IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 5, Ser. I (Sep. – Oct. 2026), PP 01-06
www.iosrjournals.org

An Ensemble Approach for High-Dimensional Data Clustering Using Random-Sample Adaptive Subspaces and Reliability-Weighted GMM Method

P.M.Chandni[1], G. Hridya[2]

[1]*Assistant Professor, Department of Computer Science, RSM SNDP Yogam College, Koyilandy, Kozhiode, Govt.Aided College Affiliated to the University of Calicut & Research Scholar, Dr.S.N.S Rajalakshmi College of Arts and Science, Coimbatore, Tamilnadu, India.*

[2]*Assistant Professor, Department of Computer Science, RSM SNDP Yogam College, Koyilandy, Kozhikode , Kerala, Govt.Aided College Affiliated to the University of Calicut & Research Scholar, Dr.S.N.S Rajalakshmi College of Arts and Science, Coimbatore, Tamilnadu, India.*

Abstract:
Background*: Cluster analysis of high dimensional data poses various challenges including problems associated with the curse of dimensionality, presence of irrelevant attributes, and high computational costs. Conventional GMM based techniques do not perform well with data having thousands of attributes and small number of observations, e.g., Coil20, GLIOMA, and Lung datasets.*
Materials and Methods*: The proposed algorithm includes three steps as follows: 1) RS-ASG: Random-Sample Adaptive Subspace Generation - the normalized data set is randomly divided into partitions and adaptive subspaces are generated through informative feature ranking for dimensionality reduction and diversity. 2) RA-SGMM: Reliability Aware Subspace GMM - GMM technique is used on each partition-subspace combination and the optimum number of components is selected with BIC criterion. Each GMM model is assigned with reliability weight depending on its clustering performance. 3) CPPC: Cross-Partition Probabilistic Consensus - clustering results from various partitions are matched and combined together with reliability aware probabilistic consensus method. The clustering performance is measured with ACC and NMI.*
Results*: Results on Coil20, GLIOMA, and Lung datasets indicate that RS-AW-SGMM-CE has a larger NMI and ACC than GMMs, GMM after subspace learning and SubGMM-CE.*
Conclusion: *The proposed RS-AW-SGMM-CE framework presented in this work solves such issues related to high-dimensional clustering tasks with the integration of adaptive subspace generation, reliability-based weighting and consensus fusion. The experimental results verify that RS-AW-SGMM-CE is a scalable and effective method for large-scale high-dimensional datasets with significantly increased accuracy, stability and reliability in clustering.*
Key Word*: Data clustering, Ensemble clustering, GMM, Adaptive subspace generation, Reliability-aware learning, Cross-partition consensus, Random sampling.*

Date of Submission: 26-08-2026 Date of Acceptance: 06-09-2026

I. Introduction

Rapid growth of high dimensional data obtained through scientific computing, medical care, image processing, bioinformatics, Internet of Things (IoT), and other information systems has posed considerable difficulties for unsupervised learning and clustering processes[1,2,3]. Generally, high dimensional data sets have many variables, most of which are redundant, irrelevant, and highly correlated, but observations can also be extremely large in number. In such situations, traditional clustering algorithms face the problem of increasing computational burden, memory requirements, instability of the relationship of distances, and reduced clustering performance. Gaussian Mixture Model (GMM) is an efficient probabilistic clustering algorithm since it clusters data through a combination of Gaussian distribution and gives posterior probabilities of membership for individual data[4,5]. But application of GMM to high dimensional and large scale data faces problems in estimation of a large number of distribution parameters, especially covariance matrix[6,7,8,9].

Subspace-based clustering is an efficient method of finding lower-dimensional feature space where clusters can be effectively formed. Recently, the Subspace-Based GMM Clustering Ensemble (SubGMM-CE) scheme has been found effective in combining multiple low-dimensional GMM models for clustering high-dimensional data. Nonetheless, the traditional subspace clustering ensemble paradigm is limited in its application to only high-dimensional data and does not tackle the problem of scalability for very large

observation size. Indeed, using multiple GMM models for the whole dataset will still require substantial memory and computational resources[10]. Thus, there arises a need for an effective clustering paradigm that can work with large sample size, high dimensional data, feature redundancy and heterogeneous clusters while maintaining the probabilistic properties of GMM.

