

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION



journal homepage: www.joiv.org/index.php/joiv

A Multi-Agent K-Means Algorithm for Improved Parallel Data Clustering

Mohammed Ahmed Jubair^{a,b}, Salama A. Mostafa^{b,1}, Aida Mustapha^c, Zirawani Baharum^{d,2}, Mohamad Aizi Salamat^b, Aldo Erianda^e

^a Department of Computer Technical Engineering, College of Information Technology, Imam Ja'afar Al-Sadiq University, Al-Muthanna, Iraq
 ^b Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
 ^c Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, 84600 Panchor, Johor, Malaysia
 ^d Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur, Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750 Johor, Malaysia
 ^e Department of Information Technology, Politeknik Negeri Padang, Sumatera Barat, Indonesia

Corresponding author: ¹salama@uthm.edu.my; ²zirawani@unikl.edu.my

Abstract—Due to the rapid increase in data volumes, clustering algorithms are now finding applications in a variety of fields. However, existing clustering techniques have been deemed unsuccessful in managing large data volumes due to the issues of accuracy and high computational cost. As a result, this work offers a parallel clustering technique based on a combination of the K-means and Multi-Agent System algorithms (MAS). The proposed technique is known as Multi-K-means (MK-means). The main goal is to keep the dataset intact while boosting the accuracy of the clustering procedure. The cluster centers of each partition are calculated, combined, and then clustered. The performance of the suggested method's statistical significance was confirmed using the five datasets that served as testing and assessment methods for the proposed algorithm's efficacy. In terms of performance, the proposed MK-means algorithm is compared to the Clustering-based Genetic Algorithm (CGA), the Adaptive Biogeography Clustering-based Genetic Algorithm (ABCGA), and standard K-means algorithms. The results show that the MK-means algorithm outperforms other algorithms because it works by activating agents separately for clustering processes while each agent considers a separate group of features.

Keywords- K-means; decision-making; clustering; multi-agent system.

Manuscript received 14 Dec. 2021; revised 29 Jan. 2022; accepted 12 Apr. 2022. Date of publication 31 May 2022. International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.



I. INTRODUCTION

Data clustering [1] is an emerging and dynamic area of data mining research with active interest due to its importance in a variety of fields of research, including data mining, statistics, spatial database technology, machine learning, web search, information retrieval, biology, marketing, and others. Clustering is a type of observation-based learning rather than learning by examples because it is an unsupervised approach to classification that lacks labeled information. Objects are partitioned into separate groups during clustering, with objects in one group being connected to objects in another but differing significantly [2]. A suitable cluster is established when there is a high level of relatedness between objects in one group but is very dissimilar from items in other groups, hence offering a technique for describing data object relationships. From some data, f: DC D=d (1,) d (2,),...,d n to cluster C=c(1,) c(2,),...,c n based on d i similarity.

K-means is an unsupervised learning system that uses unlabelled data (data that does not have clearly defined groups or categories). K-means clustering divides a dataset into k groups (a specific number of clusters or groups) [3]. However, the conventional K-means algorithm and its extensions have a significant fault in that they can only converge to local minima., which is especially problematic when dealing with the clustering issue of the minimal sum of squares. As a result, several studies have concentrated on refining. In order to maximize the chances of reaching the global optimum, the algorithm's convergence pattern must be improved [4] or at least higher-quality local minima.

On the other hand, A Multi-Agent System is a promising method employed in various disciplines (MAS). Agents [5]–[9] are a collection of numerous interacting computational elements that make up MASs. Each agent can choose how to achieve its objectives and interact with the others [10]–[13]. The ability to interact with one another, cooperate and exchange information, and act or make self-decisions are

among the characteristics of these agents. Machine learning, mobile ad hoc networks, and customizable autonomy are some of the application domains of MAS [14]–[16]. These agents are used in these fields to improve operations' speed, flexibility, efficiency, scalability, and the reusability of system modules.

The proposed technique used these qualities to create a suitable clustering solution for data-driven self-organized pattern problems in this study. According to the given solution idea, every single agent is assigned to an object in the data for collaboration with the main agent to produce a steady group. Each agent is required to sense its immediate surroundings [13]–[16] as part of the number of attributes.

