Projet de thèse

Adaptation de deux algorithmes d'indexation « big data » pour l'amélioration de la méthode local-PLS en chimiométrie

Contexte et problématique :

La spectrométrie proche infrarouge (SPIR, Siesler, 2008) est pressentie comme une technique pouvant fournir une masse considérable de données à l'agriculture numérique (Bellon-Maurel, 2015). Cette technique de mesure est étroitement liée à la chimiométrie, qui permet de transformer les spectres acquis en informations utiles (Siesler, 2008). La chimiométrie propose depuis des années des outils de régression permettant de relier des spectres (infrarouges notamment) à des grandeurs chimiques (des concentrations) ou qualitatives (des classes). Parmi ces outils, la régression Partial Least Squares (PLSR ; Helland, 1990, De Jong, 1993, Tenenhaus, 1998, Wold, 2001) et sa variante PLS - Discriminant Analysis (PLS-DA, Barker, 2003) se sont imposées car elles fonctionnent très bien sur des petites bases de données, lorsque le lien entre spectres et réponses à prédire est assez univoque. Le principe général de la PLS est de compresser la matrice spectrale (dont les colonnes sont très corrélées entre elles) en un nombre réduit d'axes orthogonaux (appelées variables latentes) corrélés avec la réponse, qui sont ensuite utilisés comme variables explicatives d'un modèle de régression linéaire multiple ou d'analyse discriminante. La PLS répond donc très bien aux problèmes posés par un grand nombre de variables ; elle d'ailleurs utilisée depuis longtemps en spectrométrie RMN (Gerbanowski, 1997), et plus récemment en génomique, protéomique et métabolomique (Eriksson, 2004).

De nouvelles demandes ont récemment vu le jour en chimiométrie pour réaliser des modèles sur des bases de données à visée « exhaustive », contenant un grand nombre d'individus et davantage de variations de grandeurs d'influence (p. ex. le type d'appareillage et de matériel biologique ou la localisation et l'année de collecte). Ainsi en agronomie, on trouve des bases de données spectrales de sols, de céréales, de fourrage pouvant contenir plus de 10 000 individus. La PLS usuelle trouve rapidement ses limites en face de l'hétérogénéité présente dans ces bases, avec des variances et des biais de prédiction souvent trop élevés.

Une réponse actuellement explorée par les chimiométriciens est la régression PLS « locale » (Shenk, 1997, Centner, 1998, Ramirez-Lopez, 2013, Allegrini, 2016) reprenant l'idée de la régression locale (Cleveland, 1979). Pour chaque spectre à traiter, la méthode consiste tout d'abord à trouver des voisins du spectre (donc à réduire l'hétérogénéité), puis à réaliser les prédictions au moyen d'une PLSR usuelle effectuée sur le voisinage déterminé. Différentes applications ont montré l'efficacité de la méthode en agronomie, par exemple dans le cadre d'analyses de sols (Clairotte, 2016), de lait et de fèces d'animaux (Tran, 2010) et de plantes annuelles vivrières (Davrieux, 2016). La même idée peut être appliquée pour des problèmes de discrimination, en remplaçant la PLSR par la PLS-DA (Bevilacqua, 2014).

Un point critique des méthodes PLS locales, aussi bien en termes d'efficacité statistique que de temps de calcul, est la sélection du voisinage. Paradoxalement, très peu de recherches ont été effectuées sur ce sujet dans le domaine de la chimiométrie. Les algorithmes actuels de PLS locales utilisent tous l'algorithme des plus proches voisins (k-NN) linéaire ou séquentiel (« brute-force method ») : pour chaque spectre à traiter, l'algorithme calcule les distances entre ce spectre et les *n* spectres de la base, ordonne les distances puis en déduit les plus proches voisins (avec quelques variantes possibles : Allegrini, 2016). La recherche linéaire a l'avantage d'être simple mais souffre d'un problème de lenteur lorsque la base contient beaucoup de spectres. Il devient extrêmement coûteux de tester les *n* points de l'espace spectral pour en

déduire le voisinage, avec un temps de calcul d'ordre $O(n)$[1]. Les temps de calcul deviennent rapidement rédhibitoires pour mener des protocoles usuels de construction et d'évaluation de modèles prédictifs (sélection de modèles et estimation des incertitudes) comme la validation croisée. Les problèmes deviendront insurmontables si on pense traiter, dans le futur, des bases de données encore plus importantes que les bases actuelles (> million d'individus). D'autres algorithmes doivent être envisagés pour assurer la pérennité de la méthode PLS locale dans un contexte de bases de données toujours plus grandes. Jusqu'à présent, toutes les optimisations de la PLS locale consistaient à réduire l'espace des spectres en un sous espace porteur de l'information utile (par une ACP ou une PLS) et de calculer des distances (K-NN) d'après ces espaces utiles. Une autre solution, déjà employée pour les séries temporelles (Yagoubi-1, 2017), consiste à opérer une autre réduction d'espace (par indexation) qui permet de placer tous les individus de la base dans des grilles ou des arbres dont la scrutation est très rapide.

