International Journal on Emerging Technologies **11**(5): 746-755(2020)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

A Strategy for Improved Data Classification using Advanced clustering with Incomplete Datasets

***P.S. Deshmukh*[1]*, M. Sivakkumar*[2] *and Varshaha Namdeo*[3]**
[1]Ph.D. Scholar, Department of CSE, SRK, University, Bhopal (Madhya Pradesh), India.
[2]Associate Professor, Department of CSE, SRK, University, Bhopal (Madhya Pradesh), India.
[3]Professor, Department of CSE, SRK, University, Bhopal (Madhya Pradesh), India.

(Corresponding author: P.S. Deshmukh)

(Received 19 October 2020, Revised 29 November 2020, Accepted 23 December 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: This study explores the application of machine learning techniques, including clustering, in various real-world domains such as cyber security, healthcare, and agriculture. It emphasizes the importance of understanding different methods like supervised, unsupervised, semi-supervised, and reinforcement learning. Clustering algorithms are particularly powerful for analyzing large volumes of data by grouping similar objects into clusters. Subspace clustering extends this concept to identify clusters in different subspaces within high-dimensional data. The study aims to address challenges like determining optimal initial cluster positions and identifying research gaps in unsupervised learning. Its findings will aid researchers in exploring new directions and comparing the effectiveness of different algorithms. Challenges in implementing improved data classification using advanced clustering with incomplete datasets may arise from difficulties in handling missing data effectively, potential biases introduced by incomplete information, and the need for robust algorithms that can adapt to diverse data patterns while ensuring accurate classification results.

Keywords: Data Mining Tools, Machine Learning, supervised learning, Unsupervised learning Clustering algorithms, Artificial Intelligence, Time Complexity, big data, Similarity Measure.

I. INTRODUCTION

Data mining is the process of extracting valuable insights from datasets, employing techniques such as clustering, classification, regression, association, and outlier detection. Clustering involves grouping similar objects together, constituting an unsupervised learning approach. Effective clustering ensures high intra-class similarity and low inter-class similarity. Clustering algorithms can be categorized broadly into hierarchical and partition algorithms [1].

Hierarchical clustering organizes data into a tree-like structure, with agglomerative (bottom-up) and divisive (top-down) approaches being its main variants. Partition clustering algorithms, on the other hand, divide data points into k partitions, each representing a cluster, based on specific objective functions [2, 3].

Machine learning, a subset of Artificial Intelligence, empowers applications to learn and enhance performance based on experience rather than explicit programming. Supervised and unsupervised learning are fundamental approaches in machine learning. Unsupervised algorithms discern hidden data structures within unlabeled datasets. Among these techniques, clustering stands out as a core component of data science [4].

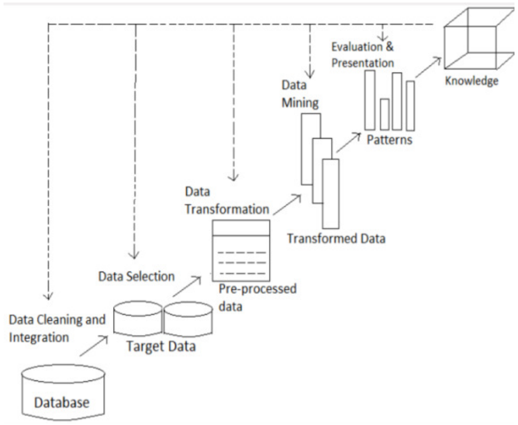


Fig. 1. Data Mining as a step in the process of knowledge discovery [16].

Clustering, an unsupervised learning method, organizes unlabeled data into clusters where data points share similarities within clusters and differences between clusters. While many clustering algorithms are adept at handling either numeric or categorical feature values, numeric features represent real values like height or distance, while categorical features categorize data into fixed groups such as colour or profession [5, 6].

Clustering algorithms rely on defining similarity, often using distance-based measures like Euclidean distance for numeric data. However, computing similarity for categorical data poses challenges due to the lack of inherent ordering. Although direct distance computation between categorical values is impractical, various methods proposed in literature offer solutions for assessing similarity among data points with categorical features [7, 8].

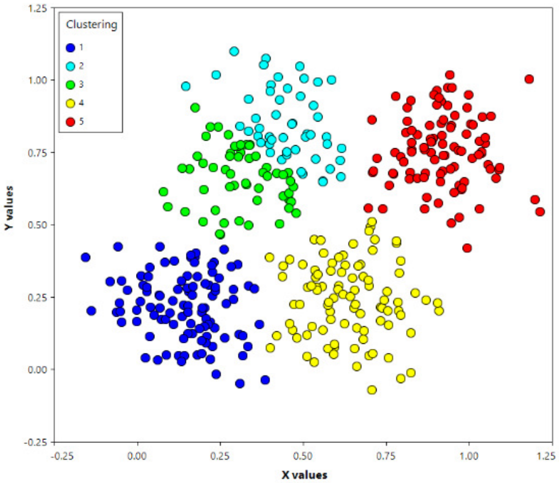


Fig. 2. Grouping in 5 clusters [18].

In the realm of data science and machine learning, K-Means clustering stands out as a powerful unsupervised technique for discerning the structure of given datasets [14]. This clustering algorithm is highly valued for its ability to effectively segment data into groups, owing to its simplicity [14]. Its versatility is evident in its application across various domains, including recommendation systems, smart city services, cybersecurity, and business data analysis [15]. Furthermore, K-Means plays a pivotal role in analyzing user behavior and facilitating context-aware services [17]. Moreover, it serves as a crucial tool for complex feature extraction tasks [17].

