doi:10.1017/S143192761900223X

Modularity Optimization for Fast Automated Detection of Solute Clusters in Atom Probe Tomography

Arpan Mukherjee[1], Scott R. Broderick[1*], Tianmu Zhang[1], and Krishna Rajan[1]

[1.] Department of Materials Design and Innovation, University at Buffalo, Buffalo, NY USA 14260
* Corresponding author: scottbro@buffalo.edu

The current work describes the use of some unsupervised Machine Learning Methods to uncover patterns in Atom Probe Tomography (APT) scans of microstructures. Data reconstructed from an experiment comprises of the following variables: position in x; y; z coordinates and mass-to-charge ratio. While the geometry of the atoms are provided by the x; y; z positions, the mass-to-charge ratio provides us information about the chemistry of the atom. The analysis performed on few datasets have been reported in the following sections.

Cluster analysis is predominant in the field of community detection, sequence analysis, image segmentation, and others. The use of completely unsupervised and parameter-free algorithm is a requirement in almost all these scientific domains. However, it is highly dependent on the use of distance functions that compute the distance between each data point. Parameter-free algorithms adopt a graph-theoretic representation of the whole dataset and attempt to identify small subgraphs that are independent of each other. They use the distance between the data points to form a weighted adjacency matrix for the purpose of analysis. Spectral Clustering [1] is a very popular methodology of graph partitioning that uses the eigenvalue and eigenvector of the transformed and scaled down adjacency matrix. The cluster structure is inferred from the eigenvalue plot and the corresponding eigenvector matrix. Non-negative Matrix Factorization (NMF) is a similar technique that iteratively performs matrix decomposition into community structure. The third approach is modularity maximization that assigns a community to a given node by maximizing the quality of a partitioned graph. Although fundamentally all these algorithms address the same problem, the accuracy of the algorithms can only be judged by application to a specific type of problem. In our problem, assessment of accuracy is even harder, since the actual cluster structure is unknown and hard to predict. Spectral clustering performs slower than the others due to its heavy reliability of eigenvalue computation. NMF even though does not require the number of clusters to be specified, has got other hyper-parameters to be set, and their sensitivity is still a topic of research. Clustering analysis of nanoscale precipitates obtained from APT is also an important topic of research. The most widely used algorithm is the Maximum-separation method that relies on a number of predefined parameters including the expected number of clusters.

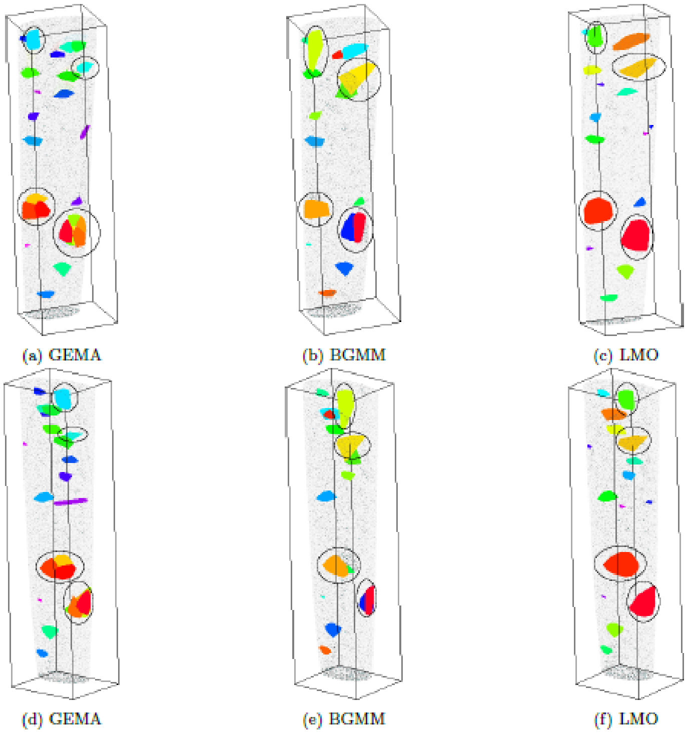
We compare the performance of three clustering algorithms: Louvain-Modularity Optimization (LMO), Gaussian Expectation-Maximization Mixture Analaysis (GEMA) and Bayesian Gaussian Mixture Model (BGMM). Comparison of unsupervised clustering algorithm is a very difficult task, since the actual cluster structure is unknown. Different measures have been developed over the years to compare the performance of clustering algorithms, but mostly they are available for supervised learning. Among, the three clustering algorithms compared, only LMO does not require the specification of the number of clusters. The GEMA computes the cluster from the BIC plot. The cluster number for BGMM is also computed using the BIC plot. The data used for clustering is the pruned point cloud of Scandium atoms obtained after performing data pruning.

https://doi.org/10.1017/S143192761900223X Published online by Cambridge University Press

Visual representation of the cluster structures obtained by the three methods of clustering is shown in Figure 1. The figure shows two different views with focus on the four major clusters as identified by LMO. Each cluster is bounded by its convex hull. Thus the presence of a sharp straight edge can signify an error in clustering. Sharp edges are seen in all three cluster structures. However, GMM and BGMM breaks the big clusters into smaller ones for a low BIC that are supposed to identify as a single cluster. Based on the Davies-Bouldin Index (DBI) of the three clustering methods, we find that LMO and GMM show comparable performance, with LMO giving slightly better performance. However, it also to be noted that DBI gives lower value if the clusters are small and compact regardless of whether they are meaningful or not. A GMM with higher number of clusters (=45) gives a much lower value of DBI (0.4471). Modularity has been chosen as the second metric for comparison of the clustering algorithms. To compute the adjacency, we have used scaled pairwise Euclidean distances between the atoms. LMO again shows better performance as compared to GMM and DGMM. It is to be noted that the modularity computation depends on whether the distances are scaled or not. A non-normalized distance matrix gives erroneous results for any of these algorithms [2].

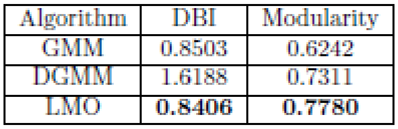
[1] U. Von Luxburg, Statistics and Computing **17** (2007) p. 395.

[2] The authors acknowledge supported by the National Science Foundation under Grant No. 1640867.



(a) GEMA (b) BGMM (c) LMO

(d) GEMA (e) BGMM (f) LMO



Algorithm	DBI	Modularity
GMM	0.8503	0.6242
DGMM	1.6188	0.7311
LMO	**0.8406**	**0.7780**

Figure 1. Visual comparison of the clustering algorithms from two different views for an Al-Sc-Mg sample with Al_3Sc precipitates, along with the accuracy in the clustering, Lower DBI an higher modularity correspond to higher accuracy in defining clusters.

https://doi.org/10.1017/S143192761900223X Published online by Cambridge University Press