Improving cluster analysis by leveraging geometric structure in data

Geoffrey Mon — gmon@princeton.edu

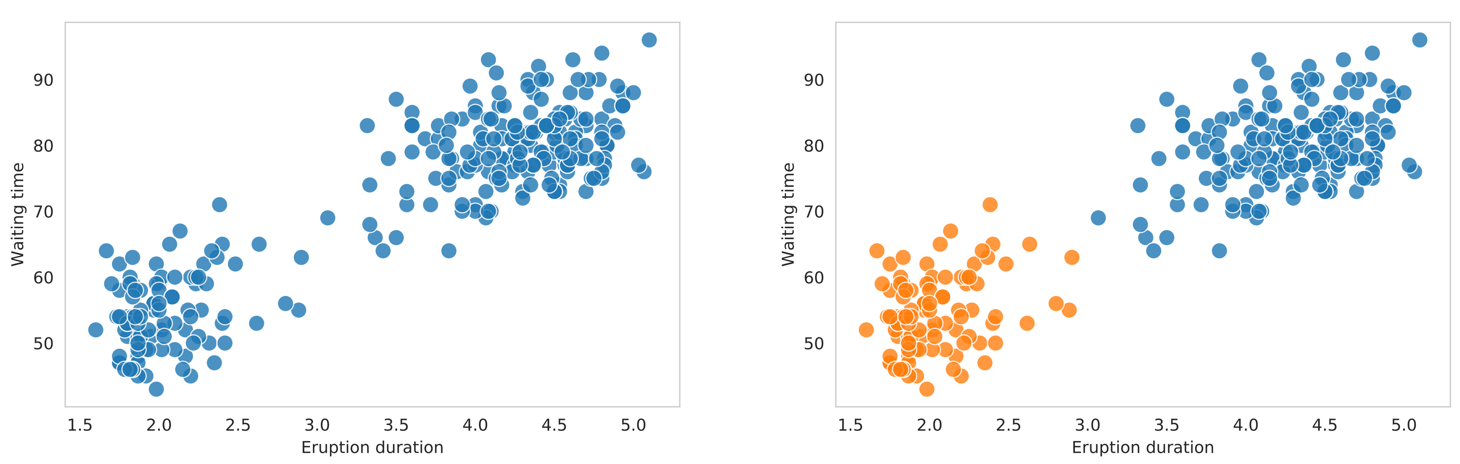
Overview

The goal of this project is to develop new data clustering approaches that take advantage of **geometric structure** in the data, with applications in tumor data analysis, etc.

- **Cluster analysis** is a widely-used data science tool.
- There are many **general-purpose** clustering algorithms, but they do not leverage **geometric structures**, such as grids, in the data.
- By developing a **novel clustering approach** that directly incorporates these **geometric structures**, we can significantly improve clustering performance.

Background

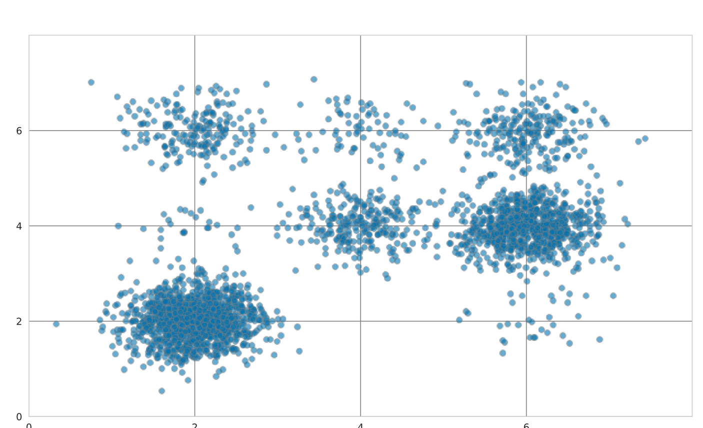
Clustering is a data science problem: given a data set *X*, divide *X* into parts (called **clusters**) so that all of the data in a given cluster are similar to each other. There are many widely-used general-purpose clustering algorithms, such as *k*-means.