A. K-Means and Fuzzy C-Means Clustering Algorithms

For the purpose of image segmentation, the most frequently implemented partial clustering algorithms are the fuzzy c-means and K-means algorithms. Fast and straightforward, K-means clustering is an algorithm for converging clusters. The detailed procedure outlined in the algorithms can be found in reference [4]. Compared to the k-means clustering algorithm, the fuzzy c-means algorithm converges more rapidly. The exhaustive description of the algorithmic procedures Bonaccorso (2017) [5]. Both clustering algorithms encounter similar challenges, including issues related to the quantity of initial centroids, dead centers, and initial centroids. Both k-means and fuzzy c-means clustering methods calculate the Euclidean distances between the image pixels and the centroids. However, when dealing with large datasets, these methods incur additional time and financial expenses. A stochastic distribution of clusters and centroids impacts the segmentation outcomes and increases the time complexity. Therefore, it is necessary to improve these algorithms so that they autonomously determine the optimal number of clusters and their respective cluster centroids. The following section examines the literature-recognized methods for determining the initial optimal number of clusters and their centers.

B. Related Work

In general, the determination of the optimal number of clusters and their centroids is a critical aspect left to the discretion of researchers Honda *et al.* (2011) [20]. Various approaches have been proposed to address this challenge, including running the clustering algorithm multiple times and selecting the desired number of clusters based on validity criteria, or automatically determining the number of clusters through meaningful methods or criteria [20]. Similarly, the selection of cluster centroids can involve random initialization followed by optimization through multiple algorithm runs [20].

Wang *et al.* (2019) propose a novel K-means based clustering algorithm that integrates clustering and imputation into a single objective function, offering enhanced optimality by balancing these two processes. They further design an alternate optimization algorithm to solve the resulting optimization problem, demonstrating its convergence theoretically. Through comprehensive experimental studies on various benchmark datasets and real-world applications, the effectiveness of their algorithm is demonstrated, outperforming several commonly-used methods for incomplete data clustering [15].

Pugazhenthi and Kumar (2020) focus on the challenges of selecting the optimal number of clusters and corresponding centroids in image segmentation using clustering algorithms such as K-means and Fuzzy c-means. They review research efforts aimed at improving the efficient isolation of clusters and explore the limitations and applications of these clustering algorithms [16].

Liu *et al.* (2019) address the issue of incomplete kernel matrices in multiple kernel clustering (MKC) algorithms, proposing two effective algorithms to handle this problem. Their algorithms integrate imputation and clustering into a unified learning procedure, directly performing multiple kernel clustering with incomplete kernel matrices. They demonstrate the superior performance of their algorithms through extensive experiments on benchmark datasets [17].

Sinaga and Yang (2020) an unsupervised learning schema for the K-means algorithm, called U-k-means, is constructed to eliminate the need for initializations and parameter selection while simultaneously determining the optimal number of clusters. They propose a novel U-k-means clustering algorithm, which automatically finds an optimal number of clusters without requiring any initialization or parameter selection. Computational complexity analysis and comparisons with existing methods validate the effectiveness of their proposed algorithm [18].

Wu *et al.* (2015) investigate K-means-based consensus clustering (KCC) as an efficient approach for finding cluster structures from heterogeneous data. They provide a systematic study of KCC, revealing necessary and sufficient conditions for utility functions and investigating factors that may affect KCC performances. Experimental results demonstrate the efficiency and robustness of KCC, particularly in handling incomplete data sets with missing values [19].

Honda *et al.* (2011) consider k-Means clustering of incomplete datasets with missing values, extending the PCA-guided k-Means procedure to address this issue. Their approach involves estimating principal component scores iteratively without imputation, deriving k-Means-like partitions through lower rank approximation of the data matrix. Experimental results show the robustness of their method to initialization problems and its effectiveness in recovering solutions even in the presence of missing values [20].

Selim and Ismail (1984) address several questions about the K-means algorithm, including its convergence properties and robustness to noisy constraints. They cast the clustering problem as a nonconvex mathematical program and provide a rigorous proof of the finite convergence of the K-means-type algorithm under certain conditions. Their study sheds light on the convergence behavior and stability of the K-means algorithm in different scenarios [21].

Yoon *et al.* (2007) propose an approach to outlier detection of software measurement data using the k-means clustering method, aiming to improve the quality of software measurement data by identifying and handling outliers. Their method leverages k-means clustering to detect outliers in software measurement data, contributing to more accurate decision-making in software project management [22].

Ye *et al.* (2017) introduce a novel unified learning method for incomplete Multiview clustering, which simultaneously imputes incomplete views and learns a consistent clustering result by explicitly modelling between-view consistency. They propose an iterative algorithm to achieve optimal clustering results while maintaining between-view consistency, demonstrating superior performance over existing methods on synthetic and real-world datasets [23].

Liu *et al.* (2020) propose the Consensus Guided Unsupervised Feature Selection (CGUFS) framework, which integrates consensus clustering to generate pseudo labels for feature selection. Their approach addresses the issue of noisy and irrelevant features by employing multiple diverse basic partitions and consensus clustering to guide feature selection, achieving superior effectiveness and efficiency compared to state-of-the-art methods [17].

Law *et al*. (2004) propose a novel expectation-maximization (EM) algorithm for feature selection in mixture-based clustering, leveraging the concept of feature saliency to drive the selection process. Their approach estimates feature saliencies using a minimum message length model selection criterion, effectively identifying relevant features while simultaneously determining the number of clusters [25].

Table 1: Summary of literature study in the field data mining in K-mean clustering and objective analysis.

