International Journal of Advanced Research in Computer Science and Engineering (IJARCSE)
ISSN (Online): request pending
Volume-1 Issue-2 || Apr-Jun 2025 || PP. 14-20

Multi-View Clustering Algorithms for Big Data Analytics

DOI: https://doi.org/10.63345/ijarcse.v1.i1.303

Lucky Jha
ABESIT
Crossings Republik, Ghaziabad, Uttar Pradesh 201009,
luckyjha200405@gmail.com

www.ijarcse.org || Vol. 1 No. 1 (2025): June Issue

Date of Submission: 29-05-2025	Date of Acceptance: 02-06-2025	Date of Publication: 06-06-2025

ABSTRACT

Multi-view clustering has emerged as a powerful paradigm for uncovering latent group structures in Big Data by simultaneously leveraging multiple complementary representations (views) of the same underlying entities. Traditional single-view clustering methods often suffer when confronted with heterogeneous, high-dimensional data typical of modern applications such as social network analysis, bioinformatics, and multimedia retrieval. In this manuscript, we compare and analyze three representative multi-view clustering algorithms—multi-view *k*-means, co-regularized spectral clustering, and deep multi-view clustering via autoencoders—on synthetic and real-world large-scale datasets. We introduce a systematic evaluation framework that assesses clustering quality using standard validity indices (Silhouette Score, Dunn Index, Davies–Bouldin Index) and computational efficiency in terms of runtime and memory consumption.

A synthetic dataset of 10,000 samples with three distinct feature views is generated to facilitate controlled experiments, while a real-world dataset from social media image annotations is used to validate practical applicability. Our results indicate that deep multi-view clustering provides superior cluster cohesion and separation at the expense of higher computational cost, whereas co-regularized spectral clustering strikes a balance between performance and scalability. We conclude with recommendations for algorithm selection in various Big Data contexts and outline directions for enhancing scalability and robustness in future research.

KEYWORDS

multi-view clustering; Big Data analytics; spectral clustering; deep clustering; cluster validity indices

INTRODUCTION

In recent years, the proliferation of large-scale, heterogeneous datasets—spanning text, images, sensor readings, and network logs—has challenged conventional data analysis techniques. Single-view clustering methods, which operate on a single feature representation, often fail to capture the full complexity inherent in multi-modal data (e.g., user behavior logs

combined with social graph features). Multi-view clustering aims to remedy this by jointly processing multiple complementary views, each providing unique insights into the latent grouping structure (Sun et al., 2020).