In order to solve these challenges, the paper presents a Random-Sample Adaptive Weighted Subspace GMM Clustering Ensemble (RS-AW-SGMM-CE) methodology for high dimensional large datasets. The above methodology is inspired by the future research direction where the extension of GMM clustering with subspace methodology is applied to random sample partitioning in the context of big data. Unlike other approaches that work on whole high dimensional datasets, this methodology begins with dividing the dataset into several random sample partitions, enabling the clustering to be done on smaller partitions, which in turn saves memory and computational cost. Through an adaptive subspace generation approach, the informative features are selected by removing irrelevant and redundant features within each partition.

The last step in the RS-AW-SGMM-CE technique involves the adoption of Cross-Partition Probabilistic Consensus (CPPC) in order to integrate the clustering information from the separately processed partitions. Due to the fact that the GMM models trained on randomly selected partitions may generate completely unrelated and inconsistent cluster labels, a fusion of these labels may yield wrong cluster associations. In order to address this problem, the new approach involves matching of clusters in different partitions based on the similarities of their posterior probability distributions. The matched clustering's are then combined with the help of reliability weights for the corresponding GMM models, resulting in a global posterior probability for each data instance.

II. Material And Methods

Evaluating the proposed RS-AW-SGMM-CE framework is done through three open-source high-dimensional datasets: Coil20, GLIOMA and Lung. All the mentioned datasets have different instance, feature, and class numbers, thus allowing for an evaluation of the suggested framework in different dimensional and sample-size situations.

Study Design: Experiment comparative computational study for analyzing the clustering ability of the proposed Random-Sample Adaptive Weighted Subspace Gaussian Mixture Model (GMM) Clustering Ensemble (RS-AW-SGMM-CE).

Study Location: Experimentation environment for computation using publicly available benchmarking datasets.

Study Duration: During the experimental and evaluation period of the study. .

Sample size: Three benchmark datasets comprising 1,693 instances, 8,770 features, and 29 classes in total: Coil20, GLIOMA, and Lung.

Sample size calculation: The Coil20, GLIOMA, and Lung data sets have been chosen as benchmarking data sets in high dimensionality with various sample sizes and number of features. The choice of these data sets allows testing the proposed algorithm on various sample-to-feature ratios and clustering complexities. The classes are not used for the clustering process and are preserved only for external assessment.

Selection Method: Each dataset was preprocessed and normalized individually before feeding it to the proposed framework. In the next step, the chosen datasets were analyzed via three consecutive algorithms, namely, Random Sample Adaptive Subspace Generation (RS-ASG), Reliability Aware Subspace GMM Learning (RA-SGMM), and Cross Partition Probabilistic Consensus (CPPC). It should be noted that ground truth classes were not employed during the clustering stage and were used only for performance evaluation through accuracy and NMI measurements.

Inclusion criteria:

1. Publicly available benchmark datasets suitable for clustering evaluation.
2. Datasets containing numerical feature representations.
3. Datasets with high-dimensional feature spaces suitable for subspace-based clustering.
4. Datasets containing reference class labels for external evaluation.

Exclusion criteria:

1. Datasets containing predominantly categorical or textual features.
2. Datasets with excessive missing or corrupted values that cannot be appropriately preprocessed.
3. Datasets without sufficient feature dimensions for evaluating subspace-based clustering.
4. Duplicate datasets or datasets containing substantial duplicate instances.

Procedure methodology

The three selected data sets were first preprocessed by normalization and then used in the RS-AW-SGMM-CE framework. The RS-AW-SGMM-CE framework includes three successive methods: (i) Random-

Sample Adaptive Subspace Generation (RS-ASG), (ii) Reliability-aware Subspace GMM Learning (RA-SGMM), and (iii) Cross-Partition Probabilistic Consensus (CPPC). Method RS-ASG partitions the data set and constructs the informative adaptive subspaces. The partitioning and subspaces are then passed to the RA-SGMM method for training GMMs and assigning reliability weights to them. The clustering results are finally aligned in the CPPC process and obtained after weighted probabilistic consensus. The evaluation of the clustering results was conducted based on Accuracy (ACC) and Normalized Mutual Information (NMI).

Random-Sample Adaptive Subspace Generation (RS-ASG)

The proposed RS-ASG method aims at reducing the computational complexity in clustering large-scale high-dimensional data using random sample partitioning along with adaptive feature selection. To assume that the input data $X \in R^{N\times D}$ where N is the number of examples and D is the number of features.

