International Conference - Innovation-2021-Innovation-2021
International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN : 2456-3307 (www.ijsrcseit.com)
Volume 8, Issue 3, May-June-2021

Exploring the Depths of K-Means Clustering: A Critical Analysis

P. S. Deshmukh[1], Dr. M. Sivakkumar[2], Dr. Varshaha Namdeo[3]
Ph. D. Scholar of Dept of CSE, SRK, University, Bhopal, Madhya Pradesh, India
Associate Professor in Dept of CSE, SRK University, Bhopal, Madhya Pradesh, India
Professor in Dept of CSE, SRK University, Bhopal, Madhya Pradesh, India

ABSTRACT

In image segmentation, clustering is the process of sub dividing the whole image into the meaningful sub images. The most commonly used image segmentation algorithms such as K-means and Fuzzy c-means clustering face the specific important problem in selecting the optimal number of clusters and the corresponding cluster centroids. Plenty of research works have been done on the limitations of the said clustering algorithms to improve the efficient isolation of clusters. This paper enumerates the works done by different researchers in selecting the initial number of clusters and the centroids using K-means and Fuzzy c-means clustering. The limitations and applications of the above-mentioned clustering algorithms are explored.

Keywords: Image Segmentation, K-Means Clustering, Fuzzy C-Means Clustering, Centroids

I. INTRODUCTION

Image segmentation [1][2] is one of the applications in digital image processing and computer vision. Image segmentation divides the digital image into meaningful region based on some criteria. There are different types of image segmentation methods available in the literature. Clustering is the most widespread method for image segmentation because of its large data handling capacity.

Clustering [2] is the process of classifying pixels of a digital image in such a way that the pixels of the same group are more identical to one another than the pixels present in remaining groups. In general, clustering techniques can be categorised as hard clustering and the fuzzy clustering. Hard clustering has a special quality feature that each pixels present in the digital image belongs to only one cluster. One of the best hard clustering techniques is K-means clustering. However, in fuzzy clustering each pixel present in the digital image belongs to all clusters. Fuzzy c-means clustering is one of the best fuzzy clustering techniques.

In general K-means and Fuzzy c-means clustering are the most commonly used clustering-based image segmentation algorithms because of their simplicity and fast convergence to optimal solution [2]. Image segmentation using clustering has the application in the field including iris recognition, health care, medical imaging, image processing, traffic image, video surveillance, pattern recognition, identifying crime-prone areas, insurance fraud detection, customer segmentation, public transport data analysis, object detection, machine vision, remote sensing, wireless sensor networks and etc [1-3]. The rest of the paper is organized as follows: In Section II, algorithm description along with application and limitations of k-

means and fuzzy c-means clustering algorithms is given. In Section III, the review on existing methods used for estimating the optimal number of clusters and the optimal cluster centroids is presented. Finally, Section IV concludes the paper with scopes.

II. K-MEANS AND FUZZY C-MEANS CLUSTERING ALGORITHMS

Most commonly used partional clustering techniques for image segmentation are K-means and Fuzzy c-means clustering algorithms. K-means clustering is the simple and fast converging clustering algorithm. The expounded steps involved in the algorithms are given in [4]. Fuzzy c-means clustering algorithm converges faster than k-means clustering algorithm. The detailed steps involved in the algorithms are given in [5]. Both the clustering algorithms face the common problems such as number of initial centroids, initial centroids and dead centers. Both k-means and Fuzzy c-means clustering determine the Euclidean distances between the pixels present in the image and the centroids, which require more time and cost for large dataset. Random number of clusters and centroids increases the time complexity and affects the segmentation results. Hence, these algorithms need to be enhanced in such a way that the optimal number of clusters and their corresponding cluster centroids are selected automatically. Next section focuses on various ways available in the literature for selecting the initial optimal number of clusters and the centroids.