Authors /year	Research Purpose	Title	Outcome Measures
Wang *et al.* [15] 2019	Real dataset and accurate data analysis and error minimization	K-means clustering with incomplete data	Copy data, more error and low accuracy and objective function
Pugazhenthi *et al.* [16] 2020	optimization clustering and error minimization Remove copy data in clustering and optimal solution	Selection of Optimal Number of Clusters and Centroids for K-means and Fuzzy C-means Clustering: A Review	More time complexity and sub optimal solution
Liu *et al.* [17] 2019	Remove copy data in clustering and optimal solution and accurate data analysis	Multiple Kernel k-Means with Incomplete Kernels	Copy data, more error and low accuracy
Sinaga and Yang [18] 2020	Machine learning database clustering analysis and Remove copy data in clustering and optimal solution	Unsupervised K-Means Clustering Algorithm	More time complexity and sub optimal solution
Wu *et al.* [19] 2015	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution	K-Means-Based Consensus Clustering: A Unified View	Copy data, more error and low accuracy
Honda *et al.* [20] 2011	Machine learning database clustering analysis and Remove copy data in clustering and optimal solution	PCA-guided k-Means clustering with incomplete data	Better results and More time complexity and sub optimal solution
Selim and Ismail [21] 1984	Remove copy data in clustering and optimal solution and accurate data analysis	K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality	Copy data, more error and low accuracy
Yoon *et al.* [22] 2007	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution and accurate data analysis	An Approach to Outlier Detection of Software Measurement Data using the K-means Clustering Method	More time complexity and sub optimal solution
Ye *et al.* [23] 2017	Real dataset and accurate data analysis and accurate data analysis	Consensus kernel-means clustering for incomplete Multiview data.	Copy data, more error and low accuracy
Liu *et al.* [28] 2018	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution	Feature selection with unsupervised consensus guidance	Effective segmentation of fish from complex background and More time complexity and sub optimal solution
Law *et al*. [25] 2004	resolves the incomplete data clustering task in feature selection and incomplete data analysis, remove copy data in clustering and optimal solution, Remove copy data in clustering and optimal solution	Simultaneous feature selection and clustering using mixture models	Copy data, more error and low accuracy

Table 1 presents a summary of the literature review pertaining to K-mean clustering and objective analysis for Kmeans in the field of data mining. K-means have a broad spectrum of applications and are utilized with various image and data types. In addition, the qualitative metrics utilized for analyses differ between applications.

Determining or proposing a universally applicable method for determining the optimal number of clusters and their centroids is thus a challenging task. The optimal method for determining the number of clusters and their centroids will be determined in consideration of both the applications and the qualitative metric requirements of the segmentation process. The enhancement in quantitative parameters such as error rate (ER) and objective function is observed with optimal cluster selection as opposed to random selection of objective function and number of centroids [4, 22]. The section examines different methodologies that have been documented in the literature to determine the optimal number of clusters and centroids to begin with.

II. PROPOSED METHODOLOGY

(a) Overview of Proposed Methodology. This methodology based on the principles of action and genetics, introduces a novel approach to optimization through the Genetic Algorithm (GA). The GA concept, first formalized in the Netherlands, mimics the principles of natural selection and evolution proposed by Darwin. It aims to tackle optimization problems represented by the objective function f(x), where $x = [x1,x2,…,xn]$ is an N-dimensional vector of optimization parameters. The GA has been recognized as one of the most efficient and powerful global optimization algorithms, particularly suitable for combinatorial optimization problems, including those with non-differentiable or discontinuous objective functions.

At the core of the proposed algorithm are chromosomes and genes, where the optimization parameters are encoded into binary strings. To facilitate the evolution towards better solutions, a fitness measure is essential to distinguish between good and poor solutions objectively. Fitness measures are relative to candidate solutions and guide the algorithm towards the emergence of better solutions. Additionally, the algorithm relies on a population of candidate solutions, where the population size significantly influences its performance and scalability [29].

The algorithm follows a series of steps to converge towards the optimal solution:

Initialization: Generating initial candidate solutions randomly within the search space.

Evaluation: Assessing the fitness values of the candidate solutions.

Selection: Favoring candidate solutions with higher fitness values for reproduction and further evolution. Various selection procedures, such as roulette-wheel selection and tournament selection, can be employed.

Recombination: Creating new candidate solutions through the combination of elements from two or more parent solutions.

Mutation: Introducing random changes to candidate solutions to explore new regions of the search space.

Replacement: Replacing the parent population with the offspring population, selectively retaining better-performing solutions [30].

Iteration: Repeating steps 2 to 6 until a termination condition is met.

The proposed algorithm's working principle involves three main steps:

Data preprocessing: Handling missing values through mean imputation and normalizing features for equal contribution to clustering and classification.

Clustering: Employing the proposed algorithm to cluster the cleaned dataset, removing outliers and irrelevant data.

Classification: Utilizing a Support Vector Machine (SVM) classifier on the reduced dataset to achieve higher accuracy compared to existing methods.

To enhance classifier performance, k-fold cross-validation is employed, and confusion matrix analysis is conducted to evaluate classification performance based on sensitivity, specificity, positive predictive value, and negative predictive value. The proposed algorithm aims to optimize clustering correctness while minimizing cluster errors in datasets.

Experimental evaluations are conducted on various datasets, including Iris, Wine, Glass, Breast Cancer, Mice Protein, Ovarian Cancer, Pen Digits, Avila, and Sensorless Drive, demonstrating the algorithm's effectiveness in clustering and optimization tasks. The efficiency of the algorithm heavily relies on the systematic selection of initial cluster centroids, which is addressed using Principal Component Analysis (PCA) to divide values into percentiles, ensuring efficient centroid initialization.

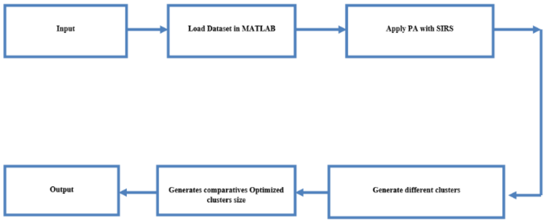

Input
Load Dataset in MATLAB
Apply PA with SIRS
Generate different clusters
Generates comparatives Optimized clusters size
Output

Fig. 3. Block diagram of PA.

