Performance of K-means algorithm based an ensemble learning

Dhurgham Kadhim Hashim, Lamia Abed Noor Muhammed
Department of Computer Science, College of Computer Sciences and Information Technology, University of Al-Qadisiyah, Al Diwaniyah, Iraq

Article Info

Article history:

Received Aug 29, 2021
Revised Dec 27, 2021
Accepted Jan 6, 2022

Keywords:

Clustering
Ensemble learning
K-means
Three-way method

ABSTRACT

K-means is an iterative algorithm used with clustering task. It has more characteristics such as simplicity. In the same time, it suffers from some of drawbacks, sensitivity to initial centroid values that may produce bad results, they are based on the initial centroids of clusters that would be selected randomly. More suggestions have been given in order to overcome this problem. Ensemble learning is a method used in clustering; multiple runs are executed that produce different results for the same data set. Then the final results are driven. According to this hypothesis, more ensemble learning techniques have been suggested to deal with the clustering problem. One of these techniques is "Three ways method". However, in this paper, three ways method as an ensemble technique would be suggested to be merged with k-mean algorithm in order to improve its performance and reduce the impact of initial centroids on results. Then it was compared with traditional k-means results through practical work that was executed using popular data set. The evaluation of the hypothesis was done through computing related metrics.

Corresponding Author:

Dhurgham Kadhim Hashim
Department of Computer Science, College of Computer Sciences and Information Technology
University of Al-Qadisiyah, 2V3J+H65, Al Diwaniyah, Iraq
Email: durgam.81.2.2@gmail.com

1. INTRODUCTION

Clustering is a popular exploratory data analysis tool for gaining and understanding of data structure. It is the task of identifying subgroups in data so that data points within the same subgroups (cluster) are extremely similar while data points within different clusters are very dissimilar. In other words, we strive to discover homogeneous subgroups within the data so that data points in each cluster are as comparable as feasible based on a similarity measure like Euclidean-based distance or correlation-based distance [1], [2]. The critical concerns in clustering are; which similarity metric should be used, how many clusters may be found in the data, which clustering method is the “best”, how should algorithmic parameters be chosen, are the individual clusters and partitions correct [3].

K-means is one of the most widely used for its characteristics such as; speed and simplicity [4]. It has been used in different fields [5], [6]. It is an iterative technique that attempts to split a dataset into k separate non-overlapping subgroups (clusters) [7], each of which contains only one data point. It attempts to make intra-cluster data points as comparable as possible while maintaining clusters as distinct (far) as possible. It distributes data points to clusters in such a way that the sum of the squared distances between them and the cluster’s centroid (arithmetic mean of all the data points in that cluster) is small as possible [8].

Within clusters, the less variance there is, and the more homogenous (similar) the data points are. If cluster have spherical-like shape, the K-means method is good at capturing data structure. It tries to build a good spherical shape around the centroid at all times. That means, as soon as the clusters have sophisticated geometric shapes, K-means fails to cluster the data [9]. In addition, it is necessary to predefine the number of cluster (k). It cannot deal with noisy data or outliers, Cluster having non-convex forms are not suited for detection [1], [8]. In addition, the final outcome is controlled by the original initial centroids.

In terms of consistency and quality, a clustering ensemble tries to integrate numerous clustering models to provide a better outcome than the individual clustering algorithms [10], [11]. It refers to a situation in which a number of different runs, as a result different clusterings have been obtained for a particular dataset, then to find a single (consensus) clustering [12]. Most of existing ensemble methods have tried to obtain the most consistent clustering result with base clusterings, “accuracy” in clustering does not have a clear meaning because it is unsupervised [13]. The term “Three-way decision” refers to a group of efficient methods and heuristics employed in human problem solving and information processing. Three-way clustering employs the core region and peripheral (fringe) region to represent a cluster as an application of Three-way decision in clustering [10], [14], [15]. Core region provide the pure clustering for objects and as a result it can be used in improving the clustering. Therefore, it was suggested to be merged with K-means algorithm in order to be improved and reduce its sensitivity problem with random initial centroids. This hypothesis was evaluated in this paper through practical work using some experiments.

2. METHOD

The work in this paper is based on two fields of methods; traditional clustering wit k-means algorithm and ensemble clustering that can be combined into proposed work in order to achieve more performance.

