Quantitative analysis with LIBS

Sixth presentation

Olivier Nicolini

06-04-2020

Summary

During this week I have explored more both the new Aluminium and Slag dataset.

The question I've tried to answer were the following:

- The cause of the worse performance on the new Slag dataset
- If the combination of multiple feature selection algorithms improve prediction
- If the wavelength selection is including some or all the element peak wavelengths

Feature selection algorithms

According to literature, the majority of linear models works best not using the whole spectra but by selecting subsets of wavelengths to be used for regression.

Feature selection algorithms can be divided into different types of methods:

- **Filter** methods
  - Variable importance in projection (VIP) filter
  - Regression coefficients (beta) filter
  - Loading weights filter
- **Wrapper** methods
  - Genetic Algorithm PLS (GA-PLS)
  - Uninformative variable elimination PLS (UVE-PLS, MC-UVE, WT-MC-UVE)
  - Multi-step variable selection based on C value (MSVC)
  - Iterative predictor weighting PLS (IPW-PLS)
  - Interval PLS (*i*-PLS)
  - Regularized elimination procedure in PLS (REP)
  - Backward variable elimination PLS (BVE-PLS)
  - Sparse Partial Least Squares (SPLS)

Data Analysis

These are the feature selection algorithms used to select wavelengths from the whole spectra together with their arguments (R programming language)

```
ipw = ipw_pls(y=y, X=x, ncomp = 20, no.iter = 100, IPW.threshold = 0.0001, filter = "RC", scale = FALSE)
ipw = ipw_pls(y=y, X=x, ncomp = 20, no.iter = 100, IPW.threshold = 0.0001, filter = "RC", scale = TRUE)
#selecta =ipw$ipw.selection
bve= bve_pls(y=y, X=x, ncomp = 15, ratio = 0.75, VIP.threshold = 1)
#selecta =bve$bve.selection
ga= ga_pls(y=y, X=x, GA.threshold = 15, iters = 100, popSize = 250)
#selecta =ga$ga.selection
rep=rep_pls(y=y, X=x, ncomp = 15, ratio = 0.75, VIP.threshold = 0.9, N = 3)
#selecta =rep$rep.selection
mcuve=mcuve_pls(y=y, X=x, ncomp = 10, N = 5, ratio = 0.75, MCUVE.threshold = NA)
#selecta =mcuve$mcuve.selection
ipls=ipls( x=x,y=y, int.ncomp = 20, int.num = 50, cv = 9, method = "forward") #method= backward or forward
#selecta =ipls$var.selected
spa = spa_pls(y=y, X=x, ncomp = 20, N = 5, ratio = 0.75, SPA.threshold = 0.05)
#selecta =spa$spa.selection
```

ipw = iterative predictor weighting
bve = backwards variable elimination
ga = genetic algorithm
rep = regularized elimination procedure

mcuve = monte carlo uninformative variable elimination
ipls = interval partial least squares
spa = sparse partial least squares

Slag - baseline

The first thing that I investigated was if the baseline correction was the cause of the bad results of prediction over Slag samples

SiO2 - **basic**	**corrected**	**raw**
ipw-f	0.92	0.86
ipw-t	0.923	0.929
bve	0.617	0.607
ga	0.76	0.67
rep	0.61	0.60
mcuve	0.84	0.64
ipls	0.85	0.73

SiO2 - **norm**	**corrected**	**raw**
ipw-f	0.939	0.925
ipw-t	0.92	0.95
bve	0.75	0.68
ga	0.83	0.76
rep	0.74	0.68
mcuve	0.12	0.74
ipls	0.95	0.86

SiO2 - **SNV**	**corrected**	**raw**
ipw-f	0.945	0.94
ipw-t	0.912	0.905
bve	0.726	0.69
ga	0.831	0.79
rep	0.73	0.69
mcuve	0.83	0.73
ipls	0.945	0.88

After some trials I concluded that baseline correction is not the cause of the bad performance of this dataset. On the contrary, baseline correction seems to improve the predictions

Slag - baseline

However, this not always the case for all elements. For MnO prediction the baseline did not help that much and often the results after correction were worse off.

MnO - **SNV**	**corrected**	**raw**
ipw-f	0.64	0.71
ipw-t	0.68	0.75
bve	0.48	0.55
ga	0.66	0.62
rep	0.49	0.54
mcuve	0.60	0.60
ipls	0.70	0.80

MgO - **SNV**	**corrected**	**raw**
ipw-f	0.44	0.05
ipw-t	0.34	0.20
bve	0.02	-0.04
ga	0.20	0.09
rep	0.04	-0.02
mcuve	0.07	-0.01
ipls	0.38	0.15

For what concerns MgO (at the moment the most difficult element to predict) the results were always improved by correction. So the baseline hypothesis was momentarily discarded.

Slag - P series and F100 samples

The other possible cause of bad results I thought about was the presence of new samples in the dataset that were not present in the previous ones (P1, P2, P3, P4, and F100 samples)

SiO2	**noPure (best)**	reduced-basic	reduced-norm	reduced-snv	reduced-sum	reduced-max	**OLD (basic)**
ipw-f	0.9569 - ipw-f	0.5874	0.8711	0.8130	0.9161	0.7857	0.9938
ipw-t	0.9582 - max	0.7904	0.9552	0.9357	0.6132	0.9371	0.9612
bve	0.7547 - sum	0.5090	0.7366	0.7353	0.8770	0.6439	0.9513
ga	0.8453 - sum	0.7126	0.8922	0.8612	0.9055	0.7913	0.9545
rep	0.7485 - sum	0.5048	0.6921	0.7372	0.8743	0.6381	0.9535
mcuve	0.8448 - basic	0.6764	0.5221	0.7417	0.5620	0.3034	0.9625
ipls	0.954 - norm	0.8295	0.9196	0.9123	0.9460	0.9060	0.9408

As you can see the removal of these samples improved a bit the results in some cases but the results were not nearly as good as those achieved on the previous dataset. This needs more thought

Slag - P series and F100 samples

Still, the results on MgO seem to improve after the removal of these samples so this could not be such a bad idea.

MgO	**noPure (snv)**	reduced-basic	reduced-norm	reduced-snv	reduced-sum	reduced-max	**OLD (basic)**
ipw-f	0.4485	0.45	0.46	0.45	0.39	0.48	0.66
ipw-t	0.34	0.34	0.50	0.48	0.44	0.462	0.6625
bve	0.02	0.2724	0.3088	0.29757	0.3342	0.2203	0.6249
ga	0.2043	0.2805	0.3986	0.3675	0.4324	0.3767	0.6400
rep	0.04	0.2709	0.3069	0.3017	0.3453	0.2143	0.5929
mcuve	0.07	0.09327	0.4013	0.3189	0.4009	0.3101	0.6601
ipls	0.38	0.3171	0.4300	0.4261	0.3005	0.4585	0.6906

Multiple feature selection

All the experiments show that no matter what combination of feature selection methods, the results are always worse off than by using just one of the algorithms on their own.

Another possible interesting combination could be an iterative approach on which I use the same feature selection algorithm repeatedly on the subset selection. This method could be very good on **filtering methods** since they take very little computational time.

Peak selection (aluminium)

The experiments ran over the aluminium dataset show that the selected wavelengths contain almost every time one or more of the element peaks.

Another interesting aspect is that the selected wavelengths improve the prediction of every method tried, sometimes even with (slightly) better performance, for example in the case of Elastic-Net model, even if the feature selection algorithms are based on PLS regression.

Future Work

- Explore more the slag datasets and try to get better results.
- Try with AUSOM partial dataset
- Try other combinations of filtering methods and preprocessing steps