Comparison of Machine Learning Methods for Spectroscopic Data Analysis

➢ Context / customer need

The computing power of computers and volumes of data to be processed are increasing significantly. This makes Machine Learning (ML) more and more popular. ML methods make it possible to analyze data in a large number of fields, with very varied applications, such as banking, marketing, or scientific research for example.

These ML algorithms can perform very well on spectroscopic data. However, it is useful to know the algorithms available, as well as how to implement them on spectral data.

➢ Ondalys solution

In order to compare the effectiveness of a few ML methods for spectroscopic issues, an example based on a NIR data set is detailed.

This database was acquired with a FOSS Tecator Infratec spectrometer over the spectral range 850-1050nm.

193 meat samples were analyzed and the fat content of each sample is measured as a reference.

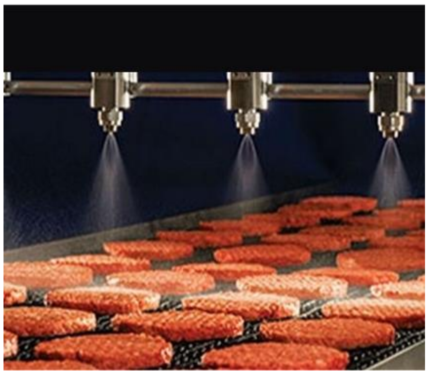
Large non-linearities are detected on this parameter, making the predictions difficult using classical linear regression methods such as PLS.

ML algorithms can be useful for modeling non-linear or complex correlations (clustered data, prediction of physical or sensory parameters, or concentrations close to detection thresholds). Three of the most popular solutions are compared: SVR - *Support Vector Machine Regression*, ANN - *Artificial Neural Networks* and CART/RF - *Classification And Regression Trees / Random Forest*.

ONDALYS S.A.S.
4, rue Georges Besse
34830 CLAPIERS - France
+33 (0)4 67 67 97 87
contact@ondalys.fr
www.ondalys.fr
Ondalys Registered Company
Montpellier RCS 447 624 024
VAT number: FR10 447 624 024

In order to demonstrate the interest of these methods compared to more « classical » chemometrics methods, data were also processed with a PLS - Partial Least Squares Regression, a PLS with prior transformation of the X-variables, and a LWR - Locally Weighted Regression.

Source dataset : http://lib.stat.cmu.edu/datasets/tecator



Results / Customer benefits

The six optimized models were compared in terms of performance (Table 1), non-linearities processing, implementation complexity and overfitting risk (Table 2).

First, regarding non-linearities, the PLS model alone was insufficient. Performing a simple X-transformation and adding combined variables (squared terms and crossed terms) were sufficient to remove the non-linearities. ML models and local model also make it possible to overcome the non-linearities of the parameter to be predicted.

Concerning the performances obtained in prediction, the best results are obtained with SVM and ANN models. The RF, local and transformed PLS models are less efficient and equivalent to each other.

With SVM and ANN, even if the final performances are equivalent, the implementation difficulty must be taken into account (Table 2). SVM method is easier to implement and gives good performances with few data. On the contrary, a large amount of data is recommended for ANN. ANN is a more difficult method to understand, and requires a larger number of training samples.

In this case study, the SVM method presents more advantages, looking at performance results and the ease of use.

Method	N LVs	Calibration				Cross Validation				Test			
		SEC	SEC (%)	R^2	RPD	SECV	SECV (%)	R^2	RPD	SEP	SEP (%)	R^2	RPD
1.PLS	5	2,6	18,5%	0,958	5	2,8	20,2%	0,950	4	2,7	19,0%	0,958	5
2.PLS with prior X-transformation	5	0,7	5,3%	0,997	17	0,9	6,2%	0,995	15	0,9	6,6%	0,996	14
3.LWR	3	0,5	3,6%	0,998	25	0,9	6,5%	0,995	14	1,0	6,9%	0,995	13
4.SVR	-	0,4	2,6%	0,999	35	0,5	3,6%	0,998	25	0,5	3,4%	0,999	27
5.ANN	-	0,5	3,9%	0,998	23	0,6	4,5%	0,998	20	0,4	2,7%	0,999	34
6.RF	-	0,6	4,3%	0,998	21	1,1	8,2%	0,992	11	1,0	7,5%	0,994	12

Table 1. Comparison of the results obtained for each Machine Learning algorithm

Method	Non-linearity handling	Performance	Implementation complexity	Overfitting risk
1.PLS	-	-	-	-
2.PLS with prior X-transformation	+	+	-	+
3.LWR	+	+	+	+
4.SVR	+	++	++	++
5.ANN	+	++	++	++
6.RF	+	+	+	+

Tableau 2. Characteristics of Machine Learning methods applied to spectral data

Contact-us

Ondalys

contact@ondalys.fr

www.ondalys.fr

☎ +33 (0)4 67 67 97 87