2.1. K-means algorithm

The unsupervised classification of patterns into groups (clusters) is clustering [16]. The most well-known and often used clustering technique is the k-means algorithm. In the literature, several k-means extensions have been proposed. K-means technique and its expansions are always impacted by initializations with a necessary number of clusters a priori [17], while being an unsupervised learning to clustering in pattern recognition and machine learning. In other words, the k-means algorithm isn’t quite an unsupervised clustering technique [1], [17]. Despite its widespread use, the algorithm has certain drawbacks. Includes issues with centroids that are randomly initialized, resulting unexpected convergence [1], [18]. Therefore, running the algorithm multiple times, different compilation results can be obtained each time, depending on initial centroid. Different solutions have been proposed to solve the algorithm problems [18], [19].

2.2. Cluster ensemble

Cluster ensemble techniques seek to develop stronger and more resilient clustering solutions by combining information from several data partitioning [20]. In another sense, it seeks to integrate various clustering models in order to create a superior outcome [18]. The ensemble technique was initially developed and extensively researched in the supervised learning domains. Because of its effectiveness in classification problems, academics have sought to adapt the similar paradigm other unsupervised learning areas during the last decade or so, specifically clustering issues, because of two aspects [11]: i) there is usually no prior information about the underlying structure or any specific features that we wish to uncover, by forcing a certain structure onto the data, various clustering algorithm might generate different clustering results for the same data; ii) there is no one clustering method that can work consistently well for various issues, and for the choice of clustering algorithms for a specific problem there are no clear rules to follow.

2.3. Three-way method

As known, hard clustering uses two-way decision in order to produce a cluster, while there is need to deal with the uncertainty world that need more representation. Three-way method is based on three decisions to give more than single region of clustering [21]. Three-way Decision state that “according to the positive, boundary, and negative regions of a set, one can make a three-way decision: accept, abstain and reject” [22]. Accordingly, it can be considered as efficient methods and heuristic methods widely utilized for the resolution and processing of decision-making problems [22]. Below some basic fundamental facts regarding three-way clustering. Suppose that $C=\{C1,\ldots, Ck\}$is a family cluster of universe $V=\{\upsilon,\ldots, \upsilon_n\}$. It uses a pair of sets to represent a Three-way cluster Ci [21].

$$Ci=(Co(Ci),Fr(Ci)) \tag{1}$$

where $Co(Ci) \subset V$ and $Fr(Ci)\subset V$ and $Tr = V - (Co(Ci) \cup Fr(Ci))$. These sets, $Co(Ci), Fr(Ci)$ and $Tr(Ci)$ are represent Core Region, Fringe Region and Trash Region [21]. The outcome of three-way clustering will be:

$$C = \{(Co(C_1), Fr(C_1)), (Co(C_2), Fr(C_2)), \dots, (Co(C_k), Fr(C_k))\}$$

Then, apply a modified three-way decision clustering algorithm using the k-means algorithm according to steps:

a. Execute original k-means algorithm multiple time.
b. Select the best performance and elicitation average performance using Davies-Bouldin index (DB hereafter), Average Silhouette index (AS hereafter) and Accuracy (ACC hereafter) [23]-[25].
c. Elicitation the core region and fringe region as:

$$Co(C_j) = \{\forall i = 1,..,m, v \in C_{ij}\} = \cap_{i=1}^{m} C_{ij},$$

$$Fr(C_j) = \{\exists i \neq p, i, p = 1,..,m, v \in C_{ij} \wedge v \notin C_{pj}\} = \cup_{i=1}^{m} C_{ij} - \cap_{i=1}^{m} C_{ij}$$

All symbols that have been used in the equations should be defined in the following text.

2.4. Measures of evaluation

Clustering assessment, also known as cluster validity, is a key procedure in assessing the efficacy of learning technique in finding important groupings. A decent cluster quality measurement will assist to compare different clustering methods and to analyze whether an approach is preferable than another [21]. For evaluating the performance of algorithm, we used:

a. Davies-Bouldin index [24], [25] (DB hereafter)

$$DB = \frac{1}{c}\sum_{i=1}^{c} \quad max_{j\neq 1}\left\{\frac{S(C_i)+S(C_j)}{d(x_i,x_j)}\right\} \tag{2}$$

Which a lower value is better.

b. Average Silhouette index [22] (AS hereafter)

$$AS = \frac{1}{n}\sum_{i=1}^{n} \quad S_i \tag{3}$$

Which a higher value is better.

c. Accuracy (ACC hereafter)

$$ACC = \sum_{c=1}^{k} \quad \frac{n_c^j}{n} \tag{4}$$

Which a higher value is better.

