International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Analysis of Single-View and Multi-view K-Means Clustering on Big Data Environment

Dr. Satish S. Banait[1], Prof. Namrata D. Ghuse[2,] Dr. Dipak D. Bage[3], Prof. Sonali N. Jadhav[4], Prof. Avinash A. Somatkar[5], Prof. Vinod B. Bhamare[6]

Submitted: 10/12/2023 **Revised**: 14/01/2024 **Accepted**: 28/01/2024

Abstract: Due to the revolutionary advancements in the signal sensing devices and its availability to civilians, the real time datasets are now having multiple views. Thus such a multi-view datasets are quite common in era of big data domain. As against learning of single-view, learning of multi-view has plenty of benefits. Clustering has been very useful technique in the machine learning and data mining. Traditional clustering techniques use only single set of features of the available dataset. However for the multi-view dataset with multiple features, how to ensemble all of these data views is a major concern. Thus problem is termed as multi-view clustering problem. The key benefits of multi-view clustering against single-view clustering are accurate description of data, reducing noises of data, and wider range of applications. This research works highlight the impact multi-view K-means clustering available in mvlearn python package with the traditional K-means clustering technique. To assess the impact of simple K-means technique and multi-view version of K-means technique, two datasets are utilized namely, nutrimouse and simulated dataset. In order to analyze the impact of multi-view clustering on clustering quality, traditional k-means technique is applied to individual views, concatenated view of the both the datasets, followed by the application of multi-view version of K-means technique on the both the datasets. We analyzed the clustering quality of multi-view K-means technique using various performance evaluation parameters such as Jaccard Coefficient (Jacc), Fowlkes Mallows Index (FM), Normalized Mutual Information (NMI), Rand Index (RI), and clustering execution times.

Keywords: *Multi-view dataset, Multi-view clustering techniques, K-means, Jaccard Coefficient (Jacc), Fowlkes Mallows Index (FM), Normalized Mutual Information (NMI), Rand Index (RI)*

1. Introduction

Multi-view datasets are widely used in various real-time data mining and clustering applications as a result of advancements in micro electro mechanical systems (MEMS). One significant class of unsupervised learning techniques is clustering. It has been used very well for market analysis, social network analysis, gene expression analysis, and heterogeneous data analysis [1]-[3]. The fundamental goal of clustering is to divide (partition) the provided dataset into numerous sub-clusters so that the data elements in one cluster have more properties in common with one another than with those in other sub-clusters. Yet, single-view data is a good fit for the currently used clustering approaches. Huge amounts of data are produced from various sources for an underlying application as a result of the rapid advancement of the Internet and computer devices. It is necessary to fully utilize the information included in numerous sources since the data connected with each of these sources includes important information, which in turn imposes the requirement for mining the intrinsically valuable hidden patterns in the data [4]. This process is termed as multi-view learning. Each data view often corresponds to a significant source of useful information. For instance, web pages may be taken into account concurrently by both page hyperlink information and the page's contents (one view) (another view).

Data clustering can greatly benefit from integrating all the information present in various data representations. Using all of this information from all data views is as simple as joining the data characteristics from each view together and using a suitable single-view clustering method. Nevertheless, this method typically fails to identify the information included in the links between various data views [5]. In other words, both critical and less important data views are handled equally in the concatenation let single-view clustering approach. As a result, the final clustering performance would deteriorate. The ideal approach to make better use of the multi-view information is to simultaneously cluster each view of the features of the data and integrate the results based on how relevant they are to the clustering process.

Department of Computer Engineering, K.K. Wagh IEER, Nashik, SPPU Pune, India[1, 2, 4]
ssbanait@kkwagh.edu.in[1,] ndghusessbanait@kkwagh.edu.in[2],
sn-jadhav@kkwagh.edu.in [4]
Department of Information Technology, Sandip Institute of Technology & research Centre, Nashik[3].
Dipak.bage@sitrc.org[3]
Vishwakarma Institute of Information Technology, Pune[5]
avinash.somatkar@viit.ac.in[5]
Department of Computer Engineering, Sandip Institute of Technology & research Centre, Nashik[6].
Vinod.bhamare@sitrc.org [6]

The research community has been obliged to use multi-view learning of multi-view data due to the availability of such multi-view data, particularly in the setting of unsupervised learning [6]. The conventional single-view clustering algorithms, however, are unable to utilise the multi-data in an unsupervised learning setup for multi-view clustering. In unsupervised clustering, simply merging the features from all the data views into a single data union and then using the single-view clustering technique is frequently ineffective. The strategy based on multi-view clustering is required to address the issue related to the clustering of the multi-view dataset.

