Journal of Science and Engineering Research
Vol. 2, No. 2, 2023, pp.37-44
eISSN: 2957-7187

An Improved K-Means Clustering Algorithm that Uses Centroid Identification based on Compactness Factor to Improve the Accuracy of Clustering

Jacky C.K. Tang

The Chinese University of Hong Kong, Hong Kong
tangchikit@emsd.gov.hk

DOI: http://dx.doi.org/10.56828/jser.2023.2.2.4
Article history: *Received (October 6, 2023); Review Result (November 10, 2023); Accepted (December 14, 2023)*

Abstract

The data collection is so big; that researchers are unable to extract information from it using conventional functions. On the other hand, k-means has many drawbacks, including time-consuming computation due to the use of cluster centres chosen at random at the start of the calculation. It affects two things: the algorithm's performance and the number of clusters produced at the start. This article presents a better k-means technique for the data clash strainer mechanism. This technique is applied as a function known as the Regional Centroid Component (RCC), which is introduced as a data conflict filter mechanism to the conventional k-means algorithm. This density-based pattern recognition method detects patterns by using collision data characteristics. By disregarding the conflict data before the data clustering procedure, the effectiveness of the clustering result is significantly improved. As a result, in contrast to other cluster algorithms that are currently in use, the enhanced method is much more accurate.

1. Introduction

In many areas, a large quantity of data is handled, and big data is managed utilising data mining methods to extract evidence. The phrase "we live in the information age" is well-known, but it is more accurate to say "we live in the data age." Terabytes or petabytes of data are uploaded by businesses every day into our computer networks, the Internet, and other data storage devices, and this data must be retrieved in a usable way to infer knowledge [1]. The data mining method includes cluster analysis, which is one of the primary foci of today's academics.

To make objects within a cluster comparable to one another but not to those in other clusters, input items are divided into groups known as clusters using the fundamental data appreciation and comprehension approach known as clustering. Using a set of points S in a Euclidean space and a parameter k, the objective of k-means is to divide S into k clusters in a way that minimizes the sum of the squared distances between each point and the cluster centre [2]. This situation leads to the development of a wide range of clustering techniques, including DBSCAN, CURE, MEANS, COBWEB, and others [3]. This paper presents a technique that prevents arbitrary choice selection at the start by detecting and eliminating

recognized far-apart data collected from clusters. Its original goal is to enhance the accuracy and decrease the complexity of the traditional k-means clustering method [4].

The remainder of the paper is structured as follows: a summary of similar works is provided in Section 2. Section 3 examines the fundamentals of the k-means clustering process and presents a suggested methodology based on the compactness-based centroid discovery technique. In Section 4, experiment results are shown and the suggested methodology is compared with other available clustering algorithms. Section 5 brings our efforts to a close.

2. Related Works

Shorab et al., (2012) explain how to utilise an empirical approach to choose suitable centroids at the first level of a k-means clustering strategy, and how to enhance the algorithm in terms of clustering precision while also focusing on how long the algorithm will take to run. The experimental findings demonstrated that the enhanced k-means clustering algorithm outperformed the conventional k-means clustering method however, as the amount of the data set expanded, so did the clustering algorithm's complexity. (see Figure 1).

Cosmin M P, et al., (2014) Consumer segmentation is accomplished via data mining to discover the customer attributes information that is buried inside. Clustering analysis is a technique used to determine which client groups a business serves. By evaluating similarities between data and other data, clustering may be used to create segments of information from a collection of information.

Patel and Prateek (2016), investigate the connections between different types of issues utilising data mining clustering mechanisms, as well as the linkages between them in this article, the K-means clustering method and the hierarchical algorithm are described in detail. During the clustering process, the performance of this method is compared, and a conclusion is drawn regarding the suitability of these techniques for diverse datasets in different types of states.

M A Syakur et al., (2018) According to the financial institution, the segmentation process places consumers in groups based on their qualities that are comparable to other customer groups. Client segmentation is a pre-processing procedure that allows you to categorise each customer into one of many customer groups. To be more successful at customer segmentation based on market research and demographics, it is necessary to grasp the characteristics of all consumers [10].

