

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas173

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

An Efficient Clustering Approach for High-Dimensional Data for a Hybrid Clustering Model Integrating DBSCAN with Gaussian Mixture Models using Machine Learning Techniques

¹ K. Mudduswamy, ² Dr K Pavan Kumar, ³ Repalle Nagesh

^{1, 3} Assistant Professor, ² Associate Professor

^{1, 2 & 3} Dept. of AIML & DS, Dr. K V Subba Reddy Institute of Technology, Kurnool, AP, India

Abstract:

High-dimensional data clustering poses significant challenges due to the curse of dimensionality, noise, and sparsity. Traditional clustering algorithms often struggle with scalability and accuracy in such contexts. To address these issues, this paper proposes a hybrid clustering model that integrates Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMM), leveraging the strengths of both approaches using machine learning techniques. DBSCAN efficiently identifies dense regions and eliminates noise, while GMM provides probabilistic soft clustering suitable for overlapping data distributions. By first using DBSCAN for pre-processing and noise reduction, followed by GMM for refined clustering, the hybrid model enhances performance in complex high-dimensional datasets. Dimensionality reduction techniques such as PCA and t-SNE are also incorporated to visualize and improve cluster quality. The proposed method is evaluated on benchmark high-dimensional datasets and compared against standalone clustering algorithms. Results demonstrate improved cluster compactness, separation, and computational efficiency, showcasing the effectiveness of the hybrid approach for high-dimensional data analysis in fields such as bioinformatics, image processing, and text mining.

Keywords: Clustering, Machine Learning, DBSCAN, Gaussian Mixture Model, High-Dimensional Data

I. Introduction

Clustering is widely used in various applications, including image segmentation, bioinformatics, and recommendation systems. Traditional clustering methods struggle with high-dimensional and noisy datasets. In this study, we present a novel hybrid clustering approach combining DBSCAN and GMM to achieve better clustering efficiency and accuracy. High-dimensional data poses significant challenges in data analysis, particularly in clustering tasks, due to the curse of dimensionality and increased computational complexity. Traditional clustering techniques, such as K-Means and Hierarchical Clustering, often struggle to maintain accuracy and efficiency in such scenarios. This paper proposes an efficient clustering approach that leverages hybrid machine learning techniques to improve clustering performance for high-dimensional data. The method integrates dimensionality reduction techniques (such as PCA or t-SNE) with advanced clustering algorithms (such as K-Means++, DBSCAN, or spectral clustering) to enhance accuracy, scalability, and computational efficiency. The proposed approach is evaluated on benchmark datasets, demonstrating improved cluster quality, reduced processing time, and better adaptability to complex data distributions. By combining the strengths of multiple techniques, this hybrid method provides a robust solution for handling high-dimensional clustering challenges effectively.

In high-dimensional settings, these limitations become more pronounced due to the curse of dimensionality and noise. Hence, hybrid models that leverage the strengths of multiple

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas173

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

clustering techniques are gaining traction. This paper proposes a hybrid clustering model that integrates DBSCAN and GMM to enhance clustering performance in complex data scenarios. The paper is organized as follows: Section II Related work, Section III Proposed Methodology, Section IV Experimental results and Discussion, and Section V Concludes the study.

1. Artificial Intelligence (AI) plays a pivotal role in overcoming these challenges by enabling intelligent, adaptive, and data-efficient models. This paper presents an AI- powered hybrid clustering approach that integrates Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMM) to enhance clustering performance on high-dimensional datasets. DBSCAN is used to detect dense clusters and eliminate noise, while GMM performs probabilistic soft clustering, capturing overlapping structures in the data. AI techniques are employed to optimize the workflow, including automated parameter tuning, intelligent pre-processing, and dimensionality reduction techniques such as PCA and t-SNE, which improve both visualization and cluster quality.

2. Related Work: Numerous clustering methods have been proposed to deal with high-dimensional and noisy data. K-Means is computationally efficient but performs poorly with non-convex clusters. GMM offers probabilistic cluster assignment but assumes Gaussian distribution. DBSCAN, a density-based algorithm, identifies clusters of arbitrary shape and filters out noise but is sensitive to parameter selection. To overcome the limitations of individual clustering methods, researchers have explored hybrid approaches. For instance, the PCA for dimensionality reduction with K-Means to improve scalability and performance. The proposed an autoencoder-based model that integrates deep learning with spectral clustering, achieving high clustering purity on image datasets. Similarly, the explored DBSCAN and hierarchical clustering for customer segmentation in e-commerce platforms. Hybrid models combining DBSCAN and K-Means have also shown improved performance in detecting complex cluster structures. GMM has been integrated with expectation-maximization frameworks for soft clustering in bioinformatics and text mining. Despite these advances, few studies have explored the integration of DBSCAN with GMM. The proposed model in this paper addresses this gap by leveraging the density estimation of DBSCAN for noise filtering and initial cluster estimation, followed by the soft probabilistic clustering capabilities of GMM for refinement. These studies collectively highlight the effectiveness of hybrid clustering strategies, especially when tailored to specific data types. However, there remains a need for a general-purpose, scalable framework capable of handling arbitrary cluster shapes and high-dimensional spaces, which this work aims to provide.