Bien que l'approche k-NN soit ancienne (Cover, 1967, Hart, 1968), l'émergence récente des problématiques big data a accru l'intensité des recherches sur les algorithmes de calcul de voisinage (point méthodologique crucial de la PLS locale). Les bases de données contenant des objets en très grand nombre et décrits dans des espaces de grande dimension engendrent des difficultés auxquelles sont confrontées toutes les techniques d'indexation pour la recherche de ressemblances. La recherche méthodologique porte sur le développement d'algorithmes d'indexation sophistiqués pour structurer les données, et d'algorithmes de recherche très performants pour y accéder efficacement (Wang, 2011). Des panoramas des techniques d'indexation multidimensionnelle pour la recherche de voisinage sont par exemple présentés dans Berrani, 2002 et Muja, 2014. On distingue les algorithmes de recherche « exacte » *vs* « approximative » des plus proches voisins. Dans le premier cas, on recherche les « vrais » plus proches voisins, dans le second cas on accepte, à l'aide d'une approche probabiliste, un certain niveau d'erreur. De manière générale, les méthodes exactes s'appuient sur l'élaboration d'une pré-structuration des données pour réduire le nombre de distances à calculer (méthodes « Tree-based » Knuth, 1997, Kim, 2010). Elles peuvent être très efficaces pour des espaces de dimension faible (2-3 descripteurs) ou modérée (< 10 à 20 descripteurs). Mais le « fléau de la dimension » rend difficile leur application aux données spectrales. Des techniques de réduction de dimension peuvent contourner ce problème, notamment celles basées sur des sélections de descripteurs pertinents (Park, 2015, García-Torres, 2016) ou sur des modèles de prédiction pré-estimés (comme dans les approches PLS de type « SIMCA » ou « clusterwise/typological » ; Vinzi, 2005, Preda, 2005). D'autres alternatives sont les méthodes de recherche approximatives pour les indexations en dimension élevée (Muja, 2014, Liu, 2005, Hyvönen, 2016). Elles engendrent une diminution de la précision du résultat mais permettent en contrepartie une forte diminution du temps de calcul. Un fort engouement s'est développé autour de ces méthodes, notamment celles utilisant des techniques aléatoires de « hachage » comme le « Local Sensitive Hashing » (Slaney, 2008, Paulevé, 2010). De nombreuses applications ont par exemple été proposées en bio-informatique pour comparer des séquences génomiques (Berlin, 2015).
Il y a d'autres méthodes de réduction de dimensions utilisées notamment pour les séries temporelles. On peut mentionner par exemple la représentation iSax «Symbolic Aggregate approXimation» (Camerra, 2014) qui permet de créer des index efficaces sur de très grandes bases de données. Il y a aussi la méthode basée sur des vecteurs aléatoires (Cole, 2005) qui permet de produire des "sketches" à partir les données d'origine, et ensuite les sketches sont utilisés pour la recherche de plus proches voisins (kNN).

Bien que les méthodes de recherche kNN basées sur l'indexation permettent des gains de temps de plusieurs ordres de grandeur par rapport au balayage séquentiel, quand elles sont centralisées leurs performances se détériorent quand la taille des données augmente. Cela pose des questions sur la capacité de ces méthodes centralisées à passer à l'échelle. Pour faire face à l'augmentation du volume des données, une solution prometteuse est d'exploiter des frameworks parallèles, tels que Spark (Zaharia, 2012), pour créer de puissantes unités de calcul et de stockage à l'aide de machines ordinaires. Il y a des différentes méthodes pour créer des index parallèles sur les grandes bases de données, notamment les techniques qui produisent des indexes arborescents pour les données représentées par iSAX (Yaghoubi-1, 2017), et aussi celles basées sur le hachage parallèle des sketches (Yaghoubi-2, 2017).

[1] Prenons l'exemple d'une base de taille n = 1 000 spectres induisant un temps de calcul d'1 sec par voisinage à calculer. Le temps de calcul par voisinage sera de 1.7 min pour une base de taille n = 100 000 spectres et de 17 min pour une taille n = 10^6 spectres. Ceci représente des temps de calcul respectivement de 2.8 h et 28 h pour 100 voisinages à calculer.

