Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Modifications of K-Means Algorithm and its Execution in Real Life Data: A Comparative Study

R. Devi
Department of Mathematics, Pachaiyappa's College, Chennai, India
devirj53100@gmail.com

Article Info
Page Number: 1115 – 1121
Publication Issue:
Vol. 71 No. 3s2 (2022)

Article History
Article Received: *28 April 2022*
Revised: *15 May 2022*
Accepted: *20 June 2022*
Publication: *21 July 2022*

Abstract
K-means algorithm and its various modifications focused on real-world applications in engineering and science. It shows a wide-ranging procedure and some familiar and some lesser-known changes. The basic structure of the K-means method has been shown to enable prototype initialization and cluster analysis with implicitly defined detachment measures. It also describes a specific application of k-means clustering for real data. The key goal of this effort is to find the importance of modifications in K-means algorithm. We will discuss about the initialization and distance-based modifications of K-means algorithm. Finally, we will conclude that which kind of modification provides better result by clustering accuracy.

Keywords: - K-means; clustering; initialization; real life data.

1. Introduction

Cluster analysis [11] is one of the many significant tools in recent data analysis. The underlying supposition is that the data has a natural tendency towards cluster structure and the aim is to be able to divulge this structure. Some proposed approaches have been discussed and its effective applications have been described. Approaches can be generally divided into dual core categories: hierarchical approaches and criteria-based approaches. The concentration of my current work is on classes of criteria-based methods, or more precisely, approaches based on K-means process.

The purpose for this is that the original K-Means and its changes, as well as possible generalizations, offer excellent, yet unexplored possibilities in modern science. K-means algorithms tend to be minimal. There are numerous articles published on the theoretical and applied aspects of the K-Means algorithm. Current work is research on recent developments in the field of K-Means clustering. The emphasis will be on the ensuing issues: Recognizing the initial cluster prototype and managing obliquely demarcated distances that can be used with the systematic K-means procedure.

It is a well-established clustering procedure in the statistics community. The prototype initialization in K-Means is the significant sensitive issue. This work aims to analyse this issue. To overcome this drawback modified methods for initializing prototypes for clusters are discussed. These modified prototype initialization methods would be simpler and easier to implement. We will expect that such initialization method deserves good scalability.

Resemblance are main mechanisms used by distance-based clustering procedures to group alike data arguments into the same groups, although disparate or detached data arguments are positioned into diverse clusters. Various k-means clustering approaches have been developed and many of them are depends upon distance measure. This paper will analyse the three reformed K-means depend on initialization and three modified K-means based on distance measures.

The work is systematized as follows. Section 2 discusses general K-Means procedure. The analysis of the variations in K-means procedure is presented in third section. Subsequently, section 4 explains the consequences of the modified methods. Finally, conclusion is summarized in Section 5.

2. K-Means Method

K-Means method intends to separate 'n' items into 'k' clusters. Individual entity goes to the group with the nearby average. First introduced by MacQueen [4]. This technique produces precisely 'k' dissimilar clusters of the prime probable derivative. The greatest amount of clusters 'k' foremost to the maximum distance is not recognized in advance and should be calculated from the entity. The aim of K-Means clustering stands to diminish totality of intra-cluster discrepancies, the square of the error function

$OF=\min\left\|d_i - P_{c(d_i)}\right\|$.

By randomly selecting an object that represents each cluster, this algorithm assigns the data entity as the initial cluster. Furthermore, every other data element is allotted to the group, then the group average is calculated using clustering condition. These measurements will be used as novel cluster arguments, and individual object will be assigned to the group centre closest to it. This will continue until the cluster is recalculated and there are no changes. The obvious limitations of K-Means were as follows: 1) The user must fix the number of groups in advance, and 2) the initial conditions are critical. Different initial conditions can lead to different cluster outcomes 3) The algorithm does not guarantee the optimal solution corresponding to the global objective function's minimum value. 4) The arithmetic mean's dimness is not strong against outliers.

The Relocation method begins with the first guess as to where the center of the cluster is. The center is then repeatedly improved by moving the connections between the clusters until stability is achieved. The resulting cluster formation depends on the initial selection of seed compounds that act as cluster centers. Therefore, the Relocation method can be affected by outlier connections. Outliers are clusters of one element (singleton or noise). It's standalone and the clustering technique isn't very similar to the others, so I didn't combine it with the others. Iterative improvements will find the best partitioning for your connection, but you will probably find a solution that is not the best because it requires analysis. Of all possible solutions to ensure that you find a global optimal. Nevertheless, the computational efficiency and mathematical basis of these methods are very popular, especially among statisticians.

3. Modifications in K-means

This section aims at briefly describing the modified k-means clustering methods used in this study and their underlying mechanisms.

Initially, Let's start discussing about these modifications.

