可参考的一些机械学习算法及其应用

## 一、k-medoids and affinity propagation：

k-medoids and affinity propagation are the most commonly used exemplar-based clustering methods. In K-medoids algorithm, as same as the famous k-means algorithm, you have to give the cluster number at first. Then K-medoids attempt to **minimize the distance between datapoints labeled to be in a cluster and a datapoint designated as the center of that cluster**.

Unlike K-medoids, affinity propagation doesn’t need the number of clusters to be determined or estimated before running the algorithm. The algorithm proceeds by alternating two message passing steps, to update two matrices, the "responsibility" matrix R and the "availability" matrix A. The iterations are performed until the exemplar decisions remain unchanged over a number of iterations. The exemplars are extracted from the final matrices as those whose 'responsibility + availability' for themselves is positive (which means it is a good cluster point).

Application in reference:

To decrease the initial training set of NN-GP pes, k-medoids and affinity propagation algorithm are used for **Partitioning the Configuration Space**. The descriptor is just distance between input vectors made of six bond length.

Examples:

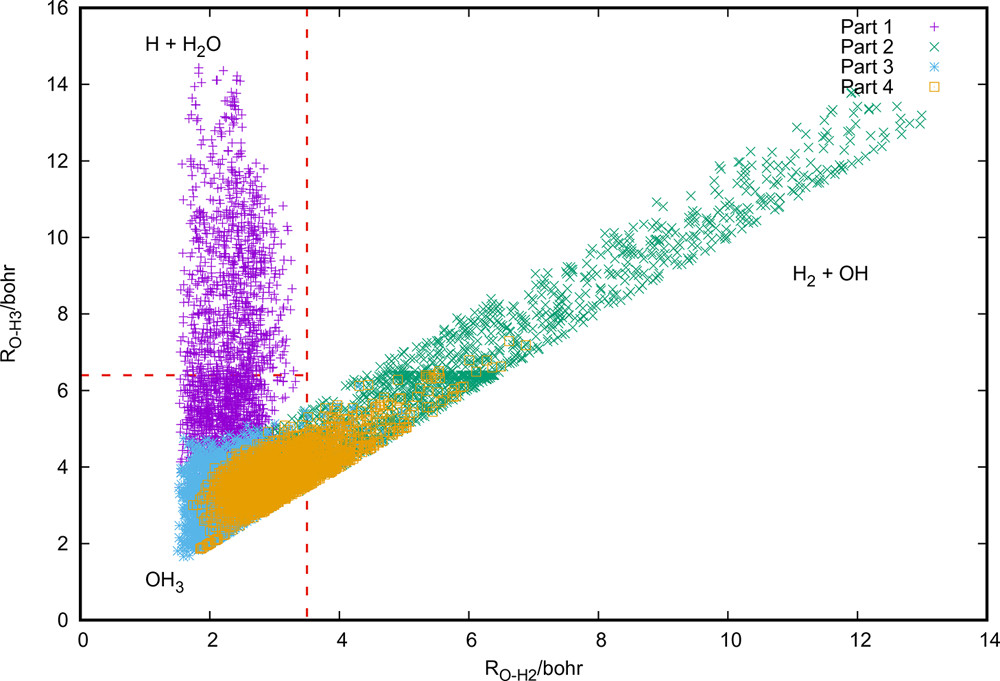


Figure 1. Spatial distribution of geometries for OH3 as a function of RO−H2 and RO−H3.