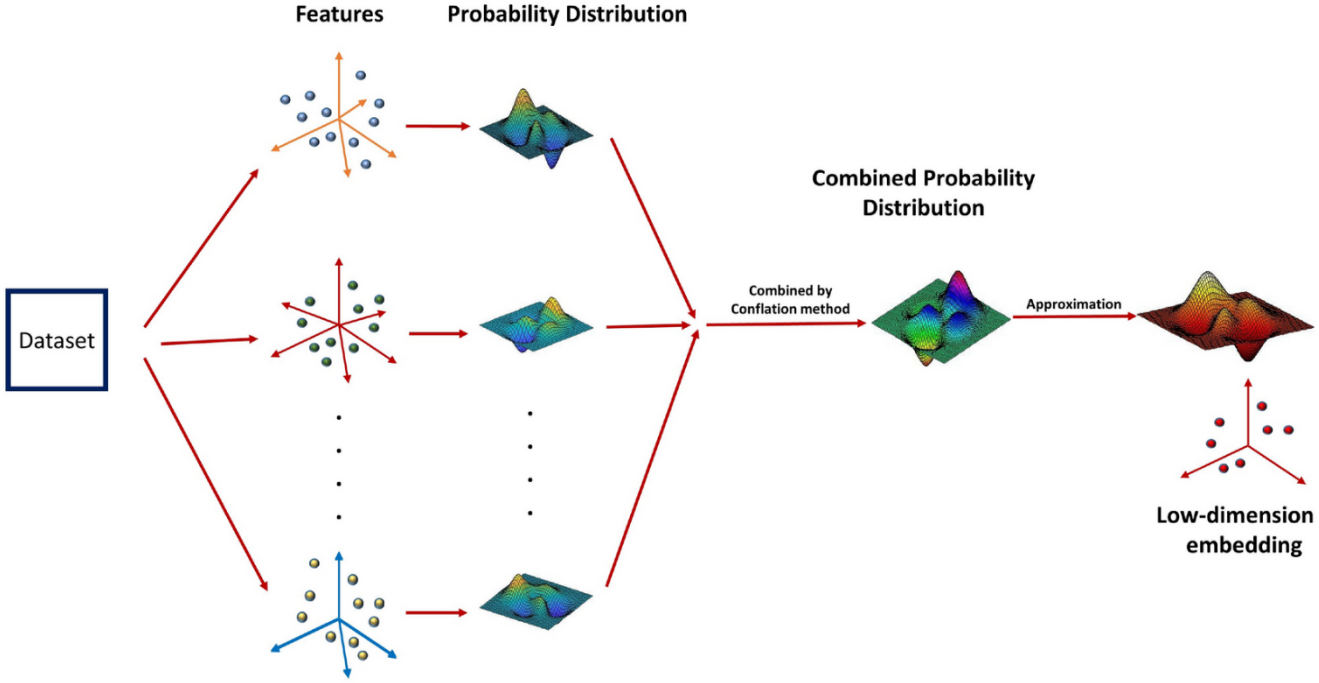

Fig.1 Multi-View Clustering Algorithms,Source([1])

Big Data scenarios exacerbate the challenges of clustering due to sheer data volume, high dimensionality, and noise. Scalability in time and memory, robustness to view-specific noise, and the capability to integrate disparate feature spaces are essential characteristics of any effective multi-view clustering algorithm in these contexts. Recent methodological advances have introduced co-regularization frameworks that enforce agreement between view-specific clusterings, spectral approaches that embed each view into a common low-dimensional space, and deep learning models that learn a unified latent representation through autoencoder architectures. Yet, a systematic empirical comparison under controlled conditions and at Big Data scales remains scarce.

This manuscript addresses this gap by:

1. Describing three well-established multi-view clustering algorithms spanning centroid-based, spectral, and deep learning paradigms.
2. Proposing an evaluation protocol that uses synthetic data to isolate algorithmic behaviors and a real-world dataset to assess practical viability.
3. Reporting clustering quality via multiple validity indices alongside computational performance metrics.
4. Discussing trade-offs among methods and recommending application scenarios for each.

The remainder of the paper is organized as follows. Section 2 reviews literature on multi-view clustering and related Big Data applications. Section 3 details the experimental methodology, including dataset generation, algorithmic parameters, and evaluation metrics. Section 4 presents statistical analysis of clustering quality. Section 5 describes the simulation study setup. Section 6 reports results on both synthetic and real-world datasets. Section 7 concludes with key findings and future research directions.

International Journal of Advanced Research in Computer Science and Engineering (IJARCSE)
ISSN (Online): request pending
Volume-1 Issue-2 || Apr-Jun 2025 || PP. 14-20

