ÉLÉMENT DE PORTFOLIO 05

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : A new coreset framework for clustering, par Vincent Cohen-Addad, David Saulpic et Chris Schwiegelshohn, article présenté à la conférence *Symposium on Theory of Computing* (STOC) 2021.

URL de l'élément : `https://hal.sorbonne-universite.fr/hal-03505350`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

L'étude des problèmes de clustering a connu ces dernières années un regain d'intérêt très fort par la communauté de l'informatique théorique, de part notamment son importance en intelligence artificielle. Une approche féconde dans la résolution de ces problèmes consiste à construire des *coresets*. Intuitivement, cela consiste à remplacer l'ensemble de points initiaux (que l'on doit partitionner) par un ensemble de points, idéalement beaucoup plus petit, équivalent en terme de résolution (cf le prochain paragraphe pour une définition précise). L'obtention de *coresets* petits permet donc de condenser la donnée du problème - et d'en améliorer la résolution.

Dans cet article, nous présentons une nouvelle technique générique pour construire des *coresets*. Cette technique tranche avec les précédentes, pour plusieurs raisons. D'abord, elle permet de retrouver, et même d'améliorer, les précédents résultats connus, de façon particulièrement simple et concise. Ensuite, en terme de techniques, notre algorithme de construction commence par partitionner l'entrée en groupes qui ont des propriétés fortes, que l'on exploite ensuite. Les précédentes analyses de l'état de l'art [1, 3, 5, 8, 9] n'essayaient pas de tirer profit d'une quelconque structure. Notre façon de partitioner l'entrée a ensuite essaimée, et plusieurs nouveaux résultats s'en inspirent [2, 6, 7]. Nous appliquons notre technique pour calculer des coresets pour les problèmes de clustering k-median et k-means quand l'espace métrique d'entrée vérifie différentes propriétés : par exemple, quand c'est un espace euclidien $(\mathbb{R}^d, \ell_2)$, quand la métrique est induite par un graphe de *treewidth* bornée ou un graphe provenant d'une famille excluant un mineur.

Par ailleurs, nous avons montré par la suite [4] que certains des résultats obtenus dans l'article sont optimaux : en plus d'être assez générale et de s'appliquer à de nombreux espaces métriques différents, notre technique est donc aussi très précise.

3 PRÉSENTATION DE CET ÉLÉMENT

Cet article présente donc une construction de *coresets* pour les problèmes de clustering k-median et k-means. Pour simplifier, concentrons-nous sur k-median dans l'espace $(\mathbb{R}^d, \ell_2)$. Un coreset est une façon de compresser l'entrée tout en préservant la fonction de coût de k-median : plus précisemment, étant donné $\varepsilon > 0$, un ε-coreset pour P est un ensemble Ω tel que, pour tout ensemble de k centres $\mathcal{S}$,

$$cost(\Omega, \mathcal{S}) = (1 \pm \varepsilon) cost(P, \mathcal{S}).$$

Pour contruire un tel Ω, la technique la plus répandue est de construire une distribution (bien choisie) de probabilité π sur P, de tirer l'ensemble Ω selon cette distribution et d'utiliser des bornes de concentrations pour montrer qu'en choisissant Ω suffisamment grand, on obtient la garantie de coreset. La distribution communément analysée est appelée *sensitivity sampling* ou *importance sampling* : la probabilité de tirer un point est proportionnel à sa sensibilité, qui est sa contribution relative maximum dans n'importe quelle solution $\mathcal{S}$.

Cette distribution est complexe à approximer et à analyser. Nous proposons à la place de décomposer l'entrée en groupes, tel que dans chaque groupe les points sont essentiellement équivalents : utiliser une distribution uniforme dans chaque groupe permet de construire un coreset (essentiellement). Cela permet de simplifier grandement les outils conceptuels utilisés par notre analyse, mais aussi de construire des coreset de taille optimale, comme mentionné précédemment.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H. C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth, 2020.

[2] Vladimir Braverman, Vincent Cohen-Addad, Shaofeng H.-C. Jiang, Robert Krauthgamer, Chris Schwiegelshohn, Mads Bech Toftrup, and Xuan Wu. The power of uniform sampling for coresets. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS*. IEEE, 2022.

[3] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA*. SIAM, 2021.

[4] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In *Proceedings of the 54th ACMSymposium on Theory of Computing STOC*, 2022.

[5] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011,*, 2011.

[6] Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. Coresets for clustering with general assignment constraints. *CoRR*, 2023.

[7] Lingxiao Huang, Jian Li, and Xuan Wu. Towards optimal coreset construction for (k, z)-clustering : Breaking the quadratic dependency on k. *CoRR*, abs/2211.11923, 2022.

[8] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces : importance sampling is nearly optimal. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC*, 2020.

[9] Christian Sohler and David P. Woodruff. Strong coresets for k-median and subspace approximation : Goodbye dimension. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, 2018.