3. PROPOSED ALGORITHM

The proposed algorithm is shown in Algorithm 1, is based on merging three-way technique with K-means algorithm. This can be done through several steps. First the traditional clustering-based k-means must be done for multiple (m) runs with different initial centroids. At each run, new initial centroids are provided, there is different results are produced. As a result, there is (m) different clustering, each object in data would be member to (m) clusters. Then these clusters would be introduced to ensemble three-way technique in order to construct "core" through intersection the objects' clusters from different runs, core region that contains the clustered objects purely and fringe region that contains other objects as shown in Figure 1.

Proposed algorithm 1:

```
Input: $m$ K-means clustering results ($C_1, C_2, \dots, C_m$)
Three-Way ensemble re-clustering results
                    $C = \{(Co(C_1), Fr(C_1)), (Co(C_2), Fr(C_2)), \dots, (Co(C_k), Fr(C_k))\}$
 for each $C_i$ in $\{C_i\}$, $i = 2, \dots, m$ do
    for $j$ to $k$ do
         get cluster j+1 from $C_1$
         for $p$ to $k$ do
              get cluster $p+1$ from $C_i$
              overlap ($j,p$)= Count ($C_{ij}, C_{1p}$);
         //overlap is a $k \times k$ matrix
           // Count ($C_{ij}, C_{1p}$) count the number of same elements of $C_{ij}$ and $C_{ij}$
     end for
     switch-tab= $\emptyset$   // switch-tab is $k \times 2$ matrix
    for $n$ to $k$ do
         $(u,v)$= argmax(overlap$(j,p)$)  // $(u,v)$ is the biggest element
         switch-tab($n,0$)= $v+1$
         switch-tab($n,1$)= $u+1$
         Delete overlap $(u,*)$
         Delete overlap $(*,v)$
    for each $C_i$ == $v$ (from switch-tab) replace with $u$ (from switch-tab)
 end for
 for j 1 to $k$ do
    Calculate $Co(C_i) = \bigcap_{i=1}^{m} C_{ij}$
    Calculate $Fr(C_i) = \bigcup_{i=1}^{m} C_{ij} - \bigcap_{i=1}^{m} C_{ij}$
 End for
 Return  $C = \{(Co(C_1), Fr(C_1)), (Co(C_2), Fr(C_2)), \ldots, (Co(C_k), Fr(C_k))\}$
```

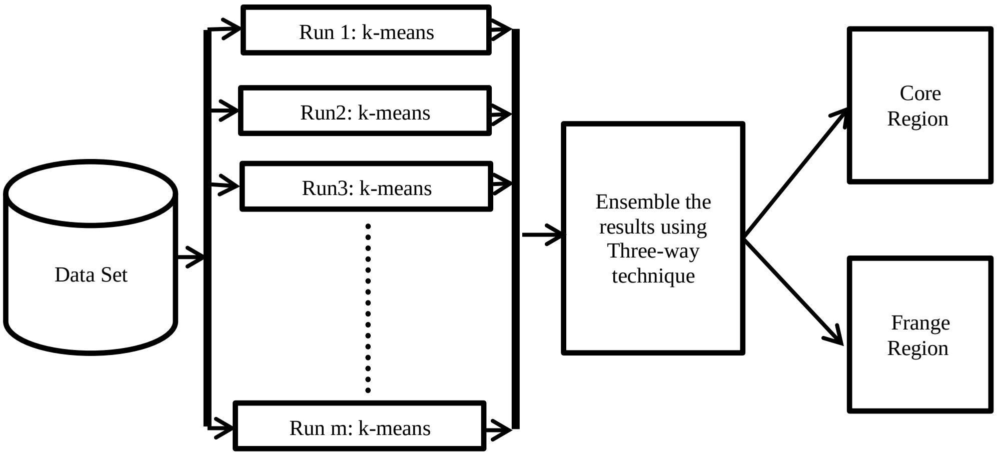

Figure 1. Proposed algorithm

4. RESULTS AND DISCUSSION

The practical test in this paper was executed using popular data sets that are extracted from "UCI machine learning repository" site. The details of these datasets are listed in Table 1, with different details (samples, attributes, and classes), they are used for clustering task. The work in testing step of proposed algorithm was achieved through experimentation of traditional k-means algorithm and ensemble k-means algorithm and then different metrics were computed for each one.