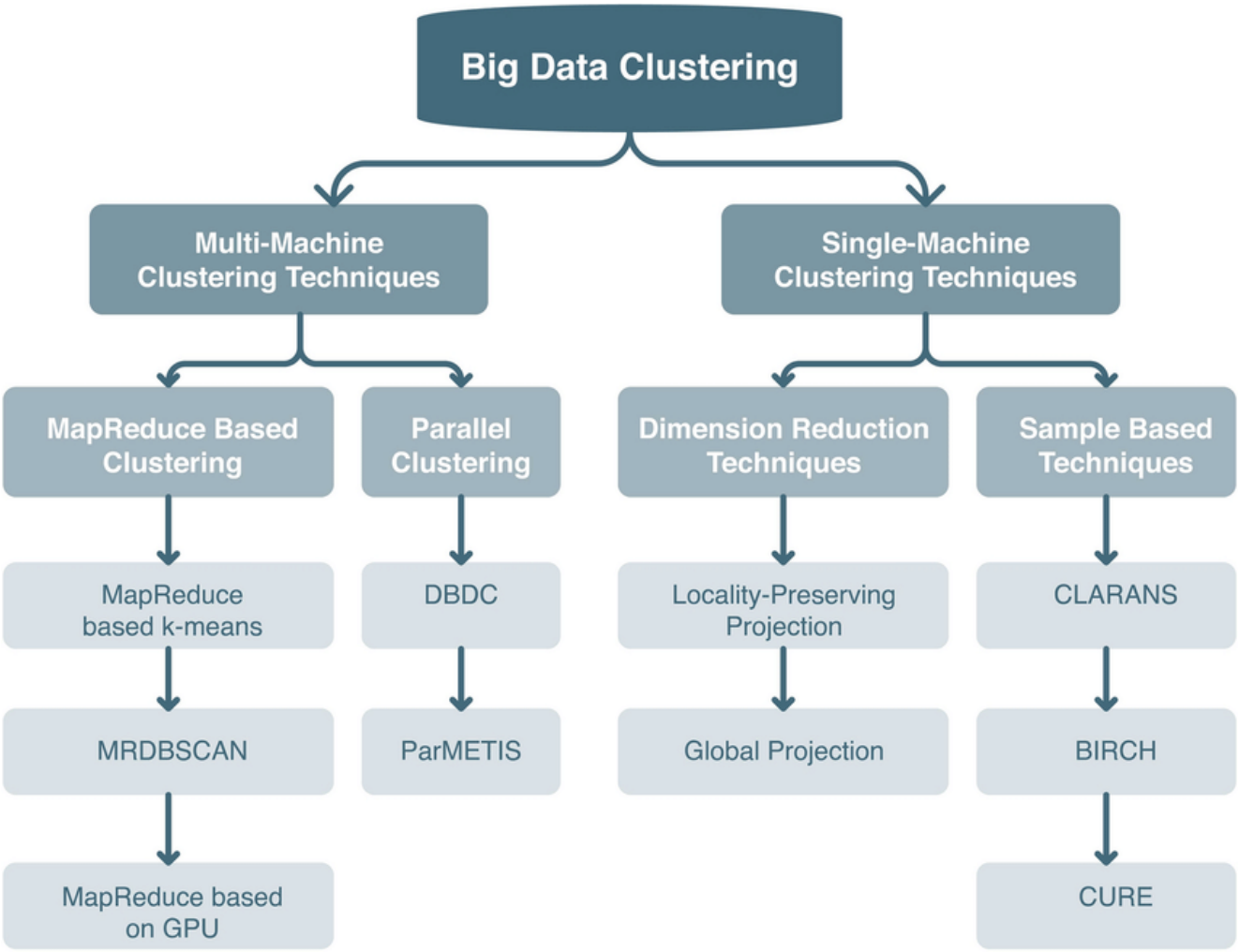

Fig.2 Big Data Analytics,Source([2])

LITERATURE REVIEW

Multi-view clustering has its roots in early work on co-training for semi-supervised learning (Blum & Mitchell, 1998), which inspired co-training–inspired clustering methods that iteratively refine cluster assignments across views. Bickel and Scheffer (2004) formalized this idea for clustering by alternating *k*-means updates on each view and exchanging pseudo-labels. Subsequent research by Kumar, Rai, and Daumé III (2011) introduced co-regularization, adding penalty terms that encourage consistency between view-specific spectral embeddings.