There are a number of classic single-view clustering algorithms that have been proposed in the literature. Some examples of these techniques include K-means [6, 7], DBSCAN [8, Fuzzy C-means (FCM) [9, 10], Spectral Clustering [11, 12], and many others. In order to place a greater emphasis on multi-view data clustering, we have made use of the multi-view form of the well-known K-means approach that is included in the mvlearn Python package. The difficulty of multi-view data clustering is attempted to be addressed by this research through the utilization of a K-means-based technique. Listed below are the most important contributions that this research work has made.

1. 1. In this investigation, we utilized the Python packages mvlearn and sklearn, respectively, in order to develop both the conventional (single-view) K-means clustering algorithm as well as the multi-view K-means clustering algorithm.
2. Phase I, which was titled "Multi-view Clustering of Nutrimouse Dataset Using K-means Method," and Phase II were the two study phases that we utilized in order to test the effectiveness of the multi-view K-means algorithm, which was referred to as "Multi-view Clustering of Simulated multi-view Dataset via K-means technique."
3. We used a number of different performance evaluation metrics to investigate the clustering performance of the multi-view K-means technique. These measures included the Normalized Mutual Information (NMI), the Jaccard Coefficient, the Fowlkes Mallows Index, and the Rand Index (RI). A comparison was also made between the speeds at which single-view and multi-view K-means clustering were carried out. This is demonstrated by the values of these performance assessment parameters for clustering quality, which indicate that the multi-view version of the K-means algorithm has a higher clustering quality than its single-view equivalent. Additionally, the execution time for the multi-view K-means algorithm is significantly less than that of its single-view equivalent in both Phase I and Phase II.

2. Related Work

Within the realm of clustering algorithms, single-view data has traditionally been the primary focus. When dealing with multi-view data, traditional clustering algorithms frequently take into account each data perspective separately. After that, conventional clustering methods apply a straightforward ensemble-based integration mechanism in order to obtain the final clustering results. In order to make use of all of the data that is accessible, all of the data perspectives are incorporated. [2]–[5] After that, an appropriate traditional single-view clustering algorithm is taken into consideration. On the other hand, this method typically fails to identify the data that can be found in the relationships that exist between the various data representations. To put it another way, the concatenation-let single-view clustering approach approaches both critical and less relevant data views in an equal manner. Due to this, the overall performance of the clustering would be negatively affected. Multi-view learning technology is necessary to tackle this problem. When compared to more traditional single-view clustering algorithms, multi-view clustering techniques are superior in terms of performance since they utilize data from several perspectives. A consequence of this is that the prominence of multi-view learning has increased within the community of machine learning. There is a revolutionary clustering method that was published in [8]. This method makes use of an approach that minimizes the number of dimensions available. We made use of real-time streaming data that was frequently unstructured and occasionally noisy while we were in the process of putting our ideas into action. In order to improve the accuracy of clustering while simultaneously minimizing the amount of time necessary to form effective clusters on massive volumes of unstructured data, hybrid clustering algorithms have been proposed. These algorithms try to achieve both those goals simultaneously.

In response to the multi-view clustering problem, members of the scientific community have provided a large variety of imaginative algorithmic solutions. These methods have been proposed by scientific community members. Developed in [13], the TW-k-means algorithm is a method for performing automatic weighted clustering using two levels of input variables. The K-means method serves as the foundation for this methodology. An innovative multi-view K-means clustering algorithm was created [14] in order to address the problem of efficiently grouping enormous volumes of multi-view data. This decision was made in order to solve the problem. Furthermore, it is immune to outliers and has the potential to learn the relative value of each viewpoint over time. This skill is a significant competitive advantage. The fuzzy clustering means (FCM) methodology has been utilized in the development of a number of multi-view clustering algorithms that have just been introduced. A number of presentations have been made regarding these methods. The Co-FC algorithm, which is a collaborative clustering algorithm, was developed by incorporating a collaborative mechanism into the traditional FCM approach [15]. This allowed for the development of the algorithm. In order to construct a multi-view fuzzy clustering methodology that was referred to as Co-FKM, the FCM method was utilized. The very first presentation of this method was made in [16]. It was possible for this method to lessen the amount of disagreement that occurred between the partitions with relation to the various perspectives. This was accomplished by integrating a penalty component into the objective function. In order to build the Co-FCM approach, which is a multi-view fuzzy clustering strategy, the conventional FCM technique was utilized through the development process. Furthermore, this approach was reported in the year [17]. As a result of the fact that it assigned various weights to a variety of different perspectives, it was eventually developed into the multi-view weighted collaborative fuzzy clustering method, which is abbreviated as WV-Co-FCM.

