Supplementary material for the paper: "An extension of Random Forest-Clustering schemes which works with partition-level constraints", Proc. Int. Conf. on Pattern Recognition (ICPR) 2024

Manuele Bicego[1][0000−0002−1008−3917] and Hafiz Ahmad Hassan[1]

Computer Science Department, University of Verona, Verona 37135, Italy
manuele.bicego@univr.it

A Code availability

The Matlab code of the proposed approach is available at https://profs.scienze.univr.it/~bicego/code.html.

B Dataset description

In this section, we briefly summarize the different datasets used in the experimental evaluation: for the sake of reproducibility, for each dataset, we report the source and the preprocessing (if any – in some cases, we slightly edited the original datasets).

- **Iris**: the Iris Data Set, taken from https://archive.ics.uci.edu/ml/datasets/iris, used as provided.
- **Glass**: the Glass Identification Data set, taken from https://archive.ics.uci.edu/ml/datasets/glass+identification; this represents a 4-class version of the original 7 categories problem. In particular, to have more balanced groups, we put together categories 4-7.
- **Ecoli**: the Ecoli Data Set, taken from https://archive.ics.uci.edu/ml/datasets/ecoli. We excluded the first feature (Sequence Name).
- **Seeds**: the Seeds Data Set, taken from https://archive.ics.uci.edu/ml/datasets/seeds, used as provided.
- **Cryotherapy**: the Cryotherapy Data Set, taken from https://archive.ics.uci.edu/dataset/429/cryotherapy+dataset. We considered the first 6 features (the last represents the class).
- **AutoMpg**: the Auto MPG Data Set, taken from https://archive.ics.uci.edu/ml/datasets/auto+mpg. We transform this problem into a classification problem by setting a threshold of 25 on the first feature; we also remove the feature with missing values.

- **Imox**: the Imox Data Set, taken from the Prtools repository `https://37steps.com/prtools-guide/overview/`, used as provided.
- **StoneFlakes**: the StoneFlakes Data Set, taken from `https://archive.ics.uci.edu/ml/datasets/StoneFlakes`. We removed objects with missing values (9 objects).
- **Pima**: the Pima Indian diabetes dataset, taken from `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`, used as provided.
- **Nose**: part of the Gas array drift dataset, taken from `http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset`. We used batches 4 and 5, which correspond to months 14, 15, and 16.
- **Microarray**: the brain tumor dataset [4]; we apply a variance filtering, keeping only the first 100 genes.
- **Flickr**: this is a part of the dataset "Favorite Images", a dataset used to study personal aesthetics, which can downloaded from `http://www.cristinasegalin.com/research/projects/phd/soft_biometrics/Dataset_IEEEForensics.zip`. Here we selected only the first 5 users, and use the features as provided.
- **Isolet**: this is a part of the Isolet Data Set, taken from `https://archive.ics.uci.edu/ml/datasets/isolet`. Here we selected only the first 4 users, and used the features as provided.
- **EEGEyestate**: a part of the EEG Eye State Data Set, taken from `https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State`. Here we used only the first 10 seconds of recording, maintaining the features as provided.
- **UAV1IntDet1, UAV1IntDet2, UAV1IntDet3 and UAV1IntDet4**: parts of the Unmanned Aerial Vehicle (UAV) Intrusion Detection Datasets, available from `http://mason.gmu.edu/~lzhao9/materials/data/UAV/`. The 4 datasets cover the different variants from the website, where we considered only the first 1100 signals.

C Comparison with the unsupervised scenario: results dataset per dataset

We reported here the comparison, dataset per dataset, of the constrained versions of RFC with the corresponding unconstrained versions. Results are displayed in Fig. 1, 2, and 3, for the RFC-Shi, RFC-Zhu and the RFC-RatioRF schemes, respectively.

