Journal of Theoretical and Applied Information Technology
15th March 2025. Vol.103. No.5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MODIFIED K-MEANS APPROACH FOR EFFECTIVE CLUSTERING USING WEIGHTED ADJACENT MATRIX

HEMA BHARDWAJ[1], DR. D. SRINIVASA RAO[2]

[1]Scholar, Department of Computer Application, Medi-Caps University, Indore, India

[2]Associate Professor, Dept. of Computer Science & Engineering, Medi-Caps University, Indore, India
E-mail: [1]hema.bhardwajj@gmail.com, [2]dsrinivasa.rao@medicaps.ac.in

ABSTRACT

K-means clustering has several limitations, such as sensitivity to initialization and determining the number of clusters. It is sensitive to outliers, especially when identifying clusters with irregular shapes or varying sizes. Handling categorical data directly in k-means can be challenging. This study aims to present methods to improve the existing k-means clustering algorithms. It proposes designing two distinct proximity matrices for this purpose. The study suggests that the new algorithm performs better than traditional clustering methods based on several evaluation metrics. Randomly chosen centroids lead to unstable outcomes. The unpredictable initialization of centroids makes it difficult to replicate clustering results. Spectral clustering begins by creating a similarity matrix, followed by eigenvalue decomposition applied to the Laplacian matrix. This decomposition results in a spectral representation. However, optimal clustering outcomes cannot be guaranteed in the initial stage of the spectral clustering algorithm. This research proposes a solution to this issue. An Initialization & Similarity approach is recommended, where both the representation and the similarity matrix are determined in a cohesive manner. Additionally, it improves clustering performance by using sum of norms regularization. Based on evaluation metrics, this clustering technique proves to be better than the original k-means algorithm. Using normalized mutual information, purity, and accuracy as measures, the proposed technique demonstrates superiority over traditional algorithms. This study presents a novel approach to K-Means clustering by integrating a weighted adjacent matrix, significantly enhancing clustering accuracy and effectively handling high-dimensional data. The proposed methods, KM-AM and KM-WAM, show improved performance metrics such as normalized mutual information, accuracy, and purity, offering a more efficient and robust solution for various data analysis applications.

Keywords: *K-means clustering, similarity matrix, spectral clustering, laplacian matrix.*

1. INTRODUCTION

The k-means technique is popular and widely used in data clustering. It falls under unsupervised machine learning. In this method, a given dataset is divided into a specified number of clusters, denoted by k, which is chosen by the user. A centroid is initially designated as the cluster centre, achieved by randomly selecting k points from the dataset. In an iterative process, each data element is assigned to a group with the nearest centroid. Following this, the cluster centre, or centroid, is recalculated iteratively by averaging all the data elements within the cluster. The k-means algorithm aims to minimize inertia, which is the sum of squared distances of points within a cluster. This method groups similar data points within a cluster. K-means is frequently used for customer segmentation, image compression, anomaly detection, and document clustering. However, this simple and efficient algorithm may struggle with issues such as non-linearly separable or overlapping clusters.

The goal of clustering in unsupervised data mining algorithms is to design and partition data points into groups where all points share certain similarities. Similar data points are placed in one group, while dissimilar points are placed in another cluster. Clustering is an unsupervised learning algorithm that analyses data without labelled datasets, in contrast to supervised learning. Therefore, labelled data is not required for k-means to function during training. It examines the data's underlying structure using only the input features. K-means divides the data into a specified number of clusters, denoted by "K," a user-defined quantity. Each data point is a member of the cluster whose mean is closest to it. Here, the cluster centre is referred to as the "centroid."

The concept of a centroid describes the central values within a cluster, derived from all the data values in that cluster. The algorithm iteratively updates the centroids of the clusters until convergence is achieved. It begins with an initial random selection of centroids, denoted by K, and aims to minimize the sum of squared distances between the data elements and the centroid of the cluster. This algorithm can efficiently handle large datasets due to its low computational complexity, making it widely used in practice. It is also simple to apply, requiring minimal effort.

The k-means algorithm involves initializing the centroids of each cluster. All data points are assigned to the cluster with the nearest centroid, and the centroids are then iteratively updated until convergence is achieved. K-means works well with numerical or continuous data, as distances between data points are calculated based on their numerical properties. This technique groups data points into clusters with the nearest centroid, making the resulting clusters interpretable. The characteristics of these clusters can be analyzed to better understand the data. Consequently, k-means clustering has gained significant attention in research. It is also popular in the field of data mining and extensively used in machine learning due to its effectiveness.

Along with these characteristics, k-means has a few drawbacks. Initial centroids have an impact on this algorithm. Varied approaches to the first centroid placement can provide varied clustering outcomes and impact algorithm performance. Various initializations can lead to diverse outcomes, such as becoming trapped in local optima. As a result, selecting the best result from among those produced after several runs using various initializations is standard procedure.
K-means clustering assumes that all clusters have equal variance, which makes it challenging to process datasets with non-spherical or unevenly sized clusters. The algorithm is particularly sensitive to outliers, as their presence can significantly impact the centroid calculation, leading to skewed clusters and suboptimal grouping.

