Woody International Publish Limited
An Multidisciplinary Academic Journal Publisher

Journal of Theory and Practice in Engineering and Technology, Volume 2, Issue 2, 2025
https://www.woodyinternational.com/

A Review of Deep Multi-View Clustering Methods

Yijian Fu*

School of Software, Jiangxi Normal University, Jiangxi 330022, China
Author to whom correspondence should be addressed.

Abstract: ***With the advent of the big data era, multi-view data has become an indispensable and crucial data format in numerous real-world applications. As an effective unsupervised learning approach, multi-view clustering (MVC) can fully exploit the information from multi-view data without requiring labels, thereby uncovering its inherent clustering structures. In recent years, advancements in deep learning technologies have significantly propelled the innovation of multi-view clustering methods. Deep multi-view clustering (DMVC) methods have garnered substantial attention due to their exceptional advantages in non-linear feature learning, unsupervised clustering, and multi-source data fusion. This paper provides a comprehensive review of the latest research advancements in DMVC methods, with a focus on different strategies such as autoencoder-based, self-representation, contrastive learning, and ensemble learning approaches. An in-depth analysis of the characteristics, strengths, and limitations of each method is also presented. Furthermore, the current challenges in the DMVC domain are summarized, and insights into future research directions are provided, aiming to offer valuable references for relevant researchers.***

Keywords: Data Mining; Deep Learning; Unsupervised Learning; Multi-View Clustering; Feature Representation.

Cited as: Fu, Y. (2025). A Review of Deep Multi-View Clustering Methods. *Journal of Theory and Practice in Engineering and Technology*, *2*(2), 1–5. Retrieved from https://woodyinternational.com/index.php/jtpet/article/view/174

1. Introduction

In the era of rapid development of information technology and the explosion of big data, data from diverse fields and perspectives emerges at an unprecedented velocity. One of the significant advantages of such massive datasets is their ability to describe data instances in multiple representations [1]. For example, in disease diagnosis, medical imaging, genomic data, and electronic health records, among other perspectives, provide comprehensive patient information and offer invaluable insights for precise diagnosis [2]. These data from various sources collectively form multi-view data, where each perspective conveys partial information and complements others in a complementary structure. Multi-view data is prevalent across multiple domains, offering abundant and comprehensive information that presents new opportunities for data mining. However, it also poses challenges in integrating and utilizing such information. The core of multi-view learning lies in extracting key information from each perspective, fusing it into a unified representation, and supporting the efficient execution of downstream tasks. Nevertheless, as the scale of data generation, measurement, and acquisition continues to expand, accurately annotating and processing such massive datasets has become increasingly difficult. In this context, multi-view clustering (MVC) [3-5], as an unsupervised learning method, can effectively extract and fuse representations from multiple perspectives while identifying the naturally formed clustering structures within the data. Consequently, multi-view clustering has emerged as a hot research direction in current studies.

Traditional MVC methods typically employ linear or mapped subspaces to learn data representations before fusing the representations across views. However, such linear constraints severely limit their ability to capture the inherent nonlinear characteristics of complex data. While kernel methods can introduce some nonlinearity to address this limitation [6-7], they still fall short in handling complex data structures effectively. Furthermore, conventional approaches exhibit poor scalability when dealing with large-scale datasets, making them inadequate for big data analysis demands. With the increasing attention paid to deep learning techniques, Deep Multi-View Clustering (DMVC) has gradually demonstrated significant advantages [8-11]. DMVC fully leverages the powerful capability of deep learning in nonlinear feature extraction to effectively capture the complex structures of data. Moreover, it capitalizes on the advancements in deep clustering techniques to simultaneously achieve deep representation

learning and data clustering in an unsupervised manner. Given its notable strengths in nonlinear feature learning, unsupervised clustering, and multi-source data fusion, DMVC is emerging as a crucial approach for solving complex multi-source data clustering problems. As research progresses, DMVC is expected to witness even broader prospects for development in both theoretical and practical dimensions.