In the proposed MK-means technique, the K-means algorithm interacts with MAS by giving agents roles such as observing the borders of each group/cluster and coordinating the corresponding measurement groups (i.e., the voting process, the distance between clusters). This study suggests a collaborative strategy between the agent system and K-means to improve the overall clustering process. This is how the rest of the paper is organized. The clustering in Section 2 is depicted. The proposed MK-means method is introduced in Section 3, and the evaluation metrics are discussed in Section 4. The research findings are discussed in Section 5, while the conclusion is presented in Section 6.

II. MATERIALS AND METHODS

A. General overview

Clustering is a method used to classify data into groups. Fig. 1 depicted a simple example of clustering [17].



Fig. 1 Clustering example

Partitional or hierarchical clustering processes exist [1]. Based on the design of the hierarchical breakdown, hierarchical clustering is further classified as divisive or agglomerative. For the agglomerative method, clusters are generated from the bottom up, whereas clusters are formed from the top down for the divisive method. Single, average, complete, median, and Ward [18] are examples of hierarchical clustering approaches. By assuming that each point belongs to just one cluster, the partitioning clustering technique allows for one-level data separation. As a result, each object must only fit into one cluster. For example, the criterion could be relaxed, as in the fuzzy separating approach [18], which separates each point into several clusters. K-means, k-means adaptive, K-modes, k-medoids, k-medians, and fuzzy C are all examples of partitioning methods.

B. Architecture of K-means Clustering

K-means is an unsupervised clustering technique that utilizes for classifying unlabelled data into a specific number of k (clusters or groups) [19]. There are three aspects to the K-means algorithm.

- We are starting by using a given K to set the center points (or first centroids).
- The data points are partitioned into K clusters based on the K current centroids.
- Updating the K centroids to take into account the newly created cluster. After several iterations of repeating steps b and c, the K-means algorithm usually converges.

Distance measurements are used to assess how similar the data pieces are in terms of regularity. The degree of interrelationship between the datasets, their similarity or dissimilarity, and the measures used to compare them must all be determined. The goal of calculating the metric in a specific problem is to find a suitable similarity or distance function [4], [8]. As a result, various distance parameters are used to appropriately find and place the data points' proximity. Cosine, City block, squared Euclidean, Correlation, and Some of the distance functions employed in this approach include Hamming. The K-means algorithm is often employed when the number of clusters is already known, and random initialization for the number of cluster centers or other user-defined parameters has been done [2], [9].

The distance between each point in data is determined by using a distance function picked from all the centers in the first iteration of the K-means technique. The data point is positioned concerning a particular center based on the least value of the distance between data points.

B. Multi-Agent System

The MAS is a branch of software agent technology in which a group of loosely linked autonomous agents collaborates to accomplish a shared objective. To do this, the agents cooperate, compete, or exchange information with one another [8-10]. MAS delivers a varied collection of agent capabilities for flexible issue resolution. The fundamental purpose of a single/group of agents is to solve the local issues, but the major goal of a MAS is to build agent groups in order to give answers to naturally dispersed problems. Each agent has different skills in a group, such as communication, coordination, denial, and collaboration [11]. In Figure 2, the MAS coordinates and communication are illustrated.



Through algorithms that facilitate coordination, such as joint intention or partial global planning, the coordination process comprises aligning and managing the set of actions for each agent in order to accomplish the assignment. Cooperation is a procedure in which agents collaborate to achieve a job to address a particular issue through exchanging knowledge [6]–[8]. The following are the fundamentals that all MASs have in common:

- The presence of distributed and complex problems that are beyond the individual agent's capabilities.
- The lack of a global control agent in the system.
- Whether or not decentralized data distribution is appropriate.
- Coordination between the agents is present.

In general, there is no such thing as a typical multi-agent architecture, and the distributed issue defines which one the system must solve [22], [24], [25]. Centralized and decentralized are two types of multi-agent architectures. The MAS is utilized to tackle complex problems like; as design issues, air traffic management, and manufacturing problems [26]–[28].