(b) Working Process:

(i) The proposed algorithm using smart information retrieval system (SIRS) for optimizing data classification performance with K-Value Selection Clustering for handling incomplete and noisy data includes the following steps:

Algorithm Development: Develop and implement the K-Value Selection Clustering algorithm with appropriate distance metrics and noise reduction techniques.

Experiments: Apply the algorithm to real-world datasets with incomplete and noisy data. Experiment with various K values within the specified range.

Evaluation: Evaluate the performance of the algorithm using clustering quality metrics such as f-score, proposed algorithm using smart information retrieval system for each K value [31].

Optimal K Selection: Determine the optimal K value that maximizes clustering quality and minimizes noise.

Data Classification Improvement:

Apply the optimal K value to cluster the data, resulting in improved data classification performance.

Validation: Validate the proposed approach's effectiveness by comparing it with traditional clustering methods on the same datasets.

III. RESULTS AND ANALYSIS

Analyses the results to demonstrate how the proposed algorithm using smart information retrieval system are optimizes data classification performance in the presence of incomplete and noisy data.

The proposed algorithm using smart information retrieval system (SIRS) aims to provide an efficient solution for data classification in challenging data environments, ensuring better data separation and classification accuracy while handling incomplete and noisy data effectively.

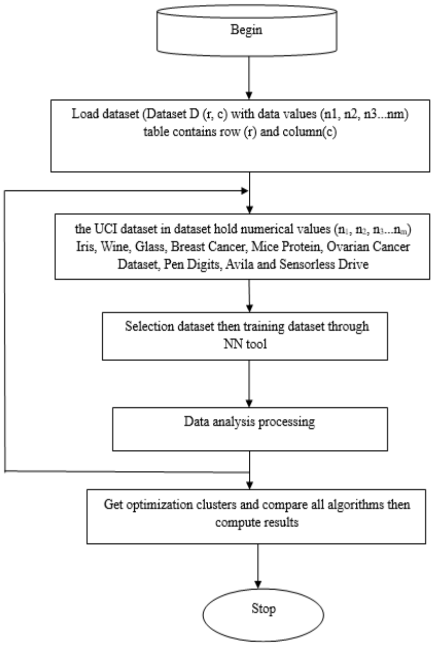

Begin
Load dataset (Dataset D (r, c) with data values (n1, n2, n3...nm) table contains row (r) and column(c)
the UCI dataset in dataset hold numerical values (n1, n2, n3...nm) Iris, Wine, Glass, Breast Cancer, Mice Protein, Ovarian Cancer Dataset, Pen Digits, Avila and Sensorless Drive
Selection dataset then training dataset through NN tool
Data analysis processing
Get optimization clusters and compare all algorithms then compute results
Stop

Fig. 4. Flow diagrams PA.

Initialization load dataset and choose dataset and set point values as numerical value are selected randomly numerical values. it's a produce the unique clustering and minimize error in dataset base on smart information retrieval system (SIRS) based clustering algorithm. A proposed algorithm improves the accuracy and efficiency as compare to the k-means clustering algorithm another clustering algorithm. It's also called minimum error rate way for assigning data points to appropriate clusters, finding the higher initial centroids but time is increase and error rate minimization. To compare the 2 initialization methods two standard datasets are utilized in analysis dataset.

(C) Proposed Optimization Clustering Algorithm: The proposed algorithm using smart information retrieval system (SIRS) in determining the optimal number of clusters (K) through the following steps: Initialization: Start by randomly assigning each data point to one of the K clusters. Intra-cluster Distance: Calculate the average Euclidean distance between data points within each cluster. This distance metric provides insight into the compactness of each cluster. Optimal K Selection: Utilize the Elbow Method, a common heuristic. Plot the K values against the corresponding average intra-cluster distances and identify the point where the rate of decrease begins to slow down. This point is typically considered the optimal K, as it balances cluster compactness and separation. Incomplete Data Handling: To address the challenge of incomplete data, a clustering algorithm is adapted to accommodate missing values: Choose Technique: Opt for a suitable incomplete data clustering technique, such as an adjusted version of proposed algorithm using smart information retrieval system (SIRS) tailored to handle missing values. Distance Metric Modification: Modify the distance calculation within the clustering algorithm to account for missing values. For instance, adjust the Euclidean distance calculation to consider only dimensions with available values in both data points.

(i) Input: Dataset with incomplete and noisy data Range of possible K values Clustering distance metric Convergence threshold Maximum number of iterations

(ii) Output: Optimally clustered data with minimized noise and maximized data classification performance Selected K value for optimal clustering

(iii) Procedure: Data Preprocessing: Handle missing or incomplete data using imputation techniques. Apply noise reduction methods to reduce the impact of noisy data points. **Initialization:** Initialize K with a value within the given range.

Repeat for Each K Value:

Step 1: Cluster Assignment Step: Randomly initialize K cluster centroids. For each data point, calculate the distance to each cluster centroid using the chosen distance metric. Assign the data point to the cluster with the nearest centroid. First load dataset Load Dataset Iris, Wine, Glass, Breast Cancer, Mice Protein,1 Ovarian Cancer Dataset, Avila and Sensorless Drive, Here Dataset is healthcare related dataset and dataset Value are numerical like v1, v2, here v1, v2 are numerical values in dataset n number of knowledge points in d dimension, and n is number of clusters. Select or Load dataset a dataset at time one in mat lab. Where Dataset N (The parameters include the amount of clusters G, v number of knowledge points in d dimension).

Step 2: Update Step: Recalculate the centroids of each cluster by taking the mean of all data points assigned to that cluster. Select a dataset1 and choose any input values numerical and randomly pick v1 number of points because the initial centres of n clusters.

Step3: Convergence Check: Check if the cluster centroids have converged, either by assessing if they have significantly changed between iterations or by reaching the maximum number of iterations Compute and dataset analysis using proposed optimization clustering algorithm and find the minimum error value and optimal vales within the given.