Objectif de la thèse :

L'objectif de la thèse est de tester et d'adapter des techniques du big data pour rendre compatibles les algorithmes de PLS locale avec les grandes (typiquement > 20 000 individus) et très grandes (>10^6 individus) bases de données. Deux méthodes d'indexation, étudiées intensivement par l'équipe Zenith du Lirmm (participant au projet de thèse), seront explorées :

- Le hachage (en particulier, le calcul de sketches)
- L'indexation arborescente des données représentées par iSAX.

Pour chacune de ces deux techniques, la méthodologie suivante sera suivie :

- Application directe des algorithmes existants (hachage et iSAX) sur les bases de données spectrales ; évaluation des gains de performances en termes de temps de calcul et des gains ou des pertes de performances en terme de précision de prédiction.
- Test de l'influence de différent prétraitements spectraux couramment utilisés en SPIR ; définition d'un prétraitement optimal pour rendre les spectres compatibles avec les algorithmes iSAX et sketches.
- Adaptation des algorithmes de recherche d'iSAX et de sketches aux particularités structurelles des spectres PIR.

Les méthodes produites seront testées sur une ou plusieurs applications directement utiles à l'agriculture numérique :

- À partir de la base de données des spectres PIR RMQS, qui contient les spectres représentatifs des sols de France (un spectre tous les 16 km), utiliser la PLS locale modifiée, pour prédire les caractéristiques de sols agricoles à une échelle plus fine, compatible avec les images de télédétection.
- L'algorithme de voisinage, développé dans le cadre de cette thèse pour la PLS locale, sera également testé sur d'autres applications de la SPIR, comme l'analyse non supervisée des séries d'images hyperspectrales, ouvrant ainsi le champ à d'autres questions de recherches qui seront présentées comme des perspectives.

Questions de recherche :

Les questions principales de recherche seront :

- Les algorithmes de hachage (sketches) et iSAX peuvent-ils être utilisés sur les spectres PIR ?
- Quelles dimensions spectrales doivent être utilisées dans les algorithmes de hachage et iSAX pour rendre optimale la recherche de voisins et les prédictions PLS ?
- Pour le hachage, comment définir au mieux de nouveaux sketches pour la SPIR ? Y a-t-il des dimensions spécifiques aux spectres PIR qui permettent un hachage plus précis ?
- Pour iSAX, comment discrétiser au mieux un spectre PIR avant l'indexation par arbre ? Y a-t-il une représentation symbolique alternative, et spécifique aux spectres PIR, qui offre de meilleures performances ?

Moyens

Cette thèse sera financée par une demie bourse Irstea et une demie bourse # DigitAg.

Encadrement

Codirection de thèse : Jean Michel Roger (Irstea - ITAP) / Matthieu Lesnoff (Cirad - SELMET).
Co-Encadrement : Nathalie Gorretta (Irstea - ITAP).

Comité de thèse rapproché pressenti : Encadrants + Florent Masseglia (Lirmm-ZENITH) + Reza Akbarinia (Lirmm-ZENITH) + Sandro Bimonte (Irstea Clermont Ferrand)

Comité de thèse élargi : Juan Antonio Fernandez Pierna (CRA Gembloux) + Dino Ienco (Irstea - TETIS) + référent ED

L'étudiant sera probablement amené à réaliser des séjours courts au Lirmm ZENITH (Montpellier) pour des travaux spécifiques sur les algorithmes d'indexation big data.

Démarche et programme de la thèse

1) Le candidat devra dans un premier temps réaliser un état de l'art :
- des différentes méthodes utilisées en chimiométrie pour mettre en oeuvre les régressions locales
- des méthodes d'indexation des bases de données et de recherche de voisinage associées dans le domaine big data (hachage, iSax)

2) les différentes méthodes identifiées seront testées sur des jeux de données synthétiques, produites à partir de jeux réels (spectres de fourrage, de sols) et simulés. La sensibilité au nombre d'individus et de longueurs d'onde sera étudiée.
3) les algorithmes identifiés comme les plus prometteurs seront modifiés pour produire une ou plusieurs méthodes adaptées à la régression et à la discrimination par PLS locale.
4) Application sur base de données sols ; évaluation des gains de performance ; projection sur une utilisation embarquée, avec ou sans utilisation de cloud.

