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 (**0.32**) Adjusted Rand Index (ARI) 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) 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] is 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] is employed as feature selection method, utilizing its default parameter settings. 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.
3) RPEClu Algorithm: A parallelized implementation of the algorithm in the package diceRplus [5] is used in place of the original (single-threaded) implementation [6]. Random projections and associated GMM fits are executed concurrently, leveraging multi-core computational resources to accelerate the process. This parallelization is feasible and highly effective because each of the B random projections—where B is the total number of random projections generated—is independent and can be processed separately.
   In each random projection, the original data matrix with p features is projected onto a lower-dimensional subspace of dimension d, yielding a new data representation of size $n \times d$. The number of clusters, G, represents the hypothesized or known number of underlying groups in the data. Empirical results further suggest that $d = 10 \cdot \log(G) + 1$ is generally effective in practice [1]. For each projection, a Gaussian Mixture Model (GMM) is then fitted with G clusters.
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 projections.

5) Evaluation: Clustering is evaluated using various internal clustering metrics, as described in Table I. External metric ARI (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	Pakhira-Bandyopadhyay-Maulik index. 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	Scatter and Dispersion Distance. 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.
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 that quantifies the similarity between a clustering result and a ground-truth classification, adjusting for the possibility of random agreements. It provides a score ranging from -1 to 1, where 1 signifies perfect agreement with the true labels, 0 indicates a clustering result that is no better than random assignment, and negative values suggest a worse-than-random classification.

EXPERIMENTAL RESULTS

The experiments were conducted on a computer running Microsoft Windows 11 Home, equipped with an Intel Core i9-14900HX processor (32 cores) and 32 GB of RAM. The main software development tools used were RStudio 2024 (R version 4.3.3) and MATLAB R2024b.

The source code for the experiments is available at [7].

RPEClu: Original vs. Parallelized Implementation

The original single-threaded RPEClu algorithm ($G = 5, d = 17, B = 1000, B^* = 100$) was applied to the meat dataset ($1,050$ features) to reproduce previously reported results [1]. To improve computational efficiency, parallelization was introduced in the random projection and Gaussian Mixture Model (GMM) fitting steps. This modification led to a significant reduction in processing time while preserving clustering quality.

TABLE III
RPECLU IMPLEMENTATION PERFORMANCE

RPEClu Implementation	Exec. Time (s)	ARI
Single-threaded as reported in [1]	6,460	0.32
Single-threaded	2,475	0.29
Parallelized	356	0.31

Impact of Unsupervised Feature Selection

Experiments were conducted using infFS unsupervised feature selection between 25 and $1,050$ variables, increasing in steps of 25. RPEClu algorithm ran with $(G = 5, d = 17)$ and 2 configurations ($B = 1000, B^* = 100$ and $B = 500, B^* = 50$). The resulting internal clustering metrics, as a function of the number of features selected by infFS, are summarized in the following figures.