K-means operates under the assumption that clusters are spherical and equal in size. However, some datasets may form asymmetric or overlapping clusters, making it difficult to accurately represent the data's structure. The number of clusters (k) must be specified in advance, and determining the optimal number of clusters can be challenging if the dataset lacks a clear clustering structure. An incorrect value for k can result in either under-segmentation or over-segmentation.

The algorithm does not account for noise and extraneous features, treating each data point equally. Noisy data points can negatively impact centroid calculation, distorting the clustering outcome. Scalability is another issue with K-means, as its computational cost increases linearly with the number of data points, making it less suitable for large datasets. Additionally, K-means suffers from the "curse of dimensionality," which makes it challenging to apply to high-dimensional data, as the significance of the distance measure diminishes in such spaces.

The solutions produced by K-means clustering are not unique, as the algorithm's results depend on the initial centroids and the order of data points. By examining the similarities between data points, the algorithm's efficiency can be improved, and its limitations can be managed. Accurate assessment of similarities between data points can be achieved through similarity measurement techniques.

To achieve an accurate similarity assessment, this study aims to develop an effective similarity matrix. We propose using the similarity matrix to enhance clustering efficiency. As part of this research, two novel feature representations are independently constructed, drawing inspiration from the spectral clustering algorithm. The first representation of the initial data points is in the form of an adjacency matrix. The second representation is a weighted adjacency matrix. K-means clustering is then applied to these output representations to improve efficiency. Consequently, the study suggests a more effective method for enhancing k-means clustering results by using two different adjacency matrices.

2. OBJECTIVE OF THE RESEARCH

Conducting research on “Modified K-Means Approach for Effective Clustering Using Weighted Adjacent Matrix” addresses core concerns related to clustering accuracy, handling complex data, and optimizing computational resources, which are crucial for advancing data analysis techniques across various domains. The study is crucial for reasons as:

2.1 Improved Clustering Precision

Traditional K-Means clustering often struggle to accurately determine the optimal number

of clusters and initialize centroids. By incorporating a weighted adjacent matrix, the modified approach can enhance clustering accuracy by capturing relationships between data points more effectively.

2.2 Managing High-Dimensional Data

Modern datasets often have a large number of features making clustering challenging. The weighted adjacent matrix helps manage high-dimensional data, leading to meaningful clusters.

2.3 Performance Enhancement

Studies have shown that modified K-Means approaches can outperform classical clustering algorithms in various evaluation metrics. This can lead to more efficient data analysis and better decision-making in applications such as image segmentation, market segmentation, and anomaly detection.

2.4 Application Across Various Fields

Effective clustering is vital in numerous fields, including data mining, machine learning, and bioinformatics. Improved clustering techniques can enhance algorithm performance in these areas, leading to technological and scientific advancements.

2.5 Optimization of Computational Resources

By improving the clustering process, the modified approach can reduce computational time and resource usage, making it more feasible to handle large datasets on standard hardware.

3. LITERATURE SURVEY

Many studies have been conducted to address the challenges associated with the k-means clustering technique. Kodinariya and Makwana developed a method based on a rule of thumb, finding that the clustering result varies with the cluster parameter. Before clustering, it is necessary to determine the number of model parameters or cluster numbers. Their research examined six alternative methods for obtaining appropriate cluster numbers, addressing the primary difficulty of cluster analysis. These techniques include the rapid clustering technique with the k-means strategy used to choose the cluster number [1].

Tibshirani, Walther, and Hastie [2] introduced the gap statistic technique, which aims to optimally select the value of k. The method estimates the number of clusters in a given dataset by comparing the actual and adjusted dispersion within the cluster, using the output of clustering algorithm, such as hierarchical or k-means algorithms. Murtagh and Contreras' study [3] surveyed hierarchical clustering methods, showing that outliers impact these techniques. Their research examined hierarchical self-organizing maps using mixture models, as well as grid-based and hierarchical density-based grouping. The study concluded with a grid-based, linear-time hierarchical clustering algorithm, which is said to be highly effective.

Zheng, Zhu, et al.'s self-paced learning technique is another noteworthy contribution to the field. They suggest that outliers should be given less weight compared to important samples. The researchers adapted the feature selection process by incorporating regularization to reduce the impact of outliers. The proposed method selects a subset of significant samples necessary for building a feature selection model, enhancing its generality by incorporating more relevant samples [4]. Similarly, Abe & Abe [5] demonstrated how this approach improves the support vector machine's ability to generalize datasets through feature selection surveys and support vector machines. Their study's main objective was to create a similarity matrix that would enable the extraction of multiple features using an algorithm.

In their research, Arora and Varshney evaluated k-means and k-medoids using dispersed data points. They compared the space complexity of overlapping clusters and the time required for cluster head selection. It was determined that k-medoids are more efficient in terms of execution time and are less affected by outliers, making them better at minimizing noise. However, this method is more complex compared to k-means [6].