D'une durée de 3 ans, la thèse se déroulera selon le calendrier prévisionnel ci-dessous

semestre	1	2	3	4	5	6
état de l'art PLS / big data	x	x				
test méthodes existantes		x	x			
article de review		x	x	x		
dévelop. test méthodes, application			x	x	x	
article méthodo				x	x	x
rédaction soutenance					x	x

Collaborations

- UMR ITAP, UMR SELMET, Equipe ZENITH (Lirmm), UR TSCF (Irstea), UMR TETIS

Valorisation envisagée

- Conférences Chimiométrie / francophones annuelles

- Conférence CAC 2020
- Un article de revue et de test de méthodes sera très apprécié par la communauté chimiométrie. Ce sera un premier article soumis à Chemometrics and Intelligent Laboratory Systems, ou a Journal of Chemometrics
- Un article méthodologique, dans Chemometrics and Intelligent Laboratory Systems, ou a Journal of Chemometrics, voire Analytica Chimica Acta.
- Elaboration d'un package R dédié aux méthodes locales ; inclusion dans le logiciel ChemFlow
- Nous examinerons aussi l'opportunité d'une communication dans la communauté du big data.

École doctorale de rattachement

ED GAIA de l'université de Montpellier

Profil du candidat recherché

Le candidat recherché devra présenter une formation en chimiométrie ou en statistique appliquée (avec une sensibilité pour l'informatique), ou une formation en informatique avec une forte sensibilité pour les applications chimiométriques. Une expérience en sciences du vivant, soit au travers de la formation initiale, soit au travers du stage de fin d'études sera un plus.

Références

V. Bellon-Maurel, 2015 : Could near infrared spectroscopy be useful to digital agriculture? Keynote lecture at NIR2017 ICNIRS Conference, Copenhagen, 11-15 June 2017.

D.-E. Yagoubi, R. Akbarinia, F. Masseglia, T. Palpanas, 2017: DPiSAX: Massively Distributed Partitioned iSAX. Proceedings of IEEE International Conference on Data Mining series (ICDM), New Orleans, USA, 18-21 November 2017.

D.-E. Yagoubi, R. Akbarinia, F. Masseglia, D. E. Shasha, 2017: RadiusSketch: Massively Distributed Indexing of Time Series. Proceedings of IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19-21 October 2017.

A. Camerra, J. Shieh, T. Palpanas, T. Rakthanmanon, E. J. Keogh, 2014: Beyond one billion time series: indexing and mining very large time series collections with i SAX2+. Knowl. Inf. Syst. 39(1): 123-151.

R. Cole, D. E. Shasha, X. Zhao, 2005: Fast window correlations over uncooperative time series. Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD), Chicago, USA, 21-24 August, 2005.

M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, I. Stoica, 2012: Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. Proceedings of USENIX Symposium on Networked Systems Design and Implementation. Chicago, USA, 3-5 April 2012.Siesler, H. W., Ozaki, Y., Kawata, S., & Heise, H. M. (Eds.). (2008). Near-infrared spectroscopy: principles, instruments, applications. John Wiley & Sons.

I. S. Helland, 1990, "Partial least squares regression and statistical models," *Scand. J. Stat.*, vol. 17, no. 2, pp. 97–114.

S. de Jong, 1993, "SIMPLS: An alternative approach to partial least squares regression," *Chemom. Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263.

M. Tenenhaus, 1998, *La régression PLS: théorie et pratique*. Paris: Editions Technip.

S. Wold *et al*, 2001, "PLS-regression: a basic tool of chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. Journal of chemometrics, 17(3), 166-173.

Gerbanowski, A., Rutledge, D. N., Feinberg, M. H., & Ducauze, C. J. (1997). Multivariate regression applied to time domain-nuclear magnetic resonance signals: determination of moisture in meat products. Sciences des aliments, 17(3), 309-323.

Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., ... & Wold, S. (2004). Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). Analytical and bioanalytical chemistry, 380(3), 419-429.

