Size characterization of clusters:

Ward's method via Lance-Williams algorithm

Aulia Tegar Wicaksono

Department of Materials Engineering, The University of British Columbia

tegar@alumni.ubc.ca

Written on: July 20, 2015

Abstract

This document describes the numerical technique implemented to characterize the size of helium clusters from molecular dynamics simulation, i.e. Chapters 7 and 8 of my thesis.

1 Background

Clustering techniques help one to group a set of points in space by categorizing them into clusters, i.e. an object consisting few of these points that are closer in space to each other than to the rest of points. A number of clustering techniques is available in the literature and generally they can be categorized into two groups: hierarchical clustering and partitional clustering. The key difference between the two is that in the latter group, the number of clusters must be known in advance before the analysis is performed.

For the type of analysis required in my thesis, the number of clusters are not known prior to the analyses, i.e. hierarchical clustering. Furthermore, I have chosen from this category the Ward's hierarchical technique [1]. While there are several other hierarchical clustering techniques (e.g. single-link, complete-link and group average [2]), Ward's technique has been shown to have the highest evaluation score among other hierarchical techniques, i.e. the correlation coefficient (see Section 3.2), when employed for analyzing several set of data.

2 Methodology

The following definition is needed in order to explain the clustering technique. A cluster k of size n is an object containing n number of atoms i. The minimum and maximum possible size of cluster k are 1 (a single point) to N (all points). The in-cluster variance of cluster k, σ_k^2, is defined as $\sigma_k^2 = \sum_{i=1}^{n} |\mathbf{r}_i - \bar{\mathbf{r}}_k|^2$ where i is the index of atom that makes up cluster k, $\mathbf{r}_i$ is the position of atom i, $\bar{\mathbf{r}}_k$ is the center-of-mass of cluster k. A cluster of size $n = 1$ (individual atoms), by definition, has $\sigma_k^2 = 0$. The clustering analysis is guided by the total variance $\sigma^2 = \sum_k \sigma_k^2$.

Each atom is initially labelled as a cluster size one ($n = 1$). A dissimilarity matrix of $N \times N$ is built, its elements being D_{kl}, a parameter that is a function of cluster k and l. In Ward's algorithm, D_{kl} is the squared-distance between the center-of-mass of cluster k and that of cluster l, taking into consideration the periodic boundary conditions. On each iteration, any two clusters k and l are agglomerated into a larger cluster. This process is repeated until there is only one cluster consisting all N atoms in the box.

The choice of agglomerating cluster k and cluster l is decided by first comparing the increase in σ^2, or $\Delta\sigma^2$, across all possible situations and choosing one that leads to the most minimum $\Delta\sigma^2$. The quantity $\Delta\sigma^2$ is a function of D_{kl}. Using Lance-Williams algorithm [3], the $\Delta\sigma^2$ and the inter-cluster distance D_{kl} can be evaluated iteratively.

- The increase in the total in-cluster variance due to agglomerating cluster k and cluster l into cluster j is
$$\Delta\sigma^2(k \cup l) = \frac{1}{2} D_{kl}$$
- The inter-cluster distance between the newly formed cluster $j = k \cup l$ and other cluster h is updated by:
$$D_{hj} = \frac{n_h + n_k}{n_h + n_k + n_l} D_{hk} + \frac{n_h + n_l}{n_h + n_k + n_l} D_{hl} - \frac{n_h}{n_h + n_k + n_l} D_{kl}$$
where n_h, n_k, n_l are the size of respective clusters.

The iteration is performed until there remains one cluster containing N atoms. A dendogram (see Figure 2) can be built to illustrate the grouping of each atom. An arbitrary cut-off value λ_c is chosen to limit the cluster size and distribution. In the analysis of helium clusters, several values of λ_c have been tested. Cluster distribution were insensitive to λ_c when λ_c is between 1.2 to 2.2 nm^2. Further analyses employed $\lambda_c = 1.5$ nm^2.

