Orange Project 3 - Ioannis Manousaridis

In this project, clustering methods were applied to the MNIST and F-MNIST datasets.

Classification, which was used in the previous projects, and Clustering are the two types of learning methods which characterize objects into groups by one or more features. These processes appear to be similar, but there is a difference between them in context of data mining. The prior difference between classification and clustering is that classification is used in supervised learning technique where predefined labels are assigned to instances by properties, on the contrary, clustering is used in unsupervised learning where similar instances are grouped, based on their features or properties and the labels are not known in advance.

The clustering methods that were applied to the MNIST dataset can be seen below. Three classes were selected, the numbers 0,3,7 and only 400 samples.

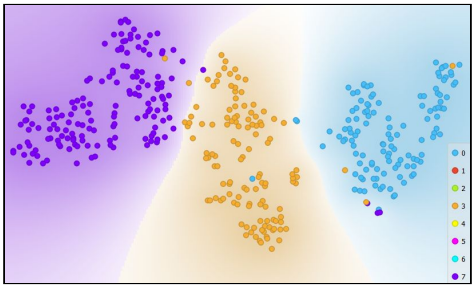

FIgure 1: The MNIST dataset with the right labels. The classes 0,3,7 are used and 400 samples.

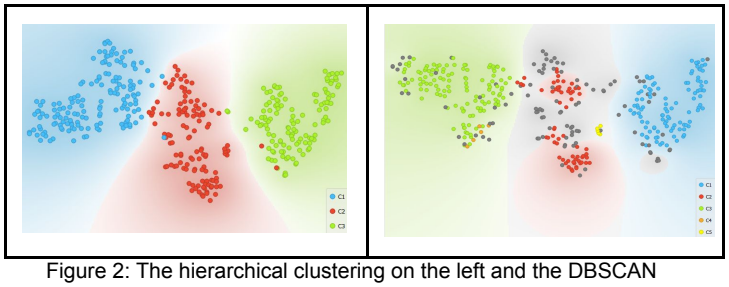

Figure 2: The hierarchical clustering on the left and the DBSCAN clustering on the right.

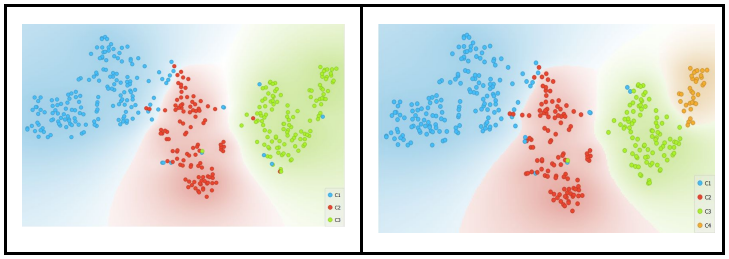

Figure 3: The k-Means method for 3 clusters on the left and for 2-7 clusters on the right.