C. MK-means Technique

The MK-means method employs distinct agent activation to conduct clustering, with each agent examining a different subset of attributes. If the total number of features is m, the total number of features in a subset is 2^{m-1} . The first feature is included in the first one. The second feature until all single feature sets have been considered, after which the other sets containing two pairs of features are considered; this process is repeated until all features have been considered. Each set is used as an input by an agent, which generates data clustering results based on the input set. Using a voting process based on the part of the labeled data, the clustering results are used to predict non-labeled data.

The MK-means give each agent a numerical index according to the rank of the candidate set. Next, calculating their accuracy using given datasets, the agents are linked to the main/central agent and asked to submit their performance score. Finally, the winning agent, defined as the one with the greatest score, is picked. The pseudocode below illustrates the suggested MK-means method and figure 3 illustrates the MKmeans method.

Input $X = \{x1, x2...xN\}$ $xi = \{f1 \ f2 \dots fm\}$ //features $Y = \{y1, y2...yn\}$ //labels n<<N Output // Predicted classes $Yp = \{yk \ y(k+1) \dots y(N-n)\}$ Features j Start 1-initiate $2^m - 1$ agents 2-initiate main agent 3-in parallel for each agent *i* from $2^m - 1$ agents 3.1 clusters(i)=call K-means with respect to selected features i 3.2 labeling of clusters based in Y using a voting process 3.3 send output Yp, i for agent i and the error ei to the main agent End 4.main agent select the decision of agent(j) with mini e_j 5. winning agent j provided the output of non-labeled data Yp = Ypj6.selected features are features *j* End

Fig. 3 MK-means algorithm

As depicted in the pseudocode, the inputs are X which represents the dataset, xi, which represents the features; and Y, which represents the benchmark part of the dataset. The output is the prediction vector Yp for the non-labeled data. The initialization of the agents is performed in Step 1 until step 3. The operations of the agents for predicting the labeled part of the data are performed in Step 3.1 until Step 3.5. Finally, the agents communicate with the main agent in Step 4, which select the winning agent from them. The winning agent is the one that achieved the least prediction error (maximum prediction accuracy) concerning the ground truth data. The selection is performed in Step 4, while its results generation for the non-labeled data is performed in Steps 5 and 6.

D. Evaluation Metrics

Numerous assessment measures, including purity, computation time, and computational complexity, are employed to assess the performance of the MK-means technique.

1) Purity: Purity refers to the percentage of items that are accurately classified. It is a measure of a cluster's quality that is calculated by allocating each individual cluster to the most relevant class within the cluster.

$$Purity(C,G) = \sum_{i} \frac{a_{i}}{n} (max_{j} \frac{n_{ij}}{a_{i}}$$
(1)

As used in previous studies [19-23], a clustering process is If the clustering is 100% pure, it is deemed ideal. process (C) =1.

2) Computation Time: The computation time is the amount of time it takes for a computer to perform a set of calculations. It is preferable to use a clustering algorithm that takes less time to compute than one that takes longer to complete the task [4], [29].

3) Computational Complexity: An algorithm's computational complexity is a measure of how many steps it takes in the worst-case scenario for a particular instance or input. The number of steps required is proportional to the object's size [4, 30].

III. RESULT AND DISCUSSION

Five datasets evaluated the MK-means algorithm's performance (Wine, Iris, Seed, Breast cancer, and Liver Disorders). The tests' output parameters are detailed in the findings, which are used to compare the MK-means algorithm's performance to the other algorithms like CGA, ABCGA, and standard K-means. The datasets are listed in the table below (Table I).

TABLE I
DATASETS DESCRIPTION

No.	Dataset Name	No. of	No. of	
		Instances	Attributes	
1	Wine	179	14	
2	Iris	150	4	
3	Seed	210	5	
4	Breast Cancer	699	10	
5	Liver Disorders	345	7	

The algorithms have been implemented, and the corresponding clusters have been found. The true label dataset is compared to these clusters. The class labels are known because the five standard datasets were used in this paper, and the quality of the obtained clusters was assessed based on the computation time and purity.

A. Comparison using purity

When the number of clusters is 8, the comparison uses purity for all testing datasets: Wine, Iris, Seed, Breast Cancer, and Liver Disorders. Table II and Fig. 4 show the purity values.

