Massively Distributed Dirichlet Process Mixture Models

Khadidja Meguelati [1], **Benedicte Fontez** [2], **Nadine Hilgert** [2], **Florent Masseglia** [1]

1. *Inria and LIRMM, Montpellier, France*
   firstname.lastname@inria.fr
2. *MISTEA, Univ. Montpellier, Montpellier SupAgro, INRA, Montpellier, France*
   benedicte.fontez@supagro.fr,nadine.hilgert@inra.fr

ABSTRACT. *Clustering with accurate results have become a topic of high interest. Dirichlet Process Mixture (DPM) is a model used for clustering with the advantage of discovering the number of clusters automatically. It is highly time consuming, which impairs its adoption and makes centralized DPM approaches inefficient. We propose DC-DPM (Distributed Clustering via DPM), a parallel clustering solution that gracefully scales to millions of data points while remaining DPM compliant, which is the challenge of distributing this process. Our experiments, on both synthetic and real life data, illustrate the high performance of our approach. The centralized algorithm does not scale and has its limit on 100K data points, where it needs more than 7 hours. In this case, our approach needs less than 30 seconds.*

RÉSUMÉ. *La classification non supervisée (ou clustering) a pour objectif d'identifier des classes pertinentes dans les données. Le mélange de processus de Dirichlet (DPM) est utilisé pour le clustering car il définit automatiquement le nombre de classes mais les temps de calculs qui l'impliquent sont généralement trop importants, nuisant à son adoption et rendant inefficaces ses versions centralisées. Dans la logique du DPM, nous proposons DC-DPM, une version parallélisée, qui s'adapte à des millions de points de données, ce qui représente un vrai défi. Nos expérimentations, tant sur des données synthétiques que réelles, illustrent la performance de notre approche. Comparativement, l'algorithme centralisé ne passe pas à l'échelle. Son temps de réponse est de plus de 7 heures sur des données de 100K points, quand notre approche prend moins de 30 secondes.*

KEYWORDS: *Dirichlet Process Mixture Model, Clustering, Parallelism*

MOTS-CLÉS : *Modèle de mélange de processus de Dirichlet, Classification non supervisée, parallélisme*

Distributed Clustering via DPM

One of the main difficulties, for clustering, is the fact that we don't know, in advance, the number of clusters to be discovered. In (Meguelati *et al.*,) we focus on the DPM approach since it allows estimating the number of clusters and assigning observations to clusters, in the same process. Unfortunately, DPM is highly time consuming. Consequently, several attempts have been done to make it distributed. However, while being effectively distributed, these approaches usually suffer from convergence issues (imbalanced data distribution on computing nodes) (Lovell *et al.*,) or do not fully benefit from DPM properties (Wang, Lin,). Furthermore, making DPM parallel is not straightforward since it must compare each record to the set of existing clusters, a highly repeated number of times. That impairs the global performances of the approach in parallel, since comparing all the records to all the clusters would call for a high number of communications and make the process impracticable. We propose DC-DPM (Distributed Clustering by Dirichlet Process Mixture), a distributed DPM algorithm that allows each node to have a view on the local results of all the other nodes, while avoiding exhaustive data exchanges. The main novelty of our work is to propose a model and its estimation at the master level by exploiting the sufficient statistics from the workers, in a DPM compliant approach. Our solution takes advantage of the computing power of distributed systems by using parallel frameworks such as MapReduce or Spark (Zaharia *et al.*,). Our DC-DPM solution distributes the Dirichlet Process by identifying local clusters on the workers and synchronizing these clusters on the master. These clusters are then communicated as a basis among workers for local clustering consistency. We modified the Dirichlet Process to consider this basis in each worker. By iterating this process we seek global consistency of DPM in a distributed environment. Our experiments, using real and synthetic datasets, illustrate both the high efficiency and linear scalability of our approach. We report significant gains in response time, compared to centralized DPM approaches, with processing times of a few minutes, compared to several days in the centralized case.

Acknowledgements: The research leading to these results has received funds from the European Union's Horizon 2020 Framework Programme for Research and Innovation, under grant agreement No. 732051.

References

Lovell D., Adams R. P. Mansingka V. (2012). Parallel markov chain monte carlo for dirichlet process mixtures. In *Workshop on big learning, nips.*

Meguelati K., Fontez B., Hilgert N. Masseglia F. (2019, April). Dirichlet process mixture models made scalable and effective by means of massive distribution. In *SAC: Symposium on Applied Computing.* Limassol, Cyprus. Retrieved from https://hal.archives-ouvertes.fr/hal-01999453

Wang R. Lin D. (2017). Scalable estimation of dirichlet process mixture models on distributed data. In *Proceedings of the 26th international joint conference on artificial intelligence*, pp. 4632–4639. AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=3171837.3171935

Zaharia M., Chowdhury M., Franklin M. J., Shenker S. Stoica I. (2010). Spark: Cluster computing with working sets. In *Hotcloud.*