If (No)

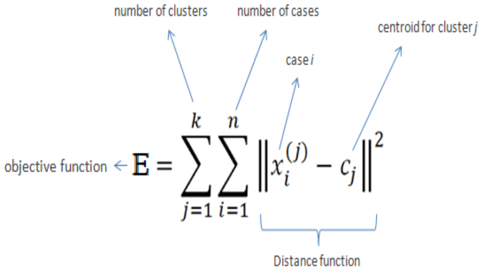
Initial set incorrect values and attend step2.

Else (Yes)

Generate cluster with output and attend next step

Step4: Evaluate Clustering Quality: Measure the clustering quality using smart information retrieval system (SIRS) Different clusters are generated and minimum error value and optimal vales within the given dataset1. All clusters are finding mean Xi and Yi mean then calculate error function**.**

$$\overline{X} = \frac{X_1 + X_2 + X_3 \ldots X_N}{N}$$



$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Here objective function E, Cj centroid (mean of objects) for cluster j, xi case i, n is number of cases, k number of cluster, distance function therefore, the target perform E tries to reduce the ad of the square distances of objects from their cluster centres.

Step 6: Select Optimal K Value: Record the clustering quality metric for the current K value. Compare K Values: Compare the clustering quality metrics for different K values. Select Best K: Choose the K value

that yields the best clustering quality based on the selected metric. Cluster Data with Optimal K: Return the clustering algorithm using the selected optimal K value.
Step7: Stop

IV. EXPERIMENTS SETUP

(a) Simulation tool: MATLAB, short for Matrix Laboratory, stands as a fourth-generation high-level programming language and interactive environment primarily tailored for numerical computation, visualization, and programming tasks. Developed by Math Works, it offers a versatile platform enabling matrix manipulations, function and data plotting, algorithm implementation, user interface creation, and interfacing with other programming languages like C, C++, Java, and FORTRAN. Additionally, MATLAB facilitates data analysis, algorithm development, and application modeling, boasting a rich repository of built-in commands and mathematical functions to aid in various mathematical calculations, plot generation, and numerical methods execution.

The computational prowess of MATLAB finds extensive utilization across a wide spectrum of mathematical tasks. Notable applications include handling matrices and arrays, conducting 2-D and 3-D plotting and graphics, performing linear algebra operations, and conducting data

(b) Dataset Discussion: The proposed algorithm is evaluated using various datasets, including several UCI and large benchmark datasets such as Iris, Wine, Glass, Breast Cancer, and Ovarian Cancer Dataset. These datasets are chosen due to their significance and representativeness in the field of incomplete data clustering. Detailed information regarding these datasets is provided. To simulate incompleteness in the original complete data matrix, missing values are randomly generated. The selected datasets, including Iris, Wine, Glass, Ovarian Cancer, and Breast Cancer, are among the most commonly used benchmarks for evaluating incomplete data clustering algorithms.

(c) Parameter:

(i) **ER:** Error rate refers to a measure of the degree of prediction error of a model made with respect to the true model. The term error rate is often applied in the context of classification models.

(ii) **objectives Function:** The objective function is one of the most fundamental components of a machine learning problem, in that it provides the basic, formal specification of the problem. For some objectives, the optimal solution parameters can be found exactly (known as the analytic solution).

V. EXPERIMENTS RESULT ANALYSIS

(a) Iirs Data Analysis:

(i) Error Rate Analysis: A comparative analysis of error rates is conducted between the proposed algorithm using Smart Information Retrieval System (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), and K-means clustering algorithm (KM)) using the Iris dataset in Case 1. The analysis reveals that the Iris dataset processed through the proposed algorithm demonstrates a higher level of correlation in clustering compared to existing algorithms. Specifically, the proposed algorithm showcases a more perfect correlation in clustering, while existing algorithms exhibit lower mutual information and correlation in clustering. Fig. 5 illustrates the results obtained from the experiments, with visually appealing graphs utilized for visualization. Normalized Mutual Information in Classification: The impact on classification performance is assessed through normalized mutual information. The results effectively demonstrate the influence of the specialized approach for handling incomplete data on enhanced cluster quality and improved classification performance.

Effect of Incomplete Data Handling: A discussion is conducted on how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance. The results showcase the effectiveness of the proposed approach.

Impact of K-values: The influence of different K-values on clustering quality is examined to understand their impact.

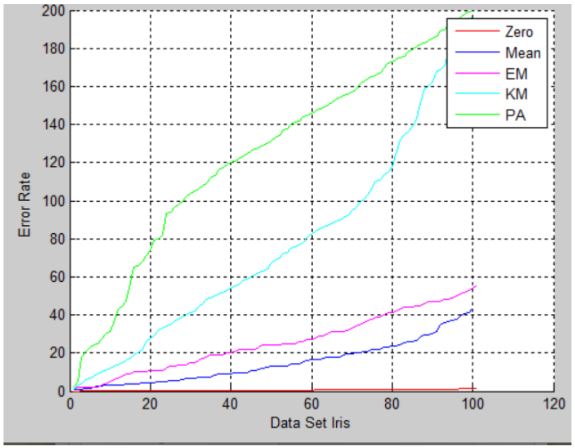


Fig. 5. Iris dataset ER analysis between existing algorithm and proposed algorithm.

(ii) objective function: The comparison of optimal parameters is conducted between the proposed algorithm using Smart Information Retrieval System (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), and K-means clustering algorithm (KM)) using Case 1 Iris dataset analysis. The analysis reveals that the Iris dataset processed through the proposed algorithm achieves an optimal solution, whereas existing algorithms yield sub-optimal solutions. Fig. 6 depicts the results obtained from the experiments, employing visually appealing graphs for visualization. Normalized Mutual Information in Classification: The resulting impact on classification performance is evaluated through normalized mutual information. The discussion highlights the effectiveness of the proposed algorithm in achieving enhanced cluster quality and improved classification performance compared to existing algorithms.

