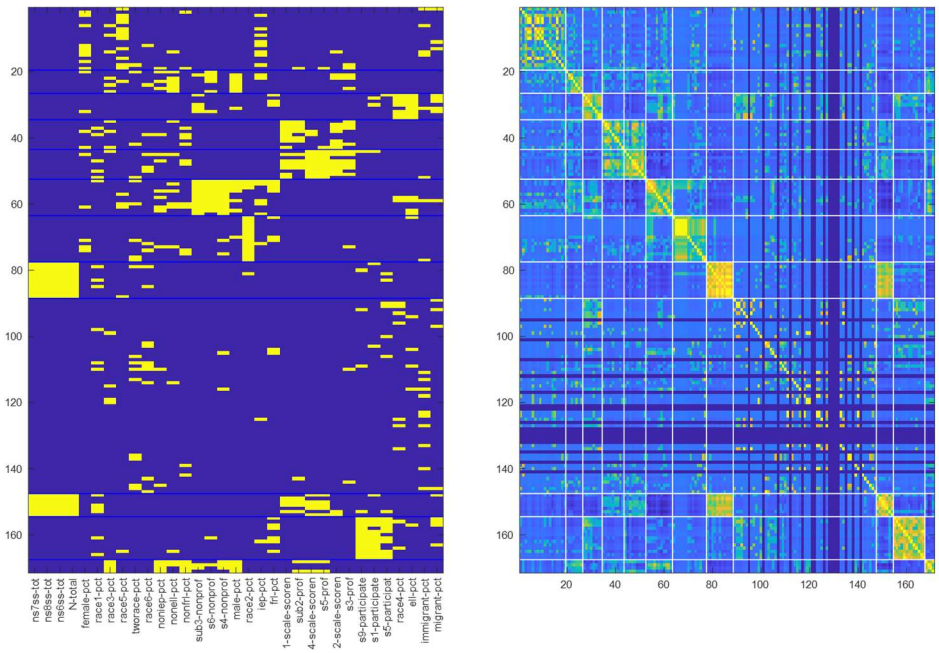
Cluster Show v_0.1

This tool helps people to quickly understand a dataset.

Written in Matlab, it combines different clustering and visualization algorithms.

Some Examples:

```
%% example 1, on a school clustering dataset
load('./demo_data/data_school.mat');
addpath('./fun/');

%%
clus_alg = 'kmeans';
dis_type = 'euclidean';
n_row_clus = 12;
n_col_clus = 5;
f_clus_n_show(X, fe_label, clus_alg, dis_type, n_row_clus, n_col_clus);
```

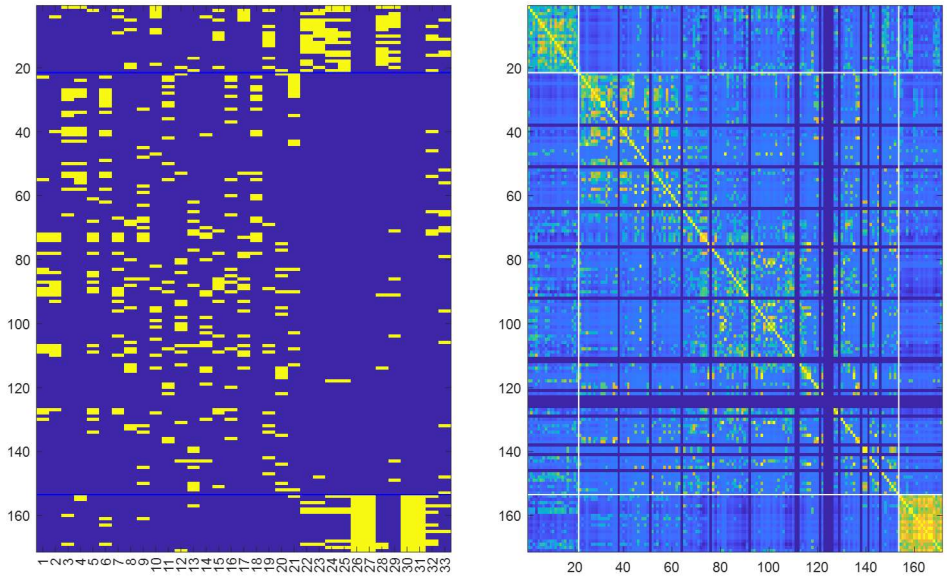


The data is a 2-dimensional array: each row is an instance - a school in this example; each column is a feature – they are the school performance on different subjects and the demographic information in this example. Here this example contains 171 schools and 33 features.