Fast K-Means Algorithm's underlying idea is: 1. Arrange the data in ascending order. Calculate variance of each attribute. Find a feature axis with uppermost discrepancy as the prime axis for separation. Compute differences among adjacent item along the item axis with the maximum discrepancy. $D_j = d\left(C_j, C_{j+1}\right)^2$ and calculate the $dsum_i = \sum_{j=1}^{i} D_j$

Calculate centroid distance of cell c: $\text{CentroidDist} = \frac{\sum_{i=1}^{n} dsum_i}{n}$, where $dsum_i$, is the sum of the discrepancies between adjacent data. Cell c should be divided into two smaller cells. A plane vertical to the main axis separates the margin then allows over a point m whose dsum i

is almost equal to centroidDist. The related list in cell c is perused and separated into two for the dual reduced cells. Calculate c delta clustering error as pre-partition entire clustering error minus dual sub-cell entire clustering error and place the cell in the vacant maximum heap. As the key, use delta clustering error. Remove the major cell from the extreme mound and replace it with a present cell. Every two non-empty sub-cells of c. Steps 3–7 must be completed for the sub-cell. Steps 3–7 must be completed for the sub-cell. Reiterate stages 8-9 till the numeral of cells ranges K.

Basic idea behind the Pseudo code approach is as follows: The first stage is to calculate the detachment among every data and entire other data in Set D to determine initial centroid. It then finds the nearest pair of data points and removes them from data set D. Then, find data point that is closest to set A1, add it to A1, and remove it from D. This procedure should be repeated until the number of elements in set A1 reaches the threshold. After that, make a new data point set A2. Repeat until you have a "k" number of such data points. Finally, through averaging entire vectors in every data element set, initial centroid is determined. Euclidean distance is used to calculate how nearby each data point is to the group centroid.

The cluster is then given points in the following stage. The main idea here is to create a cluster label and two simple data structures to keep all data elements distance to the next group constant through every repetition. It should be used in the subsequent repetition. The distance between the present data element and the centre of the new group is calculated. If the calculated distance is less than or equal to the distance to the previous centre, the data object remains in the cluster to which it was assigned. Previously, in the preceding iteration. As a result, here no prerequisite to compute the distance between this data object and the centre of another k-1 cluster, saving calculation time to the k-1 cluster centre. Then, you require compute the distance between the present data element and entire k cluster centres, locate the closest group centre, and allocate this data to the closest group centre. The label of the adjacent cluster centre and the distance to its centre are then clearly recorded. Because some data elements will persist in original group after each repetition, nearby portions of data elements will not be computed, saving an entire period of scheming the detachment and thus improving process's efficacy.

Main advantage of MP-k-means is that the range of each attribute of the data is sensibly separated in 'k' equi-sized dividers grounded on positional mean rather than arithmetic mean, wherever 'k' represents the number of clusters.

Modifications of K-means based on distance measure are discussed below.

Modified Projected K-Means approach as: randomly chooses k data elements in X as the cluster centers. Each prototype Ci is connected to a vector Wi whose components equal to 1. Assign each data element in X to the closest cluster. This consequences in a k-partition. The distance between a data element X and a cluster Xi is specified as below,

$$\mathrm{dis}(\mathrm{X},\mathrm{X_i}) = \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{d}} \mathrm{w_{ij}}\left[(1-\lambda \mathrm{j}) - \mathrm{d_l^2}\left(\mathrm{x_j},\mathrm{c_{ij}}\right) + \lambda \mathrm{j}\mathrm{d_{nl}^2}\left(\mathrm{x_j},\mathrm{c_{ij}}\right)\right] \Big/ \sum_{\mathrm{j}=1}^{\mathrm{d}} \mathrm{w_{ij}}}$$

Gr K-Means intuition is as follows: weights and groups attained through gr K-means specify that precise information is kept in gap-ratio weights for clustering process. Major gap along the y-axis is significantly larger than the average gap, while these two measures are comparable along the x-axis.

The gap-ratio for feature fj by:

$$gr_j = \frac{G_j}{\mu g_j}$$

Scaled weights are calculated using gap ratios:

$$w_j = \frac{gr_j}{\sum_{j'=1}^{N} gr_{j'}}$$

The objective of gap ratio K-means as:

$$d(x_{i,}c_k) = \sqrt{\sum_{j=1}^{N} w_j(x_{ij} - c_{kj})^2}$$

The smallest distance rule applies to enhanced k-means, with the distance being the r power of Minkowski r-metric instead of the squared Euclidean distance. The distance between the N-dimensional entities y_i and c_k as $d(y_i, c_k) = [\sum_{v=1}^{N} |y_{iv} - c_{kv}|^r]^{\frac{1}{r}}$

4. Comparative Study

The section aims to assess the performance of clustering strategies for real data in both wine and iris data. To illustrate the differences between different modifications of K-means clustering techniques. Clustering performances were assessed by Accuracy measure [2] on real data such as wine and iris dataset.

