TuijinJishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

Optimal Data Clustering through Hybrid CSO and CD-Fuzzy C means Optimization

Noor Basha[1], Ashok Kumar P S[2], *KavithaBaiA S*[3],

[1]*Research scholar, Department of computer science and engineering, Don Bosco Institute of Technology, Visvesvaraya Technological University, Belagavi -590018*

[2]*Research supervisor, Department of computer science and engineering, ACS college of Engineering, Visvesvaraya Technological University, Belagavi -590018*

[3]*Assistant professor, Department of computer science and engineering,Vemana Institute of Technology, Bengaluru - 560034*

Abstract:-Clustering is a technique that separates a given set of items into groups in such a way that objects within a cluster are both extremely similar to and dissimilar from those in other clusters The KMH-CSO algorithm is a cross between the KMH-CSO (K means harmonic clustering CSO) and the more traditional cat swarm optimisation. The second method is called Fuzzy C and it involves grouping data based on the density of the clusters. (CD-FCM). The FCM algorithm is versatile enough to be utilised for a number of different data analysis tasks. A generalised multi-objective function is employed to aggregate subsets, and this serves as a clustering criterion for the data. The DB index, the XB index, the sym-index, and the stability measure are only few of the many objectives that are included among the characteristics. In order to tackle the optimisation problem, a relatively new metaheuristic technique called cat swarm optimisation (CSO), which simulates the behaviour of cats, will be used. For the objectives of clustering, many other qualities were chosen, including petroleum oil, iris, wine, glass, cancer, and vowels. The findings have been compiled into tables, and diagrams have been drawn up to show how the data breaks down. The methods that are proposed are much more effective in comparison to other possible options.
Keywords: *cluster, optimization, fuzzy, index, KMH, CSO, C means*

1. Introduction

Clustering can be used for a wide number of purposes, including pattern recognition, picture segmentation, and the identification of anomalies. Traditional clustering strategies however are typically susceptible to noises and outliers, which significantly impair the efficiency of clustering in realistic problems [14].The Fuzzy C-Means (FCM) algorithm is considered as a technique that is frequently used for clustering data. The FCM algorithm gives each data point a membership value for each cluster, which represents the data point's degree of membership in that cluster. This value is assigned based on the cluster that the data point belongs to. The FCM has a number of limitations, some of which include the fact that it is sensitive to the number of initial clusters and that it requires prior knowledge of the cluster centre.Deep learning neural network has been brought into context for sequentially processing of big data designed based on the cat-swarm optimization algorithm, which combines the ant-lion optimization and the cat swarm optimization techniques [11]. To enhance the cluster for structures of data, where the selection of feature in structure is recognized and FAST algorithm is used as a filter approach [17]. Researchers have come up with a great deal of different FCM optimisation approaches in order to get over these limits. One of these approaches is the hybridization of different optimisation techniques. This method enhances FCM performance by combining two or more optimisation algorithms into a single solution. Cuckoo Search Optimization (CSO) and Competitive Division Fuzzy C-Means (CD-FCM) are two algorithms that have been used in recent study in this field. As a result of this research, a hybrid technique has been developed. The approach that has been suggested, which goes by the name of Hybrid CSO and CD-Fuzzy C Means Optimization, makes an effort to optimise the clustering process by determining the ideal collection of cluster centres and membership values. The CD-FCM method is used to discover ideal membership values, and the CSO algorithm is used to determine optimal cluster centres. Both of these methods are described further below. The dataset is broken up into a number of smaller datasets using the CD-FCM method, which then applies the FCM algorithm to each of the smaller datasets. The sensitivity of the FCM to the initial cluster count is reduced by using this strategy. The hybrid strategy, which brings together the best features of the two different algorithms, results in a clustering method that is more dependable and effective. The quality of the cluster is improved, and the rate of convergence is sped up, by optimising the cluster centres using the CSO algorithm and the membership values with the CD-FCM method.