Hong et al., (2019), As a result of the clustering reliability study, an improved k-means algorithm has been suggested, and the proposed method demonstrates stability and produces better results when the solidity is uneven and there is a significant difference in data clustering. The experimental findings demonstrated that the modified k-means algorithm was capable of dealing with non-uniform data collection.

3. The Proposed Methodology and the Fundamental Structure of the K-means Clustering Algorithm

The k-means technique [5] was developed as a consequence of cluster analysis using a partitioning strategy as the starting point. This technique necessitates the selection of an arbitrary number of cluster centroids at the outset (k). The process also includes the calculation of the distance between each chosen centroid and each instance of structured data collecting to determine the centroid that is closest to the user, as well as the modification of

the average distance between centroids. This procedure is continued until the function's criteria or norms are satisfied.

The following formula is used to get the mean squared deviation standard for clustering:

$$D = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \| - s_i \|^2 \quad i$$

where S_i is the class I centroid and l_{ij} is an instance. Figure 1 illustrates this process. The following phases of the clustering algorithm were provided for this K. It entails the haphazard selection of centroids, location of the data center, distance computation, and cluster formation.

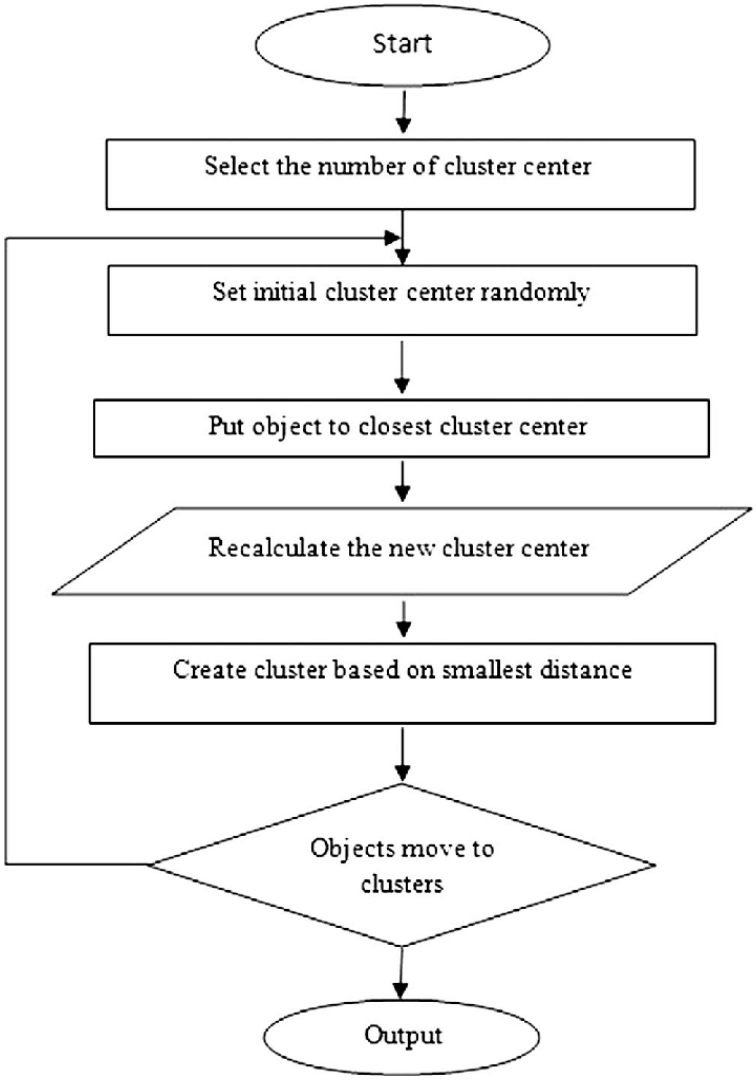

Figure 1: The K-means algorithm's basic design or scheme

Input: P instances need to be cluster {a1, a2, ……, an} and the k (no. of initial centroids)

Output: k centroids and the disagreement volume between each instance and its short-distant centroid neighbour.

K-means clustering Algorithm

The elements - number of organized data instances, number of procedure repetitions, and number of randomly picked k number of clusters - express the intricacy of this K-means technique [6].