Bachem, Lucic, and colleagues [7] explored the impact of centroid initialization on the sensitivity of k-means. They proposed a seeding strategy for k-means, replacing the D2 sampling step with an approximation based on Markov Chain Monte Carlo sampling. Their experiments demonstrated that this technique performs well with large-scale and real-world datasets, significantly reducing runtime.

Zhang [8] developed a target-resource framework with target and cost scales, creating a cost-sensitive learning model to aid in the categorization and analysis of medical data. These approaches have been shown to be effective in decision tree learning, as demonstrated and evaluated through experiments.

Zhu, X., et al. introduced a sparse low-rank subspace as a space projection of the initial data using a transformation matrix for clustering. In subspace learning with feature selection, the affinity matrix and rank constraint are utilized. This results in the creation of a matrix of intrinsic and dynamic affinity. The clustering result is represented by the affinity matrix in the low-dimensional space [9].

DBSCAN, a density-based clustering method, can be used for clusters of any shape. Sharma, A., and Sharma, A. highlighted the limitations of this clustering method, such as its reliance on user-defined parameters. They proposed a combination technique that achieves parameter-free clustering by integrating density-based clustering with k-nearest neighbour information. The data accumulation within the cluster structure helps guide parameter setting [10].

He, L., et al. proposed an efficient method for spectral clustering in large datasets. This technique represents data in kernel space using random Fourier features. This explicit mapping method accelerates eigenvector approximation, leading to improved prediction speed for spectral clustering [11].

Chen, J., et al. [12] found that traditional data clustering methods are less effective for online data streams. To address this, they proposed a grid-based clustering algorithm that performs better with hybrid data streams. They developed a model of non-uniform attenuation to increase noise resistance and introduced a similarity calculation method to achieve accurate grouping. Experiments showed that their quick clustering centre determination procedure was highly effective.

Ding, Y., and Fu, X. proposed a kernel-based algorithm for c-means clustering, specifically for pattern recognition applications. The advantage of this method is its ability to produce high-quality modelling outcomes. The study aimed to optimize fuzzy c-means clustering through a genetic algorithm to address existing issues with FCM clustering [13].

Ferreira, M. R. P., et al. presented a kernel-based clustering technique that included automatic variable weighting and metric kernelization. They evaluated all kernels for each variable and combined into a kernel function. The dissimilarity metrics in their technique were calculated as the sum of the Euclidean distances between the centroids [14].

Du, L., et al. developed a robust k-means algorithm using the 12,1-norm in feature space, and extended it to kernel space. This study proposed a multiple-kernel k-means algorithm that proved to be robust. The primary feature of this new algorithm is its ability to simultaneously determine the best clustering label with cluster membership and the optimal arrangement of multiple kernels. They also devised an alternating iterative schema to determine the ideal value [15].

4. PROBLEM STATEMENT

While the K-Means clustering algorithm is widely used and popular, it encounters several difficulties in accurately clustering data, especially with high-dimensional datasets and determining the optimal number of clusters. Traditional K-Means often experiences issues such as inadequate initialization of centroids, convergence to local optima, and struggles with varying cluster sizes and densities. These limitations impede its effectiveness in many practical applications where precise and efficient clustering is essential.

This study tries to address these challenges by developing a modified K-Means approach that incorporates a weighted adjacent matrix to enhance clustering performance. The weighted adjacent matrix will improve the algorithm's capability to capture relationships between data points, resulting in more accurate and meaningful clustering outcomes. This study will investigate the potential advantages of this modified approach in handling high-dimensional data, improving initialization, and achieving better computational efficiency.

By tackling the core concerns related to clustering accuracy and performance, this research seeks to contribute to the advancement of clustering techniques, offering a more robust and effective solution for various data analysis and machine learning applications.

5. RESEARCH HYPOTHESIS

Despite the improvements offered by incorporating a weighted adjacent matrix for enhanced clustering accuracy, the modified K-Means approach may still face challenges related to parameter sensitivity and computational complexity. These challenges could potentially limit its applicability across diverse datasets with varying characteristics, affecting the generalizability and scalability of the proposed methods.

6. ORIGINAL K-MEANS AND THE MODIFICATIONS

This study examines the original k-means and suggests a few changes. Along with the implementation, a basic understanding of spectral clustering and k-means is required.

6.1 Clustering by Original K-means

Let the data input be $X= (x_1, ..,x_n) \in R^{n \times d}$, where, xi is ith row. X is the matrix, xi_j is component of row i and column j.
K-means algorithm:
Input of initial data elements: X= {x1, ..,xn } ∈ Rn x d;
Step i) Initialisation of centroid by random selection of k data points.
Step ii) continue
Step iii) Assigning a closest centroid to the data points.
Step iv) Calculating the mean of all data elements. Updating the centre of every cluster named as centroid.
Step v) Repeat
Step vi) Centroids stop changing and algorithm converges.

