Adaptation of two Big-data indexing algorithms for LOCAL-PLS

Maxime Metz[1], Jean-Michel Roger[1], Matthieu Lesnoff[2], Reza Akbarinia[3], Florent Masseglia[3]

[1]ITAP, Univ Montpellier, Irstea, Montpellier SupAgro, Montpellier, France
[2]SELMET, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
[3]Inria & LIRMM, Univ Montpellier, France

Context: Nowadays, considerable amounts of heterogeneous data can be generated. LOCAL-PLS can provide a potential solution to this situation. However, using LOCAL-PLS with large amounts of data is almost impossible.

How to associate LOCAL-PLS with big-data algorithms?

LOCAL-PLS [1]

Selection and weighting of nearest neighbors → Optimization of PLS model → Final PLS model

Operational issues :
- Extremely long calculation time
- Choice of neighborhood not always relevant (types of compressions / metrics)

Scientific issues

? How to define the relevance of a **neighbor**?

? How to define the relevance of a **neighborhood**?

? How to adapt big-data algorithms to obtain a directly relevant neighborhood?

Envisaged directions

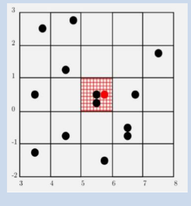
Modification of big-data algorithms for optimal direct prediction

Use of iSAX / ParCorr algorithms and adapted LOCAL-PLS

The « Big-data » algorithms

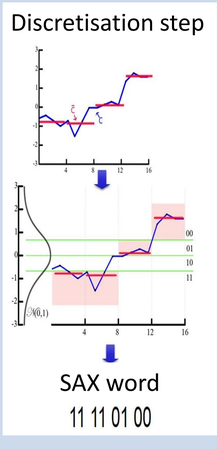
ParCorr [2]

ParCorr generates a set of random vectors, and realizes the product of vectors by the spectral matrix. This method makes it possible to reduce the size of the spectral matrix. Then, the matrix is cut in a group of columns and stored in different grids. When, searching neighbors, ParCorr looks at the grids in parallel, and the spectrums that are in the same cells (in a minimum number of grids) are good neighbor candidates.



iSAX [3]

iSAX discretizes the variables and creates a SAX word to summarize each spectrum. The system is represented by a tree that will become deeper depending on the final number of neighbors to get.



Bibliography

[1] Shenk, John S., Mark O. Westerhaus, et Paolo Berzaghi. « Investigation of a LOCAL Calibration Procedure for near Infrared Instruments ». Journal of Near Infrared Spectroscopy 5, n° 4 (octobre 1997): 223-32.

[2] Yagoubi, Djamel Edine, Reza Akbarinia, Boyan Kolev, Oleksandra Levchenko, Florent Masseglia, Patrick Valduriez, et Dennis Shasha. « ParCorr: Efficient Parallel Methods to Identify Similar Time Series Pairs across Sliding Windows ». Data Mining and Knowledge Discovery 32, n° 5 (septembre 2018): 1481-1507.

[3] Shieh, Jin, et Eamonn Keogh. « I SAX: Indexing and Mining Terabyte Sized Time Series ». In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08, 623. Las Vegas, Nevada, USA: ACM Press, (2008).

Email : maxime.metz@irstea.fr