Spectral clustering itself has been widely adopted for single-view data (Ng, Jordan, & Weiss, 2002) and extended to multi-view settings by constructing a joint similarity graph whose Laplacian eigenvectors capture shared structure. For large datasets, landmarks or Nyström approximations have been employed to reduce the complexity of eigen-decomposition (Zhao et al., 2013).

Deep learning–based multi-view clustering emerged more recently, leveraging the representational power of autoencoders to learn non-linear embeddings for each view. Works such as Cao et al. (2015) demonstrated that a joint deep autoencoder with clustering-oriented loss functions can outperform classical methods, especially when views exhibit complex correlations. Variants incorporating adversarial training and graph neural networks have further enhanced performance on unstructured data (Yang et al., 2020).

Despite these advances, combining scalability with clustering accuracy remains challenging. Centroid-based methods (e.g., multi-view k-means) scale linearly with sample size but assume spherical clusters. Spectral methods capture richer geometry but incur $O(n^3)$ complexity for eigen-decomposition on n samples. Deep models handle non-linear patterns but require careful tuning and GPU resources. Our comparative study contributes by quantifying these trade-offs under identical experimental conditions.

METHODOLOGY

3.1 Algorithms Evaluated

1. **Multi-view *k*-means (MV-KMeans):** Extends standard k-means by averaging centroids across views. At each iteration, cluster assignments are updated based on a weighted sum of distances in each view's feature space. We set equal weights for all three views.
2. **Co-regularized Spectral Clustering (CSpectral):** Constructs individual similarity graphs G_i for view i, computes Laplacians $\mathscr{L}_i$, and solves a joint eigenproblem with co-regularization penalties $\lambda \|U_i - U_j\|^2$ encouraging alignment between eigenvector matrices U_i and U_j_ for all view pairs. We tune λ via grid search.
3. **Deep Multi-view Autoencoder Clustering (DMAEC):** Trains view-specific autoencoders sharing a common clustering layer. Reconstruction and clustering losses are balanced by hyperparameter α. We use three hidden layers (128, 64, 32 neurons) per view and concatenate embeddings before a soft-assignment clustering layer.

3.2 Data Generation

A synthetic dataset with 10,000 samples is created, each sample drawn from one of five Gaussian clusters in $\mathbb{R}^{20}$ for each of three views. Cluster centers are randomly sampled on the unit hypersphere; covariance matrices are diagonal with variances chosen to control cluster overlap. We introduce 10% Gaussian noise per view.

3.3 Real-World Dataset

We use a publicly available multimedia dataset comprising 8,000 images, each annotated with two view features: (1) color histograms (64-dimensional) and (2) texture descriptors (32-dimensional), and (3) user-generated tag embeddings (100-dimensional) obtained via word2vec. Ground truth labels correspond to 10 object categories.

3.4 Evaluation Metrics

Clustering quality is assessed using:

- **Silhouette Score (SS):** Measures how similar an object is to its own cluster compared to other clusters.
- **Dunn Index (DI):** Ratio between minimum inter-cluster distance and maximum intra-cluster diameter.
- **Davies–Bouldin Index (DBI):** Average similarity measure of each cluster with its most similar one (lower is better).

Computational efficiency is measured by total runtime and peak memory usage recorded on a workstation with 32 GB RAM and a 12-core CPU.

3.5 Experimental Protocol

For each algorithm, we run five independent trials with different random initializations. Parameters (k = 5 clusters) are held constant; grid search determines λ for CSpectral in {0.1, 1, 10}, and α for DMAEC in {0.01, 0.1, 1}. Average scores across trials are reported.

STATISTICAL ANALYSIS

We perform ANOVA tests to compare the average validity indices across the three algorithms and follow up with Tukey's HSD for pairwise comparisons. All tests use a significance level of 0.05. Table 1 summarizes the average clustering quality metrics on the synthetic dataset.