III. RELATED WORK

In general, determining the optimal number of clusters and their centroids is left up to the researchers focus. There have been several attempts to find a solution for selecting the number of clusters and their centroids which gives the optimum solution in isolating a cluster. A general solution to determine the number of clusters is either to run the algorithm multiple times and select the desired number of cluster based on some validity criteria or determine automatically by some meaningful methods or criteria. In the same way the cluster centroids can be selected randomly and optimized by running the algorithm several times [21]. This paper focuses on the review works of automatic selection of optimal number of clusters and the cluster centroids along with the comparison.

Mahmoud RamzeRezaee et al. [6] proposed a segmentation technique for segmenting clinical images. Pyramidal approach is used to view the image at multilevel. Cluster validation indices such as partition coefficient and partition entropy are measured to find the number of clusters automatically. Then fuzzy c-means clustering is used for merging purpose. The proposed algorithm works well compared to the conventional algorithms.

Fahim A. M et al. [7] proposed an efficient enhanced kmeans clustering for real dataset and synthetic dataset that improves the computational speed by reducing algorithm's time complexity. In a large dataset the gravity center of the spherical shaped cluster is chosen as the centroids. As very few pixels are far away from the gravity center, the distance calculation becomes easy. The selection of best centroids leads to reduced number of iterations, which results in time complexity. The proposed algorithm is better than the existing conventional algorithms.

Anindya Bhattacharya et al. [8] proposed a divisive correlation clustering algorithm for genetic engineering. Initially the number of clusters is considered as one, then the Pearson correlation coefficient for all pair of gene within a cluster. If a gene has negative correlation then that gene is kept in another cluster. Thus, this algorithm produces the k number of cluster and centroids.

Mark Junjie Li et al. [9] proposed an agglomerative fuzzy k-means Clustering algorithm for synthetic data and real data. In [9], a negative entropy term () is introduced in the objective function of k-means

clustering. The introduced negative entropy minimizes the objective function as well as cluster dispersion. Also, different values of are used in the algorithm to find the optimal number of true clusters as well as the true centroids.The mentioned validity indices are better for the proposed algorithm than classical algorithm.

Siti Noraini Sulaiman et al. [10] proposed adaptive fuzzy k-means clustering for the consumer electronic related indoor and outdoor images taken on a digital camera. Initially the pixels are assigned to its cluster by calculating the Euclidean distance. A quantitative parameter called belongingness is introduced and the membership matrix is updated, the centroids are calculated from the updated membership matrix. Statistical evaluations prove the effectiveness of the algorithm compared with conventional algorithm.

Ujjwal Maulik et al. [11] proposed an automatic fuzzy clustering algorithm based on modified differential evolution for satellite image segmentation. A fitness function called Xie-Beni index using Euclidean distance is measured for proposer clustering. Modified differential evolution selects the optimal number of clusters based on the fitness function measured for the proper clustering. The proposed algorithm shows good qualitative and quantitative results compared to existing algorithm.

Baolin Yi et al. [12] proposed improved version of kmeans clustering for machine learning database. Density of the objects is found using Euclidean distance and Gaussian density function. The initial centroids are taken as the samples that have maximum density. The proposed algorithm shows high purity and less sensitive to initial centroids compared to existing algorithm.

Chen et al. [13] The paper proposes an incomplete high-dimensional big data clustering algorithm based on feature selection and partial distance strategy. First, a hierarchical clustering-based feature subset selection algorithm is designed to reduce the dimensions of the data set. Next, a parallel k-means algorithm based on partial distance is derived to cluster the selected data subset in the first step. Experimental results demonstrate that the proposed algorithm achieves better clustering accuracy than the existing algorithms and takes significantly less time than other algorithms for clustering high-dimensional big data.

Chau et al. [14] Hence, author define a robust and effective algorithmic framework for incomplete educational data clustering using the nearest prototype strategy. Within the framework, we propose two novel incomplete educational data clustering algorithms K_nps and S_nps based on the k-means algorithm and the self-organizing map, respectively. Experimental results have shown that the clusters from our proposed algorithms have better cluster quality as compared to the different existing approaches.

