Initial Experiments: Integrating Ensemble Clustering and Unsupervised Feature Selection

Antonio Carlos Rodríguez Bajo

INTRODUCTION

The initial experiments adopt an approach that integrates ensemble clustering using random projections with unsupervised feature selection. The underlying rationale is that eliminating irrelevant features before ensemble clustering improves both clustering accuracy and computational efficiency.

MOTIVATION AND RELATED WORK

The random projection ensemble clustering (RPEClu) algorithm, as detailed in [1], has demonstrated strong performance in the context of high-dimensional clustering. RPEClu operates by projecting high-dimensional data into lower-dimensional random subspaces, applying Gaussian mixture models (GMMs) to each projection, and aggregating the results through consensus clustering. This approach is particularly effective for datasets characterized by high dimensionality and strong inter-feature correlations, as exemplified by its application to the “meat” dataset with 1,050 variables and 231 samples.

Despite its merits, RPEClu and similar model-based clustering approaches can be computationally intensive and remain susceptible to the influence of irrelevant or noisy features. In contrast, unsupervised feature selection algorithms aim to isolate features most pertinent to the latent group structure, potentially improving both computational efficiency and clustering accuracy. The integration of feature selection with ensemble clustering can therefore be expected to yield further improvements.

METHODOLOGY

The proposed methodological pipeline comprises the following stages:

1) Data Selection: The ”meat” dataset is utilized, as in the reference study, to facilitate direct comparability of results.
2) Unsupervised Feature Selection: An unsupervised feature selection algorithm is first applied to identify and retain a subset of features most relevant to the underlying group structure. This step is intended to reduce dimensionality and mitigate the effects of noise. Auto-UFSTool [2] was utilized for unsupervised feature selection. This MATLAB toolbox provides a collection of 25 robust unsupervised feature selection approaches, most of which were developed within the last years.
3) Parallelized RPEClu Algorithm: A parallelized implementation of the RPEClu algorithm is employed. Random projections and associated GMM fits are executed concurrently, leveraging multi-core computational resources to accelerate the process.
4) Ensemble Consensus: As in the original RPEClu, the top B* random projections—selected according to the Bayesian Information Criterion (BIC)—are aggregated using consensus clustering to derive the final group assignment.
5) Evaluation: Clustering performance is evaluated using the Adjusted Rand Index (ARI), with results compared before and after feature selection and benchmarked against previously published findings.

EXPERIMENTAL RESULTS

Baseline: RPEClu on Raw Meat Data

Application of the original RPEClu algorithm to the meat dataset, in accordance with [1], reproduces previously reported results; RPEClu achieves a higher ARI (0.32) compared to alternative methods such as standard GMM, K-means, and hierarchical clustering, although the absolute ARI value indicates the task remains challenging.

Parallelization for Computational Efficiency

Parallelization of the random projection and GMM fitting steps yields a substantial reduction in computation time. The clustering outcome is preserved, while the method becomes significantly more practical for large-scale or time-sensitive applications.

Impact of Unsupervised Feature Selection

The addition of an unsupervised feature selection stage prior to RPEClu results in a marked improvement in clustering performance. An increase in ARI indicates that the selected feature subset more effectively captures the latent group structure. Furthermore, reduced dimensionality translates into additional computational savings.

ADD ARI

Metric	Row	Best Value	Ensemble Method Params	Num Features
calinski_harabasz	66	424.625	5, 500, 50	75
dunn	87	0.062	5, 500, 50	600
pbm	105	13.526	5, 500, 50	1050
tau	86	0.000	5, 500, 50	575
gamma	2	32.000	5, 1000, 100	NA
c_index	21	0.226	5, 500, 50	50
davies_bouldin	5	0.954	5, 1000, 100	40
mcclain_rao	66	0.266	5, 500, 50	75
sd_dis	12	43.149	5, 1500, 150	20
ray_turi	5	0.446	5, 1000, 100	40
g_plus	38	0.000	5, 1000, 100	425
silhouette	66	0.362	5, 500, 50	75
s_dbw	19	0.355	5, 1500, 150	90
Compactness	21	2.500	5, 500, 50	50
Connectivity	102	64.481	5, 500, 50	975

TABLE I
BEST INTERNAL METRICS FOR EXPERIMENTS

Rank	Row	Best Metric Count	Ensemble Method Params	Num Features
1	66	3	5, 500, 50	75
2	5	2	5, 1000, 100	40
3	21	2	5, 500, 50	50
4	2	1	5, 1000, 100	NA
5	12	1	5, 1500, 150	20
6	19	1	5, 1500, 150	90
7	38	1	5, 1000, 100	425
8	86	1	5, 500, 50	575
9	87	1	5, 500, 50	600
10	102	1	5, 500, 50	975
11	105	1	5, 500, 50	1050

TABLE II
RANKING OF BEST INTERNAL METRICS

DISCUSSION

The experimental findings indicate that the combination of unsupervised feature selection with ensemble clustering via random projections enhances both clustering accuracy and computational efficiency in high-dimensional settings. Feature selection effectively removes irrelevant or redundant variables, thereby amplifying the efficacy of the ensemble approach and reducing computational cost.

The parallelization of the RPEClu algorithm further contributes to scalability, enabling its application to larger datasets without compromising accuracy. These results align with prior observations that feature extraction and selection are complementary strategies: feature selection increases the signal-to-noise ratio, while random projections and consensus aggregation provide robustness and stability

CONCLUSION

A hybrid methodology for high-dimensional clustering has been proposed, leveraging both unsupervised feature selection and ensemble clustering via random projections. Experiments on the meat dataset demonstrate that this integrated approach outperforms either method alone, particularly with respect to clustering accuracy (as measured by ARI) and computational efficiency.

REFERENCES

[1] L. Anderlucci, F. Fortunato, and A. Montanari, ”High-Dimensional Clustering via Random Projections,” *Journal of Classification*, vol. 39, no. 2, pp. 191–216, 2022.

[2] F. Abedinzadeh Torghabeh, Y. Modaresnia, and S. A. Hosseini, ”Auto-UFSTool: An Automatic Unsupervised Feature Selection Toolbox for MATLAB,” *Journal of AI and Data Mining*, 2023.