4D STEM data analysis with KMeans clustering

Author: Núria Bach Siches
*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.**

Advisor: Vanessa Costa and Sònia Estradé

Abstract: When using 4D STEM methods to study various material characteristics, large amounts of diffraction images are created for each sample studied. To determine different characteristics of the material locally from the data obtained from the diffraction patterns, it has been considered to use clustering machine learning algorithms that will be able to quickly read and classify all diffraction images. A KMeans algorithm has been adapted to classify this type of data. The method has been found to work satisfactorily when applied to an experimental example.

I. INTRODUCTION

As one of the most promising techniques used in Scanning Transmission Electron Microscopy (STEM), the 4D STEM technique provides us with a 2D diffraction pattern at each pixel position of a sample scanned with a STEM. If the scanned areas measure approximately 150 x 150 pixels, we can easily have to deal with quantities of 22,500 diffraction images per 4D STEM experiment before we can extract usable information from the sample. It could be a very tedious job if we have to go through them one by one. The use of big data techniques that can divide the sample into zones with similar behaviour could allow us to get a general idea of the sample by only analysing the diffraction pattern of each of these zones.

The present work will focus on adapting the KMeans algorithms previously used in Electron Energy Loss Spectroscopy (EELS) analysis by the LENS group of the Electronic and Biomedical Engineering Department of the Physics Faculty of UB [1,2] to analyse 4D STEM diffraction data.

II. 4D STEM

In the Transmission Electron Microscope (TEM), an image is seen on the image plane of the objective lens and a diffraction pattern is seen at the focal plane of the objective lens. Then, the intermediate and projector lenses bring either of those to the observation plane at the end of the microscope. Diffraction mode is a fundamental observation mode in the TEM.

Electron diffraction in the TEM appears as a two-dimensional diffraction pattern, fulfilling Braggs Law:

$$2d(hkl)\sin\theta_B = n\lambda$$

where d(hkl) is the spacing of the hkl family of planes, θ_B is the Bragg angle, n an integer number and λ the wavelength of the incident electron, typically in the order of magnitude of the pm. At zone axis (if the crystal is observed along one of its symmetry axes) we will see a Fourier transform of the crystal along that particular axis, with the transmitted beam at the center (corresponding to those electrons not having been diffracted). The intensity of the transmitted beam will be higher for thinner samples. Additionally, as we move away from zone axis, we can reach a two-beam condition, where only the transmitted beam and one spot corresponding to one hkl are visible in the diffraction pattern.

A Scanning Transmission Electron Microscope (STEM) is capable of focusing a beam of electrons into spots of 0.05 to 0.2 nm in size always keeping the beam parallel to the optical axis. This allows for a point-by-point sweep of the sample that has been prepared thin enough to be considered 2D.

4D STEM is an electron microscopy technique that uses a pixelated electron detector that captures a diffraction pattern at each scan location. What we get is a 2D image of the reciprocal space associated with each swept spot. The data will be stored in a four-dimensional array. Each element of the array stores a number associated with the intensity received at each pixel of the detector when the electron beam is focused on each of the scan locations.

The data packaging and reading software performs a reconstruction of the sample image in real space from the diffraction images and allows us to visualize each diffraction pattern while placing it within the reconstructed image of the sample.

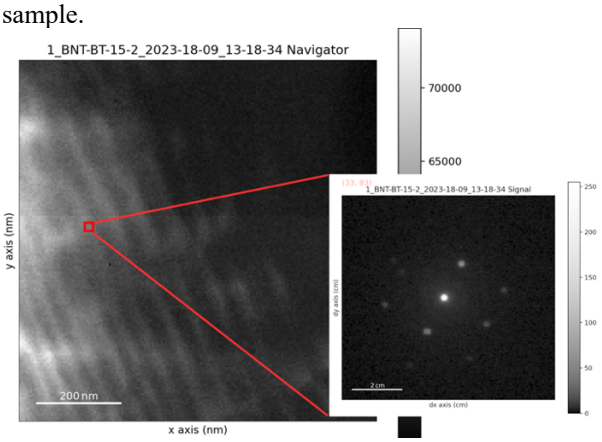

FIG. 1: Diffraction pattern of a spot located within the reconstruction of the real image of a sample.