Yan et al. [15] the clustering results use K-means algorithm as the initial scope of EM algorithm, according to the different choice of different characteristics of mining purposes, then use incremental EM algorithm (IEM) step by step EM iterative refinement repeatedly, it obtains the optimal value of filling missing data quickly and efficiently. it is concluded that the optimal value of filling missing data experimental results show that the algorithm of this paper to speed up the convergence rate, strengthened the stability of clustering, data filling effect is remarkable.

Chau et al. [16] On the other hand, many research works have recently proposed several general-purpose solutions to incomplete data clustering. The main difficulties with these solutions for reuse are how to appropriately determine a pre-specified number of the clusters for each data set in a particular application domain and the non-arbitrary shapes of the resulting clusters. Hence, we focus on two parts: the first one that resolves the incomplete data clustering task in the education domain and the second one that proposes a robust effective approach to the aforementioned

clustering task. Our resulting solution to clustering incomplete educational data is a mean shift-based clustering algorithm named iMS_nps using the nearest prototype strategy. iMS_nps can also overcome the aforementioned difficulties. In addition, experimental results have shown that the clusters from our proposed algorithm have better cluster quality as compared to some existing approachesPugazhenthi

A et al. [5] proposed a modified fuzzy cmeans clustering algorithms for background removal purpose. The co-occurrence matrixes, especially the diagonal elements are selected as the initial centroids. The qualitative results are better for the proposed algorithm than conventional algorithm.

Wang et al. [17] we propose a novel K-means based clustering algorithm which unifies the clustering and imputation into one single objective function. It makes these two processes be negotiable with each other to achieve optimality. Furthermore, we design an alternate optimization algorithm to solve the resultant optimization problem and theoretically prove its convergence. The comprehensive experimental study has been conducted on nine UCI benchmark datasets and real-world applications to evaluate the performance of the proposed algorithm, and the experimental results have clearly demonstrated the effectiveness of our algorithm which outperforms several commonly-used methods for incomplete data clustering.

Min Ren et al. [18] therefore, propose Spectral Ensemble Clustering (SEC) to leverage the advantages of co-association matrix in information integration but run more efficiently. We disclose the theoretical equivalence between SEC and weighted K-means clustering, which dramatically reduces the algorithmic complexity. We also derive the latent consensus function of SEC, which to our best knowledge is the first to bridge co-association matrix based methods to the methods with explicit global objective functions. Further, we prove in theory that SEC holds the robustness, generalizability, and convergence properties. We finally extend SEC to meet the challenge arising from incomplete basic partitions, based on which a row-segmentation scheme for big data clustering is proposed. Experiments on various real-world data sets in both ensemble and multi-view clustering scenarios demonstrate the superiority of SEC to some state-of-the-art methods. In particular, SEC seems to be a promising candidate for big data clustering.

Chau et al. [19] Hence, incomplete data clustering has been considered in many research works with many different approaches based on the well-known existing clustering algorithms such as k-means, fuzzy c-means, the self-organizing map (SOM), mean shift, etc. However, few of them have examined both effectiveness and robustness of the incomplete data clustering algorithms. Some of them are not practical due to a lot of parameters in hybrid approaches and/or cannot handle incomplete data which appear in any object at any dimension. In contrast, this paper aims at a SOM-based incomplete data clustering algorithm, iS nps, which is a robust and effective solution to clustering incomplete data in a simple but practical approach. Is nps can do clustering on incomplete data as well as estimate incomplete data using the nearest prototype strategy in an iterative manner. As compared to several different existing approaches, our proposed algorithm can produce the clusters of good quality and a better approximation of incomplete data via the experiments on benchmark data sets.

