Projet of PhD Thesis

Adaptation of two big data indexing algorithms for the improvement of the local-PLS method in chemometrics

Context and problematics:

Near-infrared spectrometry (NIRS, Siesler, 2008) is seen as a technique that can provide a considerable amount of data to digital agriculture (Bellon-Maurel, 2015). This measurement technique is closely related to chemometrics, which makes it possible to transform acquired spectra into useful information (Siesler, 2008). For many years, chemometrics has been offering regression tools to link spectra (particularly infrared) with chemical (concentration) or qualitative (class) quantities. Among these tools, the Partial Least Squares regression (PLSR, Helland, 1990, De Jong, 1993, Tenenhaus, 1998, Wold, 2001) and its PLS-Discriminant Analysis variant (PLS-DA, Barker, 2003) worked very well on small databases, when the link between spectra and answers to predict is rather univocal. The general principle of the PLS is to compress the spectral matrix (whose columns are highly correlated with each other) into a reduced number of orthogonal axes (called latent variables) correlated with the response, which are then used as explanatory variables of a multiple linear regression or discriminant analysis model. PLS therefore responds very well to the problems posed by a large number of variables; it has also been used for a long time in NMR spectrometry (Gerbanowski, 1997), and more recently in genomics, proteomics and metabolomics (Eriksson, 2004).

New applications have recently emerged in chemometrics to model on "all-in-one" databases with large numbers of individuals and more variations in influence quantities (eg type of equipment and biological material or the location and year of collection). Thus in agronomy, we find spectral databases of soils, cereals, feed that can contain more than 10 000 individuals. The usual PLS quickly reaches its limits in facingf the heterogeneity present in these bases, with variances and prediction bias often too high.

One response currently being explored by chemometricians is the "local" PLS regression (Shenk, 1997, Centner, 1998, Ramirez-Lopez, 2013, Allegrini, 2016), reflecting the idea of local regression (Cleveland, 1979). For each spectrum to be processed, the method consists first of all in finding the neighbors of the spectrum (thus reducing the heterogeneity), then in making the predictions by means of a usual PLSR carried out on the determined neighborhood. Different applications have shown the effectiveness of the method in agronomy, for example in the context of soil analyzes (Clairotte, 2016), milk and animal faeces (Tran, 2010) and annual food crops (Davrieux, 2016). The same idea can be applied to discrimination issues by replacing PLSR with PLS-DA (Bevilacqua, 2014).

A critical point of local PLS methods, both in terms of statistical efficiency and computing time, is the selection of the neighborhood. Paradoxically, very little research has been done on this subject in the field of chemometrics. The current local PLS algorithms all use the nearest-neighbor algorithm (k-NN) linear or sequential ("brute-force method"): for each spectrum to be processed, the algorithm calculates the distances between this spectrum and the n spectra of the base, order the distances then deduce the nearest neighbors (with some possible variants: Allegrini, 2016). Linear search has the advantage of being simple but suffers from a slowness problem when the database contains a lot of spectra. It becomes extremely expensive to test the n points of the spectral space to deduce the neighborhood, with a computation time of order O(n). Calculation times quickly become prohibitive for conducting standard protocols for constructing and evaluating predictive models (model selection and estimation of uncertainties) such as cross-validation. The problems will become insurmountable if one thinks to treat, in the future, databases even more important than the current bases (> million individuals). Other algorithms must be considered to ensure the sustainability of the local PLS method in a context of ever larger databases. Until now, all the optimizations of the local PLS consisted of reducing the space of the

spectra in a subspace carrying useful information (by a PCA or a PLS) and of calculating distances (K-NN) in these useful spaces. Another solution, already used for time series (Yagoubi-1, 2017), is to make another reduction of space (by indexing) which makes it possible to place all the individuals of the base in grids or trees whose scanning is very fast.

Although the k-NN approach is old (Cover, 1967, Hart, 1968), the recent emergence of big data issues has increased the intensity of research on neighborhood computing algorithms (crucial methodological point of local PLS). Databases containing large numbers of objects and described in large spaces create difficulties that are faced by all indexing techniques for the search for similarities. Methodological research focuses on the development of sophisticated indexing algorithms for structuring data, and high-performance search algorithms for effective access (Wang, 2011). For example, panoramas of multidimensional indexing techniques for neighborhood research are presented in (Berrani, 2002) and (Muja, 2014). The algorithms of "exact" vs. "approximate" search for nearest neighbors are distinguished. In the first case, we look for the "real" nearest neighbors, in the second case we accept, using a probabilistic approach, a certain level of error. In general, the exact methods rely on the development of pre-structuring of the data to reduce the number of distances to be calculated (tree-based methods Knuth, 1997, Kim, 2010). They can be very effective for spaces of small dimension (2-3 descriptors) or moderate (<10 to 20 descriptors). But the "curse of dimensionality" makes it difficult to apply them to spectral data. Dimension reduction techniques can circumvent this problem, especially those based on relevant descriptor selections (Park, 2015, García-Torres, 2016) or pre-estimated prediction models (as in "SIMCA" type PLS approaches). Or "clusterwise / typological" Vinzi, 2005, Preda, 2005). Other alternatives are the approximate search methods for high-dimensional indexing (Muja, 2014, Liu, 2005, Hyvönen, 2016). They cause a decrease in the accuracy of the result but in return allow a sharp reduction in the calculation time. A strong craze developed around these methods, especially those using random "hash" techniques such as "Local Sensitive Hashing" (Slaney, 2008, Paulevé, 2010). For example, many applications have been proposed in bioinformatics for comparing genomic sequences (Berlin, 2015).
There are other methods of reducing dimensions used especially for time series. For example, the iSax representation "Symbolic Aggregate Approximation" (Camerra, 2014) can be used to create efficient indexes on very large databases. There is also the random vector-based method (Cole, 2005) that produces "sketches" from the original data, and then the sketches are used to search for nearest neighbors (kNN).

