Learning-augmented Algorithms for Density Peaks Clustering

Yi-Fang Lee Chung-Shou Liao

National Tsing Hua University

Clustering is a fundamental task in unsupervised learning and data analysis that involves grouping a set of data points into clusters based on their similarities. The goal of clustering is to ensure that data points within the same cluster are more similar to each other than to those in different clusters, based on some predefined distance metric or criterion. Clustering finds a wide range of applications, including image segmentation, market segmentation, document classification, bioinformatics, and so on. Popular clustering methods include *k*-means, hierarchical clustering, DBSCAN, and spectral clustering, etc., each of which is suited for different data distributions and requirements. Each of them may have their own advantages and disadvantages [1].

To overcome the inherent limitations of traditional clustering methods, recent studies have begun exploring the use of learning-augmented approaches to achieve more accurate and precise clustering results. By integrating predictions from machine learning models, learning-augmented clustering can enhance the performance of conventional algorithms while retaining their theoretical guarantees. In the recent work [2], discussing learning-augmented algorithms for *k*-means and *k*-medians, Nguyen et al. proposed a deterministic *k*-means algorithm that produces centers with a better bound on clustering cost compared against previous randomized algorithms, achieving a cost of $(1 + O(\alpha))\,\mathrm{OPT}$ while preserving the $O(dm \log m)$ runtime, where d represents the ambient dimension.

However, many datasets cannot be effectively clustered using distance-based methods like *k*-means and *k*-medians. To address the limitation, we propose applying the learning-augmented paradigm to density-based clustering. In this study, we focus on the seed-and-extension-based density peaks clustering (SDP) proposed by Tung et al. [3]. We aim to pursue better clustering results by predicting the hyperparameters of SDP and leveraging the properties of the tree structure within SDP. In order to pave the way for more effective and adaptable settings for SDP, we will incorporate the conditions of the kNN (*k*th nearest neighbor) to constrain hyperparameter predictions and design learning-augmented algorithms specifically for density-based clustering.

Keywords: Density-Based Clustering, Learning-Augmented Algorithms, Prediction

References:

[1] Xu, D., Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165–193.

[2] Nguyen, T., Chaturvedi, A., & Nguyen, H. L. (2022). Improved Learning-augmented Algorithms for k-means and k-medians Clustering. In Proc. International Conference on Learning Representation (ICLR 2022), ArXiv preprint arXiv:2210.17028.

[3] Tung, M.-H., Chen, Y.-P. P., Liu, C.-Y., & Liao, C.-S. (2023). A Fast and More Accurate Seed-and-Extension Density-Based Clustering Algorithm. IEEE Trans. on Knowl. and Data Eng., 35(6), 5458–5471.