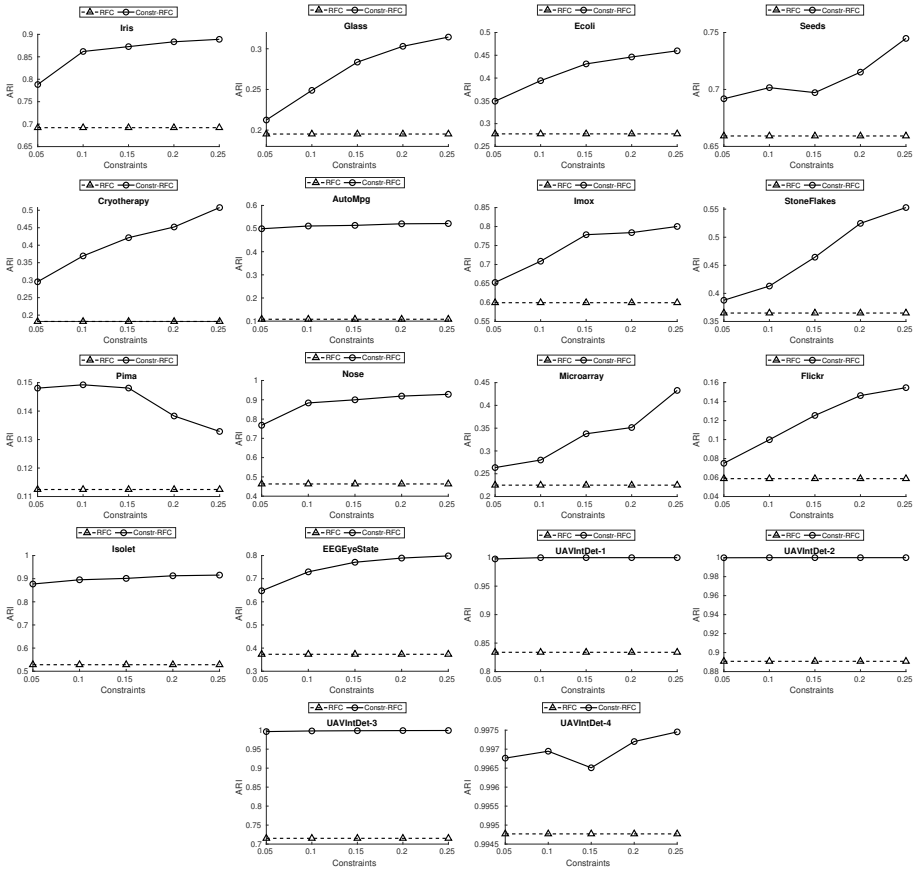

Fig. 1. Comparison between constrained and unconstrained strategies, dataset per dataset, for the RFC-Shi scheme.