2. Related Work

The CD-Fuzzy C means algorithm, on the other hand, is a hybrid clustering algorithm that is produced by combining the Fuzzy C means approach and chaos theory. This approach incorporates ideas from chaos theory into the FCM algorithm in order to attain optimal clustering, which in turn helps to reduce the size of the objective function. The effectiveness of the Hybrid CSO and CD-Fuzzy C means optimisation strategies for clustering a variety of datasets has been explored by a significant number of researchers. The results showed that the methodology was superior to both K-means and FCM when it came to the accuracy of clustering.

Clustering is a technique that is utilised extensively in the process of data analysis across a wide variety of fields, such as machine learning, data mining, pattern recognition, image analysis, along with bioinformatics. A demonstration on heart disease prediction classification using an ML approach in a big data frame work. We've indeed come a long way in terms of efficiency, but we can do better overall performance to do calculations we have to cut down on the time[10].Cluster analysis is another prominent approach of classifying data [20]. This method works by discovering groupings within a dataset based on the similarities and differences that exist between the items in those groups. In contrast to the traditional method of clustering, which allocates each data point to exactly one cluster, clustering algorithms begin with an unlabelled collection of data and divide it into a large number of groups based on similarities [21]. Data clustering is typically conceived of as an unsupervised learning process, in contrast to data classification, which needs the use of a labelled dataset for the sake of training. Since this is the case, many people believe that the performance of data clustering methods is significantly subpar [22]. Even though data classification helps improve efficiency, it can be extremely difficult and expensive to obtain tagged datasets that can be used as training data. The direct consequence of this made a wide variety of approaches have been demonstrated to be beneficial in enhancing the efficiency of clustering [23]. Clustering is a procedure that is used in the field of data science to divide big datasets into smaller groups based on the characteristics that are shared by the data in each group. The organisation of data can be accomplished by a variety of means, the most prevalent of which being the use of a hierarchical structure., a dividing scheme, and a mixed model clustering technique [24]. The choice of method is determined by a variety of considerations, including as the conclusion that is wanted, the kind of data to be analysed, the volume of data to be analysed, and the availability of suitable gear and software. Last but not least, the method of clustering data sets that makes use of Hybrid CSO and CD-Fuzzy C Means Optimization is useful. This algorithm is a well-liked option for use in data clustering applications because it is superior to FCM and other traditional clustering algorithms in terms of overcoming their constraints. In the future, research might be done to see whether or not the strategy is applicable to huge datasets and whether or not it is scalable.

3. Methods

3.1 Hybrid Cat Swarm Optimization

Cat swarm optimization algorithm has been framed based on the behaviour of the cats. It can be seen that the cats seem to rest majority of the times. But the alertness of the cats will be at its peak even when the creature is under rest. The algorithm is based on the criteria of alert looking for the next move. The tracing mode describes the target tracing process. The proposed method involves the combination corresponding to the K means harmonic clustering CSO (KMH) and Cat swarm optimization clustering (CSO).

Algorithm:

Begin
Initialize the parameters
Determine the initial population
While(condition!=terminated) do
While j<Q do

If X in cat is in mode M
Mode=seeking
Else
Mode=tracing
End if
End do
Reassignment of the cats
Output cat X
End

i. Distinct depiction:

The position of the cat j is denoted by C_j. The centres of the clusters are represented by real integers.The number of clusters in denoted by k and the attribute count is denoted by m.The solution length is given by k * m. The first m element denotes the centre of the first cluster, the second m element denotes the centre of the second cluster and so on.

ii. Initialization:

During the phase of initialization, the cats are allocated randomly between the sketching mode and the searching mode. Let the number of cats in the sketching mode be denoted as N_{skt} and the number of cats in the searching mode be denoted as $N_{src.}$

$$P_{skt} = [M_r * P]$$

$$P_{src} = P - P_{skt}$$

Mr is the ratio of the mixture that is used to control the count of cats in each mode. The cats spend most of the time resting. When they decide to move, the movement will be done in a slower means. This is denoted by the searching mode in the algorithm. The sketching mode denotes the identification of the target. Figure 1 gives the flow diagram of the cat swarm optimization.