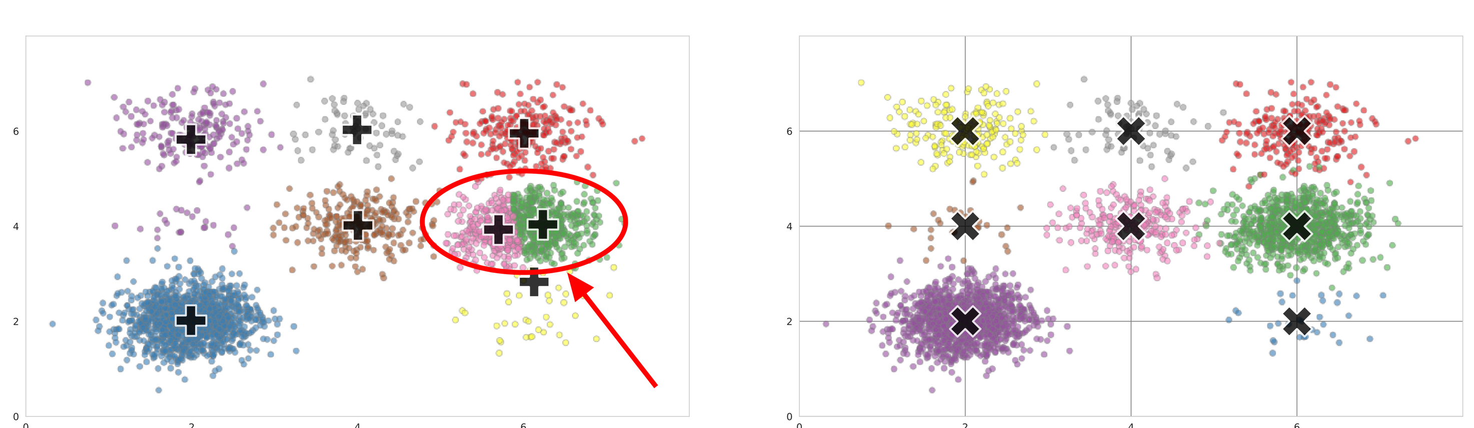
(a) Raw data set (b) Clustered data set

Figure 1. Clustering example (Old Faithful data set [1])

However, in certain applications, there are **geometric structures** such as **grids** in the data which existing algorithms do not leverage. We propose that a new clustering approach exploiting these structures will find higher-quality clusters.



(a) Consider this data, which has a grid structure.



(b) Clustered with a general-purpose clustering algorithm; notice how the intuitive cluster at the right has been divided into two inferred clusters.

(c) An algorithm that leverages the grid might produce a better clustering such as this one.

Figure 2. Clustering example with grid structure

Methods

To identify the cases where general clustering algorithms perform poorly, I generated data sets with grid structure and clustered them with a general-purpose clustering algorithm (Gaussian mixture with EM).

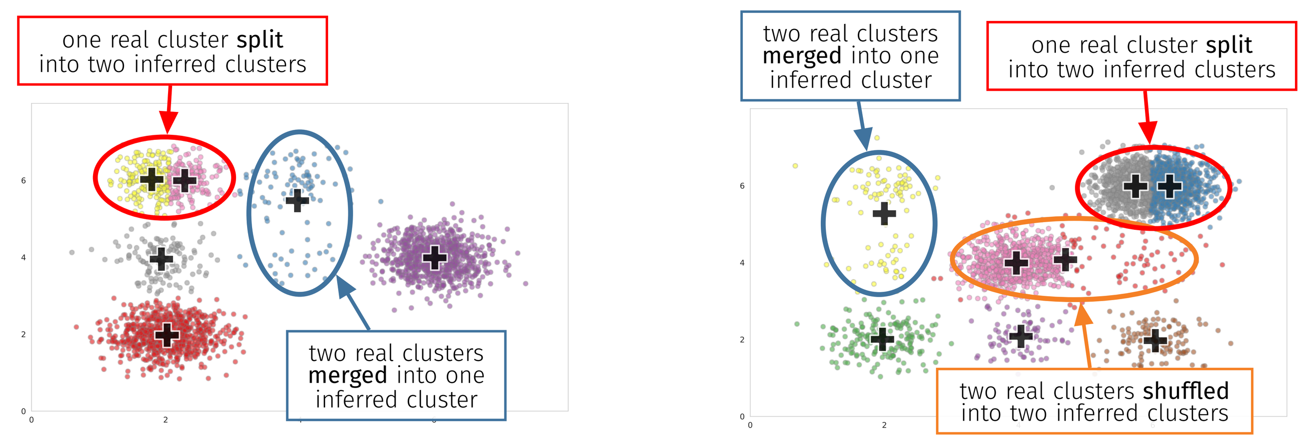


Figure 3. Characteristics of general-purpose clustering on data with grid structures

If we assume that every real cluster center has inferred cluster centers nearby, then we can use the grid constraint to find better cluster centers from the inferred cluster centers, using **neighborhood grid search**:

1. **Find the best grid that fits the inferred cluster centers.**

(a) Clusters inferred by general-purpose algorithm

(b) The best grid minimizes total distance between each inferred center and its closest grid point

Figure 4. Grid fitting example

2. **Pick the best set of grid points to use as cluster centers.** Find grid points that are close to inferred cluster centers; by assumption, the real cluster centers should be among these points. Then, try every combination of points in this set to find new cluster centers.