* Electronic address: nbachsic9@alumnes.ub.edu

III. K-MEANS

Machine learning is a branch of AI science that applies automatic data-driven learning methods to obtain accurate predictions and classification based on past data observations. It enables artificial intelligence to mimic the way humans learn, gradually improving its accuracy.

Data analysts describe the machine learning algorithms in three main stages:

1. Decision process: based on the input data, which may or may not be labelled, the algorithm makes a first guess for a pattern in the data.

2. Error function: An error function evaluates the accuracy of the predicted model. If there are known examples, uses them to make a comparison.

3. Model optimization process: After the model is compared to the data, necessary changes are made to reduce the discrepancy between the known example and the model estimate. This "evaluation and optimization" process will be iterated, updating the weights autonomously until an accuracy threshold is reached.

There are two main categories of machine learning: supervised learning and unsupervised learning. Depending on the data and feedback available one or the other will be more effective.

In supervised learning, the computer is provided with "training data" which is a set of data that contains both the inputs and the desired outputs. The goal of this process is to find a general rule that maps inputs to outputs.

In unsupervised learning no labels are given to the learning algorithm. Its goal is to find structures in data that has not been labelled, classified, or categorized. It identifies commonalities in the data and reacts based on the presence or absence of those commonalities in each new piece of data. With the kind of data we want to analyse, unsupervised learning seems to be the best choice especially clustering.

Cluster analysis consists of grouping a set of objects (in our case diffraction images) so that objects in the same group (cluster) are more similar than those in other clusters. The type of cluster analysis will be chosen according to the nature of the data and it is very common to have to pre-process this data to adjust the results with the type of classification we want it to do.

KMeans is a clustering method that divides n elements into k clusters in which each element belongs to the cluster with the closest mean (cluster centroid), serving as the cluster prototype. After running the algorithm, it provides a **label** for each element according to the cluster in which it has been placed, and k **centroids** or average objects that represent all the objects that have been placed in that group.

As an unsupervised classification algorithm, KMeans is not provided with pre-grouped and labelled data. Only the objects to be classified and the number k of clusters in which they must be placed are given.

The process is as follows: KMeans randomly choose the k centroids with which to start the algorithm. After that it begins a chain of two steps that will be repeated. First step, each element is mapped to the closest or most similar centroid creating a group of elements for the centroid. Second step, the position of the centroid of each group is updated by taking as the new centroid the average position of the objects belonging to this group.

To do the analysis with KMeans, we will work with the four-dimensional array of intensity data provided by the 4D STEM software. For a better understanding of the classified data, we will present the results emulating the concept of reconstructed real space and diffraction patterns corresponding to each spot in this space with the data obtained as a result of running the KMeans algorithm.

I. ADAPTATION OF THE ALGORITHM AND DATA PROCESSING WITH PYTHON

The Python programming language will be used for data processing. Data is collected from the microscope in .blo format and will first have to be converted into a hyperspy object, a multidimensional data analysis module specialized in reading microscopy data, to be then transformed into a four-dimensional numpy array.

KMeans only works with data in a maximum of 2 dimensions. This will force the data to be converted into a two-dimensional array. It must be done in such a way as not to lose the physical sense or the purpose of the classification. It is interesting to divide the sample into zones according to the characteristics of its diffraction patterns. The most logical and efficient thing to do in this case will be to convert the two dimensions of the reciprocal space into a single dimension (row) that will constitute each of the objects to be compared. At the same time, we will convert the two dimensions that represent the real space of the sample into one dimension (column) to be able to assign a label to each element of this column according to the classification carried out with the rows containing the reciprocal space information. It is also important that the data is normalized to ensure that it is all on the same scale and can be properly compared.

