2019-01377 - PhD Position F/M [Sub - 2019] -
Unsupervised learning of huge datasets with limited computer resources

Type de contrat : CDD de la fonction publique
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Doctorant

A propos du centre ou de la direction fonctionnelle

The Inria Lille - Nord Europe Research Centre was founded in 2008 and employs a staff of 360, including 300 scientists working in sixteen research teams. Recognised for its outstanding contribution to the socio-economic development of the Nord - Pas-de-Calais Region, the Inria Lille - Nord Europe Research Centre undertakes research in the field of computer science in collaboration with a range of academic, institutional and industrial partners.

Contexte et atouts du poste

Today’s datasets are increasingly produced by automatic processes, like specialized sensors in the real world or web data acquisition in the virtual world. As a consequence, volume of today’s datasets are at present indefinitely growing, leading irredeemably to so-called huge datasets.

This matter of fact has both a positive and a negative effect:

• The positive effect of huge datasets is the opportunity to discover unexpected information, and thus to create new knowledge for end users. This so-called unsupervised learning task (also known as clustering) is an essential component of artificial intelligence.

• The negative effect is to likely reach physical limits of computers that the end user has at hand. This limit can be for example an overflow of memory capacity, an expected maximum computing speed overflow (ex.: real time target) or an overflow of the available energy (ex.: battery of a laptop).

The scientific lock that we propose to address through this research project is to design a specific unsupervised learning approach which automatically tunes a trade-off between statistical accuracy (while limiting the accuracy loss) and the available computer resources (while strictly respecting the physical limits). Virtually, the proposed solution has to be natively scalable to any dataset volume, including current and future huge ones.

Mission confiée

Model-based clustering is now a well-recognized paradigm for addressing unsupervised learning.

First, it embeds clustering into a probabilistic well-posed environment, allowing to reformulate all the classical clustering questionings in the framework of the classical mathematical statistics (estimation theory, model selection theory). Mixture models of distributions are the keystone of this approach and many “traditional” clustering methods can be reformulated as specific mixture models (Biernacki 2017).

Second, this approach has the great advantage to be extremely flexible, allowing model extensions devoted to specific cases of new datasets like the high dimensional problem (Biernacki & Maugis 2017). Third, model-based clustering meets an uncountable quantity of empirical successes on real datasets attached to a wide variety of human activities since several decades (McLachlan & Peel 2004), having now well known packages or softwares for being used by a large variety of end users (Lebre et al. 2015, Scrucca et al. 2016).

The question we address in this PhD project is to scale up the general paradigm of model-based clustering to the case of huge datasets, with a special attention to more and more frequent high dimensional situations. In particular, co-clustering (Govaert & Nadif 2014) will be a suitable high dimensional clustering methodology as recently illustrated by Keribin & Biernacki (2018). The term “huge” has to be understood as an oversized dataset with regards to the available computer resources as explained above.

The solution we propose is to quantize the dataset, meaning the reduction of an initial dataset with say N individuals into a new dataset with say n ≤ N weighted individuals.

In the multidimensional continuous case typically, it corresponds to build multidimensional intervals (a grid), leading to what is known as binned data (Same et al. 2006, Wu 2014). Advantage of binned data over more classical solutions as subsampling is its theoretical ability to still detect low frequency groups. This property is especially important since interest to collect huge datasets is often guided by the possibility to detect new, thus infrequent but potentially significant, signal. Indeed, mathematical statistics tell us that estimate consistency (asymptotically in N) is maintained in the binned case with a regular grid, only the speed of consistency being a-ected. By reducing drastically the dataset size from N to n (with weights), it is clear that the required memory, computing energy and computing time can be drastically reduced. However, the binned grid in a high dimension space can no longer be regular because of the curse of dimensionality ("empty space" paradigm, Bouveyron & Brunet (2013)). Hence, the main question to
be addressed in this work is its automatic choice, while preserving both the finite sample ability to detect "small" clusters and the estimate consistency (until now guaranteed with a regular grid or at least a predefined one). Thus the estimation of an adaptive grid is required with a specific trade-off between the statistics theory (estimation quality) and the available computer resources. Until now, the question of the grid estimation was not addressed yet since only low or middle dimensional situations have been under study (Same et al. 2006, Wu 2014), cases where a regular grid was still admissible.

References

Biernacki, C. (2017), Mixture models, in J.-J. Droesbeke, G. Saporta & C. Thomas-Agnan, eds, ‘Choix de mod`eles et agr`egation’, Technip. URL: https://hal.inria.fr/hal-01252671


Bouveyron, C. & Brunet, C. (2013), ‘Model-Based Clustering of High-Dimensional Data: A review’, Computational Statistics and Data Analysis 71, 52–78. URL: https://hal.archives-ouvertes.fr/hal-00750909

Govaert, G. & Nadif, M. (2014), Co-Clustering, Computing Engineering series, ISTE-Wiley. URL: https://hal.archives-ouvertes.fr/hal-00933301

Keribin, C. & Biernacki, C. (2018), Co-clustering: A versatile way to perform clustering in high dimension, in ‘The 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMSStatistics 2018)’, Pise, Italy. URL: https://hal.archives-ouvertes.fr/hal-01949116


Principales activités

The question we address in this PhD project is to scale up the general paradigm of model-based clustering to the case of huge datasets, with a special attention to more and more frequent high dimensional situations. In particular, co-clustering (Govaert & Nadif 2014) will be a suitable high dimensional clustering methodology as recently illustrated by Keribin & Biernacki (2018).

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours)
- Access to vocational training
- Social security coverage

Rémunération

1st and 2nd year : 1593.50€ Net monthly salary (after taxes)

3rd year : 1676.31€ Net monthly salary (after taxes)