A correlational spectral clustering approach that is based on kernel conventional correlation estimation has been suggested as a potential solution to the challenge of grouping large-dimensional data in [18]. This technique was first presented as a potential response to the problem. The purpose of developing this approach was to find a solution to the challenge that was being faced. Initially, the multi-view data from a variety of distinct feature spaces are projected onto a low-dimensional domain that is shared by all of the participants. This technique is utilized in order to do this. Once it was complete, the data was clustered in the low-dimensional space by employing K-means or another clustering technique that was comparable. A brand-new multi-view clustering model has been developed by the authors of [19], who have used canonical correlation analysis as the basis for their work. After the multi-view data have been projected onto a shared low-dimensional subspace through the use of canonical correlation analysis, the algorithm then groups the information with low dimensions that was generated through the use of K-means or another clustering algorithm. The data are grouped after this step has been completed. For the purpose of clustering multiple views of data, academics have developed a few multi-view clustering algorithms. Additionally, non-negative matrix factorization technology has been utilized for the purpose of clustering multiple views of data. In order to turn the coefficient matrix from each perspective into a shared consistent matrix, a joint non-negative matrix factorization methodology was applied with the goal of achieving this transformation. Making use of the method allowed for the successful completion of this task. Following that, this common consistent matrix was utilized as a potential depiction of the initial data when it was used. The creation of the multi-view clustering technique that was proposed in [20] was made possible through the utilization of this technology. After that, the K-means method and several other clustering algorithms were immediately applied to the consistent matrix, and the clustering procedure was immediately started.

This foundational study addresses the well-known K-means clustering algorithm and emphasizes the significance of proper initialization, also known as seeding, of cluster centroids. It presents the K-means++ seeding approach, which offers an important advantage over random initialization by picking starting centroids that are more likely to be indicative of the information underneath the distribution. This is accomplished by the selection of initial centroids. The purpose of this study is to give theoretical insights and actual results that demonstrate the effectiveness of K-means++ in achieving superior clustering outcomes with faster convergence.

In this study, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) technique is presented. This approach is particularly useful for clustering data with various shapes and for dealing with noise. Clusters are clusters that are identified by DBSCAN as regions of high density that are divided by regions of low density. The user is not required to define the number of clusters in advance with this method. Additionally, the method is able to identify clusters of diverse sizes and forms, and it is resistant to outliers. In addition to providing a comprehensive explanation of the algorithm, the paper also includes experimental evaluations that demonstrate the functionality of the approach in a variety of applications.

The spectral clustering algorithm is presented in this study. It is a graph-based clustering algorithm that makes use of the eigenstructure of the similarity matrix that represents the data. A lower-dimensional space is created by translating the data into a lower-dimensional space using the eigenvectors that correspond to the smallest eigenvalues of the Laplacian matrix that is produced from the similarity graph. This is how spectral clustering takes place. After that, it uses conventional clustering methods, such as K-means, to divide the data that has been processed into clusters. In this study, the theoretical underpinnings of spectral clustering are discussed, along with its several advantages over more conventional approaches, particularly with regard to datasets that contain intricate geometric structures.

The authors of [21] presented the multi-view kernel k-means (MVKKM) technique, which integrated the kernels derived from the weighted views. This was accomplished by assigning a weight to each view that was determined by the degree to which it contributed to the clustering outcome. On the other hand, it does not own a particular process for feature selection; rather, it is dependent on the inner products kernel for all views. In order to address the problems that have been identified, additional study is being conducted on the topic of feature selection in multi-view data clustering. [22] presents a framework that constructs models for both feature selection and multi-source learning. This framework was originally proposed. Due to the fact that it is designed for supervised learning, this technique, on the other hand, is incapable of dealing with an unsupervised environment.

3. Proposed K-means Based Multi-View Clustering Approach

One of the methods for clustering single-view datasets that is frequently employed [6], [7], [21] is the classic K-means method. Due to its ease of use, it offers enormous potential for handling massive datasets. It has been effectively used in a variety of applications, including social network analysis, computer vision, and market segmentation. Let the dataset contains N samples, then corresponding matrix can be represented as $X = [x_1, x_2, ...x_N]$. Taking Euclidean distance as the similarity measure, data samples are clustered into $C(2 \leq C \leq N)$ clusters. The cluster centers can be presented by matrix $Z = [z_1, z_2, ...z_N]$.The objective function of the K-means algorithm is defined as

$$P(U,Z) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{i,j} \left\| x_j - z_i \right\|^2 \quad (1)$$

In order to determine the degree of similarity between the data samples, the Euclidean distance is utilized in Equation (1), as can be seen. There are a great number of data structures and data distributions that are present in the tangible world. Therefore, it is not always appropriate to apply this fundamental K-means technique in order to precisely locate the patterns that are concealed within datasets. Moreover, some datasets might not be able to be separated in a low-dimensional space. Thanks to the built-in K-means function of the mvlearn Python package for clustering multi-view datasets.

Real-world data frequently include multi-view data, which are represented by many views of different attributes for each sample. Related techniques have also gained prominence The Python package mvlearn is used to implement popular multi-view machine learning techniques. Its straightforward API closely resembles Scikit-for Learn's better usability. The package is distributed under the Apache 2.0 open-source licence and can be installed from PyPI or the conda package manager. By creating several views from a single initial data matrix using mvlearn, the use-cases for multi-view methods can be increased. This could produce better results with this data than typical single-view approaches [23]. The experimentation in this research work is split in two phases namely, Phase I and Phase II. The Phase II experimentation is further divided into two cases. Single-view and multi-view K-means technique is applied on simulated dataset with high separation and high overlapping in View 1 and View 2 in Case A, and case B respectively.

