# P20: Clustering methods for selecting representative samples in chemical databases

Felipe V. Calderan, Marcos G. Quiles

Federal University of São Paulo - emails: {fvcalderan, quiles}@gmail.com

## Definition of Cluster

In chemistry, a cluster is defined as a group of bound atoms or molecules that are held together by some force (oppositely charged ions, covalent bounds, etc.) [1] with size varying between a molecule and bulk solid. In Data Analysis, clusters are groups of items from a dataset, and the force that holds them together is given by their similarities, that is, elements within the same cluster are more similar with each other than those in different clusters.

## What is Clustering?

Clustering is a technique that, given a dataset, tries to group elements based on similarity [2]. Of course, this is very subjective in nature. For example, consider a bat, a cat and an eagle. The cat and the bat form a cluster of mammals, but this same cluster would not exist if we were interested in clustering land and flying animals, because then we'd have the bat and the eagle in one cluster, but the cat in a different one. Thus, the clustering procedure is highly dependent on the features and the similarity measure taken into account.

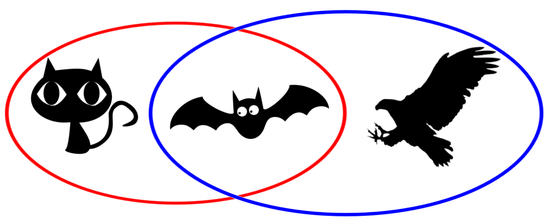

Figure: Different possibilities of clustering

## Clustering Algorithms

There are many clustering algorithms. Let's talk about one of the simplest and most commonly used one: k-means. It starts with a random initial partition and iteratively reassigns the patterns to clusters based on the similarity between the pattern and the cluster center until it converges. The figure below shows a random set of points partitioned by k-means.

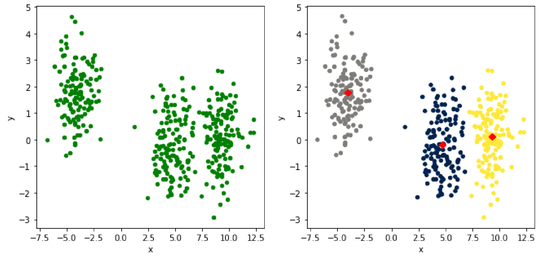

Figure: Left: random dataset; Right: k-means partitions

K-means is used very often because it is fast and easy to implement, but it has some limitations. For example, observing the next dataset we can visually separate the points in two circles, but k-means fails to cluster them in the most natural manner.

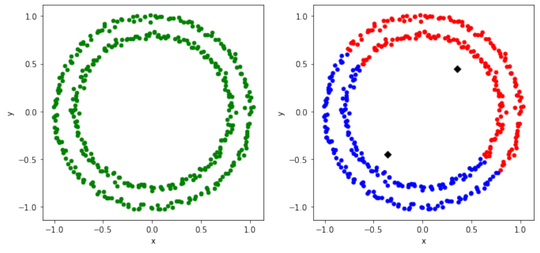

Figure: Left: rings dataset; Right: k-means partitions

## Overcoming the Problem

A fine way to deal with this situation is to use an algorithm that clusters based on a different criterion, for example: Single-Linkage (SLINK). It clusters in a agglomerative way, that is, every element starts in its own cluster and at each iteration clusters containing the closest pairs of elements are combined.

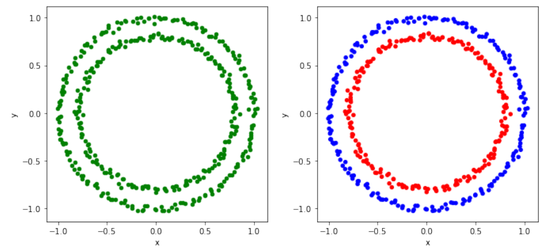

Figure: Left: rings dataset; Right: SLINK partitions

Therefore, different problems require different clustering algorithms. In the literature, several clustering algorithms based on distinct approaches are available. A taxonomy of the clustering algorithms is depicted in Figure. 5.

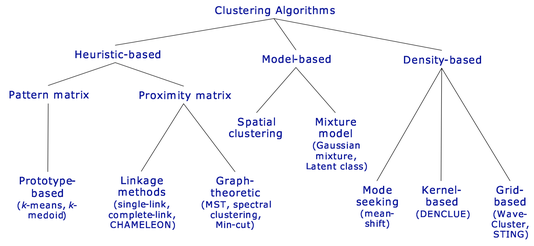


Figure: Clustering of clustering algorithms [3]

## Objectives of this Project

Here, we will investigate some clustering methods [4] for selecting representative samples from a given dataset of molecules. Mainly, this study has the following goals:

1. Perform data clustering considering distinct methods with different representation
2. Provide visual tools to facilitate the analysis of the clustering results
3. Develop a simple toolbox for data clustering and visualization customized for materials science / chemistry data

.

## References

[1] H. Harberland.
*Clusters of Atoms and Molecules: Theory, Experiment, and Clusters of Atoms*.
Springer, 2013.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn.
Data Clustering: A Review.
*ACM Computing Surveys*, 31(3):265–323, September 1999.

[3] A. K. Jain, Alexander Topchy, Martin H. C. Law, and Joachim M. Buhmann.
Landscape of clustering algorithms.
In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04)*, pages 260–263. IEEE Computer Society, 2004.

[4] A. K Jain.
Data clustering: 50 years beyond k-means.
*Pattern Recognition Letters*, 31:651–666, 2010.

## Acknowledgments

- CNPq & FAPESP
- GPAM/ICT/UNIFESP
- QTNANO/IQSC/USP