Research internship

Title : *Distributed Multi-Coclustering*

Keywords: Clustering, Coclustering, Multi-Coclustering, Bayesian Non Parametric, Mixture Models, Distributed Computing, Map-Reduce, Spark

1 Context

Clustering is an unsupervised learning method that seeks to partition data into similar groups. In the multivariate case, the clustering method only infers a row partition, whereas coclustering [6] infers simultaneously one row-partition and one column-partition. The resulting partition is composed of homogeneous blocks. To address the coclustering problem, the authors [5] introduced the Latent Block Model (LBM). This model assumes the presence of hidden block components such that all the elements belonging to the same block are drawn independently from the same distribution.

Coclustering assumes that the row-partition is shared along all the variables, when this is not the case, coclustering is not convenient. Multiple clustering omits this assumption by inferring a column-partition and a single row-partition for each column-partition. Authors in [4] proposed a functional parametric multiple clustering model designed to deal with multivariate time series. The parametric model-based methods assume that the true number of clusters is known a priori. This assumption is generally not satisfied in practice. Model selection strategies are performed to estimate the best number of clusters. However, the number of possible models is enormous even for a small number of blocks. Consequently, an exhaustive search is not feasible.

Bayesian Non-Parametric (BNP) modeling allows the automation of the components number estimation, which is performed during the inference by making a prior distribution over the model parameters. A BNP extension of the LBM (NPLBM) has been introduced in [7], this model makes two separate prior on the proportions and a prior on the block component distribution.

Multi-Coclustering [3] is a Bayesian non-parametric block structure model that combines two ideas, one idea from multi-clustering by grouping variables with the same row-partition, and a second idea from coclustering by applying a specific NPLBM structure to each column cluster. Multi-Coclustering is efficient for multiview data clustering, however, it is highly time-consuming.

Recently, more and more parallel processing techniques and frameworks are coming are implemented and used in many areas, the most popular frameworks are MapReduce [2] and Spark [8]. MapReduce is one of the most popular solutions for big data processing [1], in particular, due to its automatic management of parallel execution in computing clusters. Initially proposed in [2], it was popularized by Hadoop [1], an open-source implementation. Apache Spark [8] is also an open-source computing cluster framework that was initially developed by a research group from the University of California, Berkeley, to deal with the problems that can not be handled by MapReduce. Spark introduces multi-stage in-memory primitives that overcome disk bottlenecks and provide performance up to 100 times faster for certain applications.

2 Objectives

The main purpose of this internship is to propose a distributed implementation of Multi-Coclustering using Spark. The first research directions might be:

- Study the current state-of-the-art, with a focus on distributed model-based unsupervised learning, clustering.
- Propose an implementation for Bayesian non-parametric LBM (NPLBM) which is the second layer of Multi-Coclustering.
- Propose an implementation for the multi-clustering layer which is the first layer of Multi-Coclustering.
- Propose a complete distributed implementation of Multi-Coclustering by combining the two distributed layers.

3 Requirements and skills

- End of engineering degree, M2 in data science, statistics, artificial intelligence, or computer science.
- Excellent understanding of machine learning basics, particularly probabilistic models.
- Excellent programming skills, especially with Python and Scala/Spark.
- Comfort turning ideas into code.

4 Internship information

Internship location and duration

The internship will take place in the computer science lab (Laboratoire d'Informatique de Paris Nord - LIPN) located at Université Sorbonne Paris Nord, for a duration of 6 months.

Supervisor team

- Reda Khoufache[1,2]
- Mustapha Lebbah[1]
- Hanane Azzag[2]

How to apply

To apply, simply attach:

- Your current Curriculum Vitae (CV).
- Your motivation for the position.
- Your latest university transcripts.

Send it all by email to: mlresearch.internship@gmail.com

The internship may lead to a PHD position.

[1]DAVID Lab, UVSQ, Université de Paris Saclay
[2]LIPN UMR 7030, Université Sorbonne Paris Nord

References

[1] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling. The meaningful use of big data: Four perspectives – four challenges. *SIGMOD Rec.*, 40(4):56–60, jan 2012.

[2] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, jan 2008.

[3] E. Goffinet, M. Lebbah, G. Azzag, H., L., and A. Coutant. Multivariate time series multi-coclustering. application to advanced driving assistance system validation. *ESANN 2021-29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. ESANN.*, 2021.

[4] E. Goffinet, M. Lebbah, H. Azzag, G. Loïc, and A. Coutant. Functional non-parametric latent block model: A multivariate time series clustering approach for autonomous driving validation. *Computational Statistics & Data Analysis*, page 107565, 2022.

[5] G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36:463–473, 02 2003.

[6] G. Govaert and M. Nadif. Co-clustering: Models, algorithms and applications. 2013.

[7] T. Meeds and S. Roweis. Nonparametric bayesian biclustering. Technical Report UTML TR 2007–001, January 2007.

[8] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10)*, Boston, MA, June 2010. USENIX Association.