Comparaison de méthodes du Machine Learning pour l'analyse de données spectroscopiques

Sylvie Roussel*, Jordane Lallemand, Sébastien Preys

*Ondalys, 4 rue Georges Besse, 34830 Clapiers (France)

sroussel@ondalys.fr – jlallemand@ondalys.fr – spreys@ondalys.fr

www.ondalys.fr – Tel : +33 (0)4 67 67 97 87

Pour l'analyse de données spectroscopiques, les termes « d'analyse de données multivariées » ou de « chimiométrie » sont les plus souvent employés. Depuis quelques années, avec l'avènement des « Big Data » et autres « IoT – Internet des Objets », les termes « Machine learning » et « d'Intelligence Artificielle (IA) », sont de plus en plus employés.

Mais que recouvrent réellement ces méthodes de Machine Learning ?
Comme monsieur Jourdain, ne faisons-nous pas de la prose sans le savoir ?

Au travers d'un cas concret de spectroscopie proche infrarouge (SPIR), cette présentation a pour objectif de présenter et comparer différentes méthodes de Machine Learning permettant modéliser un paramètre non linéaire : la teneur en matières grasses dans la viande.

Les échantillons de viande sont mesurés en transmission par un instrument FOSS Tecator Infratec, sur la gamme 850-1050nm (http://lib.stat.cmu.edu/datasets/tecator). Les spectres proche infrarouge ont été convertis en absorbance et séparés en un jeu d'étalonnage et un jeu de test indépendant.

Le modèle linéaire PLS montre une non-linéarité résiduelle entre les spectres et la teneur en matières grasses. Des méthodes de Machine Learning, permettant de modéliser cette non-linéarité, ont été testées et comparées afin d'améliorer la précision des prédictions (ci-dessous par ordre de performance croissante) :

- Transformation des variables d'origine et modèles PLS
- Modèle local (Locally Weigthed Regression - LWR)
- CART / Forêt Aléatoires (Random Forest - RF)
- Support Vector Machine (SVM)
- Réseaux de Neurones Artificiels (ANN)

Méthode	Gestion non-linéarité	Performance	Complexité de mise en œuvre	Risque de sur-apprentissage
PLS	-	-	-	-
PLS sur X transformé	+	+	-	+
LWR	+	+	+	+
RF	+	+	+	+
SVM	+	++	++	++
ANN	+	++	+++	+++

Sylvie Roussel, Jordane Lallemand, Sébastien Preys. Comparaison de méthodes du Machine Learning pour l'analyse de données spectroscopiques. *24ème journées du Groupement Français de Spectroscopie Vibrationnelle GFSV 2018, Le Ventron (Vosges), France* – 16, 17 et 18 mai 2018.