K-means clustering aims for SSE, or least sum squared errors. It also denotes the lowest overall intra-cluster variance for a specific number of clusters, k. SSE is a metric used to assess a clustering solution's quality. An alternative name for it is the "within-cluster variance." Within a cluster, the square of distances between every data element and the centroid is calculated. The total of this distance is taken out.

To compute it, we follow these steps:
a. Calculating the squared Euclidean distance for each data element and the centroid. Following formula yields the squared Euclidean distance among a centroid c and a data element x:

$$\text{Squared Euclidean Distance} = \sum_{i-1}^{n} (\mathbf{xi} - \mathbf{cj})^2$$

n is referred as the number of dimensions, xi is value of data elements in dimension i. ci is value of the centroid in dimension i.

b. Add up all of the data points' squared distances inside the cluster. This is how we get the SSE for that cluster.

c. For every cluster, repeat steps 1 and 2 in turn.

d. Add the SSE in all clusters to find the total SSE for the entire dataset.

Mathematically, the SSE can be represented as:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{tj} || x_i - c_j ||_2^2 \quad (1)$$

- k refers to the number of clusters
- t_j refers to the da ta elements within the cluster j
- $|| x_i - c_j ||_2$ refers to l_2 norm of $x_i - c_j$
- c_i is centroid for cluster i

SSE is reduced using K-means clustering. A lower SSE suggests a tighter and better clustering solution because it shows that the data points are close to their centroids. Data points are assigned to clusters iteratively by K-means. In order to reduce the SSE till convergence, it updates the cluster centroids.

7. CHALLENGES WITH K-MEANS

Clustering with minimum SSE achieves the best results when the centroids are initially chosen at random. Determining the similarity metric and predicting the actual number of clusters are significant challenges. The exact number of clusters, k, is unknown, which complicates finding an effective solution. However, there are a few methods to address this issue.

One approach is manually selecting the value of k, which represents the number of clusters, in an on-demand selection algorithm. The SSE versus k graph can be used to determine the optimal value of k using the elbow method, a refined version of the gap statistic method. The process creates the given equation to determine the k's value:

$$k \approx \sqrt{n/2} \quad (2)$$

In the k-means algorithm, identifying the initial position of the centroid poses another challenge. The simplest method is to randomly select the initial centroid, but this approach can negatively affect clustering outcomes, leading to incorrect partitions. Hierarchical centroid selection [16] addresses this issue by repeatedly running the basic k-means algorithm with random initialization. The resulting centroids are then used as input data to create the final centroids. After randomly selecting

the first centroid, referred to as k1, simple cluster seeking is used to find the next element at a predetermined distance from k1. This element is identified as k2, the second centroid. This process is repeated until k centroids are obtained. Simple cluster searching (SCS) [17] is employed within the MATLAB suite for this purpose.

Another challenge in k-means clustering is defining the similarity measure between data points. As the distance between data points increases, their common features decrease. Various similarity metrics, such as the Pearson correlation coefficient, the Jaccard coefficient, and cosine similarity, are used alongside Euclidean distance. The final issue with k-means is determining the similarity measurement between data points.

8. USE OF SPECTRAL CLUSTERING

Spectral clustering pre-processes training data points by substituting high-order relationship data points for the original training data points [18]. This graph-based method is more effective for complex data structures that require dimensionality reduction techniques. The relationships between data points are represented as a similarity matrix. In spectral clustering, Euclidean distance is used as the basis for both distance metrics and similarity measures, providing an estimate of the separation between two data points.

$$W_{i,j} = \sum_{t=1}^{d}(x\,i,t - xj,t)^2$$
$$i \,\&\, j \in [1,n],\ t \in [1,i] \qquad (3)$$

W denotes similarity matrix, i,j are the ith, jth elements and t indicates t^{th} feature of the element.

Later, the Laplacian matrix L is created by transferring the similarity grid into a sparse grid taking help of a kernel function. The similarity grid is referred to as the adjacent matrix in this study. The definition of normalized Laplacian grid L is:

$$L = (1/\sqrt{D})\ (D - A)\ (1/\sqrt{D}) \qquad (4)$$

L declares Laplacian grid, 'A' is the adjacent matrix, 'D' mentions a diagonal grid. All the rows in the adjacent matrix 'A' are summed and placed as the elements of the diagonal grid.
Dimensional reduction is thus achieved using spectral grouping. This is accomplished by first choosing k eigenvectors from L, and then using this reduced matrix to perform clustering using k-means algorithm.

9. PROPOSED K-MEANS

This study introduces two clustering techniques that address the shortcomings of the original k-means clustering. We propose two types of k-means algorithms: one based on an adjacency matrix (KM-AM) and the other on a weighted adjacency matrix (KM-WM). KM-AM directly applies the k-means algorithm to the pre-constructed adjacency matrix, while KM-WM performs k-means clustering after calculating the feature weights.