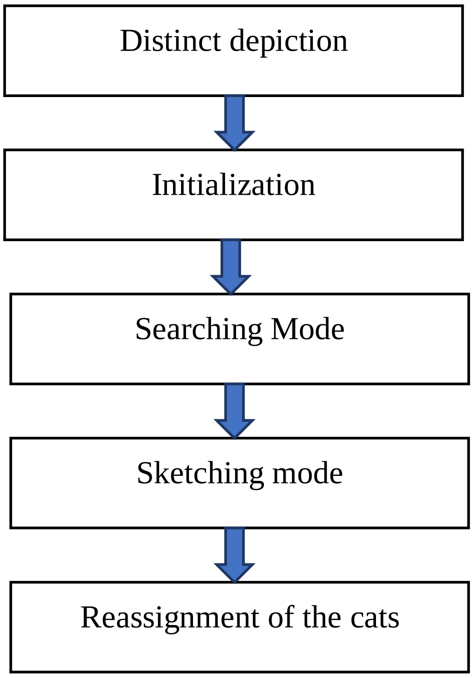

Figure 1: Flow diagram of the CSO

iii. Searching mode:

There are four factors involved in the searching mode. They are searching memory, self location flag, altered dimension and searching range. The seeking memory denotes the memory of the searching cat. It represents the number of cats in the neighbouring positions. The self locating flag is used to determine whether the current position of the cat will be it neighbouring position. If the flag value is 1, then the condition is true. Else the current position will not be the neighbouring position. The dimensions to be mutated are given by altered dimension factor. The range of the mutation is represented by the searching range.

iv. Sketching mode:

The target tracing of the cat is depicted by this mode. The cat follows the velocity of each dimension when it enters into this mode. The velocity is updated by the following equation.

$$V_{ji} = V_{ji} + R_1 \times C_1 \times (X_{ai} - X_{ji})$$

Xai and Xjidenotes the position from the individual that is known best. C1 is the velocity extension of the cat and R1 denotes a value in random that lies between 0 and 1. The updating of the cat can be done by the following equation.

$$X_{ji} = X_{ji} + V_{ji}$$

v. Reassigning the cats:

After the completion of the searching mode and the sketching mode, the cats are again assigned to one of these modes. Some of the cats are randomly selected to be in the seeking mode by using the mixing ratio R. The others are categorized into the sketching mode.

3.2 CD-Fuzzy C Means Clustering

In the fuzzy C means clustering, the datasets are grouped into clusters. Each dataset present in the cluster has a certain degree of membership. The data points that are located near the centre will have a degree of membership that is high and the data points that lie farther from the centre points will have a degree of membership that is less. The clustering can be defined by a functional equation that is given below.

$$F_{FCM}(X,Y,Z) = \sum_{j=1}^{C} \sum_{i=1}^{N} y_{ji}{}^{N} h_{ji}{}^{2}$$

In the fuzzy C means clustering the C partitions are done based on the objective function.

$$y_{ji} \in [0,1], \sum_{i=1}^{N} y_{ji} > 0, \sum_{j=1}^{C} y_{ji} = 1$$

$$y_{ji} = \frac{1}{\sum_{Q=1}^{C} \left(\frac{h_{ji}}{h_{Qi}} \right)^{2/n-1}}$$

The alternative optimization method is usually used to find the optimized value of the function. It relates the degree of membership of the fuzzy and the prototype of the clusters. The function of interest is given by the following equation.

$$z_j = \frac{\sum_{i=1}^{N} y_{ji}{}^{m} a_i}{\sum_{i=1}^{N} y_{ji}{}^{m}}$$

The distance of separation between the points of data is given by

$$d_{ji}{}^{2} = \| a_i - q_j \|^2 = (a_i - q_j) A (a_i - q_j)^T$$

In the above equation, A is a definite matrix with positive values. It is used to update the values in the iteration process.