3.1. Centroid Recognition Based on Compactness

The performance of the K-mean method is dependent on the centroid that is chosen at the outset, which has a significant impact on the outcome of the method employed. The outliers in the clusters that are far away from the data compact area lead the newly established centroid to be farther distant from the data compact region, which has a direct impact on the final clustering result, which experiences a significant divergence from the real [13]. It is preferable to remove the individual data that are collected before beginning the data clustering method to prevent such an outlier and to improve the outcome of the analysis. Using the Regional Centroid Component (RCC) is one way to find the deviating level of each instance in organized data. This method involves computing the distance between each instance and its neighbouring short-distant centroid only after the process of creating k number of centroids and k number of shortest distances between each instance and its neighbouring short-distant centroid has been finished. Based on each instance's regional centroid component, RCC then establishes the regional centroid for each instance [7].

Where sda (d) is the regional compactness of d and sda (i) is the regional compactness of the k-short-distant centroid of d. The scope of d is expressed as a centroid in RCC (d). In the data collection for compactness dispensation, the RCC has a value of roughly one. Because the regional compactness of the centroids in the data collection is significantly lower than the regional compactness of its short-distant instances, the centroid component by which the centroid is differentiated is bigger than others.

3.2. The Regional Centroid Component (RCC)-based Enhanced K-Means Clustering Algorithm

The process of eliminating distant data collecting by using the previously mentioned RCC-based recognition technique activates the mechanism [12]. It guarantees that the initial centroid computation is independent of distant data-collecting instances and eliminates them from the centroid determination process. The next stages demonstrate how the modified k-means algorithm is applied using the RCC to the newly chosen data organization.

Input: P instances need to be cluster $\{a_1, a_2, \ldots\ldots, a_n\}$ and the k (no. of initial centroids)

Output: The disagreement volume between each instance and its neighbouring short-distant centroid, denoted by k centroids.

4. Results and Discussions

The feasibility and reliability of the suggested methodology were evaluated by contrasting it with other clustering algorithms that are currently in use, such as Agglomerative Hierarchical Clustering, Mean Shift Clustering, DBSCAN (Density-based Spatial Clustering

Application Noise), and Expectation-Maximization (EM) Clustering using Gaussian Mixture Models (GMM). The researcher used the Abalone, Wine, and Iris data sets from the UCI, which is one of the most well-known neural network databases, for our investigation. These details are briefly provided in Table I. The suggested approach yields superior results and provides the best answer while avoiding clustering. The outcomes of the experiment demonstrate the superior performance of this enhanced k-means algorithm above all previous clustering techniques.

The researcher began by bringing in photographs that were sourced from three different free picture archives that dealt with histology. In the first group, there were about one million colorectal cancer histology photographs discovered [19]. (NCT-CRC-HE-100K). The United States may be reached by dialling 7181 for their country code. (CRC-VAL-HE-7K). Images are obtained when necessary and sent to the relevant parties to the requirements. People who had just been given the tragic news that they had been diagnosed with colorectal cancer are the ones who took all of the photographs that are included in the most recent collection. There are eight different types of basic tissue. The program that is responsible for saving the photographs has already given a name to each one of them. During the scanning and training phases of their operation, convolutional neural networks could use folders to help keep their data organized. When the neurons in the brain work together, the result is called "coordination." The photo collection was organized in a manner quite similar to that of three separate data stores. Seventy percent of the data was used to train the model, and fifteen percent was used to evaluate how well it performed on photographs taken in the real world. In this line of study, he effectiveness of a CNN via the use of deep learning were evaluated. The condition of the atmosphere is a factor that often has an impact on the model. First things first, think about how much it weighs and how much space it takes up. Before you get started on the project, you will need to make certain decisions about the network's dimensions, epochs, learning rates, activation function, optimizer, loss function, and architecture. The goal of this study is to identify the most effective way to use magnetic resonance imaging to segment malignant brain tumours (MRI).