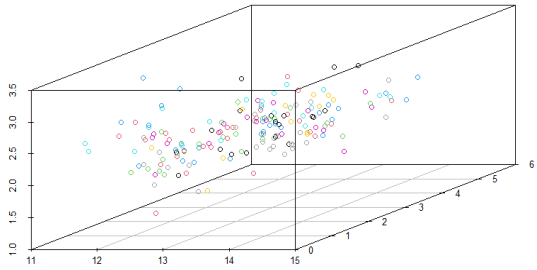

Fig1: Wine Dataset

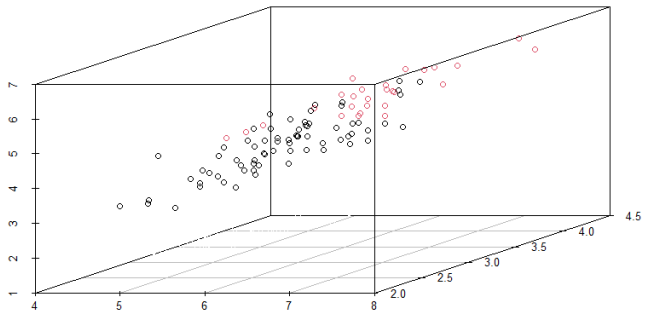

Fig2: Iris Dataset

To illustrate the differences between different modifications of K-means clustering techniques. There are several k-means clustering methods that aim to classify data points to be analysed into well separated clusters. six existing k-means clustering methods were used namely: Fast K-Means [9]; Pseudo-code approach [1]; MP-k-means [6]; Projected K-Means [7]; gap-ratio K-means [5]; Enhanced k-means [3**]**. The capacity of an algorithm to measure the accurate value is known as accuracy. It is the closeness of the obtained value to an actual or true value.

Table 1: Results of Modified Methods by Initialization

	Modification of Methods by Initialization	**Clustering Accuracy**	
		Wine	**Iris**
I1	Fast K-Means	95	97
I2	Pseudo-code approach	87.07	99
I3	MP-k-means	71.09	88.85

Table 1 summarises clustering fallouts for wine and iris datasets using all previously described algorithm executions. Pseudo-code approach executions outperform than other modified K-means on the Iris dataset. Fast K-Means attained higher clustering accuracy in the wine dataset. Table 2 shows the results of a correlative experimental investigation of the Projected K-Means, Gap-ratio K-Means, and Enhanced K-Means algorithms. The accuracy of clustering is being used to analyse the outcomes of all three methods. Table 2 displays that the Projected K-Means clustering technique outperforms the gap-ratio K-means and enhanced k-means techniques in both the wine and iris datasets.

Table 2: Results of Modified Methods by dissimilarity

	Modification of methods by dissimilarity	**Clustering Accuracy**	
		Wine	**Iris**
D1	Projected K-Means	90.249	100
D2	gap-ratio K-means	70.22	91.33
D3	Enhanced k-means	76.20	82.45

Accuracy was used to compare clustering performance on real-world data such as wine [10] and iris [8] datasets. Figures 3 and 4 compare the accuracy of clustering results from modified approaches in iris and wine datasets. The blue colour summarizes the findings of I1, I2, and I3, while the brown bar represents the algorithms D1, D2, and D3. The figures show that for the wine and iris datasets, the accuracy of grouping results using modified K-Means based on initialization is consistently good, while Modified K-Means based on distance has the higher precision but is inconsistent. This shows that enhanced modification based on initialization provides consistent result comparing with modified distance methods. The comparison revealed the dominance of modified initialization techniques are better than modified distance method. But the highest accuracy value acquired by modified distance-based method.

