technische universität dortmund

computer science 12

Master Thesis

Exploration of K-means Clustering Algorithms on Low-dimensional Datasets

Dr.-Ing. Kuan-Hsun Chen
Prof. Dr. Jian-Jia Chen
Otto-Hahn Str. 16
Technische Universität Dortmund
Email: kuan-hsun.chen@tu-dortmund.de
October 6, 2020

The k-means clustering is one of the popular problems in data mining and machine learning due to its simplicity and applicability. A given set of multi-dimensional data points should be classified into different clusters according to their similarity to enable the same treatment per cluster. However, the de facto standard solution, i.e., Lloyd's k-mean algorithm [3], likely might suffer from a large amount of time on the distance calculations. However, the bottleneck of runtime, is to identify the closest center for each input data point, which leads to significantly high time complexity, i.e., $\mathcal{O}(nkd)$, where n is the number of data points, k is the number of centers and d is the number of dimensions.

Over years several researches study how to accelerate the clustering procedure while preserving the exact results. One popular way is to adopt analytical bounds to filter data points by which unnecessary distance calculations can be omitted. Towards this, Elkan's k-means algorithm [1] firstly proposes to adopt triangle inequality to form the analytical bounds. In the same vain, the following techniques seemly perform well on low-dimenstional datasets: Hamerly in [2] proposes to solely use one lower bound on the distance between each point and its second closest center instead of keeping lower bounds. Newling and Fleuret in [4] provide an Exponion algorithm to improve Elkan's algorithm. Recently, Yu et al. [5] provide two additional filtering bounds to form a memory-optimized Elkan's clustering algorithm. Nevertheless, the dominance of the above algorithms is not decided yet.

In fact the performance of the algorithms can be influenced by many factors. One is the adopted architecture, since the underlying caching behaviours might determine the performance of implemented algorithms even more than algorithmic differences. However, the aforementioned algorithms are not studied with respect to this context yet. Specifically for low-dimensional datasets, we can vision that such an impact is much significant, which motivates us to pose this topic.

In this master thesis, the student is expected to study state-of-the-arts and an initial Elkan's based approach at first while setting up a testbed for exploring further. Afterwards, the student is free to explore various aspects on k-means clustering algorithms towards architecture-aware designs and propose advanced approaches to refine or even redesign the given ideas. Eventually, the student is supposed to extensively evaluate the proposed architecture-aware approaches with state-of-the-arts under reasonable configurations on different platforms to conclude the studied topic. Depending on the results of the thesis, the student might also experience how to write a scientific article formally.

Required Skills:

- Knowledge of C++ programming
- Clustering knowledge is beneficial
- Architecture knowledge is beneficial

Acquired Skills after the work:

- Knowledge of k-means clustering algorithms
- Knowledge of architecture-aware design
- Design, analysis, implementation of data mining software and open source development

References

[1] C. Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 147–153. AAAI Press, 2003.

[2] G. Hamerly. Making k-means even faster. In *SDM*, pages 130–140, 2010.

[3] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, Sept. 2006.

[4] J. Newling and F. Fleuret. Fast k-means with accurate bounds. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 936–944, New York, New York, USA, 20–22 Jun 2016. PMLR.

[5] Q. Yu, K.-H. Chen, and J.-J. Chen. Using a set of triangle inequalities to accelerate k-means clustering. In *Similarity Search and Applications - 13th International Conference (SISAP)*, Virtual Conference, Sep 30 - Oct 2 2020. Springer.