doi:10.1017/S1431927610054711
Microsc. Microanal. 16 (Suppl 2), 2010


# An Extremely Fast and Efficient Hierarchical Clustering Algorithm Applicable to Hyperspectral Microscopy and Microanalysis Images

C.L. Stork*

* Sandia National Laboratories, Materials Reliability Department, PO Box 5800, Albuquerque, NM 87185-0886

While hyperspectral imaging of materials generates large volumes of data (typically, greater than 10000 pixels and 100 spectral channels), a significant challenge is encountered in converting the data to useful information regarding material composition. Factor analysis is one approach that has been used to extract chemical component information from hyperspectral image data [1]. Clustering is an alternative approach that can be employed to classify the data into groups [2]. Hierarchical clustering is an important tool for understanding the similarities and relationships between samples in a data set, and is routinely used in the analysis of relatively small data sets, e.g., when the number of samples is less than 200. Hierarchical clustering organizes a set of samples into a hierarchy of clusters, based on the distances of the clusters in the variable or measurement space. Hierarchical clustering, however, is typically not applied to hyperspectral image sets due to computational and computer storage limitations. Conventional hierarchical clustering algorithms require the calculation and updating of a pair wise cluster dissimilarity matrix. A problem arises, however, in calculating and storing this cluster dissimilarity matrix for large data sets. As a case in point, for a hyperspectral image set composed of 10000 pixels, this dissimilarity matrix will initially be of dimensions 10000 by 10000, resulting in out-of-memory errors on a standard desktop computer.

This talk will describe an Extremely Fast and Efficient Hierarchical Clustering Algorithm (EFE-HCA) that reduces computational time for large data sets from days to a few minutes. EFE-HCA can be routinely applied to very large hyperspectral microscopy and microanalysis-based image sets, in contrast with commercially available software that produces out-of-memory errors. Unlike conventional hierarchical clustering algorithms, EFE-HCA does not calculate the cluster dissimilarity matrix, but instead tracks the nearest neighbor and nearest neighbor distance for each cluster [3]. Three factors contribute to the speed and efficiency of EFE-HCA: (a) Principal component analysis (PCA) is applied to the data set to reduce the number of variables that need to be processed in hierarchical clustering; (b) In the initial step of determining the nearest neighbor and nearest neighbor distance for each sample, mean values of sample vectors are calculated in order to reduce the number of required full vector distance calculations; (c) The nearest neighbor and nearest neighbor distance for each cluster are updated as the hierarchical clustering algorithm proceeds.

Figure 1 depicts the process of applying EFE-HCA to hyperspectral images of six types of wires embedded in an epoxy matrix. The material sample was imaged using a scanning electron microscope (SEM) with an attached energy dispersive X-ray spectrometer (EDS). The left hand side of Figure 1 shows two of the 1024 analyzed images. Each image contains 128 rows and 128 columns or 16384 pixels. The upper right hand portion of Figure 1 shows the dendrogram generated for the EDS data set using EFE-HCA. An inspection of this dendrogram provides information regarding the interrelationships between chemical phases within the material sample. The lower right hand portion of Figure 1 depicts the clusters identified by EFE-HCA. Notably, EFE-HCA successfully extracts clusters delineating the six wire types and the epoxy phase.

References
[1] P.G. Kotula et al., *Microsc. Microanal.* 9 (2003) 1.
[2] N.C. Wilson and C.M. MacRae, *Microsc. Microanal.* 11 (Suppl. 2) (2005) 434.
[3] M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973, 145.
[4] The author thanks Paul Kotula at Sandia National Laboratories for providing the wires EDS data set. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy (DOE) under contract DE-AC0494AL85000.

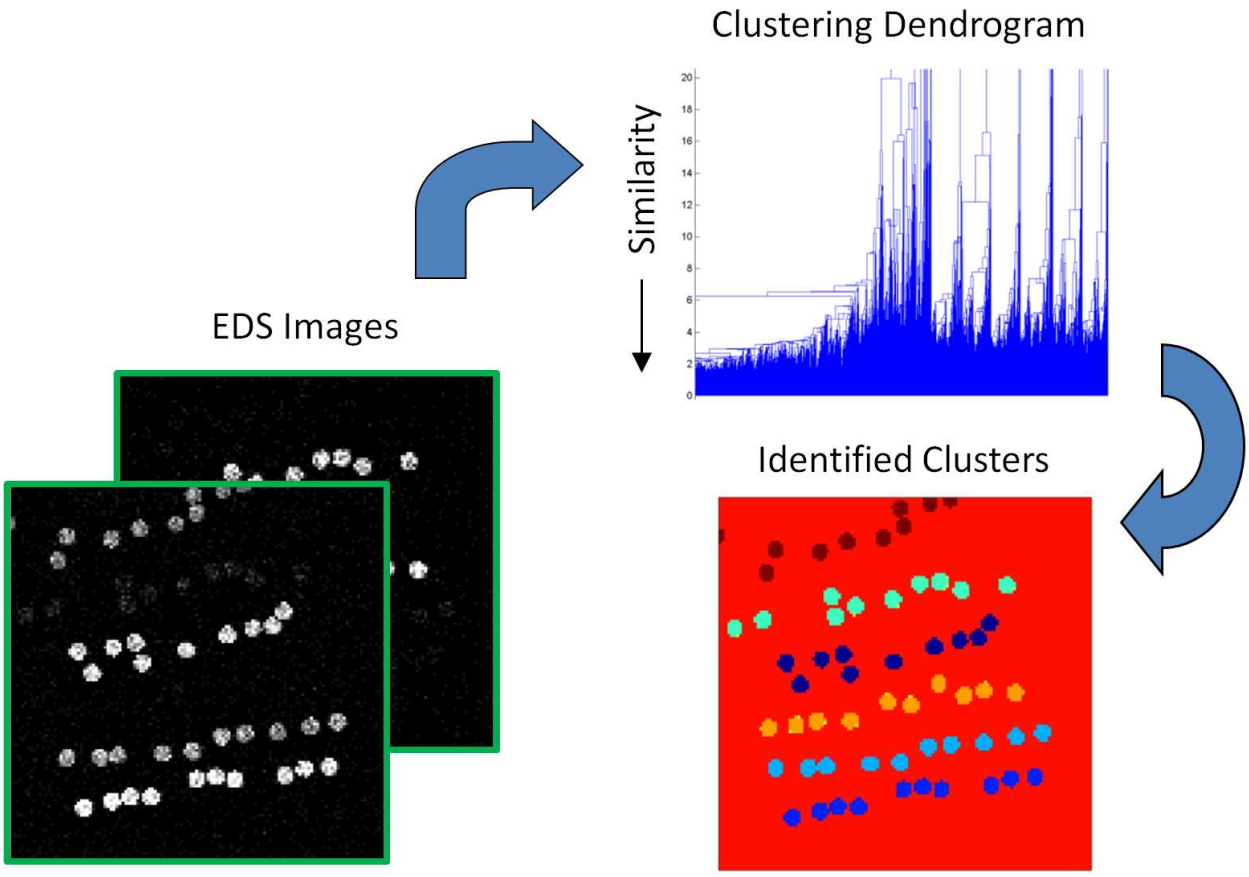


FIG. 1. Depiction of the process of applying EFE-HCA to hyperspectral EDS images of six types of wires embedded in epoxy. The left hand side shows two of the 1024 analyzed images, the upper right hand portion depicts the dendrogram generated using EFE-HCA, and the lower right hand image presents the clusters identified by EFE-HCA (each cluster is assigned a different color).

https://doi.org/10.1017/S1431927610054711 Published online by Cambridge University Press