Honda et al. [20] author proposed , the PCA-guided k-Means procedure is extended to a situation in which some observations are missing. Principal component scores, which can be identified with a rotated solution of cluster indicators of k-Means clustering, are estimated in an iterative process without imputation. Besides solving the eigenvalue problem of covariance matrices, k-Means-like partitions are derived through

lower rank approximation of the data matrix ignoring missing elements. Several experimental results demonstrate that the PCA-guided process is more robust to initialization problems even though it is based on iterative optimization, just as the k-Means procedure is.

Vauski et al. [21] this paper author examines a comparative study of different methods with advantage and drawbacks. Performing spectral ensemble cluster (SEC) via weighted k-means are not efficient to handle incomplete basic partitions and big data problems. To overcome the problems in SEC, Greedy k-means consensus clustering is combined with SEC. By solving the above challenges, named spectral greedy k-means consensus clustering (SGKCC) is proposed. The proposed SGKCC efficient to handle incomplete basic partitions in big data which enhance the quality of single partition. Extensive evaluation NMI and RI used to calculate the performance efficiency compared with existing approach proving the result of proposed algorithm.

Pugazhenthi A et al. [22] proposed improved K-means and improved fuzzy c-means clustering algorithm for cloud image segmentation of INSAT-3D satellite images. Based on a measured threshold the histogram is split and the histogram peaks are selected as the centroids. The improved algorithms segmented the cloudy satellite image into high level clouds, middle level clouds, low level clouds and no clouds. The qualitative result proved that the proposed algorithms are better than the existing algorithms. The quantitative results are better for the proposed algorithm than conventional algorithm.

Pugazhenthi et al. [23] proposed a cloud extraction algorithm based on K-means and Fuzzy c-means clustering algorithms for INSAT-3D satellite image. The optimal number of clusters is selected from the Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua sensors cloud product. The qualitative results proved that the segmentation result is comparatively similar to the MODIS Aqua sensor cloud product.

The generalized K-means and Fuzzy c-means clustering algorithms can be improved based on the recommendations given in this section to select the optimal number of clusters and their centroids automatically. The improvement can be measured qualitatively and quantitatively comparing with the conventional clustering algorithms

Table 1. Literature Review on The Selection of Optimal Number of Clusters And Centroids

Authors	Data Used / Study Purpose	Methods / Algorithms	Outcome Measures
Mahmoud Ramze Rezaee et al.[6]	Clinical images segmentation	Fuzzy c-means clustering	Detected ventricular volume in magnetic resonance images
Fahim A. M et al. [7]	Synthetic dataset and real dataset clustering	K-means clustering	Improved time complexity
Anindya Bhattacharya et al. [8]	Biological dataset (five yeast and four mammalian datasets) clustering	K-means clustering and fuzzy c-means clustering	Very high biological significance on clustering of gene expression dataset

Mark Junjie Li et al. [9]	Synthetic dataset and real dataset clustering	K-means clustering and fuzzy c-means clustering	Produced consistent clustering result with best determination of optimal number of centroids
Siti Noraini Sulaiman et al. [10]	Consumer electronic related indoor and outdoor image segmentation	K-means clustering and fuzzy c-means clustering	Good segmentation results with better qualitative and quantitative results
Ujjwal Maulik et al. [11]	Satellite image segmentation	Pyramidal image segmentation and fuzzy cmeans clustering	Better qualitative and quantitative results
Baolin Yi et al. [12]	Machine learning database clustering	K-means clustering	Clusters with high purity and less sensitive to initial assumpution
Chen et al.[13]	an incomplete high-dimensional big data clustering algorithm based on feature selection and partial distance strategy	K-means clustering	Good segmentation results with less cluster varince
Chau et al. [14]	two novel incomplete educational data clustering algorithms K_nps and S_nps based on the k-means algorithm and the self-organizing map, respectively	K-means clustering	Found better clusters in less time
Yan et al. [15]	the clustering results use K-means algorithm as the initial scope of EM algorithm, according to the different choice of different characteristics of mining purposes	K-means clustering	Effetive segmentation of fish from complex background
Chau et al. [16]	resolves the incomplete data clustering task in the education domain and the second one that proposes a robust effective	K-means clustering	Improved perfomance with better cluster accuracy