Phase I: Multi-view Clustering of Nutrimouse Dataset via K-means technique

Phase II: Multi-view Clustering of Simulated multi-view Dataset via K-means technique

A. *Performance when cluster components in both views are well separated,*

B. *Performance when cluster components in both views are highly overlapping.*

In this study, we examined two multi-view datasets—a simulated dataset [23] and nutrimouse [24]—to determine the effectiveness of the multi-view K-means approach. The mouse nutrition study is where the nutrimouse dataset is from. Pascal Martin of the Toxicology and Pharmacology Laboratory proposed it (French National Institute for Agronomic Research). It contains the following components:

- gene: data frame (40 * 120) with numerical variables
- lipid: data frame (40 * 21) with numerical variables
- diet: factor vector (40)
- genotype: factor vector (40)

In order to evaluate the clustering quality of multi-view K-means technique, various performance evaluation parameters such as Normalized Mutual Information (NMI), Jaccard Coefficient (Jacc), Fowlkes

Mallows Index (FM), Rand Index (RI), and clustering execution times, are used. The details of these metrics are given below.

1. *Normalized mutual information (NMI)* [1]-[3]:

When we are given the cluster labels, NMI provides us with the reduction in entropy of the class labels. Since we know the cluster labels, NMI sort of informs us how much the ambiguity about class labels diminishes. That is comparable to how decision trees gain knowledge. It is mathematically given by Equation (2).

$$NMI(Y,C) = \frac{2 \times I(Y;C)}{[H(Y) + H(C)]} \qquad (2)$$

Where, Y- Class labels, C- Cluster Labels, H(.)- Entropy, I(Y;C)- Mutual Information between Y and C.

2. *Jaccard Coefficient (Jacc)* [1]-[3]:

If the actual labeling details of a dataset are known, it is possible to evaluate the quality of the approach to clustering that was used by determining the difference between the real labels and the predicted labels corresponding to the dataset. This is a method that may be used to evaluate the quality of the clustering methodology. This analysis can be done in order to determine how well the technique performed. The Jacc and FM measurements are also helpful quality measurements to consider in the context of this conversation. Therefore, Jacc and FM are two approaches that can be applied to evaluate the quality of the feature subset that was produced. Both of these methods are available for use. Jacc and FM both have a range that goes from 0 to 1, with 1 indicating that there is no overlapping and 0 indicating that there is total overlap. Both variables have a range that goes from 0 to 1. The quality of the clustering technique is improved as a result of this, and the value of these two coefficients is increased when the value of the clustering technique is increased.

The Jacc is a statistical approach for determining the similarity of two sets. It is employed to contrast the corresponding set of actual labels with the corresponding set of expected labels for a sample. It is calculated by dividing the intersection's size by the union's size of the two label sets. Let K = $K_1, K_{2,..,}$ K_m and P=$P_1, P_{2,..,}$ P_n be two clustering result set, then Jacc index can be computed using Equation (3).

$$Jacc = \frac{a}{(a+b+c)} \qquad (3)$$

Where,

a- The number of point pairings that belong to the same cluster set can be determined by comparing the clustering results K and P.

b- The number of point pairings that appear in K but do not appear in P and belong to the same cluster set

c- In K, the number of point pairings that belong to various cluster sets, but in P, the number is the same.

3. *Fowlkes Mallows Index (FM)* [1]-[3]:

The method of cluster evaluation is specifically referred to as this approach. Through the utilization of this method, it is possible to mathematically ascertain the degree of similarity that exists between two clustering outcomes. In accordance with the FM index value, the degree of similarity that exists between the clusters continues to increase as the value continues to rise. The range of the score is from 0 to 1, or between the two. The fact that the number is large indicates that there is a degree of resemblance between the two groups of observations. Let K = K1, K2,.., Km and P = P1, P2,.., Pn be two clustering results then FM index can be given by Equation (4).

$$FM = \frac{a}{\sqrt{(a+b)(a+c)}} \qquad (4)$$

Where,

a- Number of point pairs belonging to same cluster set of two clustering results K and P

b- Number of point pairs belonging to same cluster set in K but not in P

c- Number of point pairs belonging to different cluster set in K but same in P.