9.1 K-means Based on Adjacent Matrix (KM-AM)

Spectral clustering process begins with the construction of the similarity grid. The elements are put in an undirected graph to create the similarity matrix. Graph G = (V, E). Here E is considered as connecting edge sets that joins with the vertices. Also E= $\{e_1,e_2,e_3\ldots e_m\}$ and (m= n * (n-1)/2) and V denotes set of vertices V=$\{v_1,v_2,v_3\ldots v_n\}$. Similarity grid W represents this unidirectional graph.

$$W = (w_{i,j})\ n_{i,j} = 1 \quad \text{where } w_{i,j} \geq 0$$

According to a specified distance matrix, it provides the similarity between xi and xj. Previous studies have created adjacent matrices such as fully linked graphs, ε-neighbourhood graphs, and k-nearest neighbour graphs. If one is the closest neighbour to the other, vi and vj will be connected in the k-nearest graph. Efforts are needed to render the graph symmetric due to the non-symmetrical nature of the neighbourhood relationship, resulting in a directed graph. As a result, a completely connected graph with a similarity scalar connecting every vertex is created.

The ε-neighbourhood graph joins neighbouring vertices (εm=1) in the condition if the distance comes under or less than a given threshold ε . If not, the value is em=0. This gives the graph's edges all approximately the same value. A graph without weight is the result. In this study. It is recommended to construct an adjacency matrix to depict the distances between data points using a similarity function, in order to create a fully connected network. The Gaussian kernel, Sigmoid kernel, and polynomial kernel are the three most often used kernel functions. When using a Gaussian kernel, the following formula can be used to define the neighbouring matrix in the condition:

$A_{ij} = \exp [-2\ \sigma\ 2\ ||\ (x_i - x_j)\ ||\ 2]$ (i, j ∈ [1,n])

(5)

$A_{i,j}$ = entry in the adjacent matrix corresponding to the connection between nodes
x_i & x_j = feature vector of nodes i & j respectively
$\| x_i - x_j \|$ = feature vectors xi & xj's Euclidean distance
σ = scale parameter of Gaussian kernel

9.1.1 Algorithm for Spectral/Graph-Based Clustering

Input – a set of data elements X is taken such that X = $\{x_1, .., x_n\} \in R^{n\ x\ d}$; the number of clusters k.
Output – Centroids C as an indicator of cluster for every data element.
Step i) Calculating the affinity matrix W for X;
Step ii) Computing the similarity/affinity matrix L;
Step iii) Computing the first k eigenvectors of L, denoted as
E = {e1, ..., ek};
Step iv) Construction of the matrix U. The formula for U is obtained as ET, U ∈ Rn x d;
Step v) Running k-means algorithm technique for the obtained U;
Step vi) Getting output in terms of C which is the cluster result.

In the second stage of the spectral clustering procedure, we calculate and compute the Laplacian matrix. It outputs first k eigenvectors. These k eigenvectors serve as initial data for the k-means algorithm. Larger datasets will require more time to process due to the computational complexity. This study looks for a solution to this issue. Instead of using a Laplacian eigenvector matrix, our first technique, KM-AM, performs k-means clustering directly on the neighbouring matrix. Utilizing KM-AM has the benefit of avoiding the Laplacian matrix's computational expense. Additionally, it avoids the eigenvalue decomposition optimization cost. In the end, there is less complexity in computers. As a result, it permits clustering on bigger datasets.

9.2 K-means Based on Weighted Adjacent Matrix (KM-WAM)

Our study proposes clustering based on an adjacency matrix. However, it is important to note that a single data point can have multiple attributes or features with varying importance. Each of these features impacts the clustering result differently. Generally, a significant attribute has a greater effect on the clustering outcome than a minor feature. This variation in the relative importance of the attributes is logical [19]. Therefore, the feature with the highest weight should be prioritized when creating the adjacency matrix.

Our paper introduces an alternative k-means clustering technique that uses a weighted adjacency matrix, based on the aforementioned rationale. The weight of a feature determines its precedence. Each data element in this adjacency matrix is identified by its features, with each feature associated with a numerical value. Consequently, this study calculates the weight of all features by dividing each feature's proportion among the other features. The process involves summing the data points for each feature, resulting in a vector with weight d. The symbol djd_j in the equation represents the sum of the matrix AA components in the jjth column. The weight vector is normalized using the following method:

$$h = d_j / \left(\sum_{j=1}^{n} d_j \right) \quad (j \in [1,n]) \qquad (6)$$

The above equation finds the summation of the elements in h. The contribution of the jth feature to each and every data point is displayed in every jth element. This is how the feature's significance can be determined. Next, each point in the neighbouring matrix A is subjected to weight vector h in order to obtain the weighted adjacency matrix Z:

$$Z_{i,j} = A_{i,j} \times h \qquad (7)$$

9.3 Algorithm for Weighted Adjacent Matrix

Input – Set of data elements X in such a way that X = { x1,..,xn } ∈ Rn x d, and k is the number of clusters.
Output – Centroids C along with the cluster indicators for every element.