3 Case study

3.1 Iterations

An example of clustering analysis using Ward's technique with Lance-Williams algorithm is illustrated. Suppose a data set containing 5 points in two-dimensional space, as shown in Figure 1.

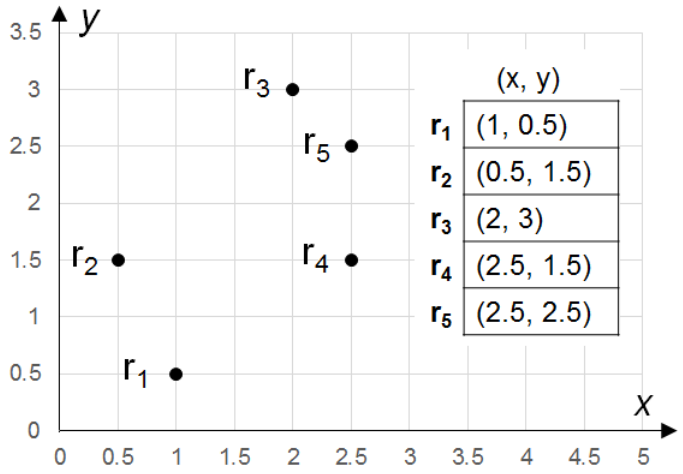

Fig. 1: A set of points in 2D space

Initially, the dissimilarity matrix is given by

Tab. 1: Dissimilarity matrix at the beginning of analysis

	r_1	r_2	r_3	r_4	r_5
r_1	0	1.25	7.25	3.25	6.25
r_2	1.25	0	4.5	4	5
r_3	7.25	4.5	0	2.5	0.5
r_4	3.25	4	2.5	0	1
r_5	6.25	5	0.5	1	0

The agglomeration process that leads to the lowest $\Delta\sigma^2$ is grouping point 3 and 5 where $\triangle\sigma^2 = \frac{1}{2}D_{35} = 0.25$. The updated inter-cluster distance is now:

- $D_{(35)1} = 8.833$.
- $D_{(35)2} = 6.167$.

- $D_{(35)4} = 2.167$.

The updated dissimilarity matrix is now:

Tab. 2: Dissimilarity matrix at the end of first iteration

	r_1	r_2	$\{r_3,r_5\}$	r_4
r_1	0	1.25	8.833	3.25
r_2	1.25	0	6.167	4
$\{r_3,r_5\}$	8.833	6.167	0	2.167
r_4	3.25	4	2.167	0

At the second iteration, the agglomeration process with the lowest $\Delta\sigma^2$ is grouping point 1 and 2, $\Delta\sigma^2 = 0.625$. The updated inter-cluster distance is now:

- $D_{(12)(35)} = 10.625$.
- $D_{(12)4} = 4.417$.

The updated dissimilarity matrix is now:

Tab. 3: Dissimilarity matrix at the end of second iteration

	$\{r_1,r_2\}$	$\{r_3,r_5\}$	r_4
$\{r_1,r_2\}$	0	10.625	4.417
$\{r_3,r_5\}$	10.625	0	2.167
r_4	4.417	2.167	0

At the third iteration, the agglomeration process with the lowest $\Delta\sigma^2$ is grouping cluster (35) with point 4, $\Delta\sigma^2 = 1.083$. The updated inter-cluster distance is now:

- $D_{(12)(345)} = 10.283$.

The updated dissimilarity matrix is now:

Tab. 4: Dissimilarity matrix at the end of third iteration

	$\{r_1,r_2\}$	$\{r_3,r_4,r_5\}$
$\{r_1,r_2\}$	0	5.142
$\{r_3,r_4,r_5\}$	5.142	0

At the final iteration, only two clusters left: (12) and (345). The $\Delta\sigma^2$ upon agglomerating these clusters into a single cluster is $\Delta\sigma^2 = 5.142$.