TABLE II
COMPARISON USING PURITY

No.	Dataset	CGA	ABCGA	K-	MK-
	Name			means	means
1	Wine	34.04	40.7	30.52	45.2
2	Iris	34.04	40.7	30.52	44.7
3	Seed	34.92	41.8	31.34	43.9
4	Breast Cancer	34.96	41.5	31.35	43.8
5	Liver Disorders	34.97	41.9	31.37	47.1



Table II indicates that the MK-means method beats the CGA, ABCGA, and regular K-means algorithms for all five datasets in terms of purity. Because the agents maintain track of each group's/borders, clusters coordinate relevant dataset groupings and make the required measurements (such as; voting process and distance between clusters).

B. Comparison using Computation Time

CGA, ABCGA, MK-means, and k-means algorithms have their computation times measured. Clusters are represented by the letter k. The computation times for the Wine, Iris, Seed, Breast cancer, and Liver Disorders datasets are shown in Table III.

 TABLE III

 COMPARISON USING COMPUTATION TIME

No.	Dataset Name	CGA	ABCGA	K- means	MK- means
1	Wine	47	16	31	12
2	Iris	47	16	16	12
3	Seed	47	15	32	11
4	Breast Cancer	47	15	31	13
5	Liver Disorders	46	16	16	11



Fig. 5 Comparison Using Computation Time

Table IV and Fig. 5 show that the MK-means algorithm outperforms the CGA, ABCGA, and standard K-means algorithms in terms of computation time for all five datasets. As a result, the MK-means algorithm distributes the work across multiple agents, and this procedure shortens the processing time and improves the clustering process.

C. Comparison using Computational Complexity

When the number of clusters is 8 for five datasets, the computational complexity of the four algorithms is calculated. Table VI shows the values for computation complexity. The comparison results for computational complexity are shown in Figure 6. When compared to other algorithms, the results show that the MK-means achieved the best result. Because the MK-means made use of key agent characteristics like the ability to act autonomously, sense the environment, and communicate with one another. In addition, The findings demonstrate that the suggested technique can cluster data in less time and with less complexity based on properly related clusters.

TABLE VI COMPARISON USING COMPUTATIONAL COMPLEXITY

No.	Dataset	CGA	ABCGA	K-	MK-
	Ivallie			means	means
1	Wine	1.6	1.2	3	0.9
2	Iris	1.6	1.2	4	0.8
3	Seed	1.6	1.2	3	1
4	Breast Cancer	1.6	1.2	2	1.1
5	Liver Disorders	1.6	1.2	3	0.8



Fig. 6 Comparison Using Computational Complexity

IV. CONCLUSION

Because of its simplicity, the standard K-Means clustering technique is widely used in data clustering. The appropriate

selection of initial cluster centroids determines the performance and accuracy of the standard K-means technique. As a result, this paper combines the MAS with the K-means method to create a parallel clustering system. The proposed technique is named MK-means (Multi-K-means). The MK-means performs clustering by separating the activating agent and taking each individual agent's subset of features into account. The main goal of the new technique is to preserve the dataset by increasing the accuracy of the clustering process. The statistical significance of the proposed method's performance was confirmed. The performance of the proposed technique was compared with CGA, ABCGA, and standard K-means algorithms. The outcomes depicted that the MK-means outperforms other techniques in terms of purity, computation time, and computation complexity. This is the case because the MK-means employs key agent characteristics such as acting autonomously, sensing the environment, communicating modules, and deliberating subsolutions. The suggested technique might be integrated with existing clustering algorithms based on the starting cluster centroids selection principle in the future study.

ACKNOWLEDGMENT

This paper is funded by the Research and Innovation, Universiti Kuala Lumpur. This paper is supported by the Center of Intelligent and Autonomous Systems (CIAS), Universiti Tun Hussein Onn Malaysia (UTHM).