A python function has been designed that will perform this reshape process, normalize the data, apply the KMeans function from sklearn.cluster and return both the cluster map in real space and the mean diffraction pattern, or centroid, corresponding to each label according to the desired number of clusters k.

It is very important to know in which shape the function returns this data and how to retrieve the information we are interested in. Given that the KMeans algorithm used is part of an external library and was not programmed by us, each piece of information had to be identified by deducing it from the data obtained after doing the first tests of the algorithm with our function. Through this procedure it has been verified that the cluster map is contained in an array of sizes m and n corresponding to the first 2 dimensions of the incoming array. The centroids are contained in a second array with a number k of rows of $o \times p$ elements each being o and p the last 2

dimensions of the incoming array and k the number of clusters indicated when calling the function. This second returned array will have to be reshaped by converting each of the k rows into a two-dimensional array of o=144 rows and p=144 columns, thus reconstructing the diffraction pattern designated as the centroid of each of the k clusters.

Once verified the entire process, the aim is to deepen the study of the possibilities of the algorithm by applying the knowledge of the physical phenomenon of diffraction patterns. The goal will be to improve the identification of different orientations and planar defects by better studying the fainter diffraction spots. It will be done by means of a manipulation of the data prior to the application of the KMeans algorithm.

In order to better study the fainter spots, a mask will be applied in each of the diffraction patterns to cover the transmitted beam. In this way, more contrast will be achieved between the fainter spots, given that the function with which KMeans is applied normalizes the data.

IV. EXPERIMENTAL DETAILS

To test the algorithm, two regions of a sample of $Bi_{0,425}Na_{0,425}Ba_{0,15}Ti\ O_3$ will be analyzed, which we will call region 1 and region 2. The 4D STEM data of region 1 corresponds to a 4D array of dimensions (181, 172, 144, 144). This indicates that the acquired images comprise a total of 181x172 (x and y axes in Cartesian coordinates) scanned spots from each of which a diffraction image of 144x144 pixels has been obtained. The 4D STEM data of region 2 corresponds to a 4D array of dimensions (131, 166, 144, 144). In this case, a total of 131x166 spots were scanned from which diffraction images of 144x144 pixels were extracted. The data was obtained in a Jeol 2100 TEM operating at 200kV.

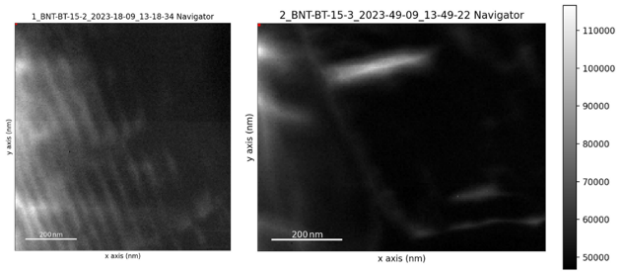

FIG. 2: Images reconstructed from the diffraction patterns of region 1 of the sample on the left and region 2 on the right.

V. RESULTS AND DISCUSSION

To find an appropriate number of clusters, we will run the algorithm on the region 1 data for various numbers of clusters starting with k=4 and then compare the results. By redistributing the data as explained and plotting properly, the following graphs have been obtained.

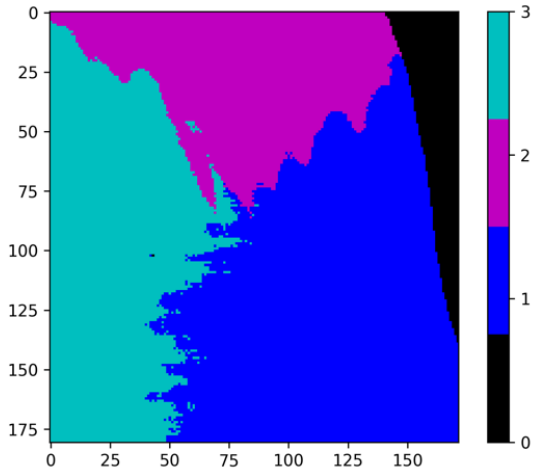

FIG. 3: Map of clusters for 4 clusters in a real-space reconstruction of region 1. Each color corresponds to each of the four labels.