3.2.1 M-FCM

In M-FCM, the distance is based on a function called Mahalanobis distance. The Mahalanobis distance can be estimated by the following equation.

$$d_{ji}^{\ 2} = \| a_i - q_j \|^2 = (a_i - q_j)^T \sum_{ji}^{-1} (a_i - q_j) - \ln |\sum_{ji}^{-1}|$$

$|\sum_{ji}^{-1}|$ represents the matrix of covariance. Mahalanobis distance is preferred over the Euclidean distance. It plays a major role in the presence of distributions that are multi normal. The objective function based on the Mahalanobis distance is given by the following mathematical expression.

$$F_{FCM}(X,Y,Z,S) = \sum_{j=1}^{C} \sum_{i=1}^{N} y_{ji}^{\ N} (a_i - q_j)^T \sum_{ji}^{-1} (a_i - q_j) - \ln |\sum_{ji}^{-1}|$$

3.2.2 σ-FCM

The fuzzy C means clustering can be regularized with the Euclidean distance. The Euclidean distance based fuzzy C means separative distance is given by the following equation.

$$d_{ji} = \frac{\| a_i - q_j \|^2}{\sigma_j}$$

The mean of weightage in the cluster is given by,

$$\sigma_j = \left\{ \frac{\sum_{i=1}^{N} y_{ji}^{\ N} \| a_i - q_j \|^2}{\sum_{i=1}^{N} y_{ji}^{\ N}} \right\}^{0.5}$$

Unlike the Mahalanobis distance, the degree of the membership is calculated from the centroid. The normalization is done based on the covariance that occurs between the features.

$$F_{\sigma-FCM}(X,Y,Z) = \sum_{j=1}^{C} \sum_{i=1}^{N} y_{ji}^{\ N} \frac{\| a_i - q_j \|^2}{\sigma_j}$$

3.2.3 CD-FCM

The existing fuzzy C means clustering techniques does not take into account the density of the points of data. It only aims to describe the features of weightage. The distance of the individual data points is alone measured by the existing methods. It fails to take into account the distribution prevailing globally. Hence a cluster density based fuzzy C means clustering has been proposed in this paper.

Using the density of the clusters, the measured distance is given by

$$d_{ji} = \frac{\| a_i - q_j \|^2}{v_j}, 1 < j < C, 1 < i < N$$

The objective function of the cluster density based fuzzy C means optimization is given by

$$F_{CD-FCM}(X,Y,Z) = \sum_{j=1}^{C} \sum_{i=1}^{N} y_{ji}^{\ N} \| a_i - q_j \|^2 \frac{\sum_{q=1}^{N} b_{jq} w_{jk}}{\sum_{q=1}^{N} b_{jq} w_{jk} v_k}$$

The update equations can be obtained by using the Lagrange's multiplication method.

$$h_j = \frac{\sum_{i=1}^{N} y_{ji}{}^{m} a_i}{\sum_{i=1}^{N} y_{ji}{}^{m}}, 1 < j < C$$

$$y_{ji} = \frac{d_{ji}{}^{-2/(N-1)}}{\sum_{q=1}^{C} d_{qj}{}^{-2/(N-1)}}$$

The first step involved in the CD-FCM is the selection of number of clusters. The index of the fuzzy, error of iterations, and the degree of membership are initialized. The centroid values are calculated. The density of the data points of every data set is determined. The membership of the fuzzy and the cluster centroids are updated. After completion, the iteration is stopped and the fuzzy membership and the cluster centroids are obtained.

4 Results

The proposed algorithms have been compared with other algorithms to identify its effectiveness. The attributes considered for the experimentation were crude oil, iris, wine, glass, cancer and vowels. K means harmonic clustering CSO (KMH) and Cat swarm optimization clustering (CSO) is compared with the proposed hybrid K means harmonic CSO (KMH-CSO).

