K-means Algorithm of Big Data In ML

Ghadeer Suleiman

July 9, 2024

K-means Algorithm Of Big Data In ML

Ghadeer Zidan Suleiman
Prince Hussein bin Abdullah College of Information Technology
Al al-Bayt University
Irbed, Jordan
2220901006@st.aabu.edu.jo

Abstract: Artificial intelligence, particularly machine learning, is crucial to emulating human cognitive processes. In the field of machine learning, unsupervised learning plays a vital role in identifying commonalities in datasets and in creating the K-means clustering algorithm constitutes a crucial method of effect. This study examines the importance of K-means clustering in machine learning, particularly in the area of big data analytics. The K-means algorithm is renowned for its simplicity and its ease of aggregation and analysis of data, allowing scientists to extract useful information and provide insights. Reviews of the literature discuss applications of K-means clustering in general data analysis, huge data clustering, travel advice, and financial fraud detection. These results highlight the need to improve K-means to achieve higher levels of accuracy and scalability in different data analysis applications.

Index Terms-Clustering, K-means Algorithm,

I.INTRODUCTION

Artificial intelligence refers to computer programs' behaviors and specific characteristics that allow them to simulate human thought processes and cognitive functions. The most crucial aspect is the system's ability to notice, infer, and respond to events not preprogrammed into it. Machine learning, a subfield of computer science and artificial intelligence, focuses on using data to make AI more accurate over time by mimicking human learning. Supervised learning and unsupervised learning are the two subtypes of machine learning. Finding clusters of related examples in the data can be the objective of unsupervised learning problems; this process is known as clustering.[1]

The k-mean method is widely recognized as one of the clustering algorithms in the field of Data analysis.

Its simplicity, scalability and efficiency make it a valuable tool for aggregating and analyzing large data sets. Using this algorithm, data scientists and researchers can extract meaningful information and make decisions based on the collected data. This paper will discuss some scientific papers on (K-means Clustering Algorithm) and its applications in machine learning.

A. *Machine Learning*

Arthur Samuel introduced machine learning as a field that enables computers to learn without explicit programming. Machine learning (ML) efficiently handles data and extracts relevant information from large datasets. With the abundance of available datasets, the demand for machine learning is increasing, and various industries use it to extract relevant data. Machine learning relies on various algorithms to solve data problems, with no single one-size-fits-all solution. The choice of algorithm depends on the problem, the number of variables, and the best model for the problem. Data scientists emphasize that there is no single one-size-fits-all algorithm for machine learning.[2]

1) *Unsupervised learning*

Unsupervised learning will be the future of technology, as it is used in product recommendation, Google translation, and other applications. The K-means cluster strategy easily finds similarities among data elements and forms clusters based on these similarities. Clusters can be created by taking the distance of each element from the other using the Euclidian distance formula. Unsupervised learning is also being used in product recommendation, where machine learning algorithms are applied to track online behavior and display advertisements for the same product. Semi-supervised machine learning algorithms may be developed in the future, which falls between supervised learning using label data and unsupervised learning using unlabeled data.[1]

Unsupervised learning methods acquire a limited number of characteristics from the data. Upon the introduction of new data, it utilizes the previously acquired characteristics to accurately identify the data's class.

Its primary use is for clustering and feature reduction.[2]

B. *K-means clustering*

1. Clustering

is a technique used in unsupervised machine learning to group similar objects together. The goal is to find similarities in a data set. Clustering algorithms are most commonly used in unsupervised learning, such as in a fruit basket containing different types of fruits. They first separate fruits based on color and other attributes like size and shape. This process creates a cluster.[1,3]

2. K-means clustering

is the most popular unsupervised learning algorithm, which divides data and objects into different clusters. The algorithm works by randomly selecting a number K, assigning each data point to the nearest barycenter, determining the change value, reallocating data points, and iterating to find the nearest barycenter for each cluster.[1,7]