Although index-based kNN search methods allow time savings of several orders of magnitude compared to sequential scanning, when they are centralized their performance deteriorates as the size of the data increases. This raises questions about the ability of these centralized methods to scale. To cope with the increasing volume of data, a promising solution is to exploit parallel frameworks, such as Spark (Zaharia, 2012), to create powerful computing and storage units using usual machines. There are different methods for creating parallel indexes on large databases, including techniques that produce tree indexes for data represented by iSAX (Yaghoubi-1, 2017), as well as those based on parallel hashing of sketches. (Yaghoubi-2, 2017).

Aim of the thesis:

The aim of the thesis is to test and adapt big data techniques to make local PLS algorithms compatible with large (typically> 20,000 individuals) and very large (> 10^6 individuals) databases. Two indexing methods, intensively studied by the Zenith team of Lirmm (participant in the thesis project), will be explored:

- Hashing (in particular, sketch calculation)
- Tree indexing of data represented by iSAX.

For each of these two techniques, the following methodology will be followed:

- Direct application of existing algorithms (hash and iSAX) on spectral databases; evaluation of performance gains in terms of calculation time and performance gains or losses in terms of prediction accuracy.
- Test of the influence of different spectral pretreatments commonly used in NIRS; definition of optimal pretreatment to make spectra compatible with iSAX algorithms and sketches.
- Adaptation of iSAX search algorithms and sketches to the structural particularities of the NIR spectra.

The methods produced will be tested on one or more applications directly useful to digital agriculture:

- From the NIR RMQS spectra database, which contains the representative spectra of the soils of France (one spectrum every 16 km), use the modified local PLS, to predict the characteristics of agricultural soils at a finer scale, compatible with remote sensing images.
- The neighborhood algorithm, developed as part of this thesis for local PLS, will also be tested on other NIRS applications, such as unsupervised analysis of hyperspectral image series, thus opening the field to other research questions that will be presented as perspectives.

Research questions:

The main research questions will be:

- Can hashing algorithms (sketches) and iSAX be used on NIR spectra?
- Which spectral dimensions should be used in the hash and iSAX algorithms to optimize neighbor search and PLS predictions?
- For the hashing, how best to define new sketches for the NIRS? Are there specific NIR spectra dimensions that allow for more accurate hashing?
- For iSAX, how best to discretise a NIR spectrum before indexing by tree? Is there an alternative symbolic representation, specific to NIR spectra, which offers better performances?

Monitoring:

Thesis direction: Jean Michel Roger (Irstea - ITAP) / Matthieu Lesnoff (Cirad - SELMET).
Monitoring: Nathalie Gorretta (Irstea - ITAP).

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

Committee : Juan Antonio Fernandez Pierna (CRA Gembloux) + Dino Ienco (Irstea - TETIS) + local

Procedure and program of the thesis

1) The candidate will first have to realize a state of the art:

- different methods used in chemometrics to implement local regressions
- associated database indexing and neighborhood search methods in the big data domain (hash, iSax)

2) the different methods identified will be tested on synthetic datasets, produced from real sets (spectra of furage, soil, ...) and simulated. Sensitivity to the number of individuals and wavelengths will be studied.
3) The algorithms identified as the most promising will be modified to produce one or more methods adapted to local PLS regression and discrimination.
4) Application based on soil data; evaluation of performance gains; projection on an embedded use, with or without use of a cloud.

For a duration of 3 years, the thesis will run according to the provisional calendar below

semester	1	2	3	4	5	6
state of the art PLS / big data	x	x				

test of existing methods		x	x			
review paper		x	x	x		
develop. test of methods, application			x	x	x	
method. paper				x	x	x
manuscript defense					x	x

Collaborations

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

Dissemination:

- Chemometrics / Francophone Annual Conferences
- CAC 2020 Conference
- An article reviewing and testing methods will be highly appreciated by the chemometrics community. This will be a first article submitted to Chemometrics and Intelligent Laboratory Systems, or a Journal of Chemometrics
- A methodological article in Chemometrics and Intelligent Laboratory Systems, or Journal of Chemometrics, or Analytica Chimica Acta.
- Development of a R package dedicated to local methods; inclusion in ChemFlow software
- We will also examine the opportunity for communication in the big data community.

Localisation

University of Montpellier

candidate profile

The candidate will be required to have a background in chemometrics or applied statistics (with sensitivity for computer science), or computer training with high sensitivity for chemometric applications. Experience in life sciences, either through initial training or through the end of studies internship will be a plus.

References

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)