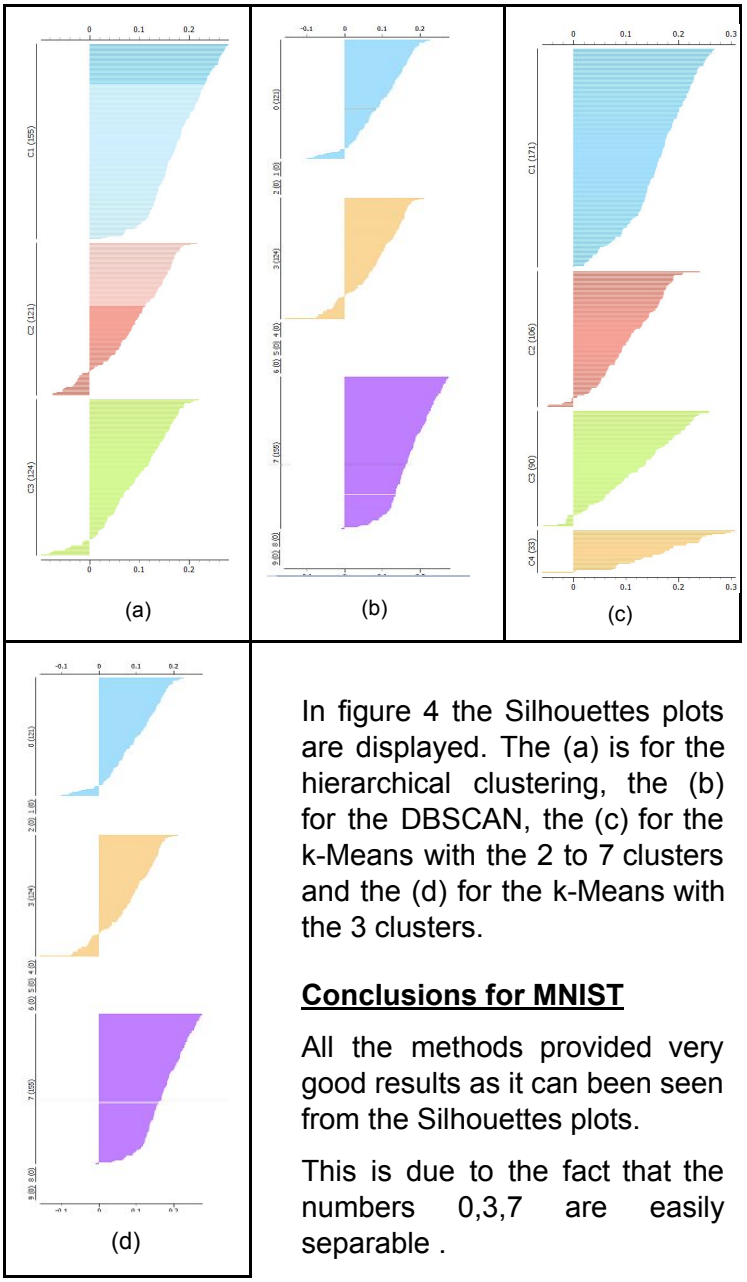

Figure 4: Silhouette plots for the clustering methods

In figure 4 the Silhouettes plots are displayed. The (a) is for the hierarchical clustering, the (b) for the DBSCAN, the (c) for the k-Means with the 2 to 7 clusters and the (d) for the k-Means with the 3 clusters.

Conclusions for MNIST

All the methods provided very good results as it can been seen from the Silhouettes plots.

This is due to the fact that the numbers 0,3,7 are easily separable .

The DBSCAN gave the poorest results. Each cluster has high values to the Silhouette plot but in general the clustering is not so well.

A lot of samples are do not belong in any cluster and are shown as gray.

The DBSCAN is for shape - clustering problems and that's why it is unable to cluster all the samples well.

The clustering methods that were applied to the F-MNIST dataset can be seen below. Three classes were selected, the labels 0,5,8 and only 400 samples.The label 0 is for the T-Shirts, the label 5 is for Sandals and the label 8 is for Bags.

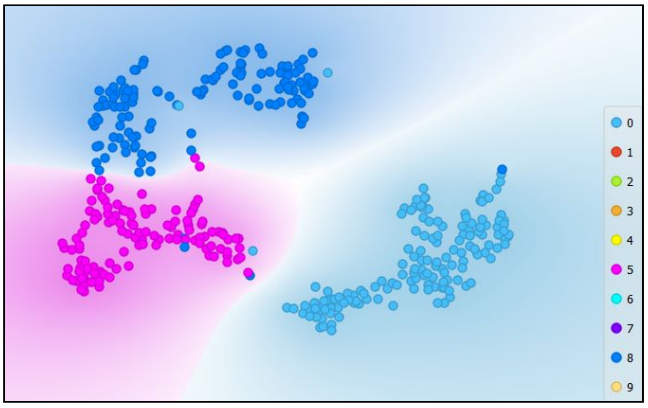

FIgure 5: The F-MNIST dataset with the right labels. The classes 0,5,8 are used and 400 samples.

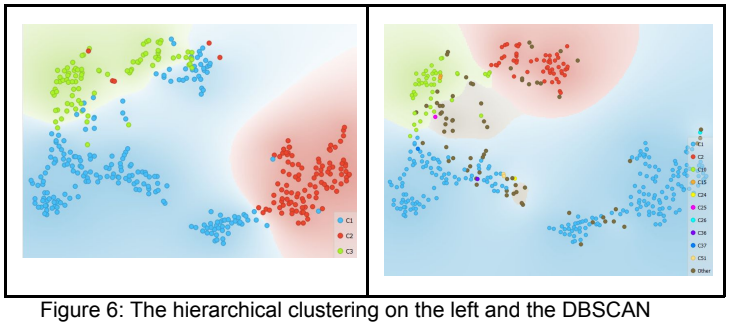

Figure 6: The hierarchical clustering on the left and the DBSCAN clustering on the right.