Table 1: Average of the attributes when q=3

Attribute	**KMH**	**CSO**	**KMH-CSO**
Crude oil	28143.25	47180.23	27752.32
Iris	186.25	284.23	186.24
Wine	1054441632.52	1325489612.22	1025568922.72
Glass	2546.52	3650.21	2542.31
Cancer	245692.54	365216.52	182.35
Vowel	27256489123.51	44563218972.54	27513649812.76

It can be seen that the average value of the crude oil by using the KHM, CSO and KMH-CSO are 28143.25, 47180.23 27752.32 and respectively. The average value of the iris by using the KHM, CSO and KMH-CSO are 186.25, 284.23and 186.24respectively. The average value of wine by using the KHM, CSO and KMH-CSO are 1054441632.52, 1325489612.22and 1025568922.72respectively. The average values produced by KHM, CSO and KHM-CSO for glass are 2546.52, 3650.21and 2542.31respectively. The average values produced by KHM, CSO and KHM-CSO for cancer are 245692.54, 365216.52and 182.35 respectively. It can be seen that the average value of the crude oil by using the KHM, CSO and KMH-CSO are 27256489123.51, 44563218972.54 and 27513649812.76 respectively.

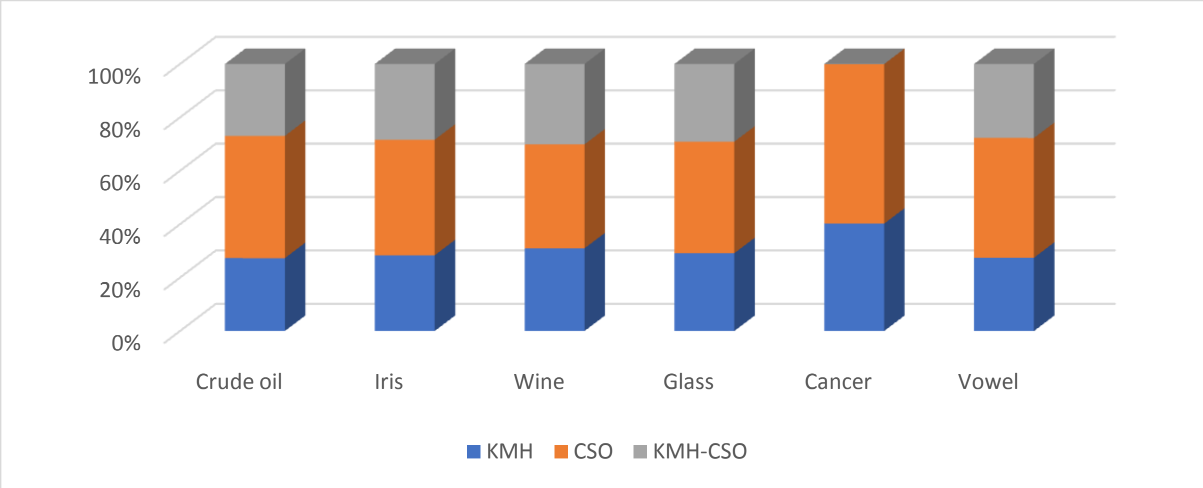

Figure 2: Graphical representation of average of the attributes when q=3

Table 1 gives the standard deviation of the attributes when q=3 and figure 2 give the graphical representation of the Standard deviation of the attributes for the cluster count of 3.

Table 2: Standard deviation of the attributes when q=3

Attribute	KMH	CSO	KMH-CSO
Crude oil	278.13	6125.34	24.32
Iris	0.14	42.61	4.50
Wine	2746.25	1.62	7664.23
Glass	55.64	290.15	26.25
Cancer	0.14	33.83	53.24
Vowel	152.21	163.11	214.32

The standard deviation of crude oil using the methods KMH, CSO and KMH-CSO are 278.13, 6125.34 and 24.32 respectively. The standard deviation of crude iris using the methods KMH, CSO and KMH-CSO are 0.14, 42.61 and 4.50 respectively. The standard deviation of wine using the methods KMH, CSO and KMH-CSO are 2746.25, 1.62 and 7664.23 respectively.

