Int. j. adv. multidisc. res. stud. 2026; 6(2):1541-1544

Received: 07-02-2026
Accepted: 07-04-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

On the Applications of K-means Algorithm for Pattern Recognition in Large Data. An Informal Approach: Pattern Recognition using K-means Clustering a Comparative Study of COVID-19 Trade Data and Employment Index Data

[1] Elhadi AA Suiam, [2] Awad H Ali

[1, 2] Faculty of Computer Science, Department of Computer Sciences, Graduate College, Al-Neelain University, Khartoum, Sudan

DOI: https://doi.org/10.62225/2583049X.2026.6.2.6134
Corresponding Author: **Elhadi AA Suiam**

Abstract

This report presents a comparative study on the use of the K-means clustering algorithm for pattern recognition in two large, real-world datasets: COVID-19 trade impact data and employment index data. The objective is to examine how data characteristics such as volatility, stability, and the presence of outliers' influence clustering quality and decision-making usefulness. Using recent literature, the study demonstrates that while K-means effectively structures both datasets into meaningful clusters, employment index data yields more stable and interpretable patterns suitable for long-term planning. In contrast, COVID-19 trade data, due to its highly dynamic nature, is more appropriate for short-term exploratory analysis. These findings align with recent advances in clustering research that emphasize the importance of data distribution and centroid stability (Awad & Hamad, 2022; Selmi *et al*., 2024) [2, 5].

Keywords: K-Means Clustering, PCA, Large Data, Pattern Recognition, Dimensionality Reduction

1. Introduction

The increasing availability of large-scale datasets in public health and economics has intensified the need for automated techniques capable of extracting meaningful insights without labeled data. Unsupervised learning methods, particularly clustering algorithms, are widely used for this purpose because they reveal hidden structures and trends in complex datasets. Recent studies highlight clustering as a key component of modern data mining and pattern recognition systems (ScienceDirect Review, 2023) [4].

Among clustering algorithms, K-means remains one of the most widely used due to its simplicity, scalability, and computational efficiency. Despite the development of more complex clustering techniques, K-means continues to be favored in large numerical datasets because of its interpretability and low computational cost (Awad & Hamad, 2022; Seyam *et al*., 2025) [2, 6]. This report applies K-means to COVID-19 trade data and employment index data to analyze how differences in data stability affect clustering performance and practical usefulness.

2. Dataset Description

2.1 COVID-19 Trade Impact Dataset

The COVID-19 trade dataset reflects the economic disruptions caused by the pandemic and includes structured numerical attributes such as year, trade value, and cumulative trade value. The dataset contains over 100,000 records, making it suitable for large-scale clustering analysis. However, pandemic-related lockdowns, policy interventions, and global supply chain disruptions introduce sharp fluctuations and extreme values, increasing data volatility (World Trade Organization, 2021) [7].

2.2 Employment Index Dataset

The employment index dataset represents labor market conditions over time through indicators such as employment values, magnitude, and time period. Employment data typically evolves gradually and exhibits fewer extreme outliers compared to pandemic-driven datasets. Recent economic studies confirm that such stable numerical indicators are well suited for

centroid-based clustering methods like K-means (OECD, 2021; Ardiansyah *et al*., 2024) [3, 1].

3. Methodology

3.1 Introduction

This chapter details the methodological framework adopted to achieve the research objectives, namely the development, enhancement, and empirical evaluation of a K-means clustering algorithm optimized for large-scale, dynamic datasets.

The study combines theoretical insight, algorithmic innovation, and empirical validation, following rigorous scientific procedures to ensure reproducibility, reliability, and interpretability. The methodology aims to uncover latent structures in complex datasets and provide a robust foundation for predictive analytics and data-driven decision-making.

Clustering is a core unsupervised learning technique, and the research focuses on extending K-means to overcome its traditional limitations: sensitivity to initial centroids, convergence to local minima, and inefficiency in large datasets. By integrating enhanced initialization, adaptive convergence, similarity-based refinement, and normalization, the proposed approach addresses these gaps while preserving computational feasibility.

3.2 Research Design.

A mixed-methods research design was adopted to ensure a comprehensive evaluation of the improved algorithm:

Quantitative experiments to measure clustering performance using multiple real-world and synthetic datasets across diverse domains.

Qualitative assessments to interpret the meaning of clusters, validate outcomes against expert knowledge, and examine practical implications.

This design allows triangulation of findings, balancing numerical accuracy with interpretability and application relevance.

3.3 Theoretical Foundation of K-Means Clustering.

K-means clustering is a partition-based algorithm that organizes n observations into k disjoint clusters. Each data point is represented as:

$$xi=(xi1,xi2,\ldots,xid),i=1,2,\ldots,nx_i = (x_{i1}, x_{i2}, \ldots, x_{id}), \quad i = 1,2,\dots,nxi=(xi1,xi2,\ldots,xid),i=1,2,\ldots,n$$

where d is the number of attributes.