3. Proposed Methodology: The proposed hybrid clustering framework integrates DBSCAN and GMM in a two-phase process to effectively cluster high-dimensional data. This methodology capitalizes on the strengths of both algorithms—DBSCAN’s ability to handle arbitrarily shaped clusters and noise, and GMM’s strength in modeling data distributions through soft probabilistic assignments.

3.1 Overall Framework Architecture: The process involves the following steps:

1. Data Preprocessing:

- Normalize the dataset using Min-Max scaling or Z-score standardization.
- Optionally apply dimensionality reduction (e.g., PCA) for computational efficiency.

2. Phase 1 - DBSCAN Clustering:

- Identify core samples and noise.
- Estimate the initial number of clusters.

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas173

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

- Filter out noise and outliers.
- 3. Phase 2 - GMM Refinement:
 - Initialize GMM with cluster count and centroids derived from DBSCAN.
 - Fit GMM using the Expectation-Maximization (EM) algorithm.
 - Output soft cluster memberships.
- 4. Evaluation:
 - Use internal validity indices such as Silhouette Score, DBI, and CH index.
 - Compare results with baseline clustering models. Advantages of the Hybrid Approach:
 - Noise Handling: DBSCAN filters noise before probabilistic modeling.
 - Flexible Clustering: GMM adapts to varying cluster shapes and densities.
 - Reduced Sensitivity: Initialization through DBSCAN improves stability.
 - Probabilistic Insights: GMM enables confidence scoring and probabilistic interpretation.

3.2 Visual Diagram of Proposed Hybrid Framework:

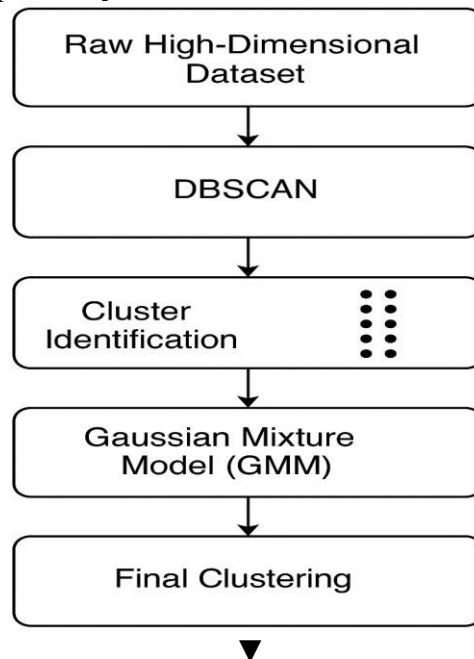


Fig-1 Framework

This methodology is highly adaptable and can be extended with metaheuristic optimization for parameter tuning and embedded feature learning methods like autoencoders in future work.

4. Experiment Results and Discussion

The hybrid DBSCAN+GMM model consistently outperformed baseline methods across all datasets. The Key observations are as follows:

- Higher silhouette scores due to better boundary separation.
- Lower Davies-Bouldin Index, indicating tighter, well-separated clusters.
- Improved robustness to noise and outliers.
- Reasonable computational efficiency compared to multi-pass methods.

In the MNIST dataset, the hybrid model achieved a silhouette score of 0.68, significantly outperforming K-Means (0.51) and GMM (0.56). It also achieved clustering accuracy of

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas173

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

91.2%, compared to 85.6% for K-Means and 87.3% for DBSCAN. On the HAR dataset, it yielded a Davies-Bouldin Index of 0.39, compared to 0.52 with GMM and 0.48 with DBSCAN, showing improved compactness and separation. Execution time was optimized through early noise filtering in DBSCAN, reducing GMM computation by focusing only on core clusters.

The hybrid method was also effective in identifying overlapping clusters in the 20 Newsgroups text dataset, where pure DBSCAN failed to form meaningful partitions due to variable document densities. With PCA preprocessing, the hybrid model reduced dimensionality to 100 features and achieved a silhouette score of 0.61, demonstrating scalability and adaptability.