4. *Rand Index (Adjusted Rand Score- RI)* [1]-[3]:

A similarity measure between two clusterings is calculated using the Rand Index, which takes into account all pairs of samples and counts the number of pairings that are assigned to the same or different clusters in both the expected and the true clusterings. This is done in order to determine the true and predicted clusterings. Let K = $K_1, K_{2,..,}$ K_m and P=$P_1, P_{2,..,}$ P_n be two clustering result set, then Rand Index (RI) can be computed using Equation (5).

$$RI = \frac{a+b}{(a+b+c+d)} \qquad (5)$$

Where,

a- The number of point pairings that belong to the same cluster set can be determined by comparing the clustering results K and P.

b- The number of point pairings in K and P that do not belong to the same cluster set as the other pair

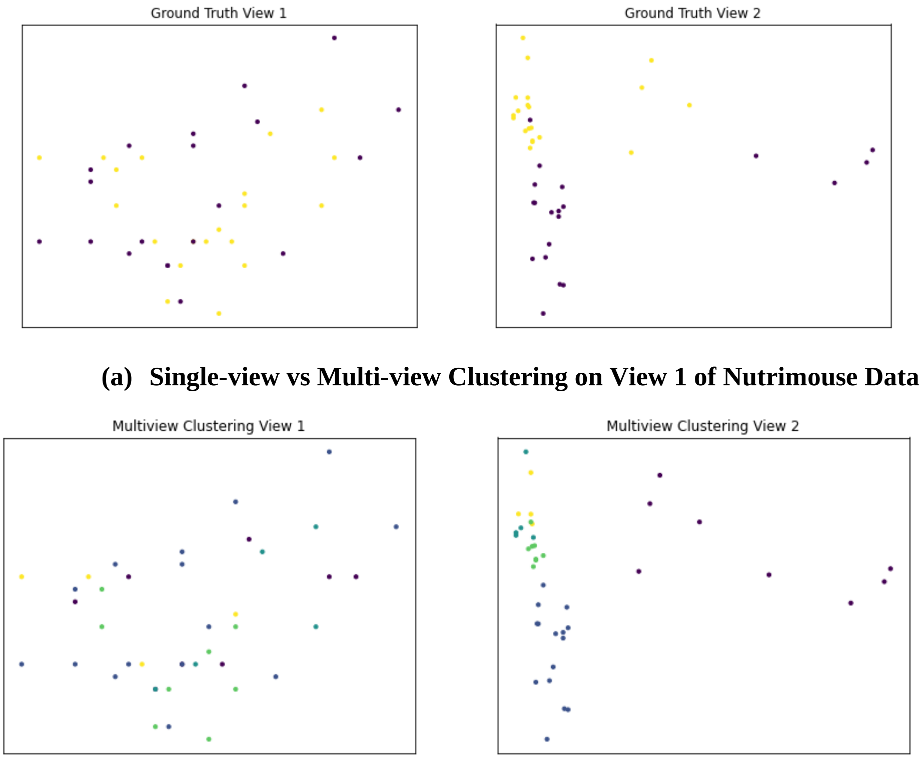
c- In K, the number of point pairs that belong to the same cluster set, whereas in P, they belong to a separate cluster.

d- The number of point pairings that belong to separate cluster sets in K but belong to the same cluster in P.

4. Discussion on Results:

Phase I: Multi-view Clustering of Nutrimouse Dataset via K-means technique:

In Phase I, the traditional k-means technique is applied to individual views, concatenated views of both datasets, followed by the application of the multi-view version of K-means technique on the multi-view Nutrimouse dataset to examine the influence of multi-view clustering on clustering quality. The clustering quality of multi-view K-means technique was examined utilising several performance assessment criteria such as Jacc, FM, NMI, RI, and clustering execution times. Fig 1 depicts the clustering findings on the multi-view Nutrimouse dataset, where we can view that the clustering quality with the multi-view strategy is somewhat better than the single-view counterpart. Table 1 values of clustering performance evaluation parameters provide a clearer view of clustering quality. We can see that the NMI, Jacc, FM, and RI values produced for multi-view clustering are higher than those obtained for single-view clustering. Although the improvement is marginal, the execution time required to run the multi-view K-means clustering method is significantly shorter than that of the single-view version.



(a) Single-view vs Multi-view Clustering on View 1 of Nutrimouse Dataset

(b) Single-view vs Multi-view Clustering on View 2 of Nutrimouse Dataset

Fig. 1 K-means technique based Clustering on Nutrimouse Dataset (Phase **I)**

Table 1 Comparison of Clustering Quality of Single-view and Multi-view K-means based approaches for Nutrimouse Dataset for Phase I

Clustering Approach	NMI	Jacc	FM	RI	Execution Time(in sec)
Single-view K-means for View 1 of Dataset	0.547	0.075	0.664	0.446	0.38