The objective is to minimize the within-cluster sum of squares (WCSS):

$$J(C)=\sum i=1k\sum x\in Ci\|x-\mu i\|2(3.1)J(\mathbf{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2 \tag{3.1}J(C)=i=1\sum k x\in Ci\sum \|x-\mu i\|2 \quad (3.1)$$

where C_i is cluster i, μ_i is the centroid of C_i, and | x - μ_i | is the Euclidean distance between a point and its centroid.

The algorithm iterates the following steps:

Initialize k centroids randomly or via enhanced seeding.

Assign each data point to the nearest centroid.

Recompute centroids as the mean of points in each cluster.

Repeat until convergence is achieved (no change in cluster assignment or minimal centroid movement).

3.4 Algorithm Selection and Justification

K-means was selected because of its simplicity, scalability in large datasets, efficiency in iterative computation, and strong empirical baseline performance. Despite alternatives (K-medoids, Fuzzy C-means, Genetic K-means), K-means allows structured enhancement and rigorous comparative evaluation across multiple real-world datasets.

Research by Awad & Hamad (2022) [2] and Seyam *et al*. (2025) [6] supports K-means as a robust choice for big data applications, particularly when optimized with enhanced initialization and convergence strategies.

3.5 Proposed Algorithm Enhancements

3.5.1 Enhanced Centroid Initialization

Traditional K-means is sensitive to initial centroids, which may result in poor convergence. To mitigate this, we implemented:

K-means++ Initialization: Probabilistic seeding that prefers distant points, reducing the likelihood of suboptimal local minima.

Genetic Algorithm-Inspired Search: Candidate centroids are evolved to explore the solution space, enhancing diversity and clustering quality.

3.5.2 Adaptive Convergence Criteria

Instead of a fixed number of iterations, dynamic stopping rules are employed:

$$\text{Convergence occurs if } \max_i \|\mu_i(t)-\mu_i(t-1)\| < \epsilon \text{ or } \Delta J < \delta \tag{3.2}$$

$$\text{Convergence occurs if } \max_i \|\mu_i(t) - \mu_i(t-1)\| < \epsilon \text{ or } \Delta J < \delta \quad (3.2)$$

where ε and δ are pre-defined thresholds.

3.5.3 Normalized Distance Metrics

Euclidean distance is sensitive to feature scales. Features are standardized:

$$xj'=xj-\bar{x}_j\sigma j,j=1,\ldots,d(3.3)x_j' = \frac{x_j - \bar{x}_j}{\sigma_j}, \quad j=1,\dots,d \tag{3.3}xj'=σjxj−x¯j,j=1,…,d \quad (3.3)$$

Normalized distance ensures equal contribution of each feature:

$$d(x,\mu)=\sum j=1d(xj'-\mu j')2(3.4)d(x, \mu) = \sqrt{\sum_{j=1}^{d} (x_j' - \mu_j')^2} \tag{3.4}d(x,\mu)=j=1\sum d(xj'-\mu j')2 \quad (3.4)$$

3.5.4 Similarity-Based Refinement Mechanism

Cluster assignment incorporates similarity evaluation: Rank points by Euclidean distance to centroid. Evaluate similarity with cluster structure. Apply majority voting to resolve ambiguous assignments. Include the point if it meets similarity threshold; otherwise, proceed to the next candidate.

3.5.5 Integration
With Other Clustering Approaches Hybrid approaches were explored, such as K-means + Fuzzy C-means and K-means + Genetic Algorithm, to leverage complementary strengths and reduce weaknesses of standard K-means.

3.6 Data Collection and Preprocessing
3.6.1 Dataset Selection
Twenty datasets were selected from multiple domains: medical, financial, biological, socio-economic, and social media, including categorical, continuous, and mixed attributes. All datasets were sourced from the UCI Machine Learning Repository.

3.6.2 Preprocessing Steps
Handle missing values via deletion or imputation. Remove noise and outliers. Standardize features (zero mean, unit variance). Reduce dimensionality for high-dimensional datasets.

3.7 Experimental Setup
3.7.1 Implementation Environment
MATLAB R2025b was used for computational efficiency. Performance Metrics Internal Validity: Silhouette Score, Davies–Bouldin Index, WCSS. External Validity: Centroid Index (CI), Success Rate (%). Computational Metrics: Execution time, memory consumption.

3.7.2 Comparative Analysis
Improved K-means was compared against Standard K-means, Bisecting K-means, Fuzzy C-means, and Genetic K-means. Optimal k values were determined via Elbow Method, Silhouette Analysis, and inter-cluster distance mapping.

3.8 Statistical Analysis and Validation
Variance analysis, significance testing, and cross-validation were conducted to ensure reproducibility and evaluate success probability:

$$E[repeats]=1p(3.5)E[\text{repeats}] = \frac{1}{p} \tag{3.5}E[repeats]=p1$$

3.9 Applications
Facility Location and Service Clustering The algorithm was applied to facility placement optimization, transportation routing minimization, and clustering service requesters based on demand and distance.