Classification based on affinity propagation

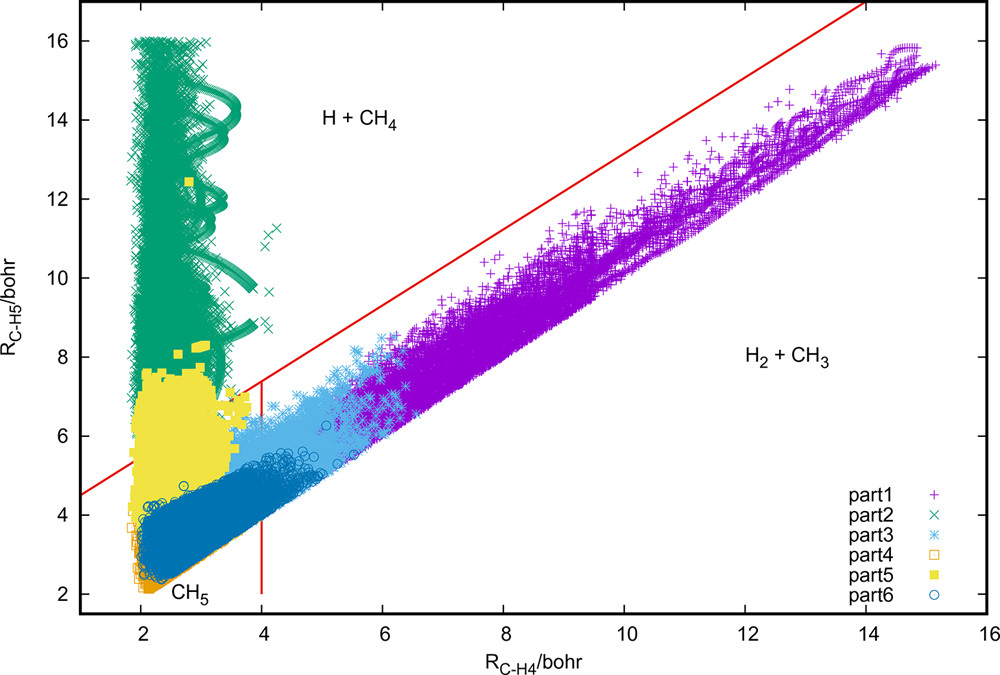


Figure 2. Spatial distribution of geometries as a function of RC−H4 and RC−H5.

Classification based on k-medoid

references：Application of Clustering Algorithms to Partitioning Configuration Space in Fitting Reactive Potential Energy Surfaces, J. Phys. Chem. A 2018, 122, 3140−3147

## 二、agglomerative hierarchical clustering：

Agglomerative hierarchical clustering is another useful approach to construct classification schemes that widely used in embedded classification problems. There are lots of special AHC methods with different time complexity and memory requirement, but most of them follows the same basic idea: let most similar point be a cluster, let point joins to the most similar cluster, let most similar cluster be a bigger class, and stop until remaining only one class(or any number you like). After that you will get a dendrogram containing all the clustering information, and you can analysis how they agglomerate together.

The most important preparation is to define what’s the similarity of point-point (Metric), point-cluster and cluster-cluster (Linkage criteria). We can use *Lance–Williams dissimilarity update formula* to understand the linkage criteria in different methods.

Application in reference:

1. cosine similarity combining the weighted pair group method with arithmetic mean (WPGMA) algorithm has been used to partition the cluster potential energy surface in terms of nuclear configurations. They **classified 37 isomers** and calculate their IR spectra by NMA and **identify which regions of the PES are populated in the electrospray ionization process.**
2. To make a remarkably informative representation between structure and property, they use SOAP-REMatch kernel method to build (dis)similarity matrix and **use the basic motifs of a particular cluster classified by normal AHC method to understand structure-property relation.**

example：

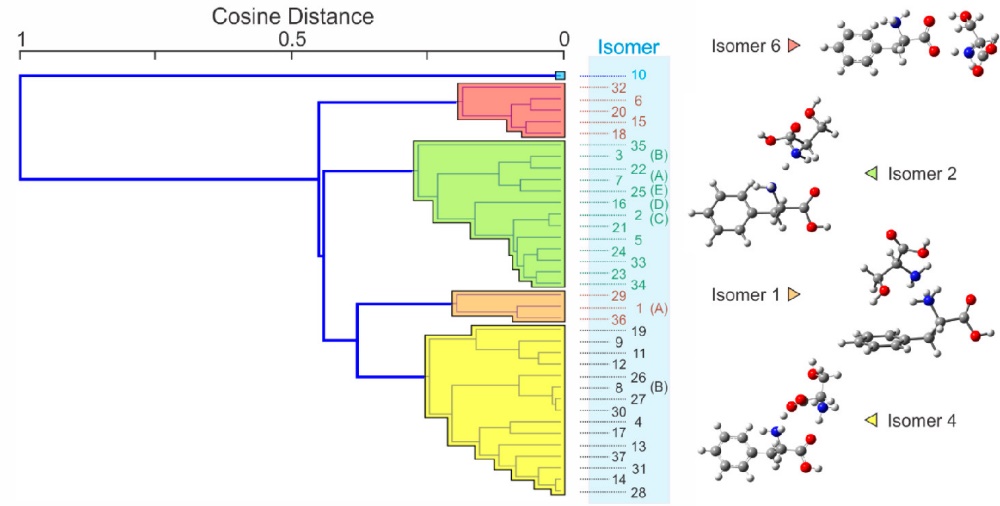


Figure 3. WPGMA dendrogram constructed from the cosine distances between the various cluster structures