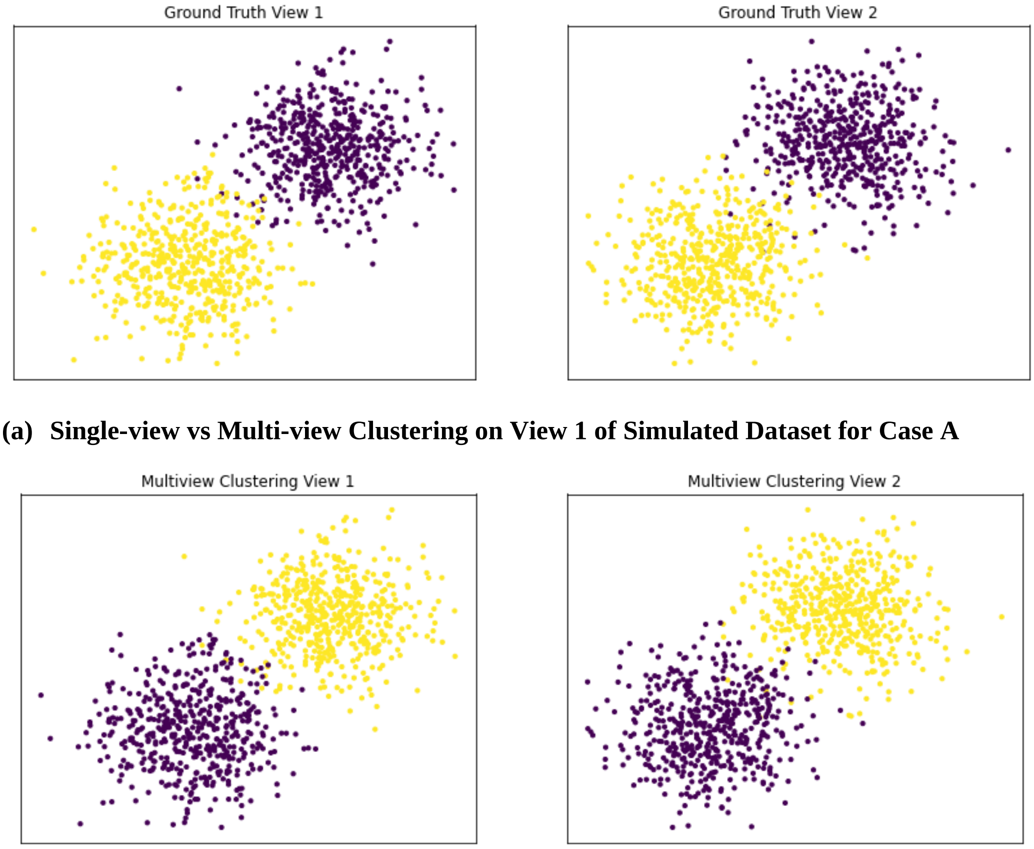
Single-view K-means for View 2 of Dataset	0.422	0.225	0.535	0.284	0.36
Single-view K-means for Concatenated Dataset	0.422	0.225	0.535	0.284	0.38
Multi-view K-means for whole dataset	0.448	0.423	0.605	0.468	0.23

Phase II: Multi-view Clustering of Simulated multi-view Dataset via K-means technique

During this phase, we use K-means algorithms with single-view and multi-view viewpoints on simulated datasets with multiple viewpoints. Within the scope of this investigation, two experiments are conducted, which are marked by the letters (A) and (B) below. In each of these studies, we carry out K-means clustering with a single view as well as that with multiple views. In order to evaluate the performance of a single view, we run the algorithm on each view separately as well as on all views collectively. Every single test involves putting every single method through a hundred random cluster initializations. Detailed information regarding the clustering results can be found below.

A. *Performance when cluster components in both views are well separated:*

We can observe that for concatenated views, multi-view K-means clustering performs nearly as well as single-view K-means clustering, and both perform better than single-view clustering for only one view. The clustering results for Case A are presented and described in Fig. 2 and Table 2.



(a) Single-view vs Multi-view Clustering on View 1 of Simulated Dataset for Case A

(b) Single-view vs Multi-view Clustering on View 2 of Simulated Dataset for Case A

Fig. 2 K-means technique based Clustering on Simulated Dataset for Phase II
(Case A: When cluster components in both views are well separated)