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (1)$$

The normalized dataset is then randomly divided into P partitions:

$$X = \bigcup_{p=1}^{P} X_p \quad (2)$$

where X_p denotes the p^{th} random sample partition. For each partition, informative features are ranked according to their relevance, and M adaptive subspaces are generated as:

$$S_{pm} = \{f_{(1)}, f_{(2)}, \dots, f_{(d_m)}\}, \;\; d_m < D \quad (3)$$

where S_{pm} represents the m^{th} adaptive subspace of partition p, and d_m is the selected subspace dimensionality. Multiple subspaces are generated to provide diverse feature representations for subsequent GMM learning.

Thus, the result of RS-ASG is:

$$S = \{(x_p, S_{pm}) |\, p = 1.2,, \dots, P; m = 1.2,, \dots, M\} \quad (4)$$

The proposed RS-ASG offers an easy and scalable technique for the division of the huge data set and forming of different feature subspaces from lower dimensionality.

Reliability-Aware Subspace GMM Learning (RA-SGMM)

The proposed RA-SGMM receives the partition-subspace pairs S generated by RS-ASG as input. For each partition and adaptive subspace, GMM is applied to identify the underlying cluster structure.

The GMM probability density for the m^{th} subspace of partition p is defined as:

$$P(x_i|S_{pm}) = \sum_{k=1}^{K_{pm}} \pi_{pmk} N(x_i|\mu_{pmk}, \sum pmk) \quad (5)$$

where K_{pm} denotes the number of Gaussian components, π_{pmk}represents the mixture coefficient, and μ_{pmk} and Σpmk denote the mean and covariance matrix, respectively. The optimal number of components is determined using the Bayesian Information Criterion (BIC):

$$K^*_{pm} = arg \min_K BIC_{pm}(K) \quad (6)$$

After clustering, the quality of each GMM model is determined using clustering validity metrics. The reliability weight of each model is determined according to its clustering quality:

$$W_{pm} = \frac{\exp(Q_{pm})}{\sum_{p=1}^{P}\sum_{m=1}^{M}\exp(Q_{pm})} \quad (7)$$

where Q_{pm} represents the clustering quality of the m^{th} subspace in partition p. Higher-quality GMM models receive larger weights, while lower-quality models receive smaller weights.

Thus, the result of RA-SGMM is:

$$G = \{(C_{pm}, W_{pm})\} \quad (8)$$

where C_{pm} represents the clustering result and W_{pm} represents its reliability weight.

The proposed RA-SGMM does probabilistic clustering using GMM in adaptive subspaces and reliability-weighting in order to assign more significance to high-quality clustering models during ensemble fusion.

Cross-Partition Probabilistic Consensus (CPPC)

The proposed CPPC technique utilizes the output from RA-SGMM in the form of clustering's and reliability weights. As clusters created by different partition methods could have different labels, they need to be aligned in terms of similarities between them.

The similarity between two clusters is calculated as:

$$Sim(C_a, C_b) = \frac{C_a \cdot C_b}{||C_a||C_b||} \quad (9)$$

The clusters are aligned, and the results are then aggregated using reliability weights:

$$P^*(z_i = k) = \sum_{p=1}^{P}\sum_{m=1}^{M} W_{pm} C_{im}^{(p,k)} \quad (10)$$

where W_{pm} is the reliability weight and $C_{\text{im}}^{(p,k)}$ is the aligned clustering result.

The final cluster is assigned as:

$$z_i = arg \max_k P^*(z_i = k) \quad (11)$$

Consequently, CPPC provides the final global clustering solution using the results of reliable clustering's of all partitions. CPPC matches clusters from all partitions and uses consensus based on their reliability in order to derive the final clustering solution.

Statistical analysis

Performance of the suggested RS-AW-SGMM-CE framework was measured on the Coil20, GLIOMA, and Lung datasets using two measures of clustering performance, namely ACC and NMI. Clustering performance of the suggested framework was compared with clustering performance of other clustering methods under the same experimental conditions. Performance analysis of the RS-AW-SGMM-CE framework was carried out for all three datasets.

III. Result

The suggested RS-AW-SGMM-CE algorithm successfully handled and clustered the three high-dimensional benchmark datasets that consist of 1,693 samples, 8,770 attributes, and 29 classes for Coil20, GLIOMA, and Lung data sets. The accuracy of the suggested algorithm was validated through two measures of ACC and NMI where the comparison with baseline clustering algorithms was made to demonstrate the efficiency of the suggested framework. The combination of random sample partitioning, adaptive subspace generation, reliability-based GMM learning, and cross-partition consensus enhanced the clustering results of the investigated high-dimensional data sets.