Glass has a standard deviation of 55.64, 290.15 and 26.25 respectively for the KMH, CSO and KMH-CSO methods respectively. Cancer has a standard deviation of 0.14, 33.83 and 53.24 for the KMH, CSO and KMH-CSO methods respectively. Vowel has a standard deviation of 152.21, 163.11 and 214.32 for the KMH, CSO and KMH-CSO methods respectively.

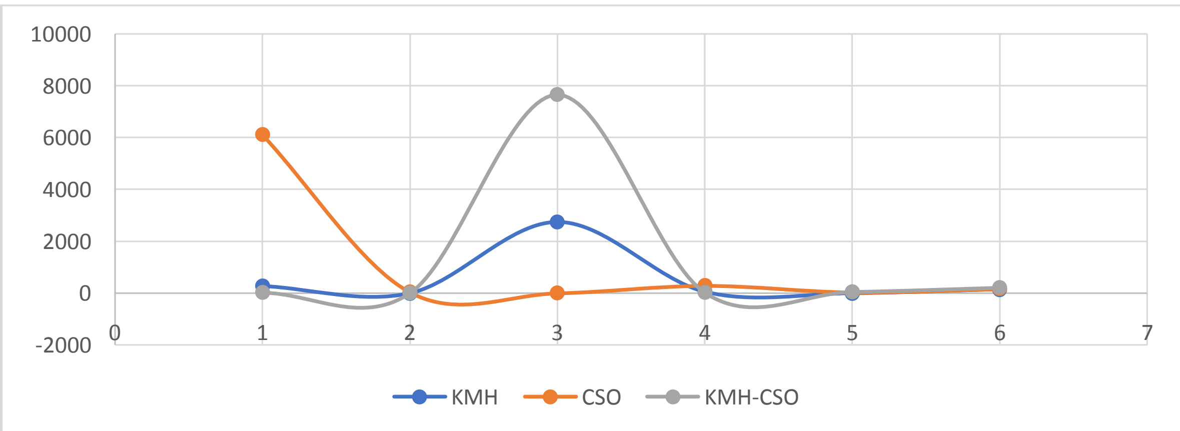

Figure 3: Graphical representation of Standard deviation of the attributes when q=3

Table 2 gives the time taken by the attributes during the clustering process in seconds when q=3 and figure 3 depicts the graphical representation of the standard deviation of the attributes when cluster count is 3.

Table 3: Time(s) taken by the attributes when q=3

Attribute	**KMH**	**CSO**	**KMH-CSO**
Crude oil	0.05	10.64	19.52
Iris	0.15	14.32	31.25
Wine	0.32	44.62	95.21
Glass	0.76	78.52	222.21
Cancer	0.14	33.82	53.65
Vowel	3.62	205.32	685.46

The time taken for clustering the crude oil through KMH, CSO and KMH-CSO are 0.05s, 710.64s and 19.52s respectively.The time taken for clustering the iris through KMH, CSO and KMH-CSO are 0.15s, 14.32s and 31.25s respectively.The time taken for clustering the wine through KMH, CSO and KMH-CSO are 0.32s, 44.62s and 95.21s respectively. Clustering of glass has consumed 0.76s, 78.52s and 222.21s using the KMH, CSO and KMH-CSO methods. Clustering of cancer has consumed 0.14s, 33.82s and 53.65s using the KMH, CSO and KMH-CSO methods. Clustering of vowel has consumed 3.62s, 205.32s and 685.46s using the KMH, CSO and KMH-CSO methods.