Table 1. Experiments' datasets details

ID	Datasets	Samples	Attributes	Class
1	Bank	1372	4	2
2	Forest	325	27	4
3	Seeds	210	7	3
4	Sonar	208	60	2
5	Wine	178	4	2

It was executed with the traditional k-means algorithm and ensemble k-means algorithm. For each data set, there are three experiments were done in order to enable the comparison between the traditional k-means and ensemble k-means through computing the metrics (DB, AS, ACC) with each experiment. The experiments contain, the best k-means performance, the average k-means performance, and then the performance of ensemble k-means. From Tables 2-4, it possible to notice an improve in the results for the performance of Core Region compared to best performance and average performance for implementation of the traditional K-means algorithms, the lower value for metrics (AS, DB) while the higher values of ACC. This is due to the exclusion of elements in the Fringe region. Then by synchronizing the results to align each result and matching the names of the clusters by uniting the clusters labels, and by intersecting the clusters, the most closely related objects in each cluster were identified (core region), and the marginal elements that are usually within the cluster boundary were isolated (fringe region). By excluding marginal objects, it became clear that the results could be improved.

Table 2. Bank and forest datasets performances

Data set	Performance metric					
Experiment type		Bank			Forest	
	DB	AS	ACC	DB	AS	ACC
Best K-means performance	0.453454718	0.000854869	0.729591837	0.060172805	0.006712123	0.710769231
Average K-means performance	0.454606287	0.00085354	0.726530612	0.082387268	0.005659233	0.603384615
Ensemble K-means (core region)	0.451699693	0.000861777	0.728205128	0.037559036	0.017986058	0.793548387

Table 3. Seeds and sonars datasets performances

Data set	Performance metric					
Experiment type		Seeds			Sonar	
	DB	AS	ACC	DB	AS	ACC
Best K-means performance	0.157134501	0.009011163	0.938095238	0.014026729	0.000343687	0.581730769
Average K-means performance	0.159051546	0.008910751	0.921428571	0.014120872	0.00021696	0.552884615
Ensemble K-means (core region)	0.147860874	0.009624134	0.94	0.013599365	0.00029952	0.553846154

Table 4. Wine data set performances

Data set	Performance metric		
Experiment type		Wine	
	DB	AS	ACC
Best K-means performance	0.157134501	0.009011163	0.938095238
Average K-means performance	0.159051546	0.008910751	0.921428571
Ensemble K-means (core region)	0.147860874	0.009624134	0.94

5. CONCLUSION

We applied the Three-way clustering re-ensemble method after modifying its algorithm to allow and improve the results obtained for the K-means algorithm after applying it several times. As the produced results that was shown from ensemble K-means, it is emergent performance. This is a good step for more related works in the future, as this method can be exploited by resetting centroids and then resetting the affiliation of the new incoming elements to the dataset without the need to repeat the process by measuring the distance between the new elements and the generated centroids.

REFERENCES

[1] R. Garcia-Dias, S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Clustering analysis," *Machine Learning Methods and Applications to Brain Disorders*, 2020, pp. 227-247, doi: 10.1016/B978-0-12-815739-8.00013-4.

[2] M. Nadif and G. Govaert, "Cluster Analysis," *Data Anal.*, pp. 215–255, 2010.

[3] A. Fred and A. Lourenço, "Cluster ensemble methods: From single clusterings to combined solutions," *Supervised and unsupervised ensemble methods and their applications*, vol. 126, pp. 3–30, 2008, doi: 10.1007/978-3-540-78981-9_1

[4] K. A. Hameed and A. I. Qays, "A novel technique for speech encryption based on k-means clustering and quantum chaotic map," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 160-170, February 2021, doi: 10.11591/eei.v10i1.2405.

[5] O. Davin and P. F. Perdana, "Quality and size assessment of quantized images using K-Means++ clustering," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 3, June 2020, doi: 10.11591/eei.v9i3.1985.