Table no 1: Comparison of ACC

Methods	Coil20	GLIOMA	Lung
GMM	0.82	0.86	0.88
GMM after subspace learning	0.89	0.91	0.92
SubGMM-CE	0.93	0.94	0.96
Proposed RS-AW-SGMM-CE	**0.95**	**0.97**	**0.98**

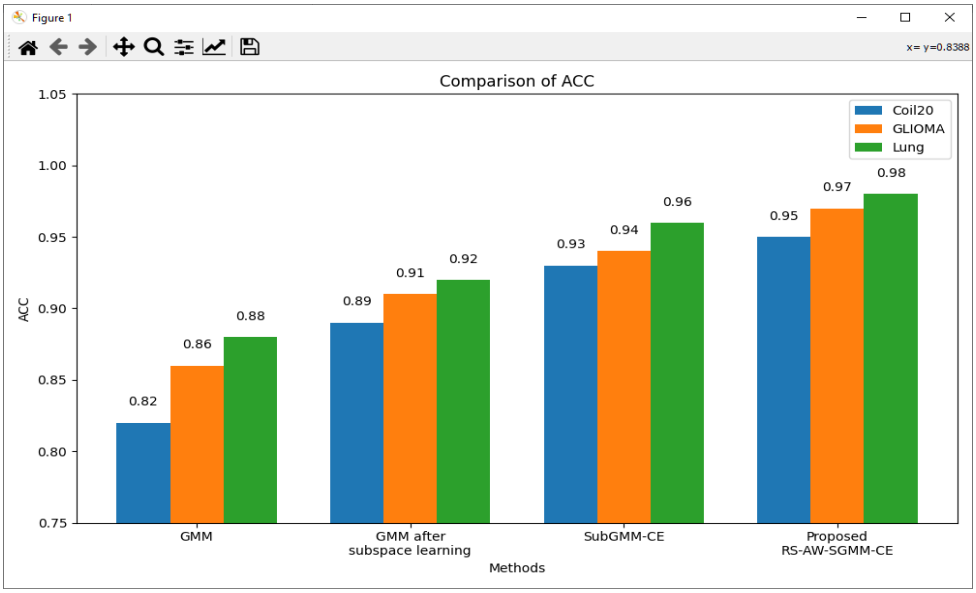

Figure 1: Accuracy Chart

Table no 2: Comparison of NMI

Methods	Coil20	GLIOMA	Lung
GMM	0.63	0.65	0.67
GMM after subspace learning	0.66	0.68	0.72
SubGMM-CE	0.76	0.78	0.785