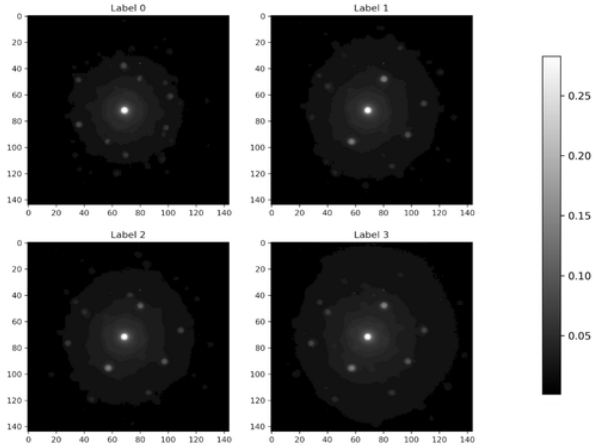

FIG. 4: Each of the 4 diffraction patterns corresponding to the centroids resulting from running the algorithm for four clusters in the data of region 1. From left to right and from top to bottom the centroids corresponding to the labels 0,1,2 and 3 in figure 3. With gray scale bar.

It has been identified and indicated in figure 4 which centroid corresponds to each label and, therefore, to each zone of the real space map of clusters (figure 3).

A 5-cluster and a 6-cluster approach have also been considered. Results are shown in figures 5 and 6 for the 5 clusters and in figures 7 and 8 for 6 clusters.

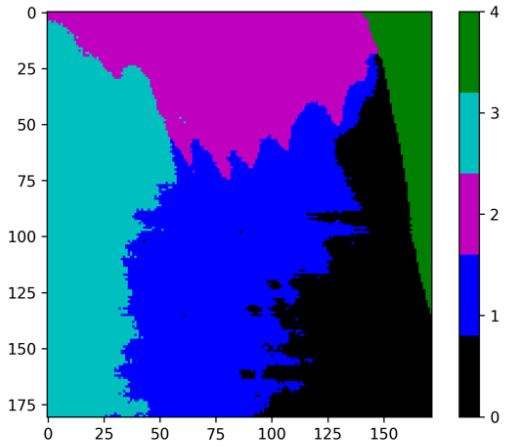

FIG. 5: Map of clusters for 5 clusters in a real-space reconstruction of region 1. Each color corresponds to each of the five labels.

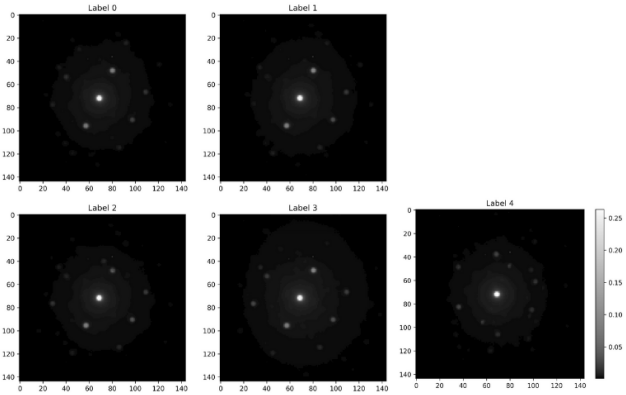

FIG. 6: Each of the 5 diffraction patterns corresponding to each centroid resulting from running the algorithm for 5 clusters on the data from region 1. From left to right and from top to bottom the centroids corresponding to labels 0,1,2,3, and 4 in figure 5.

