ondalys
Donnez du sens à vos données

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

➢ Contexte / Besoin client

Avec l'augmentation de la puissance de calcul des ordinateurs et des volumes de données à traiter, le Machine Learning (ML) est devenu de plus en plus populaire. Les méthodes de ML permettent d'analyser des données dans de très nombreux domaines, avec des applications très variées, telles que la banque, le marketing, ou la recherche scientifique par exemple.
Ces algorithmes de ML peuvent se montrer très performants sur les données spectroscopiques. Il est toutefois utile de connaître les algorithmes disponibles, ainsi que la façon de les mettre en œuvre sur des données spectrales.

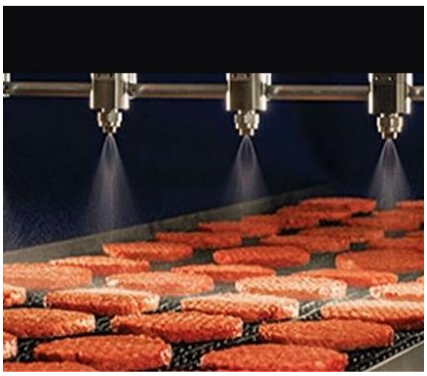
➢ Solution Ondalys

Afin de comparer l'efficacité de quelques méthodes de ML pour des problématiques de spectroscopie, un exemple est détaillé à partir d'un jeu de données spectrales proche infrarouge. Ce jeu de données a été acquis à l'aide d'un spectromètre FOSS Tecator Infratec, sur la gamme spectrale 850 – 1050nm. 193 échantillons de viande sont analysés, et le taux de matière grasse de chaque échantillon est mesuré en référence. Des non-linéarités importantes sont constatées sur ce paramètre, ce qui rend les prédictions compliquées par des méthodes de régression linéaire classique de type PLS.

Les algorithmes de ML peuvent être très utiles pour modéliser des corrélations non-linéaires ou complexes (données clusterisées, prédiction de paramètres physiques, sensoriels, ou concentrations proches des seuils de détection). Trois des solutions les plus répandues seront comparées : les SVR – *Support Vector Machine Regression* -, les réseaux de neurones (ANN – *Artificial Neural Networks*) et les arbres de régression / forêts aléatoires (CART/RF – *Classification and Regression Trees / Random Forest*).

Afin de montrer l'intérêt de ces méthodes par rapport à des méthodes de chimiométrie plus « classiques », les données ont également été modélisées à l'aide d'une régression linéaire PLS - *Partial Least Squares Regression* -, d'une PLS avec transformation préalable des X, et d'une régression locale LWR – *Locally Weighted Regression* –.

Source jeu de données : http://lib.stat.cmu.edu/datasets/tecator



➢ Résultats / Bénéfices clients

Les six modèles optimisés ont été comparés en termes de performances (Tableau 1), de gestion des non-linéarités observées, de complexité de mise en œuvre et de risque de sur-apprentissage (Tableau 2).

En termes de gestion des non-linéarités tout d'abord, il a été constaté que seul le modèle PLS était insuffisant. Une simple transformation des X, ainsi que l'ajout de variables combinées (termes au carré et termes croisés) suffit à faire disparaître la non-linéarité. Les modèles de ML et le modèle local permettent également de s'affranchir des non-linéarités du paramètre à prédire.

Concernant les performances obtenues en prédiction, les meilleurs résultats sont obtenus avec les SVM et les ANN. Les modèles RF, local, et PLS transformée sont moins performants et équivalents entre eux.

Dans les cas des SVM et des ANN, si les résultats finaux sont équivalents en termes de performances, la difficulté de mise en œuvre doit aussi être prise en compte (Tableau 2). Les SVM sont plus simples à implémenter, et performent bien avec peu de données. Il est généralement recommandé de disposer d'un grand nombre de données pour utiliser les ANN, et cette méthode est plus délicate à appréhender et nécessite un grand nombre d'essais d'apprentissage.

Dans le cas présenté ici, la méthode des SVM est plus avantageuse, en prenant en compte les performances obtenues et la facilité d'utilisation.

		Etalonnage				Validation croisée				Test			
Méthode	Nb LVs	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 sur X transformé	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

Tableau 1. Comparaison des résultats pour chaque algorithme de Machine Learning

Méthode	Gestion non-linéarité	Performance	Complexité de mise en œuvre	Risques de sur-apprentissage
1. PLS	-	-	-	-
2. PLS sur X transformé	+	+	-	+
3. LWR	+	+	+	+
4. SVR	+	++	++	++
5. ANN	+	++	+++	+++
6. RF	+	+	+	+

Tableau 2. Caractéristiques des méthodes de Machine Learning appliquées aux données spectrales

Contactez-nous

Ondalys

contact@ondalys.fr

www.ondalys.fr

☎ 04 67 67 97 87