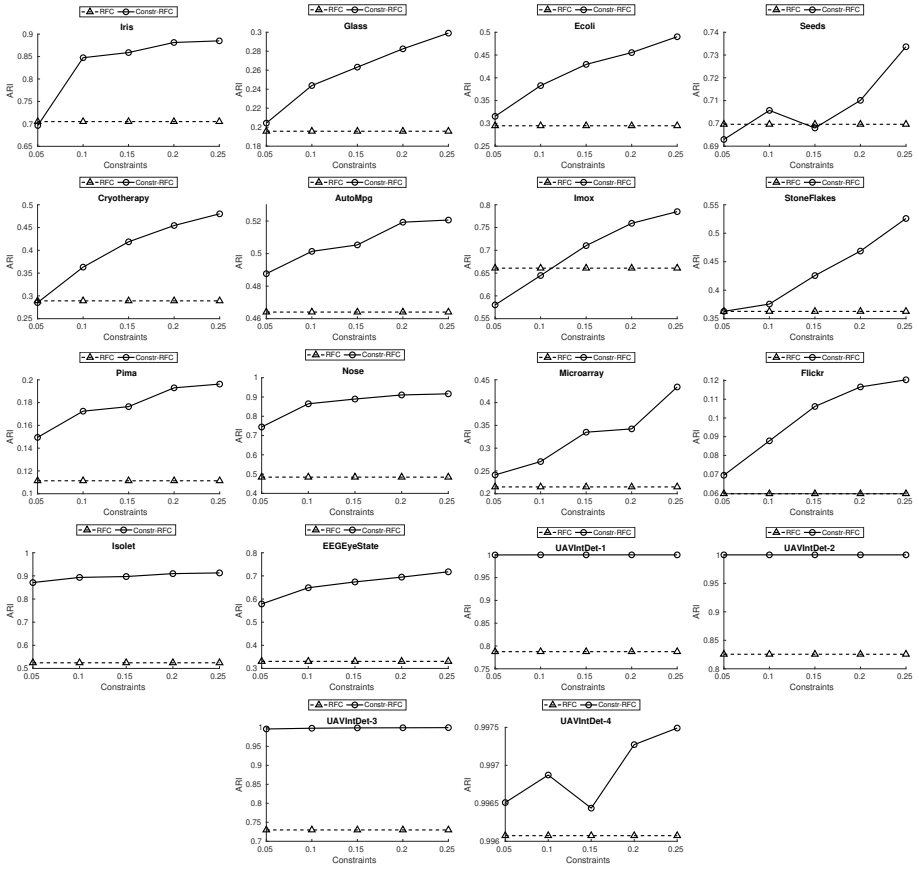

Fig. 2. Comparison between constrained and unconstrained strategies, dataset per dataset, for the RFC-Zhu scheme.

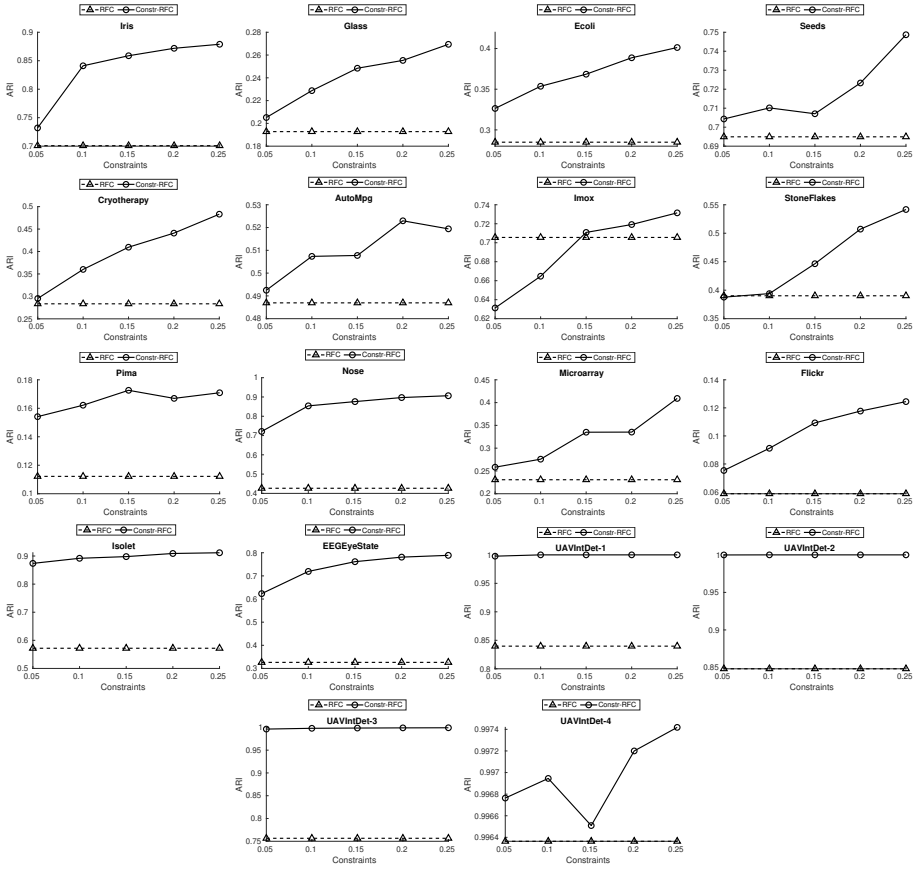

Fig. 3. Comparison between constrained and unconstrained strategies, dataset per dataset, for the RFC-RatioRF scheme.