	approach to the aforementioned clustering task		
Pugazhenthi A et al. [5]	Camera image with complex background segmentation	Fuzzy c-means clustering	Segementation of objects from complex backgrond
wang et al. [17]	a novel K-means based clustering algorithm which unifies the clustering and imputation into one single objective function.	K-means clustering	Better estimation of optimal number of clusters
Liu et al. [18]	Spectral Ensemble Clustering (SEC) to leverage the advantages of co-association matrix in information integration but run more efficiently	Fuzzy c-means clustering	Optimal or stable clustering result with less number of iterations
Chau et al. [19]	SOM-based incomplete data clustering algorithm, iS nps, which is a robust and effective solution to clustering incomplete data in a simple but practical approach	K-means clustering	Improved time complexity and number of iteration required
Honda et al. [20]	the PCA-guided k-Means procedure is extended to a situation in which some observations are missing. Principal component scores, which can be identified with a rotated solution of cluster indicators of k-Means clustering, are estimated in an iterative process without imputation.	K-means clustering	Improved efficiency along with reduced time complexity
Vauski et al. [21]	Performing spectral ensemble cluster (SEC) via weighted k-means are not efficient to handle incomplete basic partitions and big data problems	K-means clustering	Optimized accuracy in text clustering and redured time complexity

Pugazhenthi A et al. [22]	INSAT-3D satellite images segmentation	K-means clustering and fuzzy c-means clustering	Segmented the satellite image into high level clouds, middle level clouds, low level clouds and no clouds
Pugazhenthi A et al. [23]	INSAT-3D satellite images segmentation	K-means clustering and fuzzy c-means clustering	Segmented the satellite image into seven different types of clouds

Table 1 shows the literature review on the selection of optimal number of clusters and cluster centroids for K-means and fuzzy c-means clustering algorithm. The K-means and fuzzy c-means clustering algorithms have wide range of applications and applied for different types of images as well as data. The qualitative metrics used for analyses also vary with the applications. So, it is difficult to conclude or suggest a universal method to select the optimal number of clusters and their centroids. The optimal method for the selection of number of clusters and their centroids will be selected based on the qualitative metrics requirement of segmentation process and also based on applications.

The optimal choice of number clusters shows improvement in the quantitative parameters. Peak Signal to Noise Ratio, Structural Content, Means Squared Error, Structural Similarity Index, Universal Quality Index, Correlation Coefficient and Image Fidelity are some of the parameters that shown good improvement for optimal selection than random selection of number of centroids [4], [22].

IV. CONCLUSION

In this paper, a review on various works done by the researchers in selecting the optimal number of clusters for Kmeans and fuzzy c-means clustering algorithms. In addition, algorithm description, applications and limitations were discussed. It is found that lot of research works have been done on the limitations of the said clustering algorithms to improve the efficient isolation of clusters and the centroids. Most challenging task of initial optimal number of clusters by various ways has been covered and scope of future enhancement includes working on the limitations of the remaining clustering algorithms.

REFERENCE

[1] X. Cufí, X. Muñoz, J. Freixenet and J. Martí, “A review of image segmentation techniques integrating region and boundary information,” Advances in Imaging and Electron Physics, Elsevier, vol.120, 2003, pp 1-39.

[2] A. Pugazhenthi and L. S. Kumar, "Selection of Optimal Number of Clusters and Centroids for K-means and Fuzzy C-means Clustering: A Review," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9276978.

.

[3] Gonzalez R. C., Woods R. E., Digital Image Processing, 4th edition, Pearson Education, 2018.

[4] A. Pugazhenthi and J. Singhai, “Automatic centroids selection in Kmeans clustering based image segmentation,” 2014 International Conference on Communication and Signal Processing, Melmaruvathur, 2014, pp. 1279-1284.

[5] A. Pugazhenthi, G. Sreenivasulu and A. Indhirani, “Background removal by modified fuzzy C-means

clustering algorithm," 2015 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, 2015, pp. 1-3.

