Quantification of soil properties from hyperspectral data for sustainable agriculture using deep learning

Singh S.[1,*], Kasana Singara S.[1]

[1]Thapar Institute of Engineering and Technology, Patiala, Punjab, INDIA

e-mail: simranjit_singh@thapar.edu

Abstract

The characterization of soil properties is critical for optimizing farming for sustainable agriculture. All the existing techniques for soil quantification do not take advantage of the sequential nature of Hyperspectral Data. This work focuses on proposing a Hybrid Framework that can quantitatively assess the soil properties from Hyperspectral data by extracting the essential features *via* Principal Component Analysis and Locality Preserving Projections. The extracted features are combined to form the Hybrid dataset which is then given as input to Long Short-Term Memory Networks, a deep learning-based framework which is typically used for sequential problems. The effectiveness of the Hybrid Framework is shown by comparing it with the existing regression models.

Keywords: Hyperspectral data, LSTM, PCA, LPP, Nutrients.

1. Introduction

Soil is an important natural resource. It contains various metals, nutrients, *etc*. It also absorbs toxic metals to keep the environment safe. The structural and physical composition of soil changes in close proximity (Blum et al. 1993). So, it is very important to know the present nutrients and other properties of soil to keep it healthy for sustainable agriculture. Reflectance spectroscopy allows rapid estimation of soil minerals by examining all the electromagnetic regions. Notably, the Very near Infrared region is highly studied for quick prediction of soil properties [Romero *et al.* 2018]

In this paper, a Hybrid Framework (HF) is proposed using the Land Use/Cover Area Frame Survey (LUCAS) dataset [Toth *et al.* 2013]to quantify the various soil properties. It has Hyperspectral Data (HSD) which contains a lot of information in the form of high dimensional data. So, LPP and PCA are used to reduce the dimensions for producing a hybrid dataset which contains the preserved global and local information from the original data and has less dimensions. Then, to take advantage of the sequential nature of spectrums, Long Short-Term Memory Networks (LSTM), a deep learning framework is used to build the prediction model.

2. Background Of The Proposed Work

2.1 Dimensionality reduction using PCA and LPP

Principal Component Analysis and Locality Preserving Projections are one of the most popular dimensionality reduction algorithms. They are widely used in the field of Hyperspectral data analysis as it generally comprises of high dimensions. These algorithms are used to efficiently reduce the dimensions without losing the essence of the data. PCA preserves the global information and LPP conserves the local information of the input high dimensional data.

PCA- The main task of PCA is to map the high dimensional input to lower dimensional PCA space. To calculate the principal components, the objective function is defined as:

$$P = argmax \sum_{i=1}^{n} P^T.C.P$$

Where P is the transformation vector, *argmax* is the argument maximum and C=$\frac{1}{n}\sum_{i=1}^{n}(D-\overline{D}).(D-\bar{D}\,)^T$ and D is the data point.

LPP-The main task of LPP is to preserve the locality information of the input data. LPP is based on adjacency graphs and eigenmaps. The adjacency graph is built by placing an edge between those nodes that are very close to each other. Eigenmaps contains the neighbourhood information. The objective function is defined as:

$$P = argmin \sum_{ij} P^T X L X^T P$$

Where L=D - S and $D_{ii} = \sum_j S_{ij}$, *argmin* is the argument minimum.

2.2 Long Short-Term Memory Networks

LSTM [Hochreiter *et al.* 1997] is a recurrent neural network which performs better on sequential problems as shown in Figure 1. The transitions functions are defined as follows:

Input gates: $l_t = \sigma(W_l.[O_{it-1}, x_{it} + b_i])$

Forget gates: $f_t = \sigma(W_f.[O_{it-1}, x_{it}] + b_f)$

New Candidates: $\hat{C}_t = tanh(W_C[O_{it-1}, x_{it}] + b_C)$

Cell States: $C_t = f_t \circ C_{t-1} + l_t \circ \hat{C}_t$

Output gate: $m_t = \sigma(W_{output}[O_{it-1}, x_{it}] + b_{output}])$
Next hidden state: $O_{it} = m_t \times tanh(C_t)$

Where σ is the sigmoid function, $tanh$ is the tangent hyperbolic function, W is the weight, $\circ$ is the Hadamard product, b is the bias, t is the present time state.