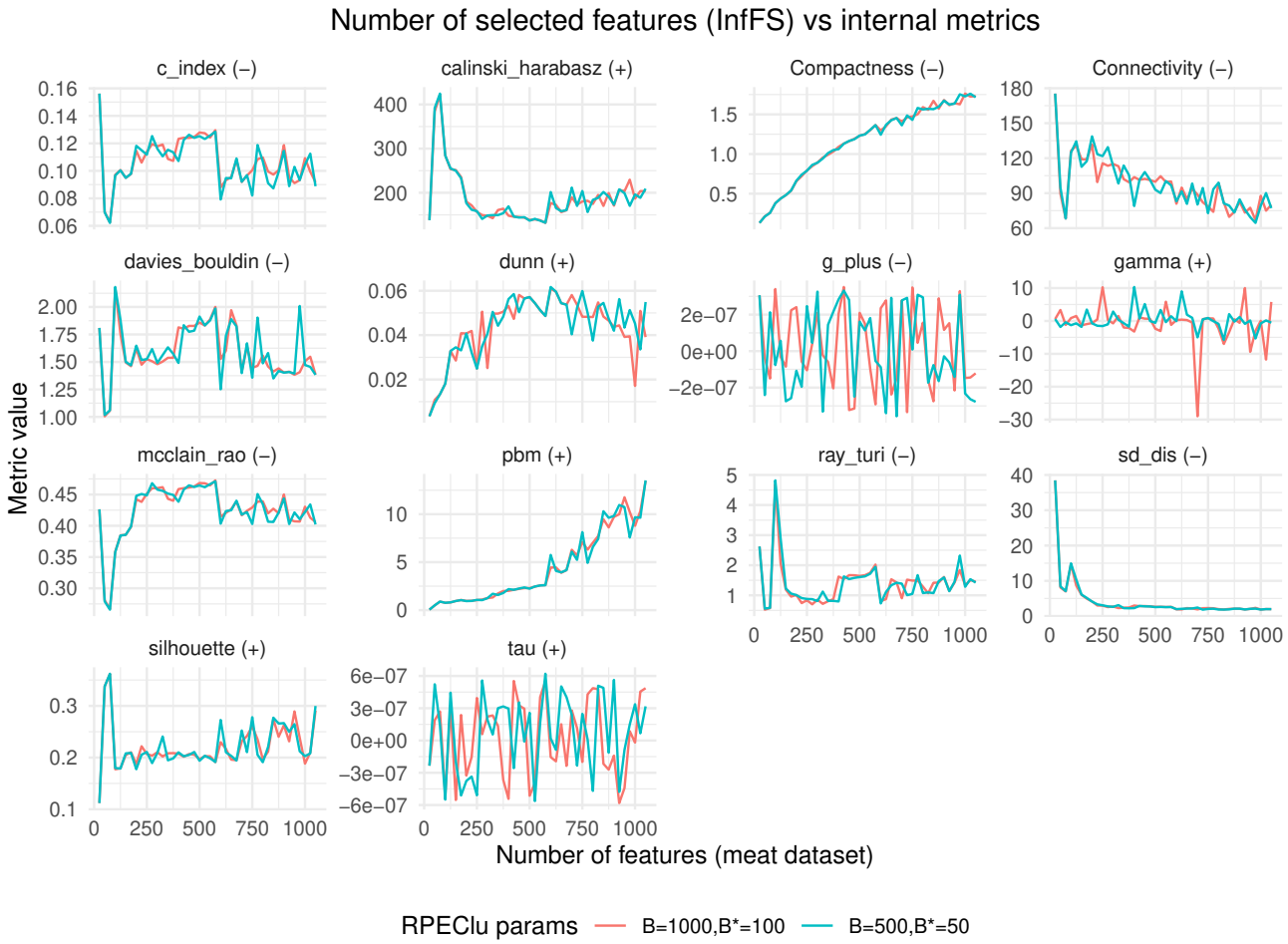

Fig. 1. Internal clustering metrics (y-axis) as a function of the number of features selected by Infinite Feature Selection (x-axis) on the meat dataset, evaluated using RPEClu with two parameter configurations. Higher (+) or lower (-) values indicate better clustering.

These plots illustrate how the performance of RPEClu varies across a range of feature subset sizes, from 25 up to 1,050 variables. Notably, several internal metrics—including silhouette, Calinski-Harabasz, and Davies-Bouldin—reach optimal or near-optimal values at intermediate feature counts (typically between 50 and 100 features), suggesting that excessive dimensionality reduction may eliminate informative variables, while retaining too many features reintroduces noise and redundancy.

Overall, these results indicate that careful tuning of the feature selection step is crucial for maximizing clustering performance: an appropriate balance between dimensionality reduction and information preservation leads to more compact, well-separated, and robust clusters. This trend is consistent across both parameter configurations tested ($B = 1000, B^* = 100$ *and* $B = 500, B^* = 50$), underscoring the stability of the ensemble clustering approach when combined with unsupervised feature selection.

Table IV summarizes the best values achieved for each internal clustering metric across the different experimental configurations. Each entry indicates the experimental parameters, the corresponding number of selected features, and the best metric value obtained. Table V presents a ranking of the experimental configurations that achieved the best values across the various internal validation indices, indicating for each configuration the number of metrics in which it attained the top result as well as the corresponding ARI value. This table allows for the identification of the most robust configurations in terms of consistent performance according to multiple clustering quality metrics. Experiment 66 ($B = 100, B^* = 50, num\_features = 75$) stands out as the preferred configuration due to its superior consistency and robustness, achieving top results across four internal metrics. This demonstrates its ability to maintain high clustering quality across diverse evaluation criteria, ensuring stable and reliable outcomes in different scenarios. Additionally, its external metric remains consistently strong (**ARI=0.68**)

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

Metric	ID Exp	Params (B, B^*)	Num Features	Best Value
c_index	66	500, 50	75	0.0621
calinski_harabasz	66	500, 50	75	424.625
Compactness	22	1000, 100	25	0.1319
Connectivity	102	500, 50	975	64.481
davies_bouldin	23	1000, 100	50	1.0057
dunn	87	500, 50	600	0.0617
gamma	79	500, 50	400	10.3333
g_plus	90	500, 50	675	$-3.5711 \cdot 10^{-7}$
mcclain_rao	66	500, 50	75	0.2658
pbm	105	500, 50	1050	13.526
ray_turi	23	1000, 100	50	0.5233
sd_dis	22	1000, 100	25	38.4742
silhouette	66	500, 50	75	0.3621
tau	86	500, 50	575	$6.2072 \cdot 10^{-7}$

TABLE V
RANKING OF BEST INTERNAL METRICS

Rank	ID Exp	Params (B, B^*)	Num Features	Best Metric Count	ARI
1	66	500, 50	75	4	0.6830
2	22	1000, 100	25	2	0.5264
3	23	1000, 100	50	2	0.6888
4	24	1000, 100	75	1	0.6882
5	79	500, 50	400	1	0.4407
6	86	500, 50	575	1	0.6096
7	87	500, 50	600	1	0.3673
8	90	500, 50	675	1	0.5369
9	102	500, 50	975	1	0.2994
10	105	500, 50	1050	1	0.3136

Figure 2 illustrates the relationship between the number of selected features by the infFS unsupervised method and the ARI in the meat dataset. The baseline ARI=0.32 reference line marks the ARI achieved without feature selection in [1], providing a benchmark to assess improvements across varying feature subsets. The plot shows substantial gains in clustering performance when feature selection is applied. Both configurations achieve comparable clustering results, but the $B = 500, B^* = 50$ setting offers a reduction in execution time (see Table VI), making it the more efficient choice.