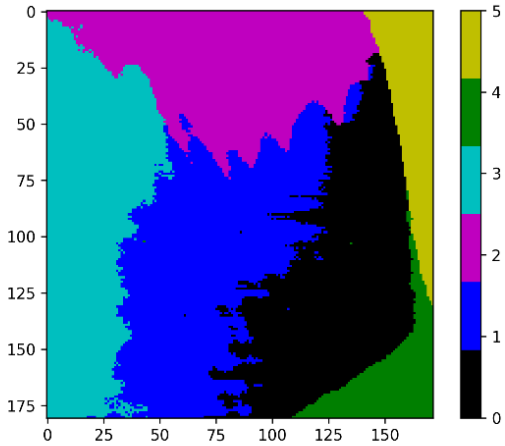

FIG. 7: Map of clusters for 6 clusters in a real-space reconstruction of region 1. Each color corresponds to each of the six labels.

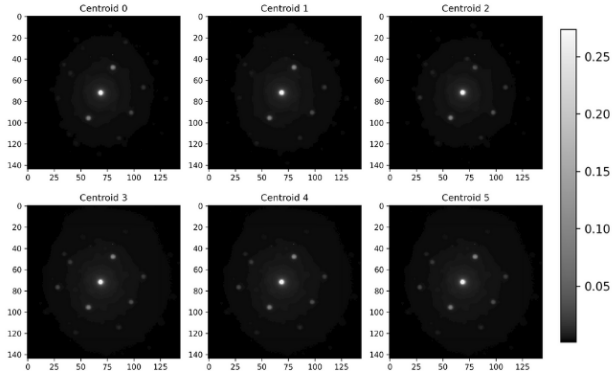

FIG. 8: From left to right and from top to bottom, diffraction patterns (labelled 0,1,2,3,4 and 5 in figure 7) that act as centroids resulting from running the algorithm for 6 clusters in the data for region 1. With gray scale bar.

The results for 6 clusters are considered satisfactory. Once k=6 is selected, the next step is to test the algorithm with the other sample region. We must consider that the dimensions of the array containing region 2 data have a different size than that of the region 1, therefore, the process to retrieve the information should be re-adjusted to region 2 data.

The graphs obtained from the analysis of the region 2 data are shown in figures 9 and 10.

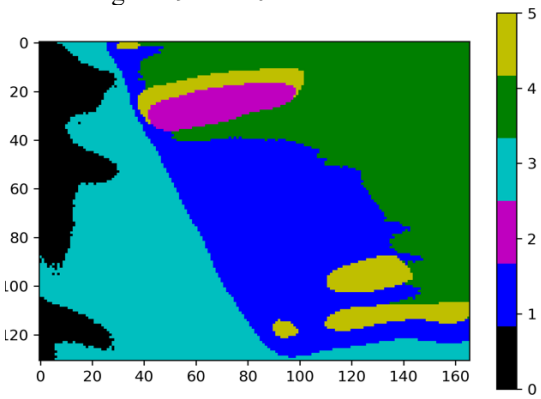

FIG. 9: Map of clusters for 6 clusters in a reconstruction of the real space of region 2. Each color corresponds to each of the 6 labels.

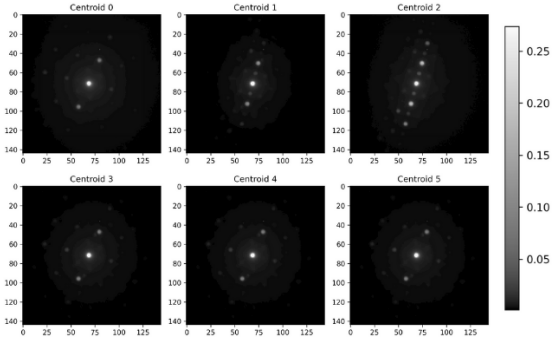

FIG. 10: From left to right and from top to bottom, diffraction patterns (labelled 0,1,2,3,4 and 5 in figure 9) that act as centroids resulting from running the algorithm for 6 clusters in the data for region 2. With gray scale bar.

We will focus on the data from the second region, which seems to have more variety, to test the application of masks and subsequent classification.

A mask covering the central beam is applied to the region 2 data as described in Section III and KMeans is run for 6 clusters. The resulting cluster map is shown in figure 11. The centroids are given in figure 12.