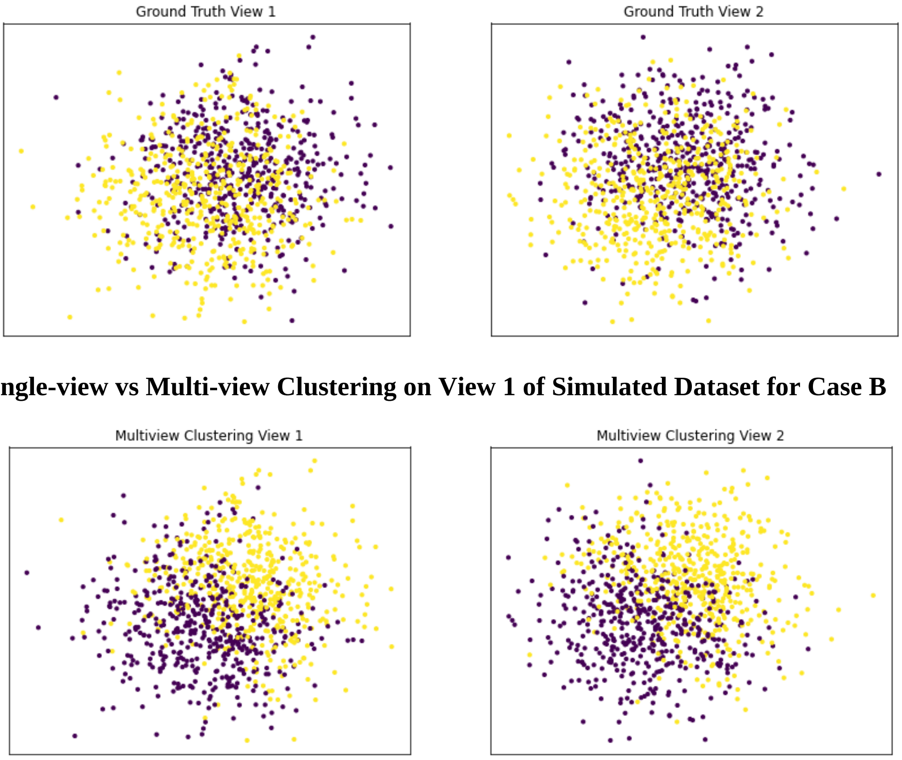
Table 2 Comparison of Clustering Quality of Single-view and Multi-view K-means based approaches for Simulated Dataset for Phase II, Case A

Clustering Approach	NMI	Jacc	FM	RI	Execution Time (in sec)
Single-view K-means for View 1 of Dataset	0.901	0.987	0.974	0.994	0.72

Single-view K-means for View 2 of Dataset	0.888	0.985	0.970	0.941	0.74
Single-view K-means for Concatenated Dataset	0.99	0.99	0.998	0.996	0.75
Multi-view K-means for whole dataset	0.99	0.993	0.999	0.997	0.275

B. Performance when cluster components are relatively inseparable (highly overlapping) in both views:

We can observe that multi-view K-means clustering performs about as poorly as single-view K-means clustering. As inputs, K-means clustering is applied to both individual and concatenated views. The clustering results for Case B are presented and described in Fig. 3 and Table 3.



(a) Single-view vs Multi-view Clustering on View 1 of Simulated Dataset for Case B

(b) Single-view vs Multi-view Clustering on View 2 of Simulated Dataset for Case B

Fig. 3 K-means technique based Clustering on Simulated Dataset for Phase II
(Case B: When cluster components are relatively overlapping)

Table 3 Comparison of Clustering Quality of Single-view and Multi-view K-means based approaches for Simulated Dataset for Phase II, Case B

Clustering Approach	**NMI**	**Jacc**	**FM**	**RI**	**Execution Time (in sec)**
Single-view K-means for View 1 of Dataset	0.062	0.445	0.541	0.083	0.72
Single-view K-means for View 2 of Dataset	0.044	0.378	0.530	0.059	0.74
Single-view K-means for Concatenated Dataset	0.098	0.318	0.566	0.132	0.75
Multi-view K-means for whole dataset	0.109	0.508	0.573	0.147	0.275

Consequently, based on the values of the clustering quality performance assessment metrics NMI, Jacc, FM, and RI acquired in Phase I and Phase II tests, it is demonstrated that the multi-view version of the K-means algorithm has higher clustering quality than its single-view counterpart. Furthermore, in both Phase I and Phase II, the execution time for the multi-view K-means algorithm is about half that of its single-view equivalent.

5. Conclusion

Clustering has shown to be an extremely valuable technique in the machine learning and data mining. Conventional clustering approaches employ only a subset of the available dataset's features. Nevertheless, how to assemble all of these data perspectives in a multi-view dataset with various features is a huge challenge. As a consequence of this, the issue is commonly referred to as the multi-view clustering problem. Multi-view clustering has a number of advantages over single-view clustering, the most important of which are an accurate representation of the data, a reduction in the amount of noise in the data, and a wider range of applications. The purpose of this work is to investigate the impact that the multi-view K-means clustering methodology, which is included in the mvlearn Python package, has on the traditional K-means clustering technique. In order to analyze the influence of the regular K-means technique and the multi-view form of the K-means technique, two datasets, namely nutrimouse and simulated, are utilized.

We evaluated the usefulness of the multi-view K-means method in two phases: Phase I (Multi-view Clustering of Nutrimouse Dataset Using K-means Technique) and Phase II (Multi-view Clustering of Nutrimouse Dataset Using K-means Technique) (Multi-view Clustering of Simulated multi-view Dataset via K-means technique). The multi-view K-means algorithm was tested for its clustering quality using a number of different performance measures. These measurements included Normalized Mutual Information (NMI), Jaccard Coefficient (Jacc), Fowlkes Mallows Index (FM), and Rand Index (RI). In addition, the execution durations of single-view K-means clustering and multi-view K-means clustering were investigated. It has been proved, on the basis of the outcomes of these clustering performance and quality measurement parameters, that the multi-view version of the K-means method has a higher clustering quality than its counterpart that only uses a single view. Furthermore, in both Phases I and Phase II, the execution time for the multi-view K-means algorithm is quite short when compared to its single-view equivalent.