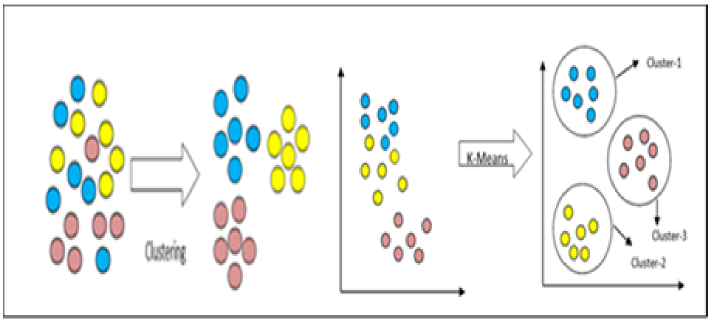
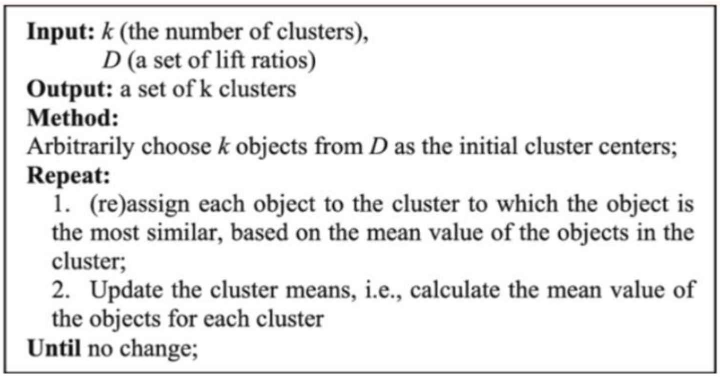


Figure 1 Clustering process using k-means algorithm [1]

3. How does a k-means algorithm work?



```
Input: k (the number of clusters),
       D (a set of lift ratios)
Output: a set of k clusters
Method:
Arbitrarily choose k objects from D as the initial cluster centers;
Repeat:
    1. (re)assign each object to the cluster to which the object is
    the most similar, based on the mean value of the objects in the
    cluster;
    2. Update the cluster means, i.e., calculate the mean value of
    the objects for each cluster
Until no change;
```

Figure 2 Pseudocode of k-means Algorithm[5]

Figure 2 shows Pseudocode of k-means clustering algorithm. At first, we select k centroids, where k is a value specified by the user, representing the desired number of clusters. Subsequently, every individual point is allocated to the nearest centroid, and a group of points assigned to a centroid is referred to as a cluster. Subsequently, the centroid of each cluster is recalculated using the points that have been allocated to that cluster. The assignment and update phases are iteratively repeated until there are no changes in the clustering of points, or until the centroids stay unchanged.

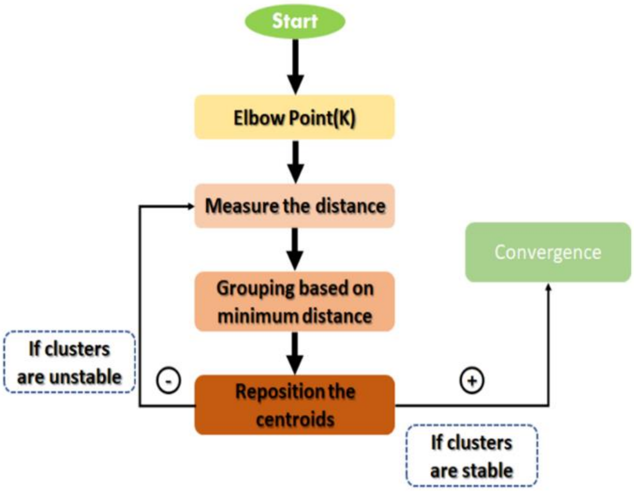


Figure 3 The flowchart below shows how k-means clustering work [6]

The flowchart shows how k-means clustering works. By evaluating various values of K, the K-means algorithm first estimates the number of clusters (K). The algorithm calculates the distance between each data point and the centroids of the clusters, usually employing the Euclidean distance metric. We categorize the data points into clusters based on their minimal distance to the centroids. We assign each point to the cluster whose centroid is closest to it. We recalculate the centroids by averaging the data points within each cluster until we achieve a stable state. If the clusters exhibit instability, the algorithm will return to the process of calculating distances and readjust the placement of the centroids. If the clusters are stable, the procedure continues. The last stage is "convergence," which indicates the effective categorization of data points into separate clusters.

II.RESEARCH METHODOLOGY

This review paper's methodology includes proposing a K-means-based big data clustering algorithm, implementing it using Spark, simulating it with large-scale data, highlighting its main contributions [3], applying K-means-based machine learning clustering in financial fraud detection, and comparing it to traditional rule-based methods. It also discusses the effectiveness of K-means clustering and emphasizes the importance of choosing the optimal K value. The study aims to improve fraud detection strategies using innovative detection frameworks [4], use social media platforms to collect tourist targets, enhance the k-means algorithm with a genetic algorithm to determine the number of clusters and select initial seeds, and apply this approach to recommend the best tourist route. We tested it in Red Sea State, Sudan [7], Use principal component analysis (PCA) and the percentile concept to efficiently estimate initial centroids in a K-means clustering algorithm to reduce the number of iterations and execution time [8].