2. Deep Multi-View Clustering

With the increasing availability of multi-source data, multi-view data has become prevalent in various domains such as image analysis, social networks, and bioinformatics. Compared to single-view data, multi-view data contains complementary information from different modalities or sources, making multi-view clustering a crucial research area. Deep Multi-View Clustering (DMVC) combines the strengths of deep learning and multi-view learning, enabling automatic feature extraction, information fusion, and clustering optimization to improve clustering accuracy and robustness. Current deep multi-view clustering methods can be broadly categorized into four major approaches: autoencoder-based methods, self-representation-based methods, contrastive learning-based methods, and ensemble-based methods. Autoencoder-based DMVC employs deep neural networks to extract low-dimensional representations and fuses multi-view information to enhance clustering performance. Self-representation-based methods construct self-representation matrices to capture intrinsic data structures, making them suitable for data with complex subspace structures. Contrastive learning-based DMVC improves the discriminative power of multi-view data by maximizing the similarity between positive pairs and minimizing the similarity between negative pairs. Ensemble-based methods integrate multiple base clustering results to enhance clustering stability and generalization ability. This article provides an in-depth review of these four categories of deep multi-view clustering methods, discussing their core ideas, representative studies, and practical advantages and challenges.

2.1 Autoencoder-based Deep Multi-View Clustering

Autoencoder is a widely used artificial neural network model in deep learning, primarily designed for unsupervised learning tasks. It consists of an encoder (Encoder) and a decoder (Decoder), aiming to learn feature representations through repeated training. The encoder maps input data into a low-dimensional latent space, while the decoder reconstructs the original data from the latent space. The primary objective of an autoencoder is to minimize the difference between the input and its reconstructed output, and this self-learning characteristic enables it to demonstrate strong capabilities in dimensionality reduction, feature extraction, and data generation. In Autoencoder-based Deep Multi-View Clustering (Autoencoder-based DMVC), autoencoders are employed to extract useful feature representations from each view. The goal is to achieve more accurate and robust clustering results by fusing multi-view information. Autoencoder-based DMVC methods typically follow the following steps: first, train an autoencoder for each view to extract its low-dimensional feature representations; then, fuse these representations, which may involve simple concatenation or more complex fusion strategies; finally, perform clustering based on the fused features. This approach fully leverages the advantages of autoencoders in feature extraction, providing an effective solution for the clustering of multi-view data. Xie et al. [12] proposed the Unsupervised for Deep Embedding Clustering Analysis (DEC) algorithm, which learns feature representations of data via autoencoders and utilizes nonlinear mappings to transform data into low-dimensional feature spaces. In this space, the clustering objective function is iteratively optimized, enabling simultaneous learning of feature representations and clustering assignments to enhance clustering performance. Lin et al. [13] addressed the problem of multi-view hierarchical clustering by proposing the Multi-View Hierarchical Clustering Network (MHCN) model. This model employs multiple hyperbolic autoencoders to encode data from different views, capturing hierarchical structures within each view, and performs multi-view representation learning and hierarchical modeling in hyperbolic space, achieving effective multi-view hierarchical clustering. Xu et al. [14] investigated the adverse effects of noisy views in multi-view clustering and proposed the MVCAN method. This method utilizes autoencoders to learn feature representations of multi-view data and introduces new clustering objectives and a two-stage multi-view iterative optimization strategy. This allows different views to have distinct clustering predictions, avoiding the negative impact of noisy views on other effective views, thereby improving the robustness of clustering.

2.2 Self-Representation-Based Deep Multi-View Clustering

Self-representation-based deep multi-view clustering methods capture the intrinsic structure of data instances through the construction of a self-representation matrix. These methods typically employ strategies such as low-rank representation (LRR) or sparse subspace learning (SSL) to discover global or local relationships between data

points and fuse information from different views. Self-representation methods are well-suited to revealing the subspace structure of multi-view data, making them applicable to tasks such as social network analysis and graph data clustering, which involve structured data. Zhang et al. [15] proposed a self-representation-based subspace clustering method that integrates multi-view data's complementary information via potential representations and explores complex non-linear relationships between potential representations and each view through neural networks. This method jointly learns potential representations and multi-view subspace representations within a unified framework and optimizes the process through an alternating direction minimization (ADM) strategy, ensuring algorithm convergence and stability. Zhao et al. [16] introduced the Aligned Self-Supervised Incomplete Multi-View Subspace Clustering Network (DASIMSC), innovatively incorporating double alignment constraints (manifold alignment and consistency alignment) and a self-supervised mechanism. This method utilizes view-specific autoencoders to learn potential representations for each view, ensuring consistency and complementarity across different views through the double alignment constraints. A self-expression module is employed to learn the self-representation of data, capturing the intrinsic structure and similarity of data, while a self-supervised clustering module uses the current clustering results to guide subspace learning, eliminating cluster-to-cluster interference. Zhang et al. [17] innovatively combined general and specific space learning and introduced an iterative self-supervision mechanism. This method captures both common and specific features through multi-view encoders and self-expression layers, and an iterative self-supervision module is used to continuously optimize the self-representation learning process, enhancing clustering performance.