References

[1] Xu, D. C. Tao, and C. Xu, “A survey on multi-view learning”, arXiv preprint arXiv: 1304.5634, 2013.

[2] C. Aggarwal and C. K. Reddy, “Data Clustering: Algorithms and Applications”, Boca Raton, FL, USA, Chapman and Hall/CRC, 2013.

[3] S. L. Sun, “A survey of multi-view machine learning”, Neural Comput. Appl., vol. 23, nos. 7&8, pp. 2031–2038, 2013.

[4] R. Xu and D. Wunsch, “Survey of clustering algorithms”, IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, 2005.

[5] Zhao J, Xie X, Xin X, Sun S, “Multi-view learning overview: Recent progress and new challenges”, pp. 43-54, Information Fusion. 2017.

[6] J. A. Hartigan, “A K-Means Clustering Algorithm,” Appl Stat, vol. 28, no. 1, pp. 100-108, 1979.

[7] L. Jing, M. K. Ng, and J. Z. Huang, “An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data,” IEEE Transactions on Knowledge & Data Engineering, vol. 19, no. 8, pp. 1026-1041, 2007.

[8] Satish S Banait, Sane S. S., Talekar S. A. - : An Efficient Clustering Technique for Mining Big Data” International Journal of Next-Generation Computing, 2022, Vol 13, Issue 3, p702-715.

[9] L. Zhu, F. L. Chung, and S. Wang, “Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions,” IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics, vol. 39, no. 3, pp. 578-591, 2009.

[10] L. O. Hall and D. B. Goldgof, “Convergence of the Single-Pass and Online Fuzzy C-Means Algorithms,” IEEE Transactions on Fuzzy Systems, vol. 19, no. 4, pp. 792-794, 2011.

[11] K. Kamvar, S. Sepandar, K. Klein, D. Dan, M. Manning, & C. Christopher, “Spectral learning,” In International Joint Conference of Artificial Intelligence Stanford InfoLab, 2003.

A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” In Advances in neural information processing systems, vol. 2, pp. 849–856, 2002.

[12] X. Chen, X. Xu, J. Z. Huang, and Y. Ye, “TW-k-means: Automated two-level variable weighting clustering algorithm for multiview data,” IEEE Transactions on Knowledge & Data Engineering, vol. 25, no. 4, pp. 932-944, 2013.

[13] X. Cai, F. Nie, and H. Huang, “Multi-view k-means clustering on big data,” In Twenty-Third International Joint conference on artificial intelligence, 2013.

[14] W. Pedrycz, “Collaborative fuzzy clustering,” Pattern Recognition Letters, vol. 23, no. 14, pp. 1675-1686, 2002.

[15] G. Cleuziou, M. Exbrayat, L. Martin, and J. H. Sublemontier, “CoFKM: A Centralized Method for Multiple-View Clustering”, pp. 752-757, Proceedings of 9th IEEE International Conference on Data Mining, 2009.

[16] Y. Jiang, F. L. Chung, S. Wang, Z. Deng, J. Wang, and P. Qian, “Collaborative fuzzy clustering from multiple weighted views”, vol. 45, no. 4, pp. 688-701, IEEE Transactions on Cybernetics, 2015.

[17] M.B. Blaschko and C.H. Lampert, “Correlational spectral clustering”, pp. 1-8, Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[18] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, “Multi-view clustering via canonical correlation analysis,” in Proceedings of the 26th annual international conference on machine learning, pp. 129-136, 2009.

[19] J. Liu, C. Wang, J. Gao, and J. Han, “Multi-view clustering via joint nonnegative matrix factorization”, pp. 252-260, Proceedings of the 2013 SIAM International Conference on Data Mining, 2013.

[20] G. Tzortzis, A. Likas, “Kernel-based weighted multi-view clustering” , pp. 675– 684, Proceedings of the 12th International Conference on Data Mining, 2012.

[21] S. Xiang, L. Yuan, W. Fan, Y. Wang, P.M. Thompson, J. Ye, “Multi-source learning with block-wise missing data for Alzheimer's disease prediction”, pp. 185–193, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013.

[22] https://mvlearn.github.io/auto_examples/cluster/plot_mv_kmeans_validation_simulated.html.

[23] P. Martin, H. Guillou, F. Lasserre, S. Déjean, A. Lan, J-M. Pascussi, M. San Cristobal, P. Legrand, P. Besse, T. Pineau - Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. Hepatology, in press, 2007.

[24] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “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, 1996, pp. 226–231.