**Can Number of Clusters Be Solved by External Index?**

**Instructions for running the codes for the experiments**

1. **Generating subsets:** Run the bat file sub2\_disjoint.bat on cs server
* The dataset and the desired size of subset should be specified.

Example: if dataset size is 5000, and you want subsets to have 20% of the original dataset points, the subset size should be set 1000.

* sub1\_disjoint.bat, sub0\_disjoint.bat, cbsubset and cbmodify are also used during the process of generating subsets.
* Currently, 100 subsets are generated. If you want to change the number, modify sub1\_disjoint.bat.
* Set the path of datasets in sub0\_disjoint.bat.
* For an example dataset s1 with 5000 objects:
	+ There should be two empty folders s1 and s1\_train beside bat files.

If you want different path for subsets, modify sub0\_disjoint.bat.

The subsets are generated in s1.

When you set the subset size 1000, other subsets with the size 4000 are also generated in the folder s1\_train, which we can use them for training in the classification-based approach. If you do not want them, just do not create ‘s1\_train’ folder, and ignore the errors happening regarding this.

* + The subsets are named like this: s1-01.ts, s1-02.ts,…, s1-100.ts.
	+ Index files are also generated: idx-01.txt, idx-02.txt, ..., idx-100.txt. An index file specifies the index of the selected objects for a subset in the original dataset.
	+ Note that the permission of bat files should allow execution.

* Generating null reference (uniform) and its subsets
	+ Use the MATLAB code generateNullReference.m. The code needs the dataset in txt format. The output for the dataset s1 will be s1\_null.txt.
	+ Convert the txt file to ts format on the server using txt2ts.bat. Edit txt2ts.bat for your dataset.
	+ Subsets can be generated similarly as s1. There is no need to create ‘s1\_null\_train’ folder.
1. **Clustering (on cs server)**
* Put the original dataset in the same folder in its subsets.
* Three clustering algorithms: k-means (KM), random swap (RS), and genetic algorithm (GA)
* Run clustering\_main.bat
* clustering\_all\_k.bat, km\_one\_k.bat, rs\_one\_k.bat, ga\_one\_k.bat, km\_one\_subset, rs\_one\_subset, and ga\_one\_subset are also used during clustering.
* clustering\_main.bat can run clustering for different datasets with the same range for number of clusters for example for s1-s4 (k=2-25). For different range of k, edit clustering\_all\_k.bat
* Check number of subsets in km\_one\_k.bat or rs\_one\_k.bat, or ga\_one\_k.bat. Currently is 100.
* Check number of iterations for RS in rs\_one\_k.bat and rs\_one\_subset.bat
* Check parameters for GA and other parameters in ga\_one\_k.bat and ga\_one\_subset.bat
* Edit the path for output pa files in km\_one\_k.bat, rs\_one\_k.bat, ga\_one\_k.bat, km\_one\_subset, rs\_one\_subset, and ga\_one\_subset
1. **Cross-validation (CV) approach for estimating the number of clusters (MATLAB)**
* Use clustering results (pa files), original dataset (txt format), and index files
* Use the code main\_CV.m
* The datasets should be here: in the “code” folder, there should be a “dataset” folder, and inside it, the folder with the name of dataset, for example s1, which includes original dataset, and the index files related to subsets.
* Specify the folder of clustering results
* Set the parameters for a specific dataset.
1. **Classification-based (CB) approach for estimating the number of clusters (MATLAB)**
* Use the code main\_CB.m
* Set datasets and clustering results folders as for CV
* Set the parameters for a specific dataset.
1. **Randomized algorithm (on cs server)**
* Do clustering on the server by running Run clustering\_main\_RA.bat. The parameter is the number of repeats for RS.
* For 100 repeat, there should be a folder with the name “s1\_rs\_RA100” for dataset s1.
* It runs RS algorithm with different iterations (1, 10, 100, 1000, 5000), 50 times.
* Check the parameters in the related bat files.
* For estimating number of clusters, use the MATLAB code main\_CV\_RA.m. Set the parameters for a specific dataset.