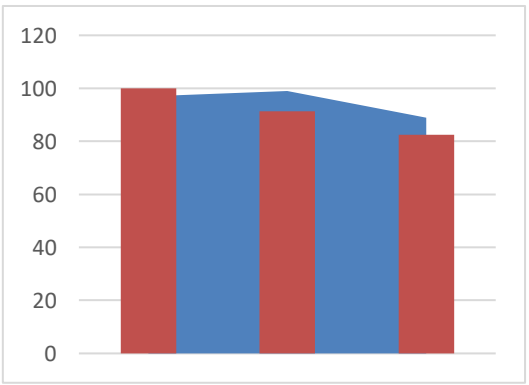

Fig3: Result by Wine Dataset

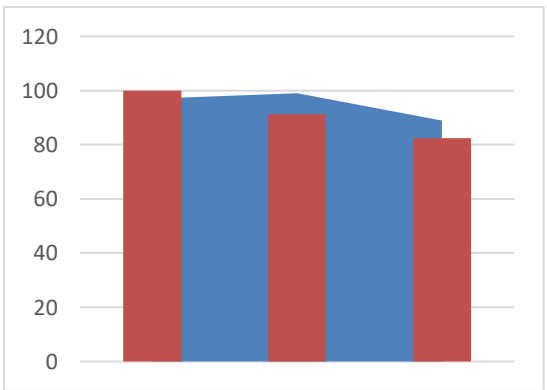

Fig4: Result by Iris Dataset

5. Conclusion and Discussion

The techniques and concepts provided in this paper are generic in nature and are applicable to clustering purpose. There are several research problems that still exist in initializing the prototypes in clustering technique. The distance measure chosen is determined by the nature of the data and the intended outcomes of the clustering process. We have made a decent attempt to compare the modifications of k-means by providing theory and its implementation into the real-life data. This paper has discussed three modified algorithms based on initialization and three processes based on distance measure, with several important improvements. In most of the tested scenarios, initialization-based methods typically performed best in the setting of real data and these modifications maintain the consistency in the result. The comparison revealed that the modified distance techniques are given highest accuracy than modified initializations. Modified Changes in k-means make sense all the way in modified initializations but the top accuracy acquired through modified distance measures.

References

1. Bhatia et al., "Experimental study of Data clustering using k-Means and modified algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013.
2. Dudek, A. "Silhouette Index as Clustering Evaluation Tool. In: Jajuga", K., Batóg, J., Walesiak, M. (eds) Classification and Data Analysis. SKAD 2019. Studies in Classification, Data Analysis, and Knowledge Organization. Springer.
3. Vanitha, D. D. . (2022). Comparative Analysis of Power switches MOFET and IGBT Used in Power Applications. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(5), 01–09. https://doi.org/10.17762/ijrmee.v9i5.368
4. Eric U. Oti et al., "New K-Means Clustering Methods that Minimizes the Total Intra-Cluster Variance", African Journal of Mathematics and Statistics Studies ISSN: 2689-5323 Volume 3, Issue 5, 2020 (pp. 42-54).
5. Jin, X., Han, J., "K-Means Clustering. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning", Springer, Boston, MA, 2011.
6. Nouby M. Ghazaly, A. H. H. . (2022). A Review of Using Natural Gas in Internal Combustion Engines. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(2), 07–12. https://doi.org/10.17762/ijrmee.v9i2.365
7. Joris Guerin et al., "Clustering for Different Scales of Measurement - the Gap Ratio Weighted K-Means Algorithm", David C. Wyld et al. (Eds) Sujatha et al., New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 2, 2013.
8. CCSEIT, AIAP, DMDB, ICBB, CNSA – 2017 pp. 35– 52, 2017. © CS & IT-CSCP 2017.
9. Manoj Kumar Gupta et al., "MP-K-Means: Modified Partition Based Cluster Initialization Method for K-Means Algorithm", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.
10. Kadhim, R. R., and M. Y. Kamil. "Evaluation of Machine Learning Models for Breast Cancer Diagnosis Via Histogram of Oriented Gradients Method and Histopathology Images".

International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 36-42, doi:10.17762/ijritcc.v10i4.5532.

11. Shanmugapriya et al., "A Modified Projected K-Means Clustering Algorithm with Effective Distance Measure", International Journal of Computer Applications, Volume 44– No.8, April 2012.
12. Chauhan, T., and S. Sonawane. "The Contemplation of Explainable Artificial Intelligence Techniques: Model Interpretation Using Explainable AI". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 65-71, doi:10.17762/ijritcc.v10i4.5538.
13. Singh, N., Srivastava, V., Komal, Iris Data Classification Using Modified Fuzzy C Means. In: Verma, N., Ghosh, A. (eds) Computational Intelligence: Theories, Applications and Future Directions - Volume I. Advances in Intelligent Systems and Computing, vol 798. Springer, Singapore, 2019.
14. Venkataramana, Boppana & Padmasree, L & Rao, M & Rekha, D & G, Ganesan., "A Study of Fuzzy and Non-fuzzy clustering algorithms on Wine Data", Journal of Communications on Advanced Computational Science with Applications, 2017.
15. Gupta, D. J. . (2022). A Study on Various Cloud Computing Technologies, Implementation Process, Categories and Application Use in Organisation. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 09–12. https://doi.org/10.17762/ijfrcsce.v8i1.2064
16. Zohaib Jan et al., "Multiple strong and balanced cluster-based ensemble of deep learners Pattern Recognition" 107 107420, 2020.
17. N. A. Farooqui, A. K. Mishra, and R. Mehra, "IOT based Automated Greenhouse Using Machine Learning Approach", Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 226–231, May 2022.