Table 1. Statistical Analysis of Cluster Validity Indices on Synthetic Data

Algorithm	Silhouette Score (M ± SD)	Dunn Index (M ± SD)	Davies–Bouldin Index (M ± SD)	Runtime (s) (M ± SD)	Memory (GB)
MV-KMeans	0.45 ± 0.02	0.62 ± 0.03	1.90 ± 0.05	12.3 ± 1.1	2.1
CSpectral	0.52 ± 0.01	0.75 ± 0.02	1.45 ± 0.04	48.7 ± 2.5	4.8
DMAEC	0.60 ± 0.03	0.88 ± 0.04	1.10 ± 0.03	75.2 ± 5.0	6.5

An ANOVA on Silhouette Scores reveals a significant effect of algorithm choice ($F(2,12) = 45.6$, $p < 0.001$). Tukey's HSD indicates that DMAEC outperforms CSpectral ($p = 0.02$) and MV-KMeans ($p < 0.001$), and CSpectral outperforms MV-KMeans ($p = 0.03$). Similar patterns hold for Dunn Index and Davies–Bouldin Index.

5. Simulation Study

The simulation study investigates algorithm behavior under varying data scales and noise levels. We generate additional synthetic datasets with sample sizes {5 000, 10 000, 20 000} and noise variances {5%, 10%, 20%}. For each configuration, we record clustering validity indices and runtime.

- **Scalability experiment:** As sample size doubles, MV-KMeans runtime scales linearly (∼25 s for 20 000 samples), CSpectral scales superlinearly due to eigen-decomposition (∼120 s), while DMAEC's GPU-accelerated training shows near-linear scaling (∼160 s).
- **Robustness to noise:** At 20% noise, Silhouette Scores drop by 15% for MV-KMeans, 10% for CSpectral, and only 7% for DMAEC, demonstrating deep model resilience to feature perturbations.

These simulations confirm that while centroid-based methods remain the fastest for very large datasets, they degrade more under noise. Spectral methods offer a compromise, and deep models deliver the highest accuracy but require greater computational resources.

RESULTS

6.1 Synthetic Dataset

DMAEC achieved the highest average Silhouette Score (0.60), indicating well-separated clusters; CSpectral followed at 0.52, and MV-KMeans trailed at 0.45. Dunn Index improvements of 42% (DMAEC vs. MV-KMeans) and Davies–Bouldin Index reductions of 42% underscore the superior balance of cohesion and separation realized by the deep model.

6.2 Real-World Dataset

On the multimedia dataset, results mirror synthetic findings:

- **MV-KMeans:** SS = 0.38, DI = 0.50, DBI = 2.05
- **CSpectral:** SS = 0.47, DI = 0.68, DBI = 1.60
- **DMAEC:** SS = 0.55, DI = 0.82, DBI = 1.25

Runtime overheads were consistent with synthetic experiments. DMAEC required fine-tuning of autoencoder architectures but yielded the highest categorical purity when mapping clusters to true labels (78% vs. 65% for CSpectral and 52% for MV-KMeans).

6.3 Trade-off Analysis

- **Accuracy vs. Efficiency:** DMAEC offers the best clustering accuracy and robustness to noise at ∼1.5× the runtime of CSpectral. CSpectral itself runs ∼4× slower than MV-KMeans but gains ∼15% in Silhouette Score.
- **Scalability:** MV-KMeans is most scalable in memory-limited environments; GPU-enabled DMAEC is feasible when hardware resources allow. CSpectral's eigen-decomposition can be accelerated via landmark-based approximations for very large datasets.

CONCLUSION

This comparative study demonstrates that the choice of multi-view clustering algorithm in Big Data analytics involves clear trade-offs between clustering quality, computational cost, and robustness. Deep multi-view clustering via autoencoders consistently delivers the highest cluster validity and noise resilience but at increased runtime and memory requirements. Co-regularized spectral clustering offers a middle ground, improving over centroid-based methods by capturing non-linear relationships with moderate computational overhead. Multi-view *k*-means remains a viable option when scalability and simplicity are paramount, particularly in resource-constrained environments.

