Quest Journals
Journal of Software Engineering and Simulation
Volume 11 ~ Issue 12 (December 2025) pp: 01-03
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Research on K-means Clustering and Classification Technology: Principles, Evolution, Challenges, and Prospects

Chu Fang
College of Economics and Management, Zhaoqing University, Zhaoqing City, Guangdong, China

Abstract: *The K-means clustering algorithm, as one of the most classical and widely used algorithms in the field of unsupervised learning, is characterized by its concise and efficient core idea, aiming to partition data into clusters with high intra-cluster cohesion through iterative optimization. Although its original design was for unsupervised data exploration and grouping, through ingenious engineering adaptations, K-means and its variants have demonstrated significant value in numerous fields such as classification, image segmentation, and information retrieval. This paper systematically elaborates on the fundamental principles and mathematical formulation of the classical K-means algorithm and delves into its inherent limitations, including sensitivity to initial centroid selection, the determination of the number of clusters K, and sensitivity to noise. Furthermore, the paper reviews the main improved algorithms developed to address these limitations, such as K-means++, ISODATA, and Kernel K-means, and discusses the extended applications of K-means in supervised and semi-supervised classification scenarios. Finally, the paper summarizes the major challenges in current research and provides prospects for future research directions, particularly regarding its integration with deep learning.*

Keywords: *K-means; Cluster analysis; Unsupervised learning; Machine learning; Classification technology; Data mining*

*Received 07 Dec., 2025; **R**evised 15 Dec., 2025; **A**ccepted 18 Dec., 2025 © The author(s) 2025. **P**ublished with open access at **www.questjournas.org***

I. Introduction

In the era of big data, automatically extracting valuable structures and patterns from massive, high-dimensional, and complex data has become a critical task. Cluster analysis in machine learning aims to partition samples in a dataset into several disjoint subsets (called "clusters"), so that samples within the same cluster have high similarity, while samples between different clusters have low similarity. As a partitional clustering method, the K-means algorithm, with its intuitive concept, simple implementation, fast convergence, and strong scalability, has become one of the most popular clustering tools in academia and industry since its proposal by J.B. MacQueen in 1967.

Although clustering itself belongs to the category of unsupervised learning, the intrinsic grouping structure of data revealed by K-means provides crucial prior knowledge for subsequent classification tasks. For example, in image classification, K-means can be used to generate a visual vocabulary; in document classification, it can be used for topic discovery; in anomaly detection, it can be used to identify outliers. Therefore, research on K-means clustering and classification technology aims not only to optimize its unsupervised clustering performance but also to explore how it can effectively serve broader classification and recognition goals. This paper aims to systematically review and discuss the core principles, developmental trajectory, application extensions, and future trends of this technology.

Principles and Process of the Classical K-means Algorithm

The core objective of the K-means algorithm is to minimize the sum of squared distances between samples within a cluster and its cluster center, i.e., to minimize the following cost function (also known as the distortion function):

$J = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} ||x - \mu_i||^2$

where k is the pre-specified number of clusters, C_i represents the set of samples belonging to the i-th cluster, μ_i

is the center of the i-th cluster (i.e., the mean of all samples in that cluster), and ||x - μ_i|| represents the Euclidean distance between sample x and center μ_i.

The classical K-means algorithm employs an iterative optimization strategy. Its standard procedure (Lloyd's algorithm) is as follows:
1. Initialization**: Randomly select k samples from the dataset as the initial cluster centers.
2. Assignment Step**: For each sample x_i in the dataset, calculate its distance to all k cluster centers and assign it to the cluster whose center is the nearest.
3. Update Step**: For each cluster C_j, recalculate its cluster center μ_j as the mean (centroid) of all samples belonging to that cluster.
4. Iteration and Termination**: Repeat steps 2 and 3 until a termination condition is met. Termination conditions typically include: cluster center positions no longer change significantly, or a pre-set maximum number of iterations is reached.
The advantage of this algorithm lies in its time complexity of O(n*k*t), where n is the number of samples, k is the number of clusters, and t is the number of iterations, providing good scalability for large-scale datasets.

Inherent Limitations of the K-means Algorithm
Despite its widespread application, K-means has several key defects that directly affect its clustering and subsequent classification effectiveness:
1. Sensitivity to Initial Cluster Centers**: Random initialization may cause the algorithm to converge to a local optimum rather than the global optimum, leading to significantly different results across different runs.
2. Requirement to Pre-specify the Number of Clusters K**: In practical applications, the optimal K value is usually unknown. An incorrect setting of K leads to unreasonable cluster structures.
3. Sensitivity to Noise and Outliers**: Since the mean is used as the center, outliers can significantly pull the center's location, causing distortion in cluster positioning.
4. Preference for Spherical Clusters**: The Euclidean distance metric inherently assumes clusters are convex and isotropic. It performs poorly on complex data structures that are non-spherical, manifold, or have uneven densities.
5. Applicability Primarily to Numerical Data**: The calculation of Euclidean distance requires data to be numerical features, making it weak in directly handling categorical data.

Major Improved Algorithms and Research Progress
To overcome the above limitations, researchers have proposed numerous improvement schemes:
1. Initialization Optimization**: The **K-means++ algorithm is a milestone improvement. It selects initial centroids in a probabilistic manner, aiming to spread them as far apart as possible. Specifically, the first center is chosen randomly, and each subsequent center is chosen with a probability proportional to the squared shortest distance to any already chosen center. This significantly improves the algorithm's stability and the quality of the final clustering.

2. Current Challenges and Future Prospects
Despite extensive research, K-means and its variants still face several cutting-edge challenges:
Ultra-high-dimensional and Large-scale Data**: As dimensionality increases, distance metrics tend to become ineffective ("curse of dimensionality"), and computational efficiency becomes a challenge. Future research needs closer integration with subspace clustering, hashing learning, and distributed computing.
Dynamic Stream Data Clustering**: Traditional K-means targets static datasets. For real-time data streams, incremental or online variants of K-means need to be developed to dynamically adapt to changes in data distribution.

II. Conclusion

As a foundational method in data mining and machine learning, the vitality of the K-means clustering algorithm stems from the simplicity of its idea and the flexibility of its extensions. This paper systematically reviews its evolution from the classical algorithm to various improved variants, which continuously broaden its application boundaries and robustness. More importantly, K-means has transcended purely unsupervised analysis and plays a significant role in classification tasks through its combination with patterns like feature engineering and semi-supervised learning.

Looking ahead, K-means technology is not obsolete but is entering a new phase of deep integration with deep learning and big data computing paradigms. Faced with ultra-high-dimensional, streaming, and complex-structured data, as well as higher demands for interpretability and reliability, the core idea of K-means will continue to inspire new algorithmic innovations and play an indispensable role in various application fields of artificial intelligence. Research in this area holds not only theoretical value but also broad practical significance.

References

[1]. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.
[2]. Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of SODA*.
[3]. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*.
[4]. Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*.
[5]. Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. *Proceedings of ICML*.