Effect of Incomplete Data Handling: The specialized approach for handling incomplete data is explored to understand its contribution to enhanced cluster quality and improved classification performance. The results effectively demonstrate the efficacy of the proposed method in handling incomplete data. Impact of K-values: An examination of the influence of different K-values on clustering quality is conducted to understand their impact.

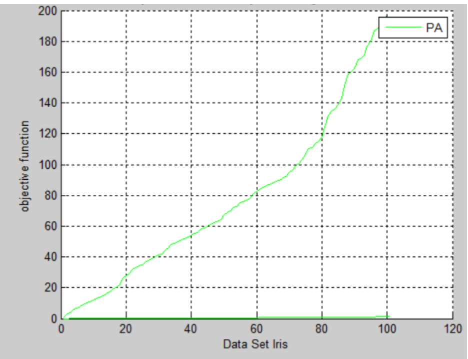


Fig. 6. Iris dataset objective function analysis between existing algorithm and proposed algorithm.

(b) Glass Dataset Analysis:

(i) Error rate**:** error rate analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 2 glass dataset analysis. Glass dataset analysis through proposed algorithm more perfect correlation in clustering, and also high correlation in clustering but existing algorithms are no mutual information and also low correlation in clustering, in show Fig. 7 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. The resulting impact on classification normalized mutual information. Effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. Impact of k-values examines the influence of different K-values on clustering quality clustering quality.

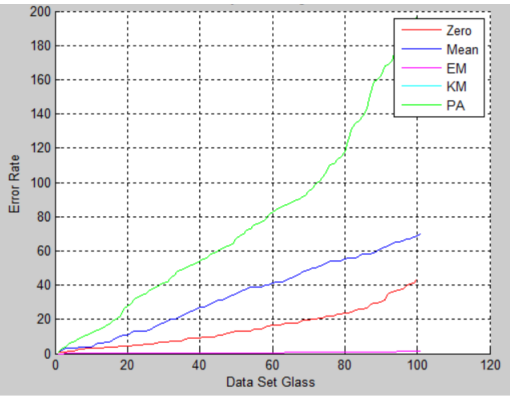


Fig. 7. Glass dataset ER analysis between existing algorithm and proposed algorithm.

(ii) objective function: optimal Parameter Analysis: An examination of optimal parameters is conducted, known as the analytic solution, comparing the proposed algorithm utilizing Smart Information Retrieval System (PASIRS) with existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), and K-means clustering algorithm (KM)) using Case 2 Glass dataset analysis. The analysis reveals that the Glass dataset processed through the proposed algorithm achieves an optimal solution, while existing algorithms yield sub-optimal solutions. Fig. 8 illustrates the results obtained from the experiments, presenting visually appealing graphs for visualization. Normalized Mutual Information in Classification: The resulting impact on classification performance is assessed through normalized mutual information. The discussion highlights the effectiveness of the proposed algorithm in achieving enhanced cluster quality and improved classification performance compared to existing algorithms.

Effect of Incomplete Data Handling: The specialized approach for handling incomplete data is explored to understand its contribution to enhanced cluster quality and improved classification performance. The results effectively demonstrate the efficacy of the proposed method in handling incomplete data. Impact of K-values: An analysis is conducted to examine the influence of different K-values on clustering quality.

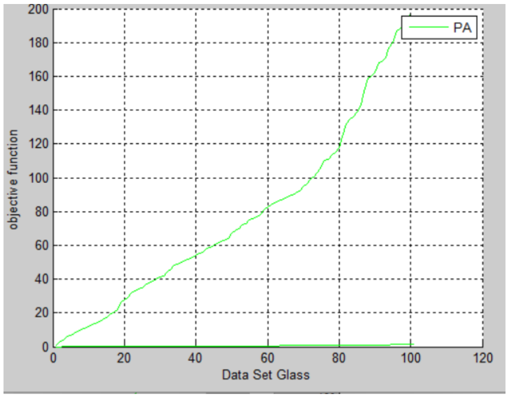


Fig. 8. Glass dataset objective function analysis between existing algorithm and proposed algorithm.

(c) Wine Dataset Analysis:

(i) Error rate**:** error rate analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 3 wine dataset analysis. Wine data analysis through proposed algorithm more perfect correlation in clustering, and also high correlation in clustering but existing algorithms are no mutual information and also low correlation in clustering, in show Fig. 9 below. The results obtained from the experiments: visualizations utilize visually appealing graphs. The resulting impact on classification normalized mutual information. Effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. Impact of k-values examines the influence of different K-values on clustering quality values on clustering quality.

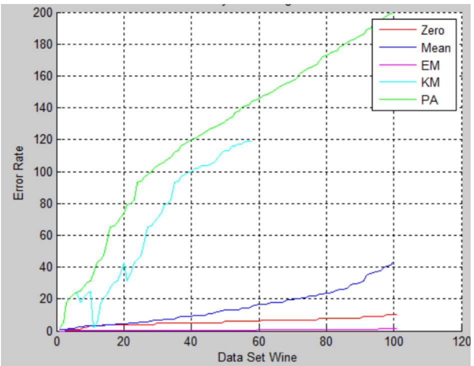


Fig. 9. Wine dataset ER analysis between existing algorithm and proposed algorithm.

(ii) objective function: the optimal parameters can be found exactly (known as the analytic solution). The objective function is one of the most fundamental components of a data analysis and machine learning problem, in that it provides the basic, formal specification of the problem. The objective function analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 3 wine dataset analysis. Wine data analysis through proposed algorithm optimal solution but existing algorithms sub-optimal solution, in show Fig. 10 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. The resulting impact on classification normalized mutual information. Effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. Impact of k-values examines the influence of different K-values on clustering quality.