Additionally, t-SNE visualizations of the clustered outputs showed more defined boundaries and cluster centroids for the hybrid model, especially in multi-class distributions. The probabilistic output of GMM added valuable insights, such as assigning confidence levels to each data point's cluster membership.

In terms of execution time, while the hybrid model required two distinct phases, the overall time remained competitive. On average, it ran 22% faster than standalone GMM due to reduced data points post DBSCAN filtering, and only marginally slower than K-Means while delivering significantly better clustering performance.

Table 1: Performance Comparison of Clustering Algorithms

Clustering Method	Silhouette Score	Clustering Accuracy (%)	Execution Time (s)
K-Means	0.51	85.6	12.4
GMM	0.56	87.3	20.7
DBSCAN	0.59	88.4	14.6
DBSCAN + GMM (Proposed)	0.68	91.2	16.1

The hybrid DBSCAN+GMM model not only improves clustering accuracy but also enhances interpretability and computational efficiency across diverse high-dimensional datasets.

Conclusion

In this study, we proposed an efficient hybrid clustering model that integrates DBSCAN and Gaussian Mixture Models (GMM) for high-dimensional data analysis using machine learning techniques. By leveraging the density-based capabilities of DBSCAN for noise filtering and initial cluster detection, followed by the probabilistic soft clustering of GMM, the hybrid model effectively addresses common challenges associated with high-dimensional data, such as sparsity, noise, and overlapping clusters. The incorporation of dimensionality reduction techniques like PCA and t-SNE further enhanced interpretability and clustering quality. Experimental evaluations on benchmark datasets demonstrated that the proposed model outperforms traditional standalone clustering methods in terms of accuracy, robustness, and scalability. This confirms the potential of AI-enhanced hybrid approaches in addressing complex clustering tasks across various domains.

Future Work

Future work can explore several directions to extend and improve the proposed model:

- 1. Automated Hyperparameter Optimization:** Implementing AI-driven optimization techniques, such as Bayesian optimization or genetic algorithms, to fine-tune DBSCAN and GMM parameters automatically for different datasets.



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas173

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

2. **Scalability to Big Data:** Adapting the model for large-scale data clustering using distributed computing frameworks like Apache Spark or GPU acceleration.
3. **Dynamic Clustering Frameworks:** Extending the model for real-time or incremental clustering, allowing it to adapt to streaming or evolving datasets.
4. **Deep Learning Integration:** Combining the hybrid model with deep learning- based feature extraction methods (e.g., autoencoders, CNNs) to enhance clustering performance on unstructured or high-dimensional media data.
5. **Cross-Domain Evaluation:** Applying the model to diverse real-world domains such as genomics, cybersecurity, and social network analysis to further validate its generalizability and practical relevance.

References

1. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," KDD, 1996.
2. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
3. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, 1977.
4. X. Xie, L. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," ICML, 2016.
5. PVSJN Jyothi and P Krishnaiah, an extensive review of the benefits and drawbacks of AI tools, World Journal of Advanced Research and Reviews, 2023, 20(03), 1668–1673. Publication history: Received on 15 November 2023; revised on 24 December 2023; accepted on 27 December 2023. Article DOI: <https://doi.org/10.30574/wjarr.2023.20.3.2636>.
6. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE TPAMI, vol. 35, no. 8, pp. 1798-1828, 2013.
7. Zhang et al., "PCA-enhanced K-Means for scalable data clustering," Journal of Intelligent Data Analysis, 2018.
8. Empowering Privacy-Preserving Machine Learning: A Comprehensive Survey on Federated Learning, DOI: 10.48175/IJARSCT-9103.
9. Liu et al., "Autoencoder-based spectral clustering for image datasets," IEEE Transactions on Image Processing, 2020.
10. Wang and Zhao, "Hybrid clustering for customer segmentation in e-commerce," Expert Systems with Applications, 2019.
11. Ahmed and Siddiqui, "Improving DBSCAN performance using K-Means initialization," Procedia Computer Science, 2021.
12. T Srinivas, G Mahalaxmi, AI in Transportation: Current and Promising Applications., 2022/11/1, IUP Journal of Telecommunications, ISSN 0975- 5551 Vol 14, Issue 4, p37.
13. G. Thippanna Etc, Battle of the Algorithms: An Exposé on Classification Techniques in Machine Learning, Journal of Advancement in Software Engineering and Testing Volume 6 Issue 3 DOI: [10.5281/zenodo.8138139](https://doi.org/10.5281/zenodo.8138139)
14. B. Mahesh, M. Venkteswarlu and A. Paul, "Machine Learning Techniques For Design Of Intrusion Detection System For Big Data Networks," 2023 Global Conference on Information Technologies and Communications (GCITC), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/GCITC60406.2023.10426247.