3.10 Data Analytics Framework
Clustering serves as preprocessing for classification, prediction, market segmentation, medical diagnosis, and text mining. Cluster-based grouping improves interpretability and accuracy.

3.11 Limitations and Scope
Focused on static datasets, distance-based clustering, primarily Euclidean distance; excludes real-time sensor and time-series data.

3.12 Ethical Considerations
All datasets are public and anonymized, adhering to ethical research standards.

3.13 Summary
The chapter presented a comprehensive methodology combining theoretical refinement, algorithmic enhancement, empirical validation, and real-world application. The proposed approach lays a robust foundation for analyzing complex datasets using improved K-means clustering.

4. Results
4.1 COVID-19 Trade Data Results
The COVID-19 trade data was grouped into three clusters representing pre-pandemic trade behavior, high trade activity periods, and COVID-affected disruption phases. These clusters clearly reflect distinct trade patterns observed during the pandemic. However, sudden spikes and drops in trade values caused noticeable centroid shifts, reducing long-term stability. Similar behavior has been reported in recent pandemic-era clustering studies (World Trade Organization, 2021) [7].

4.2 Employment Index Data Results
Clustering the employment index dataset produced compact and well-separated clusters representing strong, moderate, and weak employment conditions. Cluster centroids remained stable over time, resulting in highly interpretable patterns suitable for economic planning and policy analysis. Recent studies applying K-means to employment and unemployment data report comparable stability and interpretability (Ardiansyah *et al*., 2024) [1].

4.3 Comparative Analysis

Aspect	Employment Index Data	COVID-19 Trade Data
Data stability	High	Low
Outliers	Few	Many
Cluster compactness	Strong	Moderate
Centroid stability	High	Low–Moderate
Decision-making value	Long-term	Short-term

The comparison confirms recent findings that K-means performs best on datasets with smooth distributions and limited volatility (Selmi *et al*., 2024) [5].

4.4 Benchmarking Against Other Clustering Methods
Recent benchmarking studies show that hierarchical clustering scales poorly for large datasets, while DBSCAN is sensitive to parameter selection and struggles with variable density data. Gaussian Mixture Models and spectral clustering introduce higher computational complexity without proportional benefits for large numerical datasets. As a result, K-means remains a preferred choice for scalable pattern recognition tasks (ScienceDirect Review, 2023; Seyam *et al*., 2025) [4, 6].

5. Discussion
The results demonstrate that clustering effectiveness depends not only on algorithm choice but also on data characteristics. Employment index data aligns well with K-means assumptions, producing stable centroids and compact clusters. In contrast, the volatile nature of COVID-19 trade data leads to centroid instability, limiting long-term interpretability. These observations are consistent with

recent research on centroid-based clustering in dynamic environments (Awad & Hamad, 2022; Selmi *et al*., 2024) [2, 5].

5.1 Societal and Research Implications

From a policy perspective, employment data clustering supports long-term labor market planning, workforce development, and economic forecasting. COVID-19 clustering remains valuable for rapid assessment and situational awareness during crises. From a research standpoint, this study provides real-world evidence supporting recent claims that data stability plays a critical role in clustering reliability (Ardiansyah *et al*., 2024) [1].

5.2 Limitations

K-means requires the number of clusters to be defined in advance and is sensitive to outliers. These limitations are widely acknowledged in recent clustering literature and can be mitigated through careful preprocessing and feature scaling (Selmi *et al*., 2024) [5].

6. Conclusion

This study confirms that K-means clustering is an effective and practical approach for pattern recognition in large, unlabeled datasets. While both COVID-19 trade data and employment index data benefit from clustering, employment data produces more stable and interpretable results suitable for long-term decision-making. COVID-19 data, due to its volatility, is better suited for short-term exploratory analysis. These conclusions align with recent advances in K-means research and real-world applications from 2020–2026.

7. References

1. Ardiansyah MFH, Amany N, Anugrah CI, Syafitri UD. K-means clustering application of open unemployment during the COVID-19 period. International Journal of Applied Statistics and Data Science, 2024.
2. Awad FH, Hamad MM. Improved K-means clustering algorithm for big data environments. Electronics. 2022; 11(6):883.
3. Organisation for Economic Co-operation and Development (OECD). Employment outlook and labour market indicators, 2021.
4. ScienceDirect Review. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, 2023.
5. Selmi ATE, Zerarka MF, Cheriet A. Enhancing K-means clustering with post-redistribution. IIETA Journal, 2024.
6. Seyam TA, Hossain MS, Ghose R, *et al*. Next-generation K-means clustering: Performance analysis for big data. International Journal of Intelligent Information Systems, 2025.
7. World Trade Organization. Effects of COVID-19 on global trade, 2021.