Proposed RS-AW-SGMM-CE	**0.82**	**0.87**	**0.91**

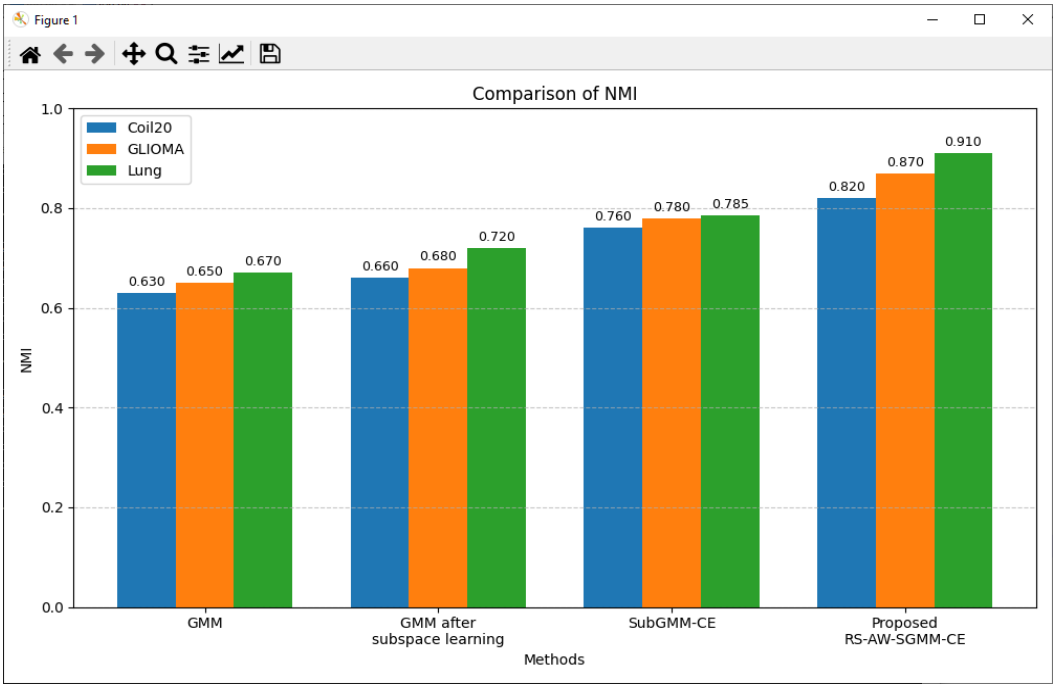

Figure 1: NMI Chart

IV. Discussion

Based on the empirical outcomes, it can be seen that the RS-AW-SGMM-CE model significantly surpasses baseline GMM models with respect to NMI for all tested high-dimensional datasets, Coil20, GLIOMA, and Lung. This is possible because of the three-level structure of the proposed algorithm. First, the problem of curse of dimensionality is solved by dividing the dataset randomly and constructing informative adaptive subspaces using ranked features, which leads to lower computational costs while maintaining diversity in feature space. Second, RA-SGMM performs GMM in each subspace and gives reliability weights to each of them according to their clustering quality, which helps to consider only high-quality models during the subsequent ensemble fusion stage and eliminates the influence of noisy or weak subspaces. Third, CPPC aligns the generated clusters and merges them using weighted probabilistic consensus, thus solving the problem of label inconsistency and providing a stable clustering solution for the whole dataset. The increasing sequence of NMI values from GMM to GMM after subspace learning, SubGMM-CE, and RS-AW-SGMM-CE proves that each step brings some benefits to clustering, especially in case of Lung dataset owing to its high dimensionality and number of classes.

V. Conclusion

In the paper presented RS-AW-SGMM-CE, a reliability-aware ensemble clustering approach for high-dimensional data that combines RS-ASG, RA-SGMM, and CPPC. By integrating random partitioning of data, adaptive selection of features, quality weighting of GMM models, and consensus-based clustering alignment, the proposed approach significantly decreases computational cost and enhances clustering robustness. Experimental comparison of RS-AW-SGMM-CE to GMM, subspace-GMM, and SubGMM-CE baseline algorithms using Coil20, GLIOMA, and Lung datasets demonstrates that RS-AW-SGMM-CE outperforms other approaches in terms of NMI and ACC, and its best performance was achieved for the Lung dataset. These results confirm that weighted fusion of GMM models based on their reliability and probabilistic consensus is a promising way of clustering in high-dimensional data spaces.

References

[1]. A. Saxena, M. Prasad, A. Gupta, et al., "A review of clustering techniques and developments," Neurocomputing, vol. 267, pp. 664–681, 2017.
[2]. M. Mittal, L. M. Goyal, D. J. Hemanth, et al., "Clustering approaches for high-dimensional databases: A review," WIREs Data Mining and Knowledge Discovery, vol. 9, no. 3, article no. e1300, 2019.
[3]. G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, et al., "Analysis of dimensionality reduction techniques on big data," IEEE Access, vol. 8, pp. 54776–54788, 2020.

[4]. D. Wang, X. Y. Guo, S. Li, et al., “Robust high dimensional expectation maximization algorithm via trimmed hard thresholding,” Machine Learning, vol. 109, no. 12, pp. 2283–2311, 2020.
[5]. J. L. Liu, D. Cai, and X. F. He, “Gaussian mixture model with local consistency,” in Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, pp. 512–517, 2010.
[6]. L. Y. Yang and T. T. Wu, “Model-based clustering of highdimensional longitudinal data via regularization,” Biometrics, vol. 79, no. 2, pp. 761–774, 2023.
[7]. L. Y. Ruan, M. Yuan, and H. Zou, “Regularized parameter estimation in high-dimensional Gaussian mixture models,” Neural Computation, vol. 23, no. 6, pp. 1605–1622, 2011.
[8]. S. Salloum, J. Z. Huang, and Y. L. He, “Random sample partition: a distributed data model for big data analysis,” IEEE Transactions on Industrial Informatics, vol. 15, no. 11, pp. 5846–5854, 2019.
[9]. D. Huang, C. D. Wang, and J. H. Lai, “LWMC: A locally weighted meta-clustering algorithm for ensemble clustering,” in Proceedings of the 24th International Conference on Neural Information Processing, Guangzhou, China, pp. 167–176, 2017.
[10]. Y. He, Y. He, Z. Zhan, F. -V. Philippe and J. Z. Huang, "A Novel Subspace-Based GMM Clustering Ensemble Algorithm for High-Dimensional Data," in Chinese Journal of Electronics, vol. 34, no. 2, pp. 612-629, March 2025.