Interestingly, these figures, upon close examination, yield information on the different regions oriented differently: Clusters 4, 3,2 and 1 correspond to different orientations of the crystal.

Also, we can see, within one of the orientations, the one corresponding to cluster 4, the presence of crystalline defects, given by cluster 5.

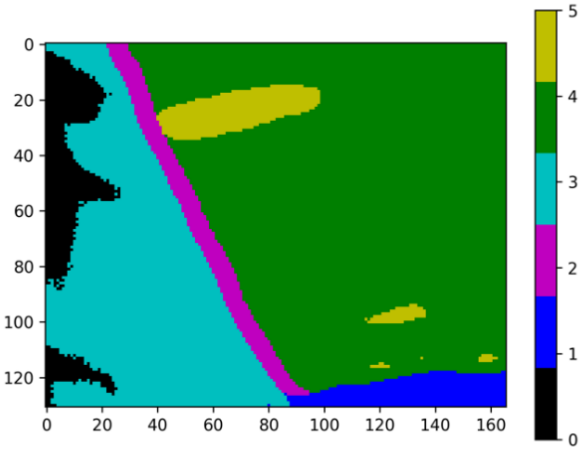

FIG. 11: Map of clusters for 6 clusters in a reconstruction of the real space of region 2 once the mask covering the transmitted beam has been applied. Each color corresponds to each of the 6 labels.

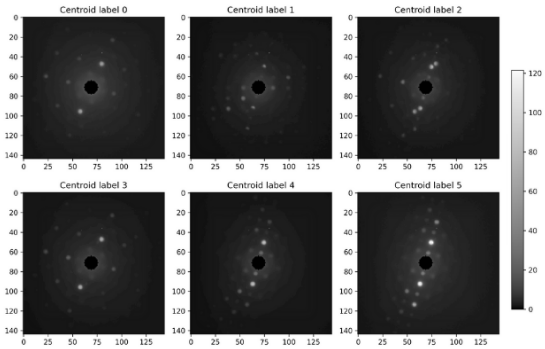

FIG. 12: From left to right and from top to bottom, diffraction patterns (labelled 0,1,2,3,4 and 5 in figure 11) that act as centroids resulting from running the algorithm for 6 clusters in the data of region 2 once the mask that covers the transmitted beam has been applied. With gray scale bar.

VI. CONCLUSIONS

In this work, the KMeans algorithm has been adapted to the classification of 4D STEM data and has been shown to soundly classify the diffraction patterns in 4D STEM. In particular, it has been tested on two 4D STEM datasets obtained from a $Bi_{0,425}Na_{0,425}Ba_{0,15}Ti\ O_3$ sample and has been able to separate regions with different crystal orientations and to identify subregions with crystalline defects within one of these regions.

Acknowledgments

I'd like to thank my advisors Vanessa Costa and Sònia Estradé for all the help and Beatriz Vargas for kindly providing the studied experimental data. I'd also want to thank my family and friends especially my mother. And last but not least, many thanks to the mental health center at the Hospital Clínic.

[1] P. Torruella, M. Estrader, A. López-Ortega, M.D. Baró, M. Varela, F. Peiró, S. Estradé. «Clustering analysis strategies for electron energy loss spectroscopy (EELS)». Ultramicroscopy, 185 (2018), pp. 42-48, 10.1016/j.ultramic.2017.11.010

[2] J. Blanco-Portals, F. Peiró, S. Estradé. «Strategies for EELS data analysis. Introducing UMAP and HDBSCAN for dimensionality reduction and clustering. » Microsc. Microanal., 28 (1) (2022), pp. 109-122, 10.1017/S1431927621013696

[3] D.B. Williams, C.B. Carter, «Transmission electron microscopy: a textbook for materials science» Springer Science. New York. 2009

[4] Ophus C, Ercius P, Sarahan M, Czarnik C, Ciston J. «Recording and Using 4D-STEM Datasets in Materials Science». Microscopy and Microanalysis. 2014,20(S3):62-63. doi:10.1017/S1431927614002037