2.3 Contrastive Learning-Based Deep Multi-View Clustering

Contrastive learning-based deep multi-view clustering (Contrastive Learning-based DMVC) methods enhance the feature representation capabilities of multi-source heterogeneous data by maximizing the similarity of positive pairs and minimizing the similarity of negative pairs. These methods leverage the advantages of contrastive learning in feature learning to capture discriminative features and improve clustering accuracy and robustness through the fusion of multi-view information. For example, Lin et al. [18] optimized information-theoretic objectives to maximize the mutual information between representations of two views while also maximizing the entropy of each view, thereby avoiding trivial solutions that only identify single clusters. Ke et al. [19] employed multi-view autoencoders to extract potential representations from each view and generated common view representations through a network, enhancing alignment and consistency across different views. Xu et al. [20] utilized autoencoders and multi-layer perceptron (MLP) to generate high-level features and clustering labels, optimizing the process using a contrastive loss function based on cosine similarity. To address the challenges of handling partial or incomplete views, Feng et al. [21] introduced view-specific contrastive encoders and applied a self-expression layer on potential representations to generate pseudo-labels, thereby optimizing the performance of contrastive encoders. Additionally, Lu et al.'s method [22] proposed a multi-step random walk approach to select neighbors, reducing errors in the selection of positive and negative samples in contrastive learning and optimizing view-specific twin encoders and cross-view decoders. Despite the remaining challenges in efficiently selecting sample pairs and handling view heterogeneity in implementation, contrastive learning provides a powerful tool for deep multi-view clustering. Its theoretical and practical prospects are promising, offering new possibilities for solving complex multi-source data clustering problems.

2.4 Ensemble-Based Deep Multi-View Clustering

Ensemble-based deep multi-view clustering (Ensemble-based DMVC) methods enhance clustering stability and robustness by integrating multiple base clustering results. Drawing inspiration from the success of ensemble clustering in single-view data, this approach was first extended to multi-view scenarios by Tao et al. [23]. Specifically, the method begins by generating multiple base clusters for each view, where the number of clusters exceeds the actual required category count. Next, an instance co-association matrix is constructed to record the average frequency of instance pairs across all base clusters. Denoising autoencoders are then trained for each view to learn low-rank representations of the co-association matrix. By modeling both consistency and inconsistency across all views, the method ultimately identifies a consensus clustering result with high consistency. This modular design can be seen as a component of a larger network structure, where nonlinear activation functions are used to learn nonlinear co-association representations of instance pairs. Subsequent research has further developed and refined ensemble methods. For instance, [24] introduced an innovative strategy that cascades multiple spectral embeddings from each view's similarity graph and inputs them into a graph autoencoder to generate more efficient representations. Simultaneously, the discrete embeddings from each view are fused to form a common graph structure, which is then used to compute a graph contrastive loss function, thereby enhancing the alignment of cross-view information. Additionally, [25] incorporated an attention-based encoder to generate potential

representations and constructed a graph structure aligned with multiple clustering partitions, further improving clustering quality and consistency. Notably, [26] employed a Transformer encoder to process fused graph representations derived from multiple view similarity graphs and optimized them using a simple decoder with Softmax activation, demonstrating efficiency and flexibility in handling complex data scenarios. These ensemble-based DMVC methods fully leverage the complementary nature of multiple base clustering results, effectively mitigating the limitations of single models when dealing with multi-view data. By combining diverse fusion strategies and advanced deep learning techniques, such as attention mechanisms and Transformer models, ensemble methods not only enhance clustering performance but also provide a variety of solutions for addressing the challenges posed by large-scale, multi-source heterogeneous data.

3. Conclusion

As an important research direction in unsupervised learning, deep multi-view clustering (DMVC) has demonstrated immense application potential across various fields such as image processing, natural language processing, and bioinformatics. While existing methods have achieved significant advancements in feature representation learning, information fusion, and clustering performance improvement, there remain numerous challenges. Firstly, the heterogeneity, redundancy, and noise interference of multi-view data pose significant challenges to clustering performance. Secondly, there is a need for further in-depth research on how to enhance the interpretability and stability of algorithms based on deep models. Moreover, the computational complexity and scalability issues of models when dealing with large-scale multi-view data remain urgent to address. Future research can further explore efficient fusion strategies, more robust optimization methods, and DMVC model designs tailored to specific tasks to advance both the theoretical and practical development of deep multi-view clustering.