3.2 Dendogram and Cophenetic Matrix

A tree diagram (dendogram) can be drawn to illustrate the $\triangle\sigma^2$-level at which two clusters are combined, see Figure 2.

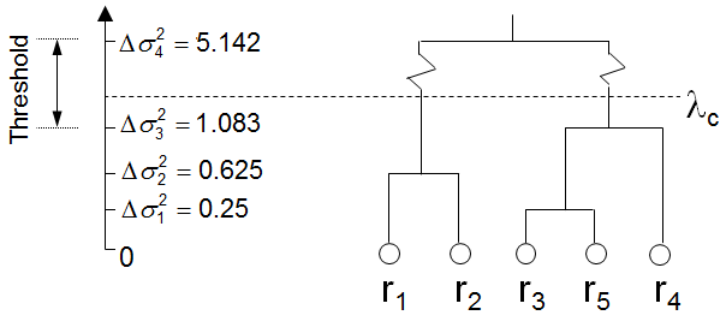

Fig. 2: Dendogram of clustering, vertical axis being $\Delta\sigma^2$

Cluster discrimination is performed by assigning an arbitrarily chosen cut-off value λ_c, a threshold for the change of the total in-cluster variance $\Delta\sigma^2$ above which two clusters will not be coalesced into one larger cluster. A good cut-off value is typically obtained from a range within which cluster distribution does not depend on the chosen cut-off value. For the case study above, a good cut-off value is between 1.1 and 5.

Additionally, a cophenetic matrix $N \times N$ can be built to evaluate the quality of clustering analysis, the matrix element being the inter-cluster distance at which two individual points belong to the same cluster for the first time, i.e. twice the $\Delta\sigma^2$ from the dendogram in Figure 2. For example, the element (1,2) below is 1.25 since this is twice the $\Delta\sigma^2$ at which atom 1 and atom 2 belong to one cluster for the first time. On the other hand, the element (1,3) is 10.283 since atom 1 and 3 belong to one cluster for the first time when $\Delta\sigma^2 = \frac{1}{2}(10.283) = 5.142$ in Figure 2.

Tab. 5: Cophenetic matrix

	r_1	r_2	r_3	r_4	r_5
r_1	0	1.25	10.283	10.283	10.283
r_2	1.25	0	10.283	10.283	10.283
r_3	10.283	10.283	0	2.167	0.5
r_4	10.283	10.283	2.167	0	2.167
r_5	10.283	10.283	0.5	2.167	0

The quality of clustering analyses is obtained by calculating correlation coefficient R between the cophenetic matrix (Table 5) and the dissimilarity matrix before clustering

(Table 1), i.e.

$$R = \frac{\sum(P_t - \overline{P})(Q_t - \overline{Q})}{\sqrt{\sum\left(P_t - \overline{P}\right)^2}\sqrt{\sum\left(Q_t - \overline{Q}\right)^2}}$$

where

- Index t represents the element index in the dissimilarity and cophenetic matrix, i.e. t = (1,2), (1,3), ... , (4,5);
- P_t represents the element in the dissimilarity matrix (Table 1), i.e. $P_{(1,2)} = 1.25$, $P_{(1,3)} = 7.25$, ..., and
- Q_t represents the element in the cophenetic matrix (Table 5), i.e. $Q_{(1,2)} = 1.25$, $P_{(1,3)} = 10.283$, ...

The correlation coefficient for the case study here is $R = 0.858163$.

References

[1] J. Ward Jr, *Hierarchical grouping to optimize an objective function*. J. Am. Stat. Assoc. **58**, 236 (1963). doi:10.1080/01621459.1963.10500845

[2] G. Gan, C. Ma, J. Wu, *Data Clustering: Theory, Algorithms and Applications* (SIAM, 2007)

[3] G. Lance, W. Williams, *A general theory of classificatory sorting strategies I. Hierarchical systems*. The Computer Journal **9**, 373 (1967). doi:10.1093/comjnl/9.4.373