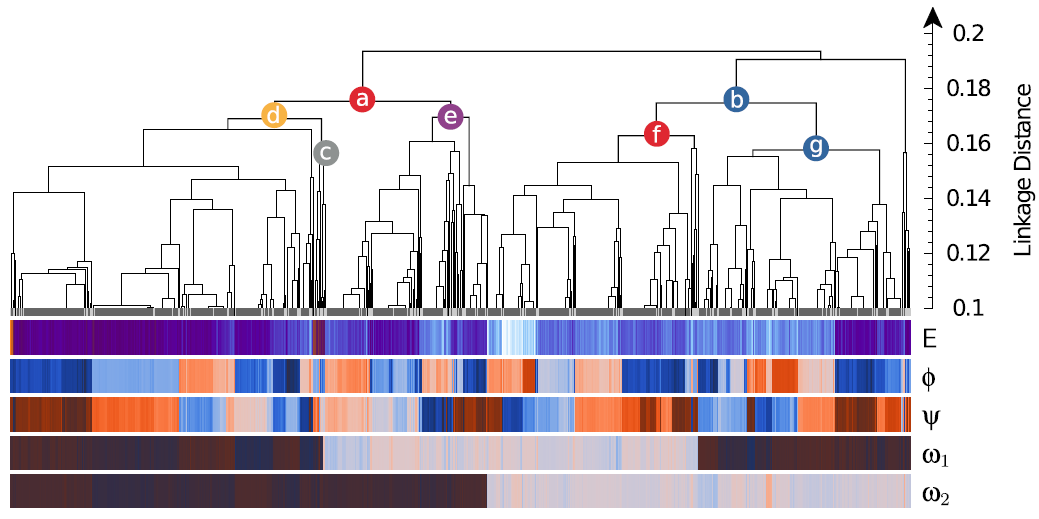


Fig. 4Representation of the similarity matrix corresponding to the lysine dipeptide dataset using the agglomerative clustering algorithm

reference:

1、Applying Machine Learning to Vibrational Spectroscopy, J. Phys. Chem. A 122, 1, 167-171

2、Mapping and classifying molecules from a high‑throughput structural database，De *et al. J Cheminform (2017) 9:6*

## 三、density-based model and mean shift algorithm:

All of the algorithms mentioned above are similarity-based algorithm, while there is another classifying model that based on data density. The most famous density-based clustering algorithm is DBSCAN, which seems rarely used in chemistry field (maybe due to the less of noise point). In most case, they focus more in the density model and use other way, such as mean shift method, to classify molecular configuration. A widely used density estimate algorithm is Kernel density estimation method with Gaussian probability function.

Mean shift algorithm is aim to locate the maxima of a [density function](https://en.wikipedia.org/wiki/Density_function), a so-called [mode](https://en.wikipedia.org/wiki/Mode_(statistics))-seeking algorithm. Its basic idea is like k-means, an iterative method that start with an initial estimate {\displaystyle x}point. In MSA, we calculate the weighted mean of the density in the window as centroid in k-means, and update the center point to the neighborhood until the density unchanged.

Application in reference:

1. In PAMM program, we can use quick shift algorithm with Gaussian kernel density model to **Recognizing molecular patterns from molecular dynamic trajectory.**
2. They uses the HDBSCAN clustering technique to **identify automatically the main structural motifs**.

Example:

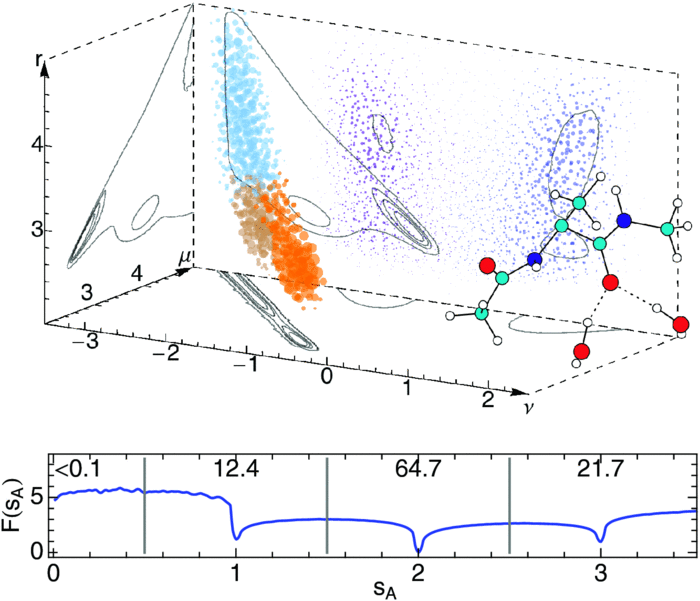


FIG. 5. Distribution of (ν, μ, *r*) configurations forOw–H···OC in a simulation of alanine dipeptide in water

Reference:

1. Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond, THE JOURNAL OF CHEMICAL PHYSICS **141**, 174110 (2014)

2. Machine learning for the structure–energy– property landscapes of molecular crystals, Chem. Sci., 2018, 9, 1289–1300 | 1289

## 四、Discussion about similarity:

Similarity is one of the most important part in clustering program. There are millions of paper focus on the best similarity form and how to use them in specific problems. For example, the smooth overlap of atomic positions (SOAP) similarity kernels is a invariant to symmetry operations of the atoms of each environment, which could measure the similarity between atomic environment and help us build a configuration space of complex materials. This similarity method has been successful in molecular crystals and atomization energies prediction for a database of small organic molecules. Can we find some descriptors to measure the similarity of our scaling problem? Or can we use this idea in our protein/ polypeptide systems? Maybe the most difficult work is not find a clustering method but choose a good similarity criteria.