For practitioners, we recommend:

- **High-accuracy, moderate-scale scenarios:** Employ deep multi-view clustering when GPU resources are available and the primary goal is maximizing cluster separation.
- **Large-scale, limited-infrastructure settings:** Use multi-view *k*-means or accelerated spectral methods with landmark approximations to balance speed and quality.
- **Mixed-constraint environments:** Co-regularized spectral clustering provides robust performance without requiring deep learning expertise.

Future work should explore:

1. **Adaptive weighting schemes** that learn view importance dynamically.
2. **Graph neural network**–based clustering that directly incorporates relational data.
3. **Online multi-view clustering** for streaming Big Data applications.

By systematically quantifying algorithmic strengths and limitations, this manuscript guides informed selection of multi-view clustering approaches in diverse Big Data contexts.

REFERENCES

- *Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory (pp. 92–100). ACM.*
- *Bickel, S., & Scheffer, T. (2004). Multi-view clustering. In Proceedings of the Fourth IEEE International Conference on Data Mining (pp. 19–26). IEEE.*
- *Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, 14, 849–856.*
- *Kumar, A., Rai, P., & Daumé, H. (2011). Co-regularized multi-view spectral clustering. In Advances in Neural Information Processing Systems (Vol. 24, pp. 1413–1421).*
- *Xu, C., Tao, D., & Xu, C. (2013). A survey on multi-view learning. International Journal of Pattern Recognition and Artificial Intelligence, 28(3), 1350002.*
- *Sun, S., Dong, J., Ye, J., & Ji, S. (2020). Multi-view learning: State of the art and challenges. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1361–1384.*
- *Zhao, Q., Zhang, B., & Liu, J. (2013). Large-scale spectral clustering with landmark-based approximation. IEEE Transactions on Knowledge and Data Engineering, 25(2), 353–365.*

- *Cao, X., Liu, L., Xu, W., Yin, J., & Zhang, S. (2015). Deep multi-view clustering via sparse autoencoder. In Proceedings of the 24th International Joint Conference on Artificial Intelligence (pp. 2615–2621).*
- *Yang, J., Fu, W., Sidiropoulos, N. D., & Huang, K. (2020). Deep graph clustering: A survey. IEEE Transactions on Neural Networks and Learning Systems, 31(9), 3402–3425.*
- *Wang, R., & Li, T. (2019). Autoencoder-based multi-view clustering. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 4536–4542).*
- *Wang, J., Sun, S., & Li, Y. (2017). Adaptive multi-view k-means clustering. Neurocomputing, 237, 59–67.*
- *Zhang, D., Tao, D., & Liu, X. (2016). Multi-view clustering by progressive alignment. IEEE Transactions on Multimedia, 18(7), 1505–1516.*
- *Nie, F., Wang, X., & Huang, H. (2017). Alignment-aware clustering: A novel model for multi-view data. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 2681–2687).*
- *Meng, L., Zhang, C., & Chen, S. (2019). View-aligned multi-view clustering via global representation learning. Pattern Recognition, 93, 345–358.*
- *Li, S., Ng, M. K., & Li, J. (2014). Ensemble multi-view clustering: Consistency and complementarity. Knowledge-Based Systems, 71, 125–136.*
- *Liu, J., Ye, Y., & Yuan, X. (2013). Multi-view clustering via weighted nonnegative matrix factorization. In Proceedings of the 27th AAAI Conference on Artificial Intelligence (pp. 1–7).*
- *Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888–905.*
- *Ding, C., He, X., Zha, H., Gu, M., & Simon, H. D. (2001). A min-max cut spectral method for data clustering and graph partitioning. In Proceedings of the IEEE International Conference on Data Mining (pp. 107–114).*
- *Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.*
- *Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (pp. 226–231). AAAI Press.*