J. Shenk *et al*, 1997, "Investigation of a LOCAL calibration procedure for near infrared instruments," *J. Infrared Spectrosc.*, vol. 5, no. 1, p. 223.
V. Centner and D. L. Massart, 1998, "Optimization in Locally Weighted Regression," *Anal. Chem.*, vol. 70, no. 19, pp. 4206–4211.
L. Ramirez-Lopez *et al*, 2013, "The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets," *Geoderma*, vol. 195–196, pp. 268–279.
F. Allegrini *et al*, 2016, "Regression models based on new local strategies for near infrared spectroscopic data," *Anal. Chim. Acta*, vol. 933, pp. 50–58.
W. S. Cleveland, 1979, "Robust Locally Weighted Regression and Smoothing Scatterplots," *J. Am. Stat. Assoc.*, vol. 74, no. 368, p. 829.
M. Clairotte *et al.*, 2016, "National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy," *Geoderma*, vol. 276, pp. 41–52.
H. Tran *et al*, 2010, "'Global' and 'local' predictions of dairy diet nutritional quality using near infrared reflectance spectroscopy," *J. Dairy Sci.*, vol. 93, no. 10, pp. 4961–4975.
F. Davrieux *et al.*, 2016, "LOCAL regression algorithm improves near infrared spectroscopy predictions when the target constituent evolves in breeding populations," *J. Infrared Spectrosc.*, vol. 24, no. 2, p. 109.
Bevilacqua, M., & Marini, F. (2014). Local classification: Locally weighted–partial least squares-discriminant analysis (LW–PLS-DA). Analytica chimica acta, 838, 20-30.
T. Cover and P. Hart, 1967, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27.
P. Hart, 1968, "The condensed nearest neighbor rule (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 515–516.
Kim, Y. J., & Patel, J. (2010). Performance Comparison of the R*-Tree and the Quadtree for kNN and Distance Join Queries. IEEE Transactions on Knowledge and Data Engineering, 22(7), 1014-1027.
Wang, X. (2011, July). A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality. In Neural Networks (IJCNN), The 2011 International Joint Conference on (pp. 1293-1299). IEEE.
S.-A. Berrani *et al*, 2002, "Recherche par similarités dans les bases de données multidimensionnelles : panorama des techniques d'indexation," *Ingénierie Systèmes Inf.*, vol. 7, no. 5–6, pp. 9–44.
M. Muja and D. G. Lowe, 2014, "Scalable Nearest Neighbor Algorithms for High Dimensional Data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240.
D. E. Knuth, 1997, *The art of computer programming*, 3rd ed. Reading, Mass: Addison-Wesley.
C. H. Park and S. B. Kim, 2015, "Sequential random k-nearest neighbor feature selection for high-dimensional data," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2336–2342.
M. García-Torres *et al*, 2016, "High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach," *Inf. Sci.*, vol. 326, pp. 102–118.
V. E. Vinzi *et al*, 2005, "PLS Typological Regression: Algorithmic, Classification and Validation Issues," in *New Developments in Classification and Data Analysis*, H.-H. Bock, W. Gaul, M. Vichi, P. Arabie, D. Baier, F. Critchley, R. Decker, E. Diday, M. Greenacre, C. Lauro, J. Meulman, P. Monari, S. Nishisato, N. Ohsumi, O. Opitz, G. Ritter, M. Schader, C. Weihs, M. Vichi, P. Monari, S. Mignani, and A. Montanari, Eds. Berlin/Heidelberg: Springer-Verlag, pp. 133–140.
C. Preda and G. Saporta, 2005, "Clusterwise PLS regression on a stochastic process," *Comput. Stat. Data Anal.*, vol. 49, no. 1, pp. 99–108.
T. Liu *et al*, 2005, "An investigation of practical approximate nearest neighbor algorithms.," *Adv. Neural Inf. Process. Syst.*, pp. 825–832.
V. Hyvönen *et al.*, 2016, "Fast k-nn search," *arXiv*, p. preprint arXiv:1509.06957.
M. Slaney and M. Casey, 2008, "Locality-Sensitive Hashing for Finding Nearest Neighbors [Lecture Notes]," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 128–131.
L. Paulevé *et al*, 2010, "Locality sensitive hashing: A comparison of hash function types and querying mechanisms," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1348–1358.
K. Berlin *et al*, 2015, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nat. Biotechnol.*, vol. 33, no. 6, pp. 623–630.
Camerra, A., Palpanas, T., Shieh, J., & Keogh, E. (2010, December). iSAX 2.0: Indexing and mining one billion time series. In Data Mining (ICDM), 2010 IEEE 10th International Conference on (pp. 58-67). IEEE.
Camerra, A., Shieh, J., Palpanas, T., Rakthanmanon, T., & Keogh, E. (2014). Beyond one billion time series: indexing and mining very large time series collections with iSAX2+. Knowledge and information systems, 39(1), 123-151.

Romain Perriot, Jérémy Pfeifer, Laurent d'Orazio, Bruno Bachelet, Sandro Bimonte, Jérôme Darmont. “Cost Models for Selecting Materialized Views in Public Clouds”. 2014 International Journal of Data Warehousing and Mining (IJDWM) 10(4)

Sandro Bimonte, “Current Approaches, Challenges, and Perspectives on Spatial OLAP for Agri-Environmental Analysis.” IJAEIS 7(4): 32-49 (2016)