Table 1: Showing the accuracy of the models

ALGORITHMS			Accuracy Rate
Adam	Ramsprop	Sgdm	Model
93.34%	93.45%	92.90%	RM
95.03%	95.89%	94.30%	LG
94.56%	65.11%	95.30%	K-Means
90.89%	93.77%	95.99%	LG
89.65%	91.56%	93.56%	LG
86.26%	91.89%	90.87%	LR
82.56%	89.54%	86.56%	LR
97.56%	97.35%	97.63%	Proposed Classifier

If the strategy that has been suggested does not work, there may be a significant issue. In the course of the study, eleven cutting-edge optimizers, including the following, were investigated: There are three distinct varieties of SGD: Adagrad, AdaDelta, and AdaDelta (SGD). The following are the three aspects that makeup Adam's ability to change: The concepts of CLR, adaptive max pooling, and adaptive max pooling with CLR (Adamax), as

well as the concepts of Root Mean Square (RMS) and CLR (Adamax) with RMS and Adaptive Momentum, can also be expressed using the terminology of RMS and Adaptive Momentum. This is in addition to the terminology of CLR and adaptive max pooling (Nadam) the cultivation of plants as well as the dissection of their genetic make-up are both well within Nesterov's skill set. CNN employs the usage of it. Adam was the most significant asset the team had. Are MRI scans reliable, or are they more of a placebo than anything else? During this time, there was a significant amount of research being conducted. This research is geared toward those who have the goal of making the world a more pleasant environment for others to live in. Only SGDM, RMSProp, and Adam have merited membership in this exclusive club as a result of their remarkable accomplishments. It is recommended that you look into if the architecture of your network is unique. An open-access dataset including 5000 individual items is the subject of the author's discussion here. The photographs are then performed on a total of eight different types of tissue, and the results of the experiment are shown thereafter. Figure 6 depicts a confusion matrix in Table 2.

There was a significant disparity in the metastatic disease rates between the two groups. There is not a single one that is not correct. Comparatively speaking with M0, which is the stage of colon cancer that has not yet spread, it has a greater number of fibroblasts present (colorectal cancer without metastases). When compared to M0, M1 has a greater number of epithelial cells as well as inflammatory cells in its composition (colorectal cancer without evidence of distant metastasis). The body included a greater number of "other" cells, such as fibroblasts, due to the presence of a tumour. There was a significant disappearance of ectoderm as well as pro inflammatory cells. Fibroblasts are cells that enter the body and help with the healing process. They do this by travelling via the veins and arteries. A deficiency in the number of epithelial cells has been linked to vein invasion. The proliferation of pro-inflammatory cells, which was seen in non-mucinous carcinomas but not in mucinous carcinomas, was absent in mucinous carcinomas. Within the TCGA dataset, the researcher discovered 12 people who had lower levels of inflammatory cells than those who were still alive at the time of their death. These individuals were confirmed to have died from cancer.

There was no correlation between the age of the women, their T or N stage, or a previous history of colon polyps and their clinical characteristics. Estimating the likelihood of developing cancer may be done using a variety of factors, including the kind of anatomic neoplasm present, the number of previous malignancies, and others (left vs. right side tumour location).

Table 2: Showing the performance evaluation of the model.

S.NO	Authors	Architectures	ACC (%)	SEN (%)	PV (%)
1	Kumari et al.,	CNN	81.22±22.02	79.89±24.85	84.89±22.89
2	Muja.et.al	RNN	78.84±12.52	84.45±13.56	77.32±14.45
3	Tan.Z.et.al	LG	79.22±12.36	91.69±13.78	77.94±14.35
4	E.S.N.et.al	RF	80.23±23.02	77.88±24.85	79.89±22.89
5	Vardhan et al.,	LR	77.84±21.79	78.98±24.53	82.54±21.55
6	Liu et al.,	CNN	79.22±22.02	79.89±24.85	81.89±22.89
7	M A Syakur .et.al	CNN	78.22±22.02	79.89±24.85	82.89±22.89
8		Algorithms	97.89±11.71	97.24±13.15	97.92±17.89