Step 1: Use equation 3) to form similarity matrix W of X;
Step 2: Use equation 5) to compute the adjacent matrix A;
Step 3: Calculate the output C by running k-means clustering algorithm on A;
Finally, we obtain the weighted adjacency matrix Z. Now, K-means clustering is then applied to Z to produce the clustering output. This represents the final clustering result for the original dataset. The following are the phases in our suggested strategy, which utilizes k-means clustering based on a weighted adjacency matrix.

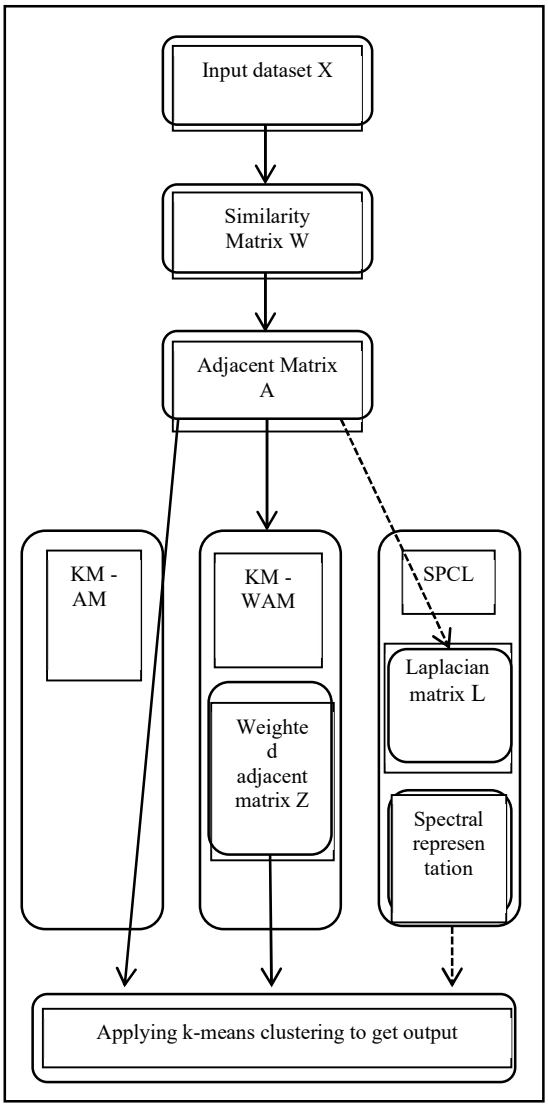

Figure 1: Proposed Method – Graphical representation

9.4 Algorithm for K-means Clustering Based on a Weighted Adjacency Matrix

Input- data points X={x1,..,xn } ∈ Rn x d; cluster number k.
Output – The centroids C along with the cluster indicators for all data points.

Step 1: Use equation 3) to form similarity matrix W of X;
Step 2: Use equation 5) for calculating adjacent matrix A;
Step 3: Use equation 6) for computing weight vector h;
Step 4: Use equation 7) to calculate and form the weighted adjacent matrix Z;
Step 5: Compute the output C by running k-means clustering on Z;

10. TRIAL AND ANALYSIS

In addition to comparing the three clustering algorithms using the three clustering assessment criteria, this study assesses our two clustering techniques.

10.1 Data Declaration

We have selected datasets from data mining centre website and also from UCI ML Repository. The chosen data are from various categories. This contains a range of features to evaluate the effectiveness and reliability of the proposed technique. The datasets are mentioned in the given table.

10.2 Algorithm for comparison

The following algorithms are compared in the paper:

a. K-means clustering is a popular and widely used algorithm for clustering. It groups elements or points into distinct clusters, with the number of clusters denoted as 'k'. Data points with different features are expected to be in different clusters, while those with similar attributes are expected to be in the same cluster. The commonly used built-in MATLAB function typically has parameters for the "selection algorithm for initial centroid positions" and "distance" set to "cluster" and "Euclidean distance," respectively.

b. Another variation of the standard k-means algorithm is the k-means++ algorithm. This algorithm finds centroids using a heuristic approach, which usually results in a lower sum of SSE and faster convergence.
c. Normalized spectral clustering (SPCL) is a well-known variation of the spectral clustering algorithm [20]. In this method, the normalized eigenvector matrix is used for k-means clustering. To achieve a norm of 1, the row sum must be normalized.

10.3 Set of Parameters

The 10-fold cross validation approach was employed in the study to assess each of the algorithms that were discussed. Adjusting the parameter σ is necessary for the paper's suggested techniques because it significantly impacts the kernel function's performance and the clustering outcome [21]. The parameter σ has been tested on the datasets within the range σ ∈ [10-5, … 1014]. To

be evaluated, Similarity matrix W's mean value is chosen as σ:

$$\sigma = \text{mean}(W) \quad (8)$$

Similarity/Affinity matrix W is calculated using equation 3)

Table 1: Evaluation of the Technique using Dataset.