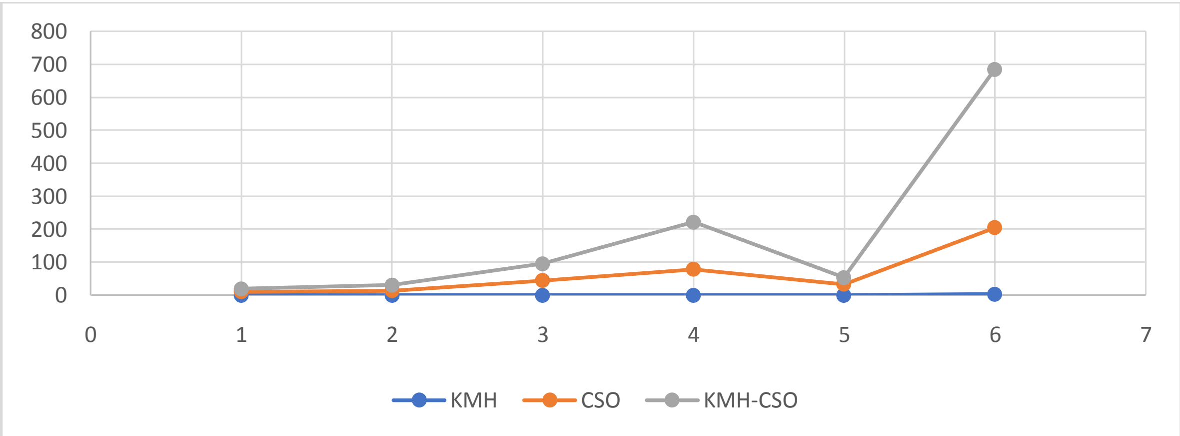

Figure 4:Graphical representation of Time(s) taken by the attributes when q=3

Table 3 gives the indices of different types of fuzzy C means clustering. Figure 4 depicts the graphical representation of the indices of different types in FCM.

Table 4: Indices of different types of FCM

Clustering technique	**Partition coefficient**	**Partition index**	**Xie and Beni's index**
M-FCM	0.51	1.63	0.18
σ-FCM	0.24	1.52	0.01
CD-FCM	0.34	0.42	0.02

The partition coefficient of M FCM, σ-FCM and CD-FCM are 0.51, 0.24 and 0.34 respectively. The partition index of M FCM, σ-FCM and CD-FCM are 1.63, 1.52 and 0.42 respectively. The Xie and Beni's index of M FCM, σ-FCM and CD-FCM are 0.18, 0.01 and 0.02 respectively.

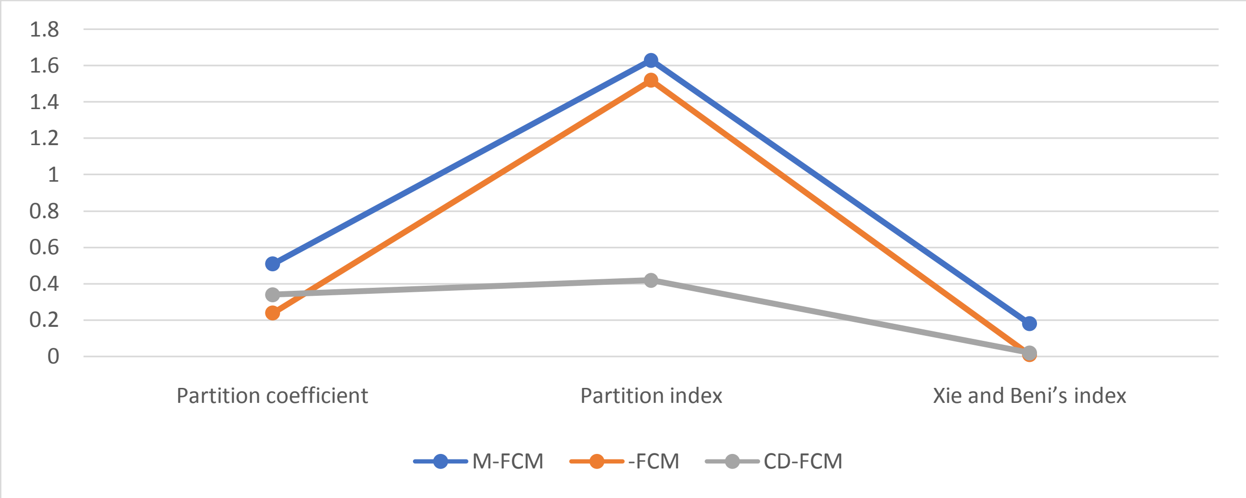

Figure 5: Graphical representation of indices of different types of FCM

Table 4 gives the time complexity comparison of fuzzy C means clustering and K means clustering with respect to iteration count. Figure 5 depicts the graphical representation of Time complexity comparison of fuzzy C means clustering and K means clustering with respect to iteration count.