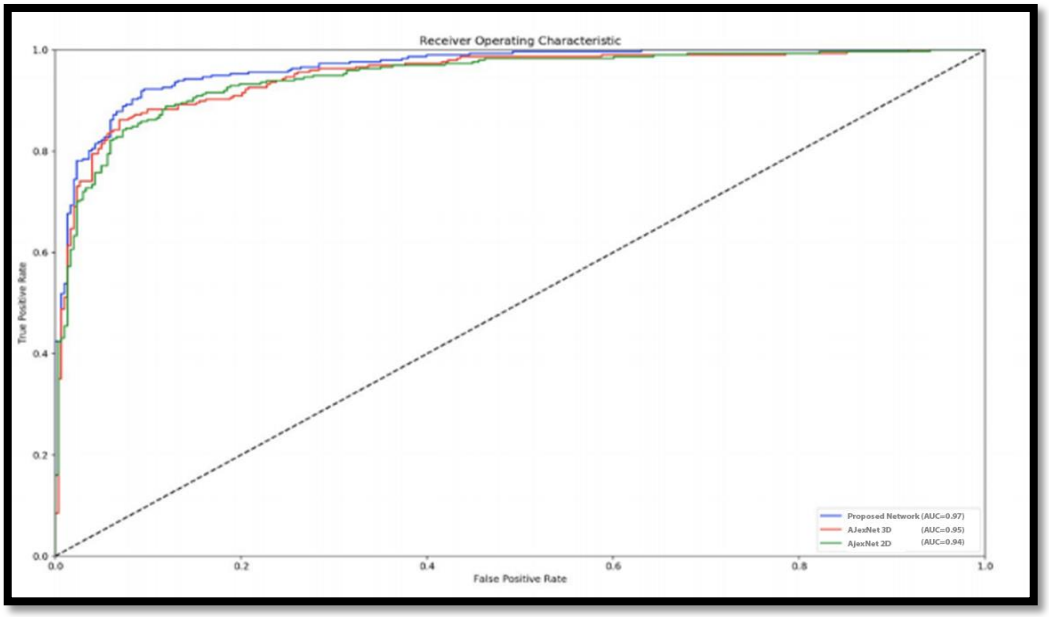

Figure 2: Showing the curve of the proposed classifier.

5. Conclusion

As a consequence of the arbitrary selection of a "k" number of starting choices, traditional K-means algorithms become unstable and difficult to use. As a consequence, overall accuracy and overall performance suffer as a result of this. The issue was addressed by including a new Regional Centroid Component Recognition (RCCR) function into the K-means clustering method, which was based on compactness, to recognize regional centroid components. During testing, the method demonstrated superior accuracy to all other known clustering algorithms in almost the same amount of time as previous approaches, which was consistent with predictions.

References

[1] Kumari, N. M. J. & K. K. Krishna. (2018). Prognosis of diseases using machine learning algorithms: A survey. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 1-9, IEEE.

[2] Muja, M. & D. G. Lowe. (2009). Fast approximate nearest neighbours with automatic algorithm configuration. *VISAPP*, 2(1), 331-340.

[3] Tan, Z., Jia, W., & W. R. Jin. (2018). Adaptive beam forming using K-means clustering: A solution to the high complexity of the reconstruction-basedalgorithm. *Radioengineering*, 27(2), 595-601.

[4] Joshua, E. S. N., Bhattacharyya, D., Chakkravarthy, M., & Y. –C. Byun. (2021). 3D CNN with visual insights for early detection of lung cancer using gradient-weighted class activation. *Journal of Healthcare Engineering*, Article ID 6695518, 11 pages. DOI:10.1155/2021/6695518.

[5] Bhattacharyya, D., Kumari, N. M. J., Joshua, E. S. N., & Rao, N. T. (2020). Advanced empirical studies on group governance of the novel coronavirus, MERS, SARS and EBOLA: A systematic study. *Int J Cur Res Rev*, 12(18), 35.

[6] Vardhan, K. A., Sudha, K., Rao, N. T., Bhattacharyya, D., & T. Kim (2009). IoT model-based smart advisory health system using ontology mechanism. DOI: 10.14257/astl.2017.147.47

[7] Nazeer, K. A. & M. P. Sebastian. (2009). improving the accuracy and efficiency of the k-means clustering algorithm. In *Proceedings of the World Congress on Engineering*, 1, 1-3, London: Association of Engineers.

[8] Liu, B. (2006). A fast density-based clustering algorithm for large databases. In *2006 International Conference on Machine Learning and Cybernetics*, 996-1000, IEEE.

[9] Qi, J., Yu, Y., & Wang, L. (2017). An effective and efficient hierarchical k-means clustering algorithm. In *International Journal of Distributed Sensor Networks*, 13(8).

[10] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S. & B. D. Satoto. (2017). Integration K-means clustering method and elbow method for identification of the best customer profile cluster IOP conference series: Materials science and engineering. *The 2nd International Conference on Vocational Education and Electrical Engineering (ICVEE*, 336.