REFERENCES

- [1] P. I. Dalatu, A. Fitrianto, and A. Mustapha, "Hybrid distance functions for K-Means clustering algorithms," Stat. J. IAOS, vol. 33, no. 4, pp. 989-996, 2017.
- [2] M. A. Jubair, S. A. Mostafa, A. Mustapha, M. H. Hassan, M. A. Salamat, and M. S. Jawad, (2021, September). Exploring the Role of Multi-Agent Systems in Improving K-Means Clustering Method. In 2021 4th International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR) (pp. 59-63). IEEE.
- Y. Wang, M. Lees, W. Cai, S. Zhou, and M. Y. H. Low, "Cluster based [3] partitioning for agent-based crowd simulations," In Proceedings of the 2009 Winter Simulation Conference (WSC), pp. 1047-1058, 2009.
- P. Fränti and S. Sieranoja, "K-means properties on six clustering [4] benchmark datasets", Applied Intelligence., vol. 48, no. 12, pp. 4743-4759.2018.
- P. Belsis, A. Koutoumanos, and C. Sgouropoulou, "PBURC: A [5] patterns-based, unsupervised requirements clustering framework for distributed agile software development," Requirements engineering, vol. 19, no. 2, pp. 213-225, 2014.
- N. Kaur and S. Aggarwal, "Designing a New Hybrid K-Means [6] Optimization Algorithm," Int. J. of Adv. Res. in Com. Sci., vol. 8, no. 5, pp. 1567-1573, 2017.
- [7] M. A. Mahmoud, M. S. Ahmad, A. Ahmad, A. Mustapha, M. Z. M. Yusoff, and N. H. A. Hamid, "Building norms-adaptable agents from potential norms detection techniques (PNDT)," Int. J. of Int. Inf. Tech, 9(3), 38-60, 2013.
- M. A., Mahmoud, M. S. Ahmad and, M. Z. M. Yusoff, (2016, March). [8] A norm assimilation approach for multi-agent systems in heterogeneous communities. In Asian Conference on Intelligent Information and Database Systems (pp. 354-363). Springer, Berlin, Heidelberg.
- [9] S. A. Mostafa, M. S. Ahmad, M. Annamalai, A. Ahmad, S. S. Gunasekaran. "A conceptual model of layered adjustable autonomy," In Advances in information systems and technologies 2013 (pp. 619-630). Springer, Berlin, Heidelberg.
- S. A. Mostafa, S. S. Gunasekaran, M. S. Ahmad, A. Ahmad, M. [10] Annamalai. and A. Mustapha, (2014, June). Defining tasks and actions complexity-levels via their deliberation intensity measures in the layered adjustable autonomy model. In 2014 International Conference on Intelligent Environments (pp. 52-55). IEEE.

