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. This dataset contains homogenized raw meat samples from 5 animal species (Beef, Chicken, Lamb, Pork, Turkey), measured by near infrared spectroscopy [2]. Application of the RPEClu algorithm to the meat dataset achieves a higher Adjusted Rand Index (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.

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) Parallelization of 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.
3) 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 [3] 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. In these initial experiments, Infinite Feature Selection (infFS) [4] was employed as feature selection method. infFS is a filter-based feature selection technique that ranks features by evaluating the importance of each feature in the context of all possible feature subsets, using a graph-based approach. The method constructs a feature affinity graph and utilizes the concept of all possible feature paths of arbitrary length (hence "infinite") to determine the relevance and redundancy of each feature. The final ranking reflects both how informative and how non-redundant each feature is with respect to the entire dataset. Experiments are conducted using between 25 and 1,050 variables, increasing in steps of 25.
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. Consensus clustering in RPEClu combines multiple clusterings by aligning and averaging them (accounting for label permutations), resulting in a single, robust final partition that reflects the most consistent grouping structure across the best random projection
5) Evaluation: Clustering is evaluated using various internal clustering metrics, as described in Table I. ARI external metric (see Table II) is used as benchmark against previously published findings.

TABLE I
CLUSTERING INTERNAL METRICS SUMMARY

Metric	Better if	Explanation
Calinski-Harabasz	Higher	Measures the ratio of between-cluster dispersion to within-cluster dispersion. Maximizing this indicates well-separated, dense clusters.
Dunn	Higher	Quantifies the ratio of the smallest inter-cluster distance to the largest intra-cluster distance. Higher values imply better-defined clusters.
PBM	Higher	A composite metric balancing compactness and separation. Higher scores indicate superior clustering.
Tau	Higher	A rank correlation metric where higher values reflect better agreement between distance rankings and cluster assignments.
Gamma	Higher	Measures the correlation between pairwise distances and cluster membership. Higher values denote stronger cluster structure.
C-index	Lower	Compares intra-cluster distances to a hypothetical ”perfect” clustering. Lower values indicate more compact clusters.
Davies-Bouldin	Lower	Averages similarity between clusters and their closest neighbors. Lower values imply better separation.
McClain-Rao	Lower	Ratio of within-cluster to between-cluster distances. Minimizing this improves clustering quality.
SD_Dis	Lower	Measures cluster dispersion. Lower values indicate tighter, more cohesive clusters.
Ray-Turi	Lower	Combines within-cluster compactness and between-cluster separation. Lower scores are optimal.
G_plus	Lower	Derived from Goodman-Kruskal’s gamma. Lower values suggest fewer discordant pairs in clustering.
Silhouette	Higher	Ranges from -1 to 1. Higher values indicate better-defined clusters with minimal overlap.
S_Dbw	Lower	Evaluates compactness and separation density. Lower scores reflect superior clustering.
Compactness	Lower	Directly measures intra-cluster variance. Lower values mean tighter clusters.
Connectivity	Lower	Counts violations in nearest-neighbor assignments. Fewer violations (lower scores) are desirable.

TABLE II
CLUSTERING EXTERNAL METRICS SUMMARY

Metric	Better if	Explanation
Adjusted Rand Index (ARI)	Higher	Measures the similarity between clustering results and a ground-truth classification, adjusted for chance. Higher values indicate better alignment with the true labels.

EXPERIMENTAL RESULTS

Baseline: RPEClu on Meat Dataset

Application of the original RPEClu algorithm to the meat dataset to reproduce previously reported results in [1].

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.

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

TABLE III
BEST INTERNAL METRICS (INFFS FEATURE SELECTION - MEAT DATASET)

Metric	ID Exp	Best Value	Params (g, B, B*)	Num Features
calinski_harabasz	66	424.625	5, 500, 50	75
dunn	87	0.0617	5, 500, 50	600
pbm	105	13.526	5, 500, 50	1050
tau	86	6.207e-07	5, 500, 50	575
gamma	2	32.000	5, 1000, 100	N/A
c_index	66	0.0621	5, 500, 50	75
davies_bouldin	5	0.9535	5, 1000, 100	40
mcclain_rao	66	0.2658	5, 500, 50	75
sd_dis	12	43.149	5, 1500, 150	20
ray_turi	5	0.4456	5, 1000, 100	40
g_plus	5	-3.627e-07	5, 1000, 100	40
silhouette	66	0.3621	5, 500, 50	75
s_dbw	19	0.3555	5, 1500, 150	90
Compactness	12	0.1183	5, 1500, 150	20
Connectivity	102	64.481	5, 500, 50	975

TABLE IV
RANKING OF BEST INTERNAL METRICS

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

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] J. McElhinney, G. Downey, and T. Fearn, ”Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats,” *Journal of Near Infrared Spectroscopy*, vol. 7, pp. 145–154, 1999.
[3] 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.
[4] G. Roffo, S. Melzi, and M. Cristani, ”Infinite feature selection,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4202–4210, 2015.