Dataset	No of Samples	No of Features	No of Classes
Dexter	300	20000	2
20news	3970	8014	4
Binalpha	1404	320	9
Coil20Data	1440	1024	20
SolarFlare	1066	12	6
WebsitePhishing	1353	9	3
Cardiotocography	2126	41	3
ParkinsonSpeech	1040	28	2
GermanCredit Data	1000	23	2

10.4 Evaluation Measurement

In order to examine the many facets of the clustering outcomes, our study suggests using the assessment measurements, namely AC, NM and PR. Accuracy can be checked with the help of AC, purity is demonstrated by PR, and NM is normalized mutual information.

$$AC = N_{corg} / N \quad (9)$$

N_{corg} = the no. of data points placed accurately to the respective group or clusters

NM illustrates how quality and cluster number are traded off:

$$NM = 2\,[M(x_i, x_j) / E(x_i) + E(x_j)] \quad (10)$$

$sM(x_i, x_j)$ = relationship between two variables, E( ) = variable's entropy

PR provides an overview of each cluster's percentage of categorized data points:

$$PR = \sum_{i=1}^{k} \left(\frac{Si}{n}\right) Pi \quad (11)$$

Where, k = no of clusters, S_i = no. of data elements from i^{th} class,

P_i = correct segregation of data points divided into clusters

11. COMPARISON FROM PRIOR WORK

The proposed research stands apart from traditional K-Means clustering algorithms and prior modifications in several key aspects:

a. Integration of Weighted Adjacent Matrix: Unlike standard K-Means and its variants, this study introduces a weighted adjacent matrix to improve clustering accuracy. Previous research has mainly focused on centroid initialization and alternative distance metrics but has not thoroughly explored the advantages of weighted adjacency.

b. Dual Adjacency Matrix Approach: This research utilizes two distinct types of adjacency matrices—adjacent and weighted adjacent matrices. Most prior studies have employed a single type of adjacency or similarity matrix, limiting the exploration of multidimensional data relationships.

c. Enhanced Performance Metrics: By incorporating the weighted adjacent matrix, the proposed approach aims to outperform classical clustering algorithms in terms of normalized mutual information (NM), accuracy (AC), and purity (PR). This represents a significant departure from traditional methods that primarily focus on improving SSE (sum of squared errors).

d. Optimization for High-Dimensional Data: While many previous studies have addressed clustering challenges, this research specifically targets the complexity of high-dimensional datasets, offering a more robust solution for managing large feature spaces.

12. DESCRIPTIVE ANALYSIS OF THE RESEARCH

By addressing these pros and cons, the present research aims to advance the field of clustering techniques, offering innovative solutions while acknowledging areas that require careful consideration and further research.

12.1 Pros

12.1.1 Improved Clustering Accuracy

The use of a weighted adjacent matrix enhances the ability to capture intricate relationships between data points, resulting in more precise clustering outcomes.

12.1.2 Effective Management of High-Dimensional Data

The proposed method effectively handles high-dimensional datasets, ensuring meaningful cluster formations even with a large number of features.

12.1.3 Reduced Computational Complexity

By eliminating the need for eigenvalue decomposition in the Laplacian matrix, the proposed approach reduces computational overhead, making it feasible for larger datasets.

12.1.4 Flexibility and Scalability

The dual adjacency matrix approach provides greater flexibility in managing diverse datasets and can be easily scaled to handle extensive data points.

12.1.5 Comprehensive Evaluation Metrics

The focus on metrics such as NM, AC, and PR ensures a thorough evaluation of clustering performance, providing a clearer understanding of the algorithm's effectiveness.

12.2 Cons

12.2.1 Parameter Sensitivity

The performance of the proposed approach may be sensitive to the choice of parameters, particularly the σ parameter in the Gaussian kernel. Fine-tuning these parameters is essential for optimal results.

12.2.2 Complexity in Implementation

The introduction of weighted adjacency and dual matrices may increase the complexity of the implementation, requiring careful design and debugging.

12.2.3 Dependence on Data Characteristics

The effectiveness of the proposed method may vary depending on the nature of the datasets. It might not perform equally well on all types of data, particularly those with unique or highly irregular structures.

12.2.4 Computational Overhead in Weight Calculation

While the approach reduces some computational costs, the process of calculating feature weights and constructing the weighted adjacent matrix can introduce additional overhead.

13. CONCLUSION

This research effectively addresses the significant limitations of traditional K-Means clustering, particularly regarding high-dimensional datasets and determining the optimal number of clusters. By incorporating a weighted adjacent matrix, the proposed modified approach enhances clustering performance, offering a more accurate and robust solution for various clustering tasks.

The newly developed methods, namely K-Means based on an adjacent matrix (KM-AM) and K-Means based on a weighted adjacent matrix (KM-WAM), exhibit substantial improvements over traditional techniques. The dual adjacency matrix approach enables more effective handling of complex data structures and high-dimensional datasets, resulting in clusters that are more meaningful and representative of the underlying data. Comprehensive evaluation metrics, including normalized mutual information (NM), accuracy (AC), and purity (PR), demonstrate that the modified approach significantly outperforms classical clustering algorithms. The reduction in computational complexity and the enhancement in clustering accuracy highlight the practical feasibility of these methods for large-scale applications.