D Comparison with literature alternatives: description of the methods and results dataset per dataset

In this section, we will provide the description of the methods used for comparison, with the implementation details, together with the detailed results dataset per dataset. In particular, in our comparison, we considered the following alternatives:

- **COP-KM**: the COP-Kmeans algorithm of [5], one of the first and the most applied constrained clustering algorithm. In order to increase the robustness, we run the algorithm 20 times, starting from random initialization, keeping the result leading to the minimum of the optimization function. This algorithm performs hard constraint satisfaction, meaning it fails if all constraints are not satisfied. If none of the 20 runs returns a viable solution, we simply adopt the K-means solution. Here we use the Matlab implementation available in Github[1].
- **Adv-COP-KM**: a variant of COP-Kmeans, recently introduced in [1], which permits the violation of some low-priority constraints, so that a solution can always be found. In this case, we used the Matlab implementation available from the authors's Github page[2].
- **LCVQE**: another well-known and widely applied constrained clustering method, proposed in [2]. Here we used the Matlab implementation available in internet[3].
- **Boost-COP-KM**: a recent constrained clustering algorithm based on boosting, introduced in [1]. We used the Matlab implementation available from the authors's Github[4], keeping the parameters set to their default values.
- **WSSR+**: a very recent constrained spectral clustering algorithm, proposed in [3]. We employed the Matlab code available from the author's Github page[5], setting all parameters as suggested in the manuscript.

For all those methods which work with pairwise constraints (Must Link and Cannot Link), we simply transform partition level constraints into instance level constraints by considering all pairs of objects and assigning a MustLink constraint if the two objects have the same label, a CannotLink constraint otherwise.

In figure 4 and 5 we will provide the comparison dataset by dataset between our approaches and all alternatives, in terms of ARI and NEC ratio, respectively.

[1] `https://github.com/kemalty`
[2] `https://github.com/mokabe1567/bckm/blob/master/bckm`
[3] `https://danielkhashabi.com/files/2015_constrained_clustering/lcvqe.m`
[4] `https://github.com/mokabe1567/bckm/tree/master/bckm`
[5] `https://github.com/hankuipeng/WSSR/tree/master/WSSRplus`

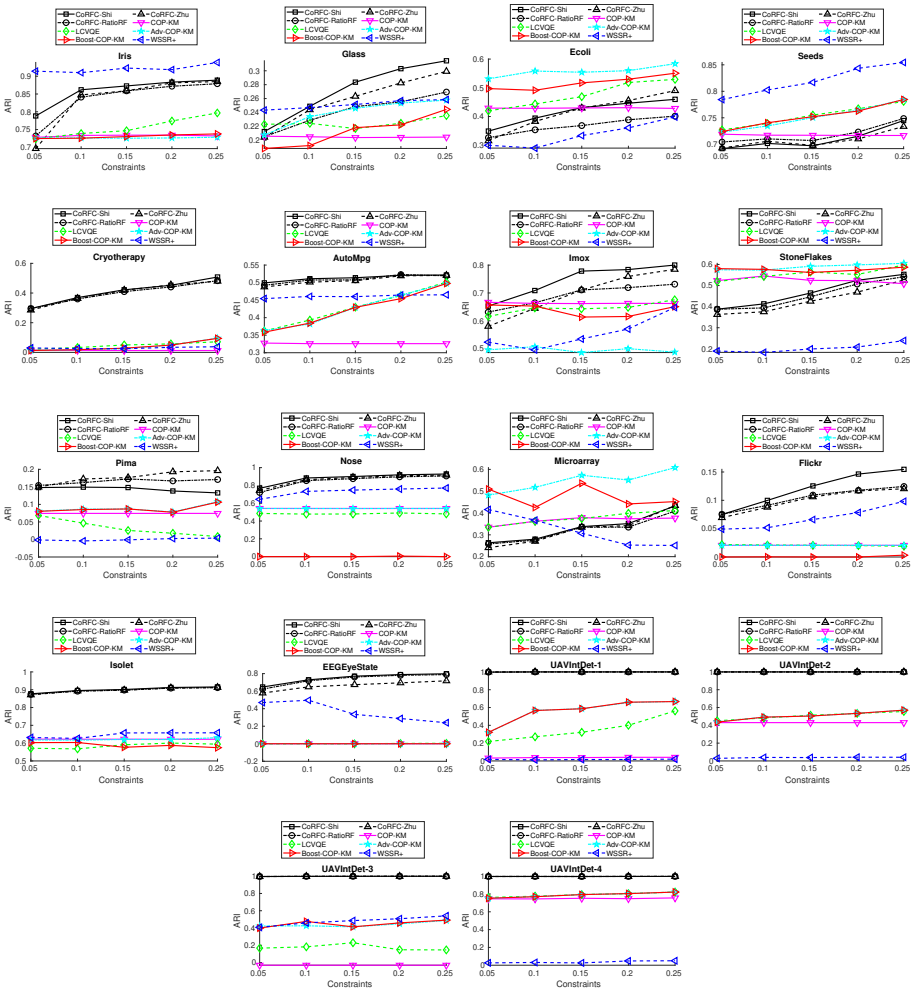

Fig. 4. Comparison between our proposed approaches and some alternatives, dataset per dataset, in terms of ARI.

