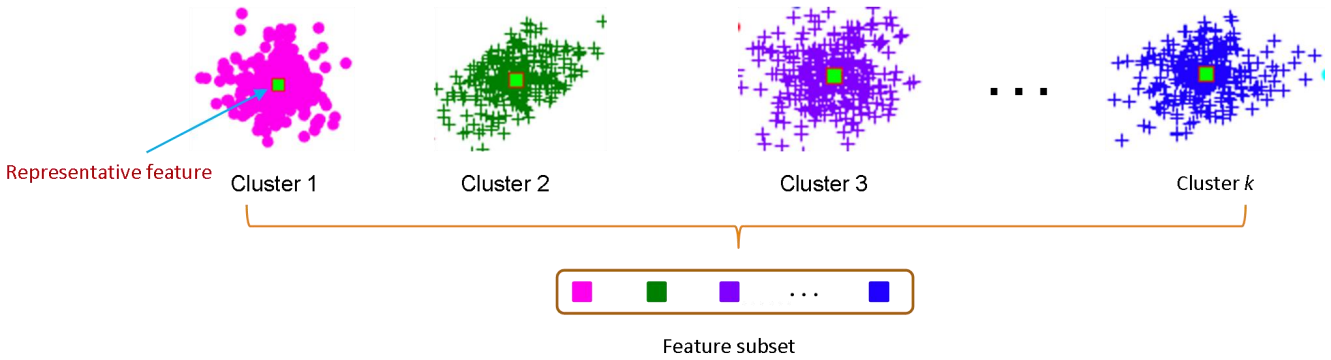
Unsupervised Spectral Feature Selection Algorithms for High Dimensional Data

Mingzhao WANG, Henry HAN, Zhao Huang , Juanying XIE

Frontiers of Computer Science, DOI: 10.1007/s11704-022-2135-0

Problems & Ideas

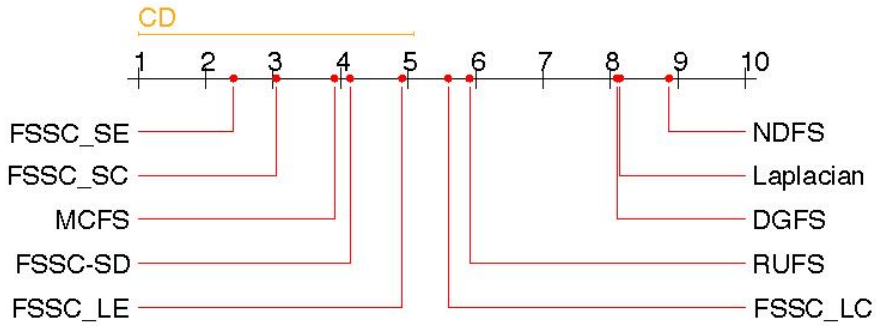
- Problems of carrying out explainable analysis for high dimensional data:
  - Biomedical data often comprise a small number of exemplars while having tens to thousands features for each sample.
  - It is very challenging to detect the informative to carry out explainable analysis for this type of high dimensional data.
  - How to evaluate the representative of a feature is the first thing to be considered seriously.
- Ideas: Feature selection especially the unsupervised ones are the right way to deal with the challenge and realize the task.



The idea of the unsupervised feature selection algorithms. Features are grouped with respect to the similarity between them, then the representative feature from each cluster will be detected out to comprise the feature subset.

Main Contributions

- Contributions:
  - Two unsupervised spectral feature selection algorithms referred to FSSC_SE and FSSC_SC are proposed, due to the advantages of spectral feature selection;
  - They group features using the advanced Self-Tuning spectral clustering algorithm based on local standard deviation, so as to detect the global optimal feature clusters as far as possible;
  - Two feature ranking techniques, including cosine-similarity-based feature ranking and entropy-based feature ranking, are proposed, so that the representative feature of each cluster can be detected to comprise the feature subset on which the explainable classification system will be built.



The significance comparison results of the 10 feature selection algorithms against each other using the Nemenyi's test. The proposed FSSC_SE and FSSC_SC are superior to the other compared unsupervised feature selection algorithms.