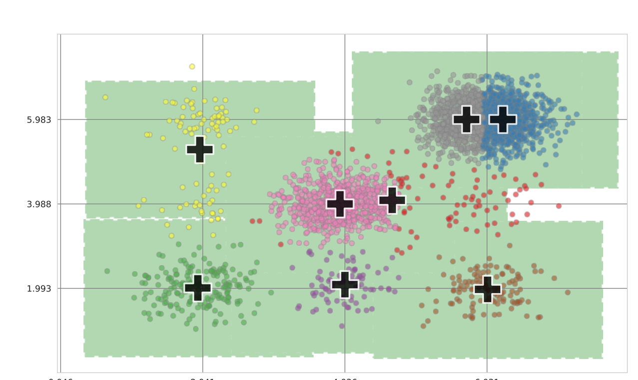
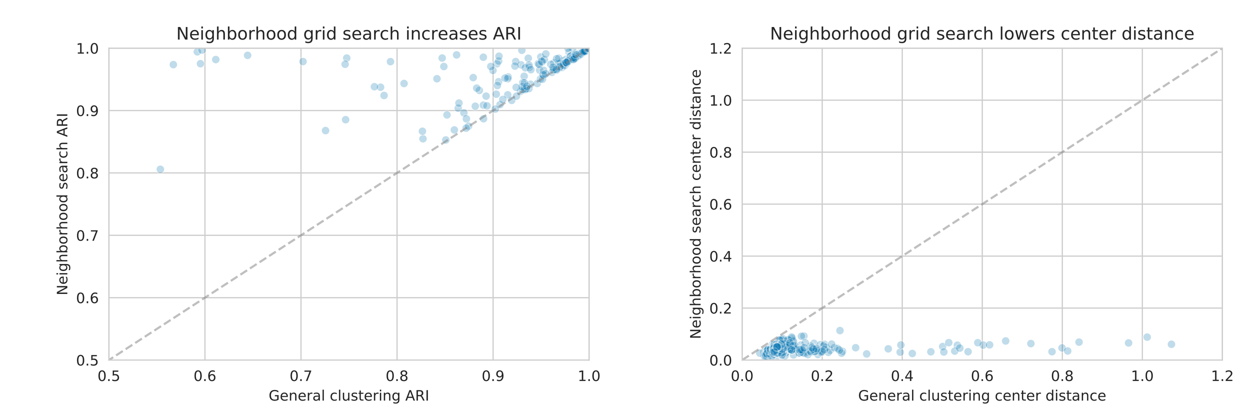


Figure 5. Grid points within the green region are considered "close" to at least one inferred cluster center; note that in this example, all real cluster centers are in this search region

Results

In my generated data, applying neighborhood grid search improved clustering performance in virtually every case.



(a) Cluster assignment accuracy (higher is better)

(b) Distance between inferred and real cluster centers (lower is better)

Figure 6. Neighborhood grid search improves clustering quality

These results are very promising, and we are working on adapting the neighborhood grid search technique for use with DNA sequencing data from tumors.

Acknowledgments

I would like to thank **Gryte Satas**, **Dr. Simone Zaccaria**, and **Prof. Ben Raphael** for their mentorship and guidance throughout this project. I would also like to thank the members of the **Raphael Lab** for their comments and feedback, and the **Office of Undergraduate Research** for supporting my research through OURSIP.

References

[1] Adelchi Azzalini and Adrian W. Bowman. "A Look at Some Data on the Old Faithful Geyser". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 39.3 (1990), pp. 357–365. ISSN: 00359254, 14679876. DOI: 10.2307/2347385.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN: 978-0387-31073-2.

[3] Phillip Compeau and Pavel Pevzner. *Bioinformatics Algorithms: An Active Learning Approach*. 2nd. Vol. 2. Active Learning Publishers, 2015. ISBN: 0990374629.

[4] Arthur P. Dempster, Nam M. Laird, and Donald B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38. ISSN: 00359246.

[5] Stuart P. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (Mar. 1982), pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489.