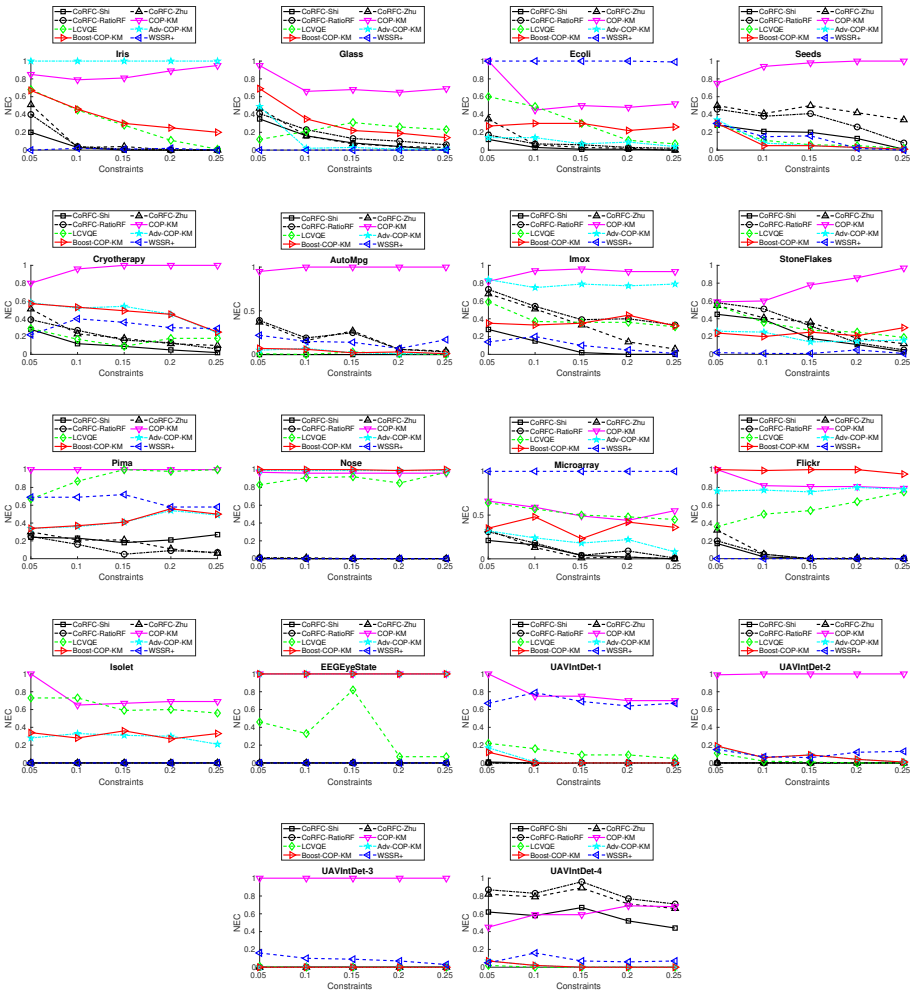

Fig. 5. Comparison between our proposed approaches and some alternatives, dataset per dataset, in terms of NEC ratio.

References

1. Okabe, M., Yamada, S.: Clustering using boosted constrained k-means algorithm. Frontiers in Robotics and AI **5**, 18 (2018)
2. Pelleg, D., Baras, D.: K-means with large and noisy constraint sets. In: Proc. European Conference on Machine Learning. pp. 674–682 (2007)
3. Peng, H., Pavlidis, N.G.: Weighted sparse simplex representation: a unified framework for subspace clustering, constrained clustering, and active learning. Data Mining and Knowledge Discovery **36**(3), 958–986 (2022)
4. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C., et al.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature **415**(6870), 436–442 (2002)
5. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: Proc. Int. Conf. on Machine Learning. vol. 1, pp. 577–584 (2001)