[6] M. R. Rezaee, P. M. J. van der Zwet, B. P. E. Lelieveldt, R. J. van der Geest and J. H. C. Reiber, "A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering," IEEE Transactions on Image Processing, vol. 9, no. 7, July 2000, pp. 1238-1248.

[7] A. M. Fahim, A. M. Salem and F. A. Torkey, "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang UniversitySCIENCE A, vol. 7, no.10, 2006, pp. 1626–1633.

[8] A. Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," Bioinformatics,vol.24, no.11, 2008, pp. 1359 1366.

[9] M. J. Li, M. K. Ng, Y. Cheung and J. Z. Huang, "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, Nov. 2008, pp. 1519-1534.

[10] S. N. Sulaiman and N. A. Mat Isa, "Adaptive fuzzy-K-means clustering algorithm for image segmentation," IEEE Transactions on Consumer Electronics, vol. 56, no. 4, Nov 2010, pp. 2661-2668.

[11] U. Maulik and I. Saha, "Automatic Fuzzy Clustering Using Modified Differential Evolution for Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 9, Sep 2010, pp. 3503-3510.

[12] B. Yi, H. Qiao, F. Yang and C. Xu, "An Improved Initialization Center Algorithm for K-Means Clustering," 2010 International Conference on Computational Intelligence and Software Engineering, Wuhan, 2010, pp. 1-4.

[13] F. Bu, Z. Chen, Q. Zhang and X. Wang, "Incomplete Big Data Clustering Algorithm Using Feature Selection and Partial Distance," *2014 5th International Conference on Digital Home*, Guangzhou, China, 2014, pp. 263-266. doi: 10.1109/ICDH.2014.57

[14] V. T. N. Chau, N. H. Phung and V. T. N. Tran, "A robust and effective algorithmic framework for incomplete educational data clustering," *2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, Ho Chi Minh City, Vietnam, 2015, pp. 65-70.doi: 10.1109/NICS.2015.7302224\

[15] S. Hua-Yan, L. Ye-Li, Z. Yun-Fei and H. Xu, "Accelerating EM Missing Data Filling Algorithm Based on the K-Means," *2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, Wuhan, China, 2018, pp. 401-406. doi: 10.1109/ICNISC.2018.00088

[16] V. T. N. Chau, P. H. Loc and V. T. N. Tran, "A Robust Mean Shift-Based Approach to Effectively Clustering Incomplete Educational Data," *2015 International Conference on Advanced Computing and Applications (ACOMP)*, Ho Chi Minh City, Vietnam, 2015, pp. 12-19.doi: 10.1109/ACOMP.2015.14

[17] S. Wang *et al.*, "K-Means Clustering With Incomplete Data," in *IEEE Access*, vol. 7, pp. 69162-69171, 2019. doi: 10.1109/ACCESS.2019.2910287

[18] H. Liu, J. Wu, T. Liu, D. Tao and Y. Fu, "Spectral Ensemble Clustering via Weighted K-Means: Theoretical and Practical Evidence," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 5, pp. 1129-1143, 1 May 2017.doi: 10.1109/TKDE.2017.2650229

[19] V. T. Ngoc Chau, "A Robust Self-Organizing Approach to Effectively Clustering Incomplete Data," *2015 Seventh International Conference on*

Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnam, 2015, pp. 150-155.doi: 10.1109/KSE.2015.11

[20] K. Honda, R. Nonoguchi, A. Notsu and H. Ichihashi, "PCA-guided k-Means clustering with incomplete data," *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, Taipei, Taiwan, 2011, pp. 1710-1714.doi: 10.1109/FUZZY.2011.6007312

[21] M. Vasuki and S. Revathy, "Efficient Handling of Incomplete basic Partitions by Spectral Greedy K-Means Consensus Clustering," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2020, pp. 299-305.doi: 10.1109/ICCMC48092.2020.ICCMC-00056