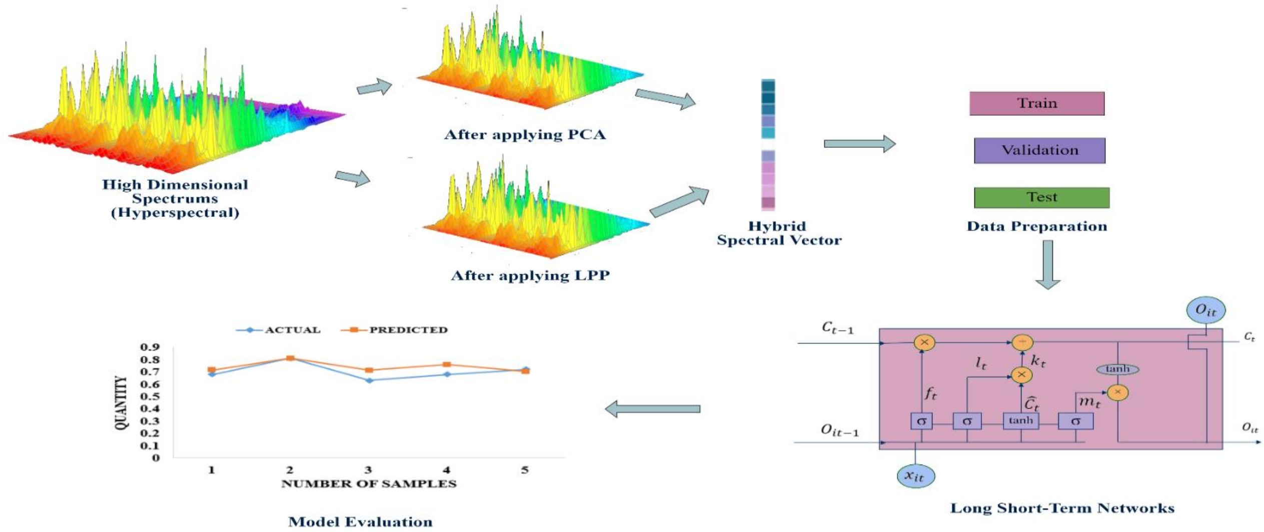

Figure 1. Hybrid Framework

3 Experimental Results And Analysis

Experiments are performed on the HSD obtained from LUCAS with 19036 samples using Hybrid framework shown in Figure 1. Firstly, the high dimensional HSD is passed through PCA and LPP to preserve global and local information. Then the outputs of both are combined to form the new hybrid spectrum which contains less dimensions than the LUCAS. Afterward, it is given to LSTM for training. LSTM is a deep learning framework which is suitable for the hybrid spectrums due to the sequential nature. LSTM uses 65% and 10% of the spectral data for training and validation respectively. In the end, the trained prediction model is obtained which predicts the corresponding quantity of soil properties on 25% testing data.

Table 1. Performance comparison of Hybrid Framework with other popular models.

Parameters		HF	PLSR	ANN	SVR	MLR
Clay	R^2	0.80	0.64	0.75	0.44	0.64
	RMSE	6.07	8.21	6.89	10.30	8.20
CEC	R^2	0.69	0.26	0.38	0.46	0.26
	RMSE	5.96	9.25	8.48	7.94	9.25
pH-CaCl2	R^2	0.88	0.77	0.76	0.76	0.77
	RMSE	0.54	0.74	0.75	0.74	0.74
CaCO3	R^2	0.84	0.71	0.73	0.11	0.71
	RMSE	60.46	81.81	79.03	160.23	81.55
N	R^2	0.83	0.44	0.71	0.68	0.45
	RMSE	1.64	2.97	2.14	2.27	2.96

The predicted results are noted and compared with the actual results for evaluation of the model. The evaluation is based on the co-efficient of determination (R^2) and Root Mean Square Error (RMSE). The performance of the Hybrid Framework (HF) is compared with the popular models like Partial Least Square Regression (PLSR), Artificial Neural Network (ANN), Support Vector Regression (SVR) and Multiple Linear Regression (MLR). It can be seen be seen from Table 1 that the parameters chosen for quantification are Clay, Cation Exchange Capacity (CEC), pH of Calcium Chloride ($CaCl_2$), Calcium Carbonate ($CaCO_3$) and Nitrogen (N). HF outperforms all the models and shows the least error of 0.54 and best R^2 of 0.88 in case of pH of $CaCl_2$.

4 Conclusion

In this paper, a novel Hybrid Framework to quantify the soil parameters is proposed. Firstly, the hyperspectral input is transformed to form hybrid spectral features with the help of LPP and PCA. The hybrid features are processed by LSTM for efficient learning. The trained model is used to predict results. There is at most 763% improvement in R^2 and at most 2.6 times less error in RMSE for all cases. This proves the effectiveness of using deep learning in spectral studies.

References

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Romero, D. J., Ben-Dor, E., Demattê, J. A., e Souza, A. B., Vicente, L. E., Tavares, T. R., ... & Gallo, B. C. (2018). Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. Geoderma, 312, 95-103.

Tóth, G., Jones, A., & Montanarella, L. (2013). The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environmental monitoring and assessment, 185(9), 7409-7425.

Blum, W. E. H. (1993). Soil protection concept of the Council of Europe and integrated soil research. In Integrated soil and sediment research: A basis for proper protection (pp. 37-47). Springer, Dordrecht.