PhD Thesis; Irstea Montpellier France 2018-2021

Adaptation of two big data algorithms of indexation to improve the local-PLS chemometrics method

Near infrared spectrometry can provide huge amounts of data to digital agriculture. The main tool of chemometrics, used to analyze NIR spectra, is Partial Least Squares (PLS) regression. PLS allows building efficient predictive models from a large number of variables even if these variables are highly correlated. The method has proved its relevance for small homogeneous databases. Its extension to medium-sized bases (<10,000 individuals) is the "local-PLS": it determines a neighborhood of the individual to be predicted, and then realizes a usual PLS on this neighborhood. This method combines the power of the k nearest neighbors' method (k-NN) and the PLS. However, it is is not able to process large databases (e.g. >50,000 individuals) or even >1 million of individuals that will appear in the near future to digital agriculture. The current local-PLS algorithms all use sequential k-NN algorithms for which calculation times become unrealistic; other algorithms must be considered. Paradoxically, very little research has been done on this challenge in chemometrics. Our idea is that algorithms of indexation used in big data, integrated in the local-PLS method, could lift this methodological lock. We propose to consider two algorithms of dimension reduction and fast neighborhood searches used by the Zenith Team of Lirmm-Montpellier for processing large data sets of time series (that have a similar data structure as the NIR spectra): the hashing (calculation of sketches) and the iSax (Symbolic Aggregate approXimation). The work will consist in two steps: (1) a "business as usual" integration of the two algorithms in the local-PLS algorithm, (2) an optimisation of the algorithms taking into account the chemometric specificity of the NIR spectra. The new algorithms developed in this thesis will improve the ability to predict physico-chemical variables from large heterogeneous NIRS data bases, and will find direct applications in many domains (plants, feed, soils, etc.).

Contact: jean-michel.roger@irstea.fr, matthieu.lesnoff@cirad.fr, nathalie.gorretta@irstea.fr

Application form: https://pasi.irstea.fr/en/campagne/1