- How can machine learning-based K-means clustering improve financial fraud detection in the digital financial environment, given rising fraudulent activities?[4]
- How can we automate the determination of the number of clusters in the K-Means algorithm to handle big data clustering effectively?[3]
- How may social media information improve the k-means algorithm's recommendation of the best tourist routes?[7]
- How can we effectively choose the starting centroids of the conventional K-means clustering method to minimize the number of iterations and execution time, thereby improving its performance?[8]

III.LITERATURE REVIEW

The study by Zengyi Huang and his coworkers, 2024 suggests a machine learning-based K-means clustering method for financial fraud detection. To increase detection accuracy and efficiency, our method clusters a lot of financial transaction data. It detects odd patterns and behaviors, making it possible to quickly identify fraudulent activities. Compared to conventional rule-based approaches, this strategy improves detection accuracy and flexibility. By concentrating monitoring and preventative efforts on high-risk regions, it also aids financial institutions in more effectively allocating their resources. The strategy is to give the financial industry a more reliable and secure transaction environment. Nevertheless, K-means is greatly influenced by the selection of K value. The project's goal is to build a fraud detection classifier using supervised learning techniques while taking class imbalance and data quantity into account. According to the study, clusters 1 and 2 had almost no fraud cases, whereas cluster 3 had most of them. The financial security sector may become more adaptable and nimbler by using the K-means clustering approach to detect financial fraud.[4]

Another research by Ankita Sinha and Prasanta K Jana, 2016 Presents a K-Means-based clustering technique for big data that automates the number of clusters to manage large-scale data sets. When implemented using Spark, the approach performs better on large-scale data sets than the K-Means algorithm in the Spark Machine Learning Library. Large-scale synthetic data sets and real-world data are used to extensively run the method on a 4-node cluster, showing superior performance compared to the scalable K-Means++ implemented in Spark's MLLIB package. The study aims to create a vast data clustering method based on K-Means, automate the number of clusters, use Spark to apply the algorithm and solve the over-resolution issue. The method also considers the sampling strategies used in the initial and final MR operations, which can cause problems when used on large data sets. In the future, the authors want to expand the system to accommodate substantial real-time streaming data sets.[3]

A study conducted by Damos and colleagues 2024 The k-means algorithm for tourist path suggestions is improved in this work by combining survey and social media tourism data. The primary contribution is the application of the genetic algorithm (GA) to choose the first seeds, count the clusters (k), and, using social media tourism data, suggest the optimal tourist itinerary. This technique optimizes in 0.01 seconds after five iterations. Along with tourism goals from national and international agencies and regional polls, the data covers popular social media platforms like Facebook, WhatsApp, WeChat, and TripAdvisor. By tackling issues including data overlap, massive dataset management, and k-means algorithm execution time reduction, the GA helps to overcome the drawbacks of previous methods. Through clustered and optimized tourism objectives, the improved GA solves the traveling salesman problem (TSP) and suggests the optimal tourist route. Future research should focus mostly on the number of tourist goals considered and the integration of internal and external factors in evaluating web users' behavior in tourism data analysis. The purpose of this work is to gather trip goals from social media platforms, propose the optimal timing to use the genetic algorithm, and enhance the k-means algorithm using the genetic algorithm.[7]

Research done by Md. Zubair et al, 2022 An enhanced K-means clustering technique is presented in this work that effectively locates the best starting centroids to shorten execution times and iterations. The approach minimizes iterations and gives the ideal number of constant iterations for implementing the algorithm by using Principal Component Analysis (PCA) and percentile notions. The technique is examined on a synthetic dataset of 10 million instances with 8 dimensions as well as real-world datasets like COVID-19 and patient records. The suggested method is shown by experimental findings to be more efficient in terms of computing time and iterations than conventional kmeans++ and random centroids initialization techniques. The paper presents an algorithmic overview of the proposed method, comparing it with random centroid selection and kmeans++ centroid selection methods, and comparing it more generalized in different fields. Easy to use and requires no additional setup, the approach is appropriate for a wide range of real-world applications, such as security, IoT, smart city services, and personalized services.[8]

IV.RESULTS

The studies reviewed in this paper demonstrate the effectiveness and versatility of the K-means clustering algorithm in various applications:

Fraud Detection Financial

Financial fraud detection accuracy and efficiency were increased over traditional rule-based methods using the machine learning-based K-means clustering method.

The discovery of unusual patterns and behaviors made possible by clustering financial transaction data made it possible to quickly identify fraudulent activity.

Financial institutions may be able to better distribute resources, as seen by the study, which revealed that clusters 1 and 2 had virtually no fraud instances while cluster 3 had most of them.[4]

big data clustering

When applied with Spark, the K-Means-based clustering method for huge data outperformed the K-Means algorithm in the Spark Machine Learning Library and automated the number of clusters.