[6] Jason Xu and K. Lange, "Power k-Means Clustering," *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6921-6931, 2019.

[7] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 219, no. 1, pp. 103–119, 2005, doi: 10.1243/095440605X8298.

[8] A. Batra, "Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms," *5th IEEE International Conference on Advanced Computing & Communication Technologies*, no. 274, pp. 274–279, 2011.

[9] I. Dabbura, *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*, Towar. Data Sci., 2018. [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[10] P. Wang and X. Chen, "Three-Way Ensemble Clustering for Incomplete Data," in *IEEE Access*, vol. 8, pp. 91855-91864, 2020, doi: 10.1109/ACCESS.2020.2994380.

[11] T. Alqurashi and W. Wang, "Clustering ensemble method," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 1227–1246, 2019, doi: 10.1007/s13042-017-0756-7.

[12] Li Zheng, Tao Li, and Chris Ding, “Hierarchical Ensemble Clustering,” *2010 IEEE International Conference on Data Mining*, 2010, pp. 1199-1204, doi: 10.1109/ICDM.2010.98.

[13] B. Liang, L. Jiye, and C. Fuyuan, "A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters," *Information Fusion*, vol. 61, pp. 36–47, 2020, doi: 10.1016/j.inffus.2020.03.009.

[14] P. Wang, Q. Liu, X. Yang, and F. Xu, “Ensemble Re-clustering: Refinement of Hard Clustering by Three-Way Strategy,” In: Sun Y., Lu H., Zhang L., Yang J., Huang H. (eds) *Intelligence Science and Big Data Engineering. IScIDE 2017. Lecture Notes in Computer Science*, vol. 10559, 2017, doi: 10.1007/978-3-319-67777-4_37.

[15] P. Wang, H. Shi, X. Yang, and J. Mi, “Three-way k-means: integrating k-means and three-way decision,” *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 10, pp. 2767–2777, 2019, doi: 10.1007/s13042-018-0901-y.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999, doi: 10.1145/331499.331504.

[17] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[18] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means algorithm: A comprehensive survey and performance evaluation,” *Electron*, vol. 9, no. 8, pp. 1–12, 2020, no. 10.3390/electronics9081295.

[19] D. Alvincent E. and R. Edjie De Los, “eHMCOKE: an enhanced overlapping clustering algorithm for data analysis,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2212-2222, August 2021, doi: 10.11591/eei.v10i4.2547.

[20] R. Ghaemi, M. N. Sulaiman, H. Ibrahim, and N. Mustapha, “A survey: Clustering ensembles techniques,” *Proceeding of World Acad. Sci. Eng. Technol.*, vol. 38, pp. 644–653, 2009, doi: doi.org/10.5281/zenodo.1329276.

[21] P. Wang, Q. Liu, G. Xu, and K. Wang, “A three-way clustering method based on ensemble strategy and three-way decision,” *Information*, vol. 10, no. 2, pp. 1–13, 2019.

[22] Y. Yao, “Three-way decision: An interpretation of rules in rough set theory,” In: *Rough Sets and Knowledge Technology. RSKT 2009. Lecture Notes in Computer Science*, vol 5589. pp. 642–649, 2009, doi: 10.1007/978-3-642-02962-2_81.

[23] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 3, pp. 301-315, June 1998, doi: 10.1109/3477.678624.

[24] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650-1654, Dec. 2002, doi: 10.1109/TPAMI.2002.1114856.

[25] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, 1987, pp. 53-65, doi: 10.1016/0377-0427(87)90125-7.

BIOGRAPHIES OF AUTHORS

Dhurgham Kadhim Hashim received the degree in Computer Science from University of Al-Qadisiyah in 2007, Iraq. He is working a teacher in primary school in Al-Diwaniya city. He joined for Master study in 2020. He interested in machine learning and big data fields. He can be contacted at email: durgam.81.2.2@gmail.com.

Lamia Abed Noor Muhammed is an Assistant Professor at the Department of Computer Sciences, Universiti of Al-Qadisiyah, Iraq, where she has been a faculty member since 2002. She completed her Ph.D. in computer sciences from higher institution for graduate studies, Iraq, in 2008. Her research interests are primarily *in* the area of machine learning, big data, data science as well as biomedical computing, where she is the author/co-author of over 220 research publications. She can be contacted at email: lamia.abed@qu.edu.iq.