Solution for the AfSIS competition

CodiLime

1 Summary

This note describes our approach to the Africa Soil Property Prediction Challenge competition. We have used the prototype of `DeepSense.io` (CodiLime's Machine Learning engine) to visualise the data, extract the features, suggest algorithms and tune parameters. At the end we used manual and very basic model averaging to obtain the final solution.

2 Features Selection / Extraction

Most of the decisions were based on the location-wise CV. The data was grouped into top/bottom soil pairs. We ended up using only spectral features. For most of the targets we discarded the subintervals of spectra the prototype of `DeepSense.io` determined to contribute to overfitting. Finally, wavelet transforms were used as a final feature set. See further sections for more details.

3 Modelling Techniques and Training

Our final submission was based on averaging three different solutions (with weights):

3.1 First solution - 50%

Ca

Target transformation: $f(x) = \sqrt{x + 0.536}$.

Algorithm: Gaussian process regression (gausspr) with $scaled = False$, $\sigma = 0.01$ and $Var = 0.01$.

Features: Subintervals of spectrum $[1100; 1500]$, $[1700; 1850]$, $[2150; 3200]$ transformed with discrete wavelet transform (level 4 scaling coefficients form dwt with filter `la12`).

P

Target transformation: Outliers were removed by dropping all observations with **P** > 1.0.

Algorithm: Neural Networks (nnet) with $size = 5$, $decay = 0.004$, bagged 200 times with train set sampled ×1.1.

Features: Full spectrum transformed with discrete wavelet transform (level 4 scaling coefficients form dwt with filter `la8`).

pH

Algorithm: Neural Networks (nnet) with $size = 8$, $decay = 0.004$, bagged 200 times with train set sampled ×1.1.

Features: Subintervals of spectrum $[1400; 1850], [2150; 3200]$ transformed with discrete wavelet transform (level 4 scaling coefficients form dwt with filter `la8`).

SOC

Algorithm: Neural Networks (nnet) with $size = 7$, $decay = 0.004$, bagged 200 times with train set sampled ×1.1.

Features: Subintervals of spectrum $[1400; 1850], [2150; 3200]$ transformed with discrete wavelet transform (level 4 scaling coefficients form dwt with filter `la8`).

Sand

Algorithm: Neural Networks with $size = 3$, $decay = 0.04$, bagged 200 times with train set sampled ×1.5.

Features: Full spectrum transformed with discrete wavelet transform (level 4 scaling coefficients form dwt with filter `la8`).

3.2 Second solution - 25%

Outliers in **P** were removed by setting **P** value to 1.5 for all observations where it would be higher. We used bagged (sampling the training set ×1.5) Neural Nets (nnet) with 7 hidden units, 0.01 decay. Spectral features were wavelet transformed via `la8`, then ran though "MillCheck's smoothing" — we won't elaborate, because there is little reason to believe it had an effect on the final result (it's still included in the code).

3.3 Third solution - 25%

Here we used SVR on unscaled spectral features with $cost = 10000$, as in Abishek's "beat the benchmark" post. See `https://www.kaggle.com/c/afsis-soil-properties/forums/t/10351/beating-the-benchmark`.

4 Code Description

We used R packages (nnet, kernlab, e1071, wavelets). A small fix in the kernlab library was needed to allow $scaled = False$ in gausspr package. The amended version is included in the solution code.

5 Dependencies

To run the code you need R with the following packages:

- nnet
- e1071
- wavelets
- kernlab (with the fix)

6 How to Generate the Solution

A README file containing the instructions in included.

7 Additional Comments and Observations

We would consider Ca, pH and SOC quite well optimised and if significant improvement to the model was to be made, we believe it should be on P (obviously) or Sand.

8 Simple Features and Methods

Using selected subintervals of spectra has improved our results a lot (on the CV). Removing outliers for P has also improved our predictions.

9 References

See `DeepSense.io`.