Using both large-scale synthetic data sets and real-world data, the method outperformed the scalable K-Means++ included in Spark's MLLIB package.[3]

Notes on Tourist Routes

The best travel schedule was recommended in 0.01 seconds after five iterations using the improved K-means algorithm, which employed genetic algorithms to maximize initial seed selection and the number of clusters.

Overcoming the limitations of earlier techniques, the approach addressed problems like data overlap, enormous dataset management, and k-means algorithm execution time reduction.[7]

Effective Setting of Centroids

Principal Component Analysis (PCA) and percentile concepts were employed in the improved K-means clustering method to find the optimal beginning centroids, which reduced iterations and provided the optimal number of constant iterations for algorithm implementation.

The proposed approach was shown by experimental results to be more time and iteration-efficient than traditional kmeans++ and random centroids initialization methods.

These findings demonstrate the ability of the K-means clustering technique to optimize performance by appropriate centroid initialization, manage huge data clustering, and suggest the best tourist itineraries.[8]

V.CONCLUSION

The K-means clustering algorithm is well recognized as a robust and adaptable tool in the realm of machine learning, exhibiting a multitude of applications across many areas. The papers examined in this paper emphasize many significant progressions and uses of the K-means algorithm:

By using machine learning algorithms, the study conducted by Huang and colleagues in 2024 Improved identification of financial fraud The K-means method utilizes machine learning techniques to cluster financial transaction data, allowing for the identification of aberrant patterns and behaviors. This enables a more efficient and precise detection of fraudulent actions, beyond the capabilities of standard rule-based systems. Financial organizations can efficiently allocate their resources and a more secure transaction environment for their customers. The unique K-means variation created by Sinha and Jana 2016 includes an automated method to determine the number of clusters. This feature enables rapid processing of large-scale datasets, thus enabling efficient clustering of large data on Spark. These methods outperform previous scalable K-means algorithms in performance, making them very suitable for real-world applications that include huge amounts of data.

In 2024, Damos and his colleagues enhanced the K-Means method by utilizing a genetic algorithm to propose the most efficient routes based on social media and data. This technique tackles issues such as data overlap, management of big datasets, and minimizing execution time, leading to the production of the best routes within a brief timeframe.

In 2022, Zubair et al. proposed a technique known as Efficient Initialization of the by employing techniques like component analysis and percentile concepts, effective approaches for initializing centroids in K-means can significantly decrease runtime and the number of iterations compared to random and K-means++ methods. These enhancements optimize the effectiveness and flexibility of the algorithm for many real-world applications, including security, Internet of Things (IoT) services in smart cities, and customized services.

The K-means method is continuously developing and adjusting to fulfill the increasing demands of data analysis and clustering in the age of big data and artificial intelligence. The achievements highlighted in this study illustrate the algorithm's ability to effectively and actively use developing domains for continued growth and development. As research advances in this area, we expect to observe more revolutionary applications of K-means clustering that expand the limits of what can be achieved with machine learning.

VI.REFERENCES

[1] M. Suyal and S. Sharma, “A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning,” Journal of Artificial Intelligence and Systems, vol. 6, no. 1, pp. 85–95, 2024, doi: 10.33969/AIS.2024060106.

[2] B. Mahesh, “Machine Learning Algorithms-A Review,” International Journal of Science and Research, 2018, doi: 10.21275/ART20203995.

[3] J. (Telecommunications engineer) Wu, LNM Institute of Information Technology, IEEE Communications Society, M. IEEE Systems, and Institute of Electrical and Electronics Engineers, 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) : September 21-24, 2016, the LNM Institute of Information Technology (LNMIT), Jaipur, India.

[4] Z. Huang, H. Zheng, C. Li, and C. Che, “Application of Machine Learning-Based K-means Clustering for Financial Fraud Detection,” 2024.

[5] Fritz, “Understanding the mathematics behind K-means clustering,” Fritz ai, https://fritz.ai/mathematics-behind-k-means-clustering/ (accessed May 9, 2024).

[6] P. K. Anwla, “K-means,” TowardsMachineLearning, https://towardsmachinelearning.org/k-means/ (accessed May 3, 2024).

[7] M. A. Damos et al., “Enhancing the K-Means Algorithm through a Genetic Algorithm Based on Survey and Social Media Tourism Objectives for Tourism Path Recommendations,” ISPRS International Journal of Geo-Information, vol. 13, no. 2, Feb. 2024, doi: 10.3390/ijgi13020040.

[8] M. Zubair, M. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, “An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling,” Annals of Data Science, 2022, doi: 10.1007/s40745-022-00428-28.