References

[1] Christoudias C, Urtasun R, Darrell T. Multi-view learning in the presence of view disagreement[C]. Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence. 2008: 88-96.
[2] Salvi M, Loh H W, Seoni S, et al. Multi-modality approaches for medical support systems: A systematic review of the last decade [J]. Information Fusion, 2024, 103:102134.
[3] Zhou L, Du G, Lü K, et al. A survey and an empirical evaluation of multi-view clustering approaches[J]. ACM Computing Surveys, 2024, 56(7): 1-38.
[4] Chao G, Sun S, Bi J. A survey on multiview clustering[J]. IEEE transactions on artificial intelligence, 2021, 2(2): 146-168.
[5] Fang U, Li M, Li J, et al. A comprehensive survey on multi-view clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(12): 12350-12368.
[6] Liu J, Cao F, Gao X Z, et al. A cluster-weighted kernel k-means method for multi-view clustering[C]//Proceedings of the aaai conference on artificial intelligence. 2020, 34(04): 4860-4867.
[7] Chen Y, Xiao X, Zhou Y. Jointly learning kernel representation tensor and affinity matrix for multi-view clustering[J]. IEEE Transactions on Multimedia, 2019, 22(8): 1985-1997.
[8] Chen J, Mao H, Woo W L, et al. Deep multiview clustering by contrasting cluster assignments[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 16752-16761.
[9] Du G, Zhou L, Li Z, et al. Neighbor-aware deep multi-view clustering via graph convolutional network[J]. Information Fusion, 2023, 93: 330-343.
[10] Fard M M, Thonet T, Gaussier E. Deep k-means: Jointly clustering with k-means and learning representations[J]. Pattern Recognition Letters, 2020, 138: 185-192.
[11] Law M T, Urtasun R, Zemel R S. Deep spectral clustering learning[C]//International conference on machine learning. PMLR, 2017: 1985-1994.
[12] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C]//International conference on machinelearning. PMLR, 2016: 478-487.
[13] LinF, Bai B, Guo Y, et al. Mhcn: A hyperbolic neural network model for multi-viewhierarchical clustering[C]//Proceedings of the IEEE/CVF internationalconference on computer vision. 2023: 16525-16535.
[14] Xu J, Ren Y, Wang X, et al. Investigating and mitigating the side effects of noisyviews for self-supervised clustering algorithms in practical multi-viewscenarios[C]//Proceedings of the IEEE/CVF Conference on Computer Vision andPattern Recognition. 2024: 22957-22966.

[15] Zhang C, Fu H, Hu Q, et al. Generalized latent multi-view subspace clustering[J]. IEEEtransactions on pattern analysis and machine intelligence, 2018, 42(1): 86-99.
[16] Zhao L, Zhang J, Wang Q, et al. Dual alignment self-supervised incomplete multi-viewsubspace clustering network[J]. IEEE Signal Processing Letters, 2021, 28:2122-2126.
[17] Zhang Y,Huang Q, Zhang B, et al. Deep multiview clustering via iterativelyself-supervised universal and specific space learning[J]. IEEE Transactions onCybernetics, 2021, 52(11): 11734-11746.
[18] Lin Y, Gou Y, Liu Z, et al. Completer: Incomplete multi-view clustering via contrastive prediction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11174-11183.
[19] Ke G, Hong Z, Zeng Z, et al. CONAN: contrastive fusion networks for multi-view clustering[C]//2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021: 653-660.
[20] Xu J, Tang H, Ren Y, et al. Multi-level feature learning for contrastive multi-view clustering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16051-16060.
[21] Feng W, Sheng G, Wang Q, et al. Partial multi-view clustering via self-supervised network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(11): 11988-11995.
[22] Lu Y, Lin Y, Yang M, et al. Decoupled contrastive multi-view clustering with high-order random walks[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(13): 14193-14201.
[23] Tao Z, Liu H, Li S, et al. Marginalized multiview ensemble clustering[J]. IEEE transactions on neural networks and learning systems, 2019, 31(2): 600-611.
[24] Zhao M, Yang W, Nie F. Deep multi-view spectral clustering via ensemble[J]. Pattern Recognition, 2023, 144: 109836.
[25] Hao Z, Lu Z, Li G, et al. Ensemble clustering with attentional representation[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(2): 581-593.
[26] Zhao M, Yang W, Nie F. MVCformer: a transformer-based multi-view clustering method[J]. Information Sciences, 2023, 649: 119622.