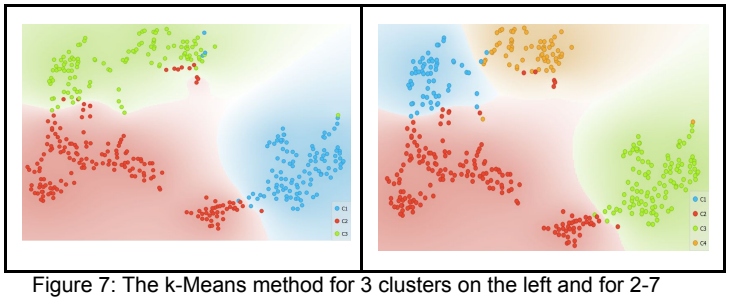

Figure 7: The k-Means method for 3 clusters on the left and for 2-7 clusters on the right.

In figure 4 the Silhouettes plots are displayed. The (a) is for the hierarchical clustering, the (b) for the DBSCAN, the (c) for the k-Means with the 2 to 7 clusters and the (d) for the k-Means with the 3 clusters.

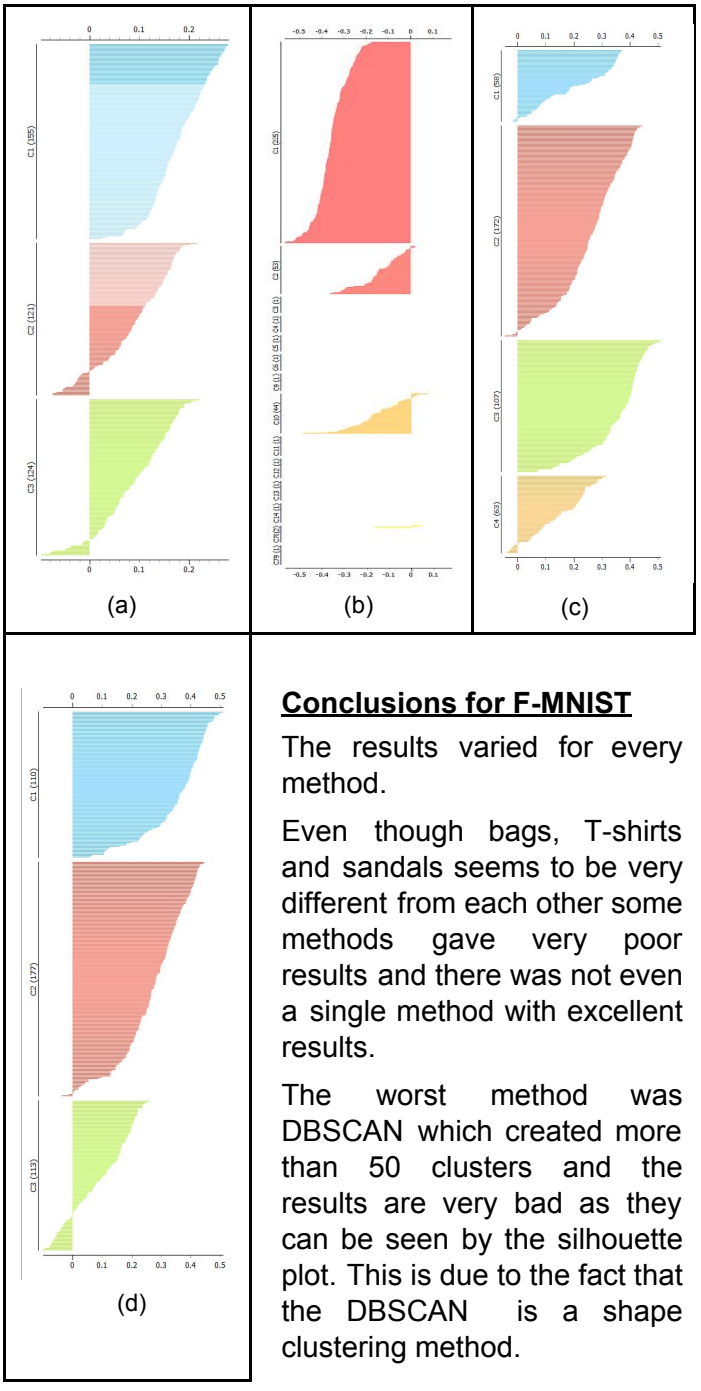

Figure 8: Silhouette plots for the clustering methods

Conclusions for F-MNIST

The results varied for every method.

Even though bags, T-shirts and sandals seems to be very different from each other some methods gave very poor results and there was not even a single method with excellent results.

The worst method was DBSCAN which created more than 50 clusters and the results are very bad as they can be seen by the silhouette plot. This is due to the fact that the DBSCAN is a shape clustering method.

The hierarchical clustering gave good results but not perfect. A big number of data that belongs to the t-shirts and to bags were included in the scandals cluster.

The k-Means method gave again the highest results but this time the results were not excellent.