variam seu local de coleta de 1 a 28 e o seu nível em relação ao solo de A até I. O atributo meta é o Teor de Carbono (TOC) e as variáveis são de r1100 a r1648 (somente para esta Tabela em específico). Na coluna r1100, o “r” indica que o índice em questão é a refletância e 1100 indica o comprimento de onda aplicado, 1100 nm.

Para os algoritmos desenvolvidos será utilizado somente a coluna TOC e as colunas de refletância em função do comprimento de onda (Tabela 1).

2.2 Expansão da Base de Dados

Foram aplicadas transformações na base de dados original para gerar novas bases. Isso feito para testar se transformações nos dados iriam melhorar ou facilitar a tarefa de predição. As transformações feitas foram a transformação Neperiana e a transformação polinomial de grau 4. Também foi utilizado a base sem nenhuma transformação.

2.3 Seleção de Atributos

O maior desafio dessa base de dados seria sua alta dimensionalidade, visto que cada faixa do espectro é uma variável e, somente a tabela de refletância NIRS já possui 275 faixas, portanto 275 atributos. O primeiro passo foi selecionar as variáveis mais importantes visando reduzir a dimensionalidade dos dados. Para isso foram utilizados diversos algoritmos de seleção de atributos, sendo eles: Kbest, melhor coeficiente de correlação de postos de Spearman, melhor correlação de Pearson, eliminação recursiva de atributos e algoritmo das projeções sucessivas.

O algoritmo Kbest é um algoritmo de seleção de atributos que escolhe os melhores ‘k’ atributos a depender de uma função de avaliação escolhida pelo usuário. Em nosso caso escolhemos as melhores 10 faixas espectrais e as funções de avaliação escolhidas foram a regressão por valor-F e a regressão de informação mútua.

O algoritmo de melhor correlação de Pearson basicamente mede por meio de métodos estatísticos a correlação entre as variáveis e o atributo meta e escolhe as variáveis que melhor pontuam nesse teste.

Semelhante ao anterior, o algoritmo de melhor coeficiente de correlação de postos de Spearman também compara a correlação entre duas variáveis, porém este não está limitado a testes com relações lineares, o que é o caso da correlação de Pearson.

O algoritmo de eliminação recursiva de atributos treina o modelo com uma função

específica usando um subconjunto dos dados que possui todos os atributos exceto um. Ele faz isso com todos os subconjuntos n-1 da base de dados até que se ache a variável que menos contribui para o modelo e a remove da base. O algoritmo repete esse processo até que se reste apenas o número desejado de atributos (dez atributos em nosso caso).

Por fim o algoritmo das projeções sucessivas interpreta os atributos como vetores em um espaço multidimensional e através de transformações matemáticas e projeções desses vetores ele consegue achar um subconjunto que possui o menor grau de correlação entre si, diminuindo assim a redundância do modelo.

Parte dos algoritmos para a seleção de atributos foram escolhidos com base nas pesquisas de Xueying, et al (2024) e Ugulino Araújo et al (2001).

2.4 Treinamento do Modelo

Todos os novos subconjuntos gerados foram alimentados algoritmos de deep learning e a métodos estatísticos convencionais. Os algoritmos escolhidos foram a regressão de mínimos quadrados parciais, a rede neural convolucional de uma dimensão (CNN-1D) e a rede neural de memória de curto longo prazo (LSTM).

O algoritmo de regressão de mínimos quadrados parciais foi escolhido pois ele é eficaz em casos onde existem um número maior de variáveis ao número de observações, o que era nosso caso.

A rede neural convolucional de uma dimensão é semelhante a rede convolucional de duas dimensões, porém esta é mais recomendada para dados bidimensionais com imagens enquanto a de uma dimensão funciona bem para dados sequenciais como faixas de áudio ou, em nosso caso, faixas do espectro luminoso.

Por fim, a rede neural de memória de curto longo prazo, é um tipo de rede neural recorrente e, semelhantemente a rede convolucional de uma dimensão, é recomendado seu uso em dados sequenciais.

Os algoritmos de deep learning (CNN-1D e LSTM) foram escolhidas para essa tarefa com base nas pesquisas de Li et al (2024) e Bai et al (2023).

3. Resultados e Discussões

A métrica utilizada para a medição da eficácia dos algoritmos foi a raiz quadrada do erro médio (RMSE).

Como esperado, o algoritmo de regressão por mínimos quadrados parciais se mostrou a pior solução para o problema, possivelmente por se tratar de um método estatístico e não uma rede neural. Seu erro RMSE variou entre 33 e 35 pontos.

Ao utilizar algoritmos de deep learning foi notável a diferença favorável. Utilizando o algoritmo de memória de curto longo prazo a métrica variou entre 9.2 e 9.3 pontos nos melhores casos. Com o algoritmo de rede neural convolucional a métrica ficou em 8.9 pontos.

4. Conclusões e Perspectivas de Pesquisas Futuras

Foi concluído que os algoritmos de deep learning se mostraram superiores ao método convencional estatístico na predição do COT e que a redução da dimensionalidade do problema através de algoritmos ajudou a melhorar a acurácia da rede final. Vale também ressaltar que as transformações feitas na base não tiveram uma resposta conclusiva, ajudando em alguns casos, porém piorando em outros.

Para pesquisas futuras seria interessante a aplicação da tanto da transmitância quanto da absorbância na rede neural. Também é possível a utilização de novos modelos de deep learning.

Agradecimentos

Os autores agradecem à CAPES - Comissão de Aperfeiçoamento de Pessoal de Nível Superior - e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, pelo financiamento da pesquisa por meio de bolsas de estudo.

Referências

ARAÚJO, M. C. U., et al. **The successive projections algorithm for variable selection in spectroscopic multicomponent analysis.** Chemometrics and Intelligent Laboratory Systems, 20 de janeiro de 2001. 57, 65-73.

BAI, Z. et al. **Estimation of soil inorganic carbon with visible near-infrared spectroscopy coupling of variable selection and deep learning in arid region of China.** Geoderma, 7 de julho de 2023. 437.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. Centro Nacional de Pesquisa de Solos. **Sistema Brasileiro de Classificação de Solos.** 3ª ed. Brasília- DF, 353p., 2013.

IAPAR. Instituto Agronômico do Paraná. **Médias históricas em estações do IAPAR.** Disponível em http://www.iapar.br/arquivos/Image/monitoramento/Medias_Historicas/Ponta_Grossa.html. Acesso em: 01/09/2023

LI, X. et al. **Optimizing soil carbon content prediction performance by multi-band feature fusion based on visible near-infrared spectroscopy.** Journal of Soils and Sediments, 17 de janeiro de 2024. 24, 1333-1347.

MARINHO, A. S. G. **Fatores e elementos climáticos usando a classificação de Köppen-Geiger**. UFAL, Maceió, 2022.