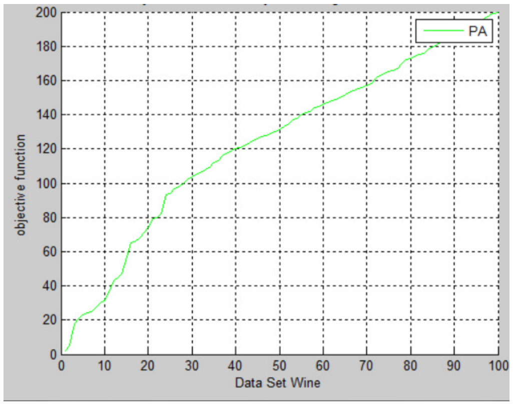


Fig. 10. Wine dataset objective function analysis between existing algorithm and proposed algorithm.

(d) Breast Cancer Dataset Analysis:

(i) Error rate**:** error rate analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 4 breast cancer dataset analysis. Breast cancer data analysis through proposed algorithm more perfect correlation in clustering, and also high correlation in clustering but existing algorithms are no mutual information and also low correlation in clustering, in show Fig. 11 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. The resulting impact on classification normalized mutual information. Effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. Impact of k-values examines the influence of different K-values on clustering quality.

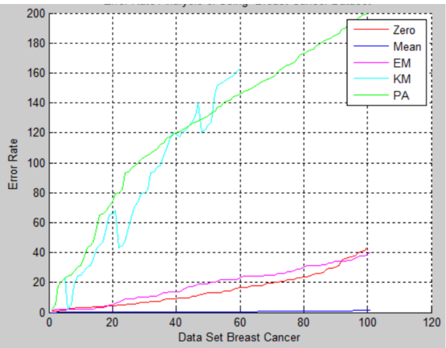


Fig. 11. Breast cancer dataset ER analysis between existing algorithm and proposed algorithm.

(ii) Objective function: The optimal parameters can be found exactly (known as the analytic solution) The objective function is one of the most fundamental components of a data analysis and machine learning problem, in that it provides the basic, formal specification of the problem. The objective function analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 4 breast cancer dataset analysis. Breast cancer data analysis through proposed algorithm optimal solution but existing algorithms sub-optimal solution, in show Fig. 12 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. The resulting impact on classification normalized mutual information. Effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. Impact of k-values examines the influence of different K-values on clustering quality.

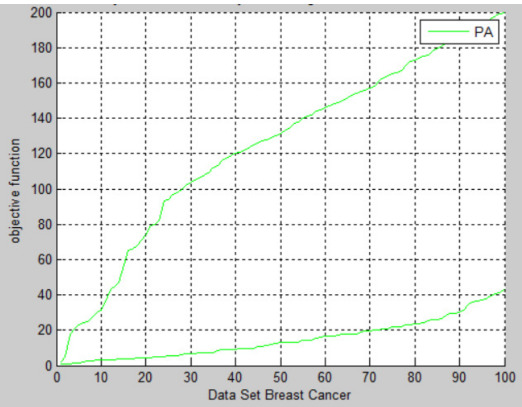


Fig. 12. Breast cancer dataset objective function analysis between existing algorithm and proposed algorithm.

(e) Ovarian cancer dataset Analysis:

(i) Error rate**:** Error rate analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 5 Ovarian cancer dataset analysis. Ovarian cancer data analysis through proposed algorithm more perfect correlation in clustering, and also high correlation in clustering but existing algorithms are no mutual information and also low correlation in clustering, in show Fig. 13 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. The resulting impact on classification normalized mutual information. Effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. Impact of k-values examines the influence of different K-values on clustering quality.

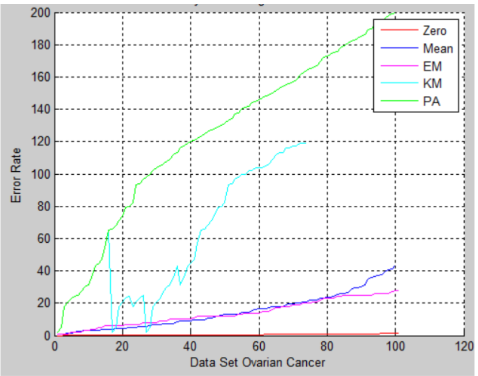


Fig. 13. Ovarian cancer dataset ER analysis between existing algorithm and proposed algorithm.

(ii) Objective function:

Optimal Parameter Analysis: In the examination of the Ovarian cancer dataset, PASIRS outperforms existing algorithms (ZF, MF, EM, KM) by achieving an optimal solution, while others yield sub-optimal results. Normalized Mutual Information: PASIRS significantly enhances cluster quality and classification performance compared to existing algorithms, as depicted in Fig. 14. Effect of Incomplete Data Handling: PASIRS' specialized approach for handling incomplete data greatly contributes to improved cluster quality and classification performance.

Impact of K-values: The analysis investigates the influence of different K-values on clustering quality, providing valuable insights into algorithm performance.

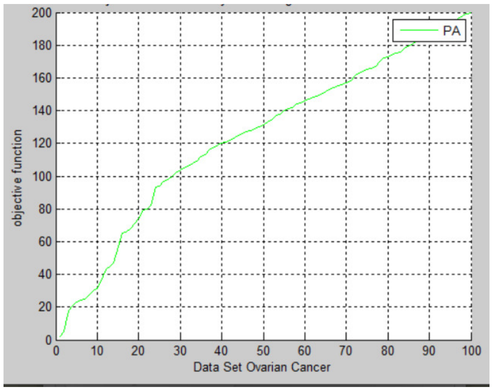


Fig. 14. Ovarian cancer dataset objective function analysis between existing algorithm and proposed algorithm.