The advantages of the proposed approach, such as improved clustering accuracy, effective management of high-dimensional data, reduced computational overhead, and comprehensive evaluation metrics, underscore its potential for advancing data analysis techniques across various domains. However, it is crucial to acknowledge the limitations, including parameter sensitivity, complexity in implementation, and dependence on data characteristics, which require careful consideration and further research.

In conclusion, this research makes a valuable contribution to the field of clustering techniques by providing innovative solutions that address core concerns related to clustering accuracy and performance. The findings support the initial problem statement and emphasize the importance of ongoing exploration and refinement of clustering methodologies to meet the evolving needs of data analysis and machine learning.

14. DECLARATION

No funding was received for conducting this study.

15. CONFLICT OF INTEREST

Authors have no conflict of interest for publication of this paper in this journal.

REFERENCES:

[1] Kodinariya, T. M., & Makwana, P. R., "Review on determining number of Cluster in K-Means Clustering". *International Journal*, 1(6), 2013, pp. 90-95.

[2] Tibshirani, R., Walther, G., & Hastie, T., "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 2001, pp. 411-423.

[3] Murtagh F, Contreras P, "Algorithms for hierarchical clustering: an overview". *Wiley Data Mining Knowledge Discoverv* 2(1), 2012, pp. 86-97.

[4] Zheng, W., Zhu, X., Wen, G., Zhu, Y., Yu, H., & Gan, J. "Unsupervised feature selection by self-paced learning regularization", *Pattern Recognition Letters*, 132, 2020, pp. 4-11.

[5] Abe, S., & Abe, S., "Feature selection and extraction", *Support Vector Machines for Pattern Classification*, 2010, pp. 331-341.

[6] Arora, P., & Varshney, S, "Analysis of k-means and k-medoids algorithm for big data", *Procedia Computer Science*, 78, 2016, pp. 507-512.

[7] Bachem, O., Lucic, M., Hassani, H., & Krause, A., "K-mc2: approximate k-means++ in sublinear time", *AAAI,* 2016.

[8] Zhang, S. "Multiple-Scale Cost Sensitive Decision Tree Learning", *World Wide Web*, 21, 2018, pp. 1787-1800.

[9] Zhu, X., Zhang, S., Li, Y., Zhang, J., Yang, L., & Fang, Y., "Low-rank sparse subspace for spectral clustering", *IEEE Transactions on knowledge and data engineering*, 31(8), 2018, pp. 1532-1543.

[10] Sharma, A., & Sharma, A., "KNN-DBSCAN: Using k-nearest neighbor information for parameter-free density-based clustering". *IEEE, International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT),* July 2017, pp. 787-792.

[11] He, L., Ray, N., Guan, Y., & Zhang, H., "Fast large-scale spectral clustering via explicit feature mapping", *IEEE transactions on cybernetics*, 49(3), 2018, pp. 1058-1071.

[12] Chen, J., Lin, X., Xuan, Q., & Xiang, Y, "FGCH: a fast and grid based clustering algorithm for hybrid data stream", *Applied Intelligence*, 49, 2019, pp. 1228-1244.

[13] Ding, Y., & Fu, X., "Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm", *Neurocomputing*, 188, 2016, 233-238.

[14] Ferreira, M. R., de Carvalho, F. D. A., & Simões, E. C., "Kernel-based hard clustering methods with kernelization of the metric and automatic weighting of the variables", *Pattern Recognition*, 51, 2016, pp. 310-321.

[15] Du, L., Zhou, P., Shi, L., Wang, H., Fan, M., Wang, W., & Shen, Y. D., "Robust multiple kernel k-means using l21-norm", *Twenty-fourth international joint conference on artificial intelligence,* June 2015.

[16] Zahra, S., Ghazanfar, M. A., Khalid, A., Azam, M. A., Naeem, U., & Prugel-Bennett, A., "Novel centroid selection approaches for KMeans-clustering based recommender systems", *Information sciences*, 320, 2015, pp. 156-189.

[17] Pavan, K. K., Rao, A. A., Rao, A. D., & Sridhar, G. R., "Single pass seed selection algorithm for k-means", *Journal of Computer Science*, 6(1), 2010, pp. 60.

[18] Tremblay, N., Puy, G., Gribonval, R., & Vandergheynst, P., "Compressive spectral clustering", *International conference on machine learning,* PMLR, June 2016, pp. 1002-1011.

[19] Gebru, I. D., Alameda-Pineda, X., Forbes, F., & Horaud, R., "EM algorithms for weighted-data clustering with application to audio-visual scene analysis", *IEEE transactions on pattern analysis and machine intelligence*, 38(12), 2016, pp. 2402-2415.

[20] Von Luxburg, U., "A tutorial on spectral clustering", *Statistics and computing*, 17, 2007, 395-416.

[21] Souza, C. R., "Kernel functions for machine learning applications", *Creative commons attribution-noncommercial-share alike*, 3(29), 2010, 1-1.