Finding the Curse of Dimensionality Sweet Spot Between Traditional Clustering and Deep Clustering

TU/e supervisor: dr. W. Duivesteijn
Additional supervision by: dr. S. C. Hess, TU/e

The Curse of Dimensionality in Clustering

Informally, the task of clustering seeks to partition a dataset into coherent groups, such that similar objects are grouped together and dissimilar objects are in separate groups. There are many variants to this task and paramaters within these variants that can be set to turn this into disparate concrete learning tasks, but almost every clustering task defines *similar* and *dissimilar* in terms of some *distance measure*. Objects that have a small distance in a data space should be more likely to end up in the same cluster than objects that have a big distance in a data space.

This sounds all very intuitive and logical. A problem here, however, is that clustering quickly suffers from the *curse of dimensionality*: when the number of dimensions in a dataset increases, distances start to lose all meaning. A formal definition of this concept is:

$$\forall_{\varepsilon>0} \lim_{d\to\infty} P\left[\text{dist}\left(\frac{\text{Dmax}_d - \text{Dmin}_d}{\text{Dmin}_d}, 0\right) \leq \varepsilon\right] = 1$$

where:

- Dmin_d = distance to the nearest neighbor in d dimensions;
- Dmax_d = distance to the farthest neighbor in d dimensions.

Reformulating this back into informal terms again: when the number of dimensions keeps increasing, the difference between the distance to my nearest neighbor and the distance to my farthest neighbor becomes negligible. This holds for a wide range of data distributions and distance functions; it spells doom for the task of clustering. Often, one does not need to let d explode to make this problem appear: half a dozen to a dozen dimensions often already suffices.

Salvation in Deep Clustering?

Deep Clustering methods [1, 2, 3] leverage the power of deep learning models to learn informative representations of the original data space, which enables clustering even in (originally) high-dimensional data. Rejoice! We have beaten the Curse of Dimensionality!

But hang on a minute... isn't it true that deep learning models require enormous amounts of data to perform well? Do we not need massive amounts of dimensions from which to extract a representation that is informative enough, and do we not need many observations in the dataset?

The Project

We seem to naturally run into a conclusion that there is still a research gap to be bridged. For lower-dimensional datasets, traditional clustering can be used. For higher-dimensional datasets, deep clustering methods will work. But there may be a sweet spot inbetween (the moderately-dimensional datasets, if you will), where neither traditional nor deep clustering satisfies. The goal of this project is to empirically investigate whether such a sweet spot exists, and if so, what its boundaries are. In this project, we want you to:

- investigate the curse of dimensionality for wide range of data distributions (seed clusters or no, how many, Gaussian/Poisson/Zipf/...; outliers!!?) and distance functions;
- devise a protocol to determine the dimensionality upper bound beyond which traditional clustering methods start to fail;
- devise a protocol to determine the dimensionality lower bound below which deep clustering methods cannot yet function;
- identify a space of distributions/dimensionalities/dataset sizes where a research gap still remains for moderately-dimensional clustering methods.

Current Status

We must investigate whether papers exist saying something sensible about the curse of dimensionality in clustering and what shapes and forms it can take. But we think that there is a clear gap in the literature here; to the best of our knowledge, noone has taken this sandwiching approach pitting traditional and deep clustering against each other (and finding out whether there is actual space for filling the sandwich).

Requirements

The main task of a student in this project is to derive an experimental protocol and execute it. You will need to dive into methods for generating artificial datasets. When done well, we expect this project to lead to a publication at a top-level data mining or machine learning conference, or perhaps a journal.

References

[1] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. A survey of clustering with deep learning: From the perspective of network architecture. IEEE Access, 6:39501–39514, 2018.

[2] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S Yu, and Lifang He. Deep clustering: A comprehensive survey. arXiv preprint arXiv:2210.04142, 2022.

[3] Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, Martin Ester, et al. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. arXiv preprint arXiv:2206.07579, 2022.