VI. CONCLUSION

In conclusion, this paper presents a novel multilayer data clustering framework that integrates feature selection and a modified K-Means algorithm, demonstrating superior performance over existing methods on gene data. Moreover, it highlights the significance of addressing noisy or uncertain information for clustering and classification tasks, as evidenced by the enhanced classification accuracy achieved through quadratic discriminant analysis. Moving forward, future research endeavors should focus on exploring additional databases, algorithms, and statistical distributions to further improve clustering and classification outcomes. Comparative studies among diverse algorithms and investigations into semi-supervised classification techniques could provide valuable insights for advancing the field. Furthermore, examining the stability and accuracy of ensembles comprising single clustering algorithms versus those comprising multiple clustering algorithms would be a promising avenue for future exploration. Overall, these endeavors aim to enhance the robustness and efficacy of unsupervised learning techniques in handling complex, real-world datasets.

REFERENCES

[1]. Elavarasi, S. A., Anitha, & Akilandeswari, J. (2014). Survey on clustering algorithm and similarity measure for categorical data. *ICTACT Journal on Soft Computing, 4*(2), 715-722.

[2]. Han, J., & Kamber, M. (2000). Data Mining Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann.

[3]. Elavarasi, S. A., Akilandeswari, J., & Sathiyabhama, B. (2011). A Survey on Partition Clustering Algorithms. *International Journal of Enterprise Computing and Business Systems, 1*(1), 1-14.

[4]. Sarker, I. H. (2022). AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science, 3*(2), 1–20.

[5]. Bonaccorso, G. (2017). Machine learning algorithms 3. Sarker, I. H. (2021). Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science, 2*(5), 1–22.

[6].Han, J., Pei, J., & Kamber, M. (2011). Data mining: Concepts and techniques. McGraw-Hill/Irwin.

[7]. Olson, D. L., Shi, Y., & Shi, Y. (2007). Introduction to business data mining (Vol. 10). McGraw-Hill/Irwin.

[8]. Sarker, I. H., Colman, A., Han, J., & Watters, P. A. (2021). Context-aware machine learning and mobile data analytics: Automated rule-based services with intelligent decision-making. Springer Nature.

[9].Khan, S. S., & Ahmad, A. (2004). Cluster center initialization algorithm for kmeans clustering. *Pattern Recognition Letters, 25*(11), 1293–1302.

[10].Jain, A., & Dubes, R. (1988). Algorithms for clustering data. Prentice Hall.

[11]. Bishop, C. M. (2008). Pattern recognition and machine learning. Springer.

[12]. Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques (2nd ed.). Morgan Kaufmann.

[13]. Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 2008 SIAM International Conference on Data Mining (pp. 243–254).

[14]. Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access, 7*, 31883–31902.

[15]. Wang, S., Li, M., Hu, N., Zhu, E., Hu, J., Liu, X., & Yin, J. (2019). K-means clustering with incomplete data. *IEEE Access, 7*, 69162–69171.

[16]. Pugazhenthi, A., & Kumar, L. S. (2020). Selection of Optimal Number of Clusters and Centroids for K-means and Fuzzy C-means Clustering: A Review. In 2020 *5th International Conference on Computing, Communication and Security (ICCCS)* (pp. 1-4). IEEE.

[17]. Liu, X., Zhu, X., Li, M., Wang, L., Zhu, E., Liu, T., & Gao, W. (2019). Multiple kernel $ k $ k-means with incomplete kernels. *IEEE transactions on pattern analysis and machine intelligence*, *42*(5), 1191-1204.

[18]. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access, 8,* 80716–80727.

[19]. Wu, J., Liu, H., Xiong, H., Cao, J., & Chen, J. (2015). K-Means-Based Consensus Clustering: A Unified View. *IEEE Transactions on Knowledge and Data Engineering, 27*(1), 155–169.

[20]. Honda, K., Nonoguchi, R., Notsu, A., & Ichihashi, H. (2011). PCA-guided k-Means clustering with incomplete data. In 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011) (pp. 1710-1714). IEEE.

[21]. Selim, S. Z., & Ismail, M. A. (1984). K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(1), 81–87.

[22]. Yoon, K. A., Kwon, O. S., & Bae, D. H. (2007). An Approach to Outlier Detection of Software Measurement Data using the K-means Clustering Method. In First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007) (pp. 443-445). IEEE.

[23]. Ye, Y., Liu, X., Liu, Q., & Yin, J. (2017). Consensus kernel-means clustering for incomplete multiview data. Computational Intelligence and Neuroscience.

[24]. Liu, H., Shao, M., & Fu, Y. (2018). Feature selection with unsupervised consensus guidance. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2319–2331.

[25]. Law, M. H. C., Figueiredo, M. A. T., & Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(9), 1154–1166.

[26]. Li, Z., Liu, J., Yang, Y., Zhou, X., & Lu, H. (2013). Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Transactions on Knowledge Data Engineering, 26*(9), 2138–2150.

[27]. Zhao, Z., & Liu, H. (2007). Semi-supervised feature selection via spectral analysis. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 641-646). Society for Industrial and Applied Mathematics.

[28]. Liu, H., & Fu, Y. (2015). Clustering with Partition Level Side Information. In 2015 IEEE International Conference on Data Mining (pp. 877-882). IEEE.

[29]. Tao, Z., Liu, H., & Fu, Y. (2018). Partition Level Constrained Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(10), 2469–2483.

[30]. Li, J., Tang, J., & Liu, H. (2017). Reconstruction-based Unsupervised Feature Selection: An Embedded Approach. *In IJCAI* (pp. 2159-2165).

[31]. Feng, Y., Xiao, J., Zhuang, Y., & Liu, X. (2013). Adaptive unsupervised multi-view feature selection for visual concept recognition. In Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November

How to cite this article: P.S. Deshmukh, M. Sivakkumar and Varshaha Namdeo (2020). A Strategy for Improved Data Classification using Advanced clustering with Incomplete Datasets. *International Journal on Emerging Technologies*, *11*(5): 746–755.