- [11] M. A. Mahmoud, M. S. Ahmad, A. Ahmad, M. Z. M. Yusoff, A. Mustapha, and N. H. A. Hamid, "Obligation and Prohibition Norms Mining Algorithm for Normative Multi-agent Systems," KES-AMSTA, pp. 115-124, 2013.
- S. A. Mostafa, , M. S. Ahmad, , M. Annamalai, , A. Ahmad, , and S. [12] S. Gunasekaran, "A dynamically adjustable autonomic agent framework," In Advances in information systems and technologies, Springer, Berlin, Heidelberg, pp. 631-642, 2013.
- S. A. Mostafa, M. S. Ahmad, A. Mustapha, and M. A. Mohammed, [13] "Formulating layered adjustable autonomy for unmanned aerial vehicles," International Journal of Intelligent Computing and Cybernetics, 2017.
- [14] S. A. Mostafa, A. Mustapha, A. A. Hazeem, S. H. Khaleefah, and M. A. Mohammed, "An agent-based inference engine for efficient and reliable automated car failure diagnosis assistance," IEEE Access, 6, pp. 8322-8331, 2018.
- S. A. Mostafa, M. S. Ahmad, A. Ahmad, M. Annamalai, and S. S. [15] Gunasekaran, (2016, August). A Flexible Human-Agent Interaction model for supervised autonomous systems. In 2016 2nd International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR) (pp. 106-111). IEEE.
- M. H. Hassan, , S. A. Mostafa, , H. Mahdin, , A. Mustapha, , A. A. [16] Ramli, , M. H. Hassan, and M. A. Jubair, "Mobile ad-hoc network routing protocols of time-critical events for search and rescue missions," Bulletin of Electrical Engineering and Informatics, 10(1), 192-199, 2021.
- [17] M. K. Abd Ghani, M. A. Mohammed, N. Arunkumar, S. A. Mostafa, D. A. Ibrahim, M. K. Abdullah, and M. A. Burhanuddin, "Decisionlevel fusion scheme for nasopharyngeal carcinoma identification using machine learning techniques," Neural Computing and Applications, 32(3), 625-638, 2020.
- S. A. Mostafa, M. Aida, M. A. Mohammed, M. S. Ahmad, and M. A. [18] Mahmoud, "A fuzzy logic control in adjustable autonomy of a multiagent system for an automated elderly movement monitoring application," International journal of medical informatics 112, 173-184.2018.
- [19] A. deep and A. Gupta, "A Novel Fuzzy-K Means based Support Vector Machine for Software Quality Prediction," Int. J. Eng. Trends Technol., vol. 37, no. 2, pp. 80-89, 2016.
- [20] M. Mezghani, J. Kang, and F. Sèdes, "Using k-means for redundancy and inconsistency detection: Application to industrial requirements, In International Conference on Applications of Natural Language to *Information Systems*, pp. 501–508, 2018. [21] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering
- Algorithms," Annals of Data Science, vol. 2, no. 2, pp. 165-193, 2015.
- P. Lin, Y. Wang, H. Qi, and Y. Hong, "Distributed Consensus-Based [22] K-Means Algorithm in Switching Multi-Agent Networks," Journal of Systems Science and Complexity, vol. 31, no. 5, pp. 1128–1145, 2018.
- V. Bhatnagar, R. Majhi, and P. R. Jena, "Comparative Performance [23] Evaluation of Clustering Algorithms for Grouping Manufacturing Firms," Arabian Journal for Science and Engineering, vol. 43, no. 8, pp. 4071-4083, 2018.
- N. Mesbahi, O. Kazar, S. Benharzallah, M. Zoubeidi, and S. [24] Bourekkache, "Multi-agents approach for data Mining based K-Means for improving the decision process in the ERP systems," International Journal of Decision Support System Technology (IJDSST), vol. 7, no. 2, pp. 1–14, 2015.
- [25] S. A. Mostafa, S. S. Gunasekaran, S. H. Khaleefah, A. Mustapha, M. A. Jubair, and M. H. Hassan, "A fuzzy case-based reasoning model for specifications quality software requirements assessment." International Journal on Advanced Science Engineering and Information Technology, 2019.
- A. K. Dubey, U. Gupta, and S. Jain, "Comparative study of K-means [26] and fuzzy C-means algorithms on the breast cancer data," International Journal on Advanced Science, Engineering and Information Technology, 8(1), 18-29, 2018.
- [27] D. Alexi, G. Francisco, C. Juan, V. Tom, and C. Chiara, "Water Quality Analysis in Mantaro River Peru, Before and After the Tailing's Accident Using the Grey Clustering Method," International Journal on Advanced Science, Engineering and Information Technology, vol. 11, no. 3, pp. 917-922, 2021.
- M. F. A. Saputra, and T. Widiyaningtyas, A. P. Wibawa, "Illiteracy [28] classification using K means-Naïve Bayes algorithm," International Journal on Informatics Visualization (JOIV), 2(3), 153-158, 2018.
- A. Satar, A. Mohamed, and A. M. Ali, "Data Mining Techniques for [29] Pandemic Outbreak in Healthcare," International Journal on Informatics Visualization (JOIV), 5(2), 162-169, 2021.

- [30] I. T. R. Yanto, R. Setiyowati, and N. Azizah, "A Framework of Mutual Information Kullback-Leibler Divergence based for Clustering Categorical Data," *International Journal on Informatics Visualization* (JOIV), 5(1), 11-15, 2021.
- [31] S. Y. Tan, H. Arshad, and A. Abdullah, "An efficient and robust mobile augmented reality application," *International Journal on*

Advanced Science, Engineering and Information Technology, 8, 1672-1678, 2018.

[32] S. Francesca, C. G. Carlo, F. D. N. Luca, and R. Marco, "Comparison of low-complexity algorithms for real-time QRS detection using standard ECG database," *International Journal on Advanced Science*, *Engineering and Information Technology*, 8(2), 307, 2018.