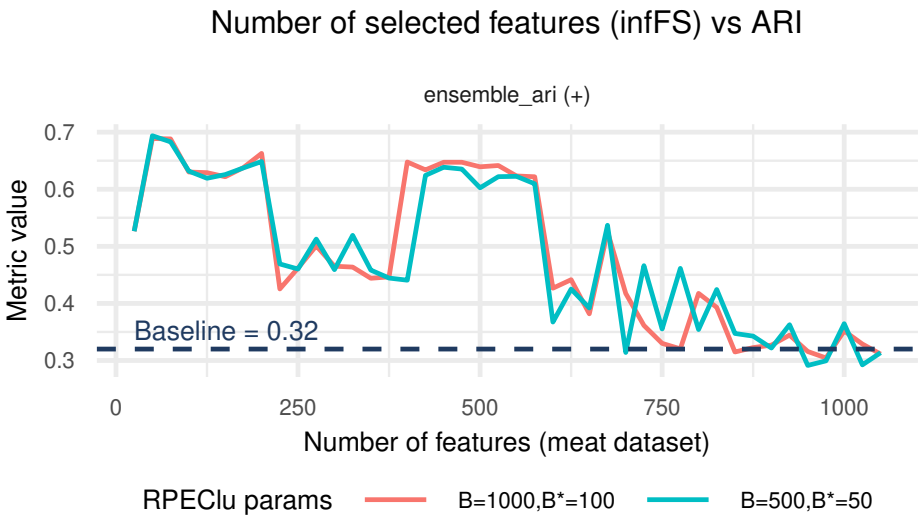

Fig. 2. Number of selected features (infFS) vs ARI. Higher ARI values suggest better agreement between the clustering structure and ground-truth labels.

TABLE VI
COMBINED INFFS AND RPECLU EXECUTION TIME

Execution Time (s)			
Params (B, B^*)	infFS Feature Ranking	Parallel RPEClu	Total
500, 50	1	5,399	5,400
1000, 100		12,318	12,219

The execution time for the combined approach is largely driven by repeated RPEClu clustering runs, while the infFS feature selection step is computationally negligible, requiring just one second. To optimize this process, rather than exhaustively evaluating feature subset sizes in 25-feature increments, internal metrics could be monitored at the end of each iteration. If these metrics plateau or deteriorate, further evaluations can be halted. Additionally, experimenting with a reduced number of random projections (i.e., lowering the values of B and B^*) could further decrease computation time, provided clustering quality remains acceptable. This strategy may be especially beneficial in exploratory phases or with much larger datasets, where computational resources are a limiting factor.

CONCLUSION

The experimental findings demonstrate that integrating unsupervised feature selection with ensemble clustering via random projections improves clustering accuracy in high-dimensional settings. By eliminating irrelevant variables, feature selection strengthens the ensemble approach and lowers computational cost. In this study, feature selection was performed using infFS, which effectively ranked features based on their relevance and redundancy. This improvement is reflected in the clustering performance metrics for the meat dataset, where the ARI increases from 0.32 (without feature selection in [1]) to 0.68 when using the optimized feature selection and ensemble clustering approach.

To enhance computational efficiency further, it is worthwhile to consider not only monitoring internal clustering metrics to identify the optimal feature subset size (and thus halt unnecessary evaluations), but also to explore reducing the number of random projections. Provided that clustering performance remains robust, running the ensemble with fewer projections can significantly shorten processing time, which is especially advantageous during preliminary analyses or with very large datasets where resources are constrained. Also, it would be valuable in future work to compare other feature selection algorithms to assess whether the observed gains are consistent across different techniques.

In summary, the proposed combined methodology—leveraging both unsupervised feature selection and ensemble clustering via random projections—outperforms the approach with the full set of features, particularly in terms of clustering accuracy (as measured by ARI), at least within the context of the meat dataset. These findings highlight the value of combining feature selection and ensemble methods for robust, scalable high-dimensional clustering. However, as these results are based on a single dataset, additional experiments across diverse datasets and feature selection methods are necessary to assess the broader applicability of this approach.

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.
[5] A. C. Rodriguez Bajo, "diceRplus," GitHub repository, 2025. [Online]. Available: https://github.com/antoniocarlosrodriguezbajo/diceRplus
[6] L. Anderlucci, F. Fortunato, and A. Montanari, "RPEClust: Random Projection Ensemble Clustering Algorithm," R Package Documentation Repository, 2022. [Online]. Available: https://rdrr.io/cran/RPEClust/
[7] A. C. Rodriguez Bajo, "Meat Experiments," GitHub repository, 2025. [Online]. Available: https://github.com/antoniocarlosrodriguezbajo/diceRplus/blob/master/experiments/meat_experiments.R