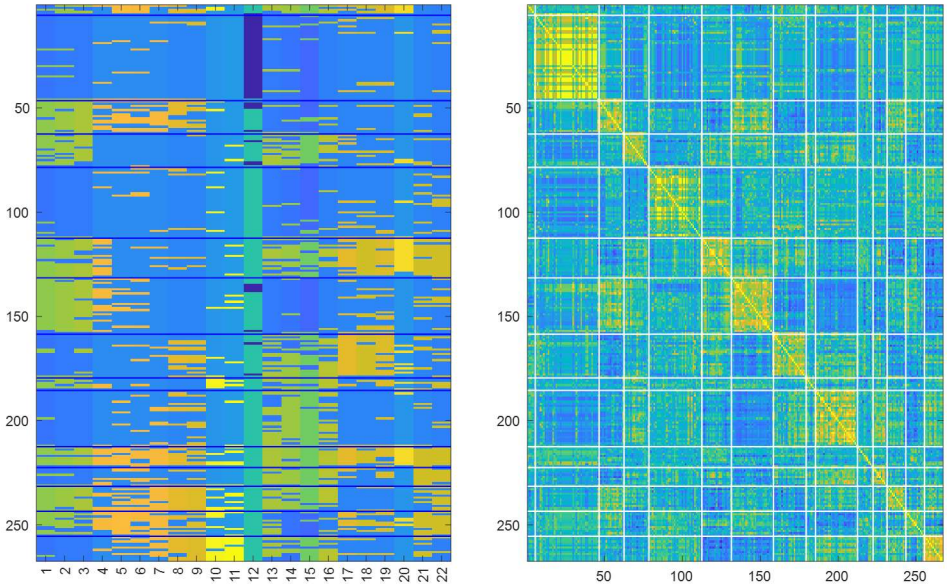
We can specify the clustering algorithm and the distance type to be used. If we choose ‘kmeans’ clustering, we also need to specify the desired number of clusters. If we do not want to specify the number of clusters, we can choose to use 'Qcut' or 'HQcut'. Current version support 3 clustering algorithms, 'kmeans', 'Qcut' and 'HQcut'. More clustering algorithm will be added. The current supported distance types are 'euclidean', 'cosine' and 'corr'.

The visualization contains two sub-figures - the figure on the left side is the clusters view on the original dataset, which shows different clusters of schools. Looking down from the highlighted block within a cluster, we can see the feature names, and these common features of a cluster might show some interesting insights of the dataset. On the right side, it shows the clustered similarity matrix of the instances. Each element (pixel) shows the similarity level between two instances. The brighter (yellow) the color, the more similar between two instances; the darker (blue) the color, the more different between the two instances. Each rectangle area shows the similarities of the elements between two clusters. A bright rectangle means the two clusters are very similar. Here on the above figure, most of rectangles (square actually) on the diagonal are much brighter than the other rectangle not on the diagonal, which means that the clustering result is very good – the elements inside each clusters are highly similar, while they are very different from all the result elements.

```
%% example 2, on the iris dataset
load('./demo_data/data_school.mat');

%%
clus_alg = 'kmeans';
dis_type = 'euclidean';
n_row_clus = 3;
n_col_clus = 2;

X = normalize(X);
[X] = f_discrete_data(X);

f_clus_n_show(X, '', clus_alg, dis_type, n_row_clus, n_col_clus);
```




```
%% example 3, on data planning
load('./demo_data/data_planning.mat');
n_row_clus = 15;
n_col_clus = 5;

X = normalize(X);
%[X] = f_discrete_data(X);

f_clus_n_show(X, '', clus_alg, dis_type, n_row_clus, n_col_clus);
```


50
100
150
200
250
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
50
100
150
200
250
50 100 150 200 250