Table 5: Time complexity comparison of fuzzy C means clustering and K means clustering with respect to iteration count

S.No	Iteration Count	K means	Fuzzy C means
1	10	5000	5000
2	20	8000	12000
3	30	11000	19000
4	40	14000	26000
5	50	17000	32000

When the iteration count is 10, the time complexity of K means and fuzzy C means are the same, (i.e) 5000. When the iteration count is 20, the time complexity of K means and C means are 8000 and 12000 respectively. The K means and the fuzzy C means has a time complexity of 11000 and 19000 when the count of the iteration is 30. When the iteration count is 40, the time complexity of K means and C means are 14000 and 26000 respectively. The K means and the fuzzy C means has a time complexity of 17000 and 32000 when the count of the iteration is 50.

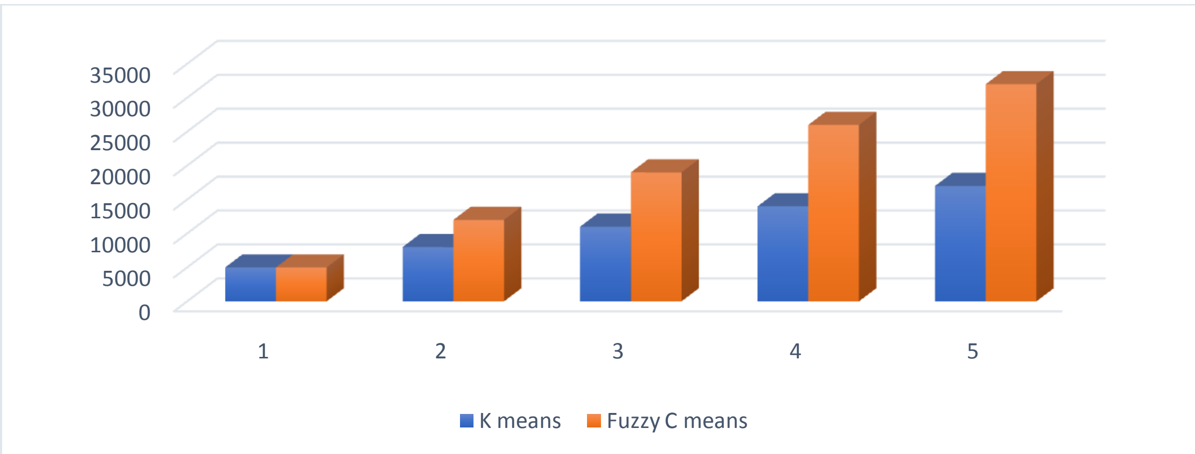

Figure 6: Graphical representation of time complexity of fuzzy C means and K means clustering with respect to iteration count

Table 5 shows the Time complexity comparison of fuzzy C means clustering and K means clustering with respect to cluster count. Figure 6 depicts the graphical representation of time complexity comparison of fuzzy C means clustering and K means clustering with respect to cluster count

Table 7: Time complexity comparison of fuzzy C means clustering and K means clustering with respect to cluster count

S.No	Iteration Count	K means	Fuzzy C means
1	1	6000	8000
2	2	13000	26000
3	3	18000	58000
4	4	27000	84000
5	5	31000	95000

When the cluster count is 1, the time complexity of K means and fuzzy C means are the same, 6000 and 8000 respectively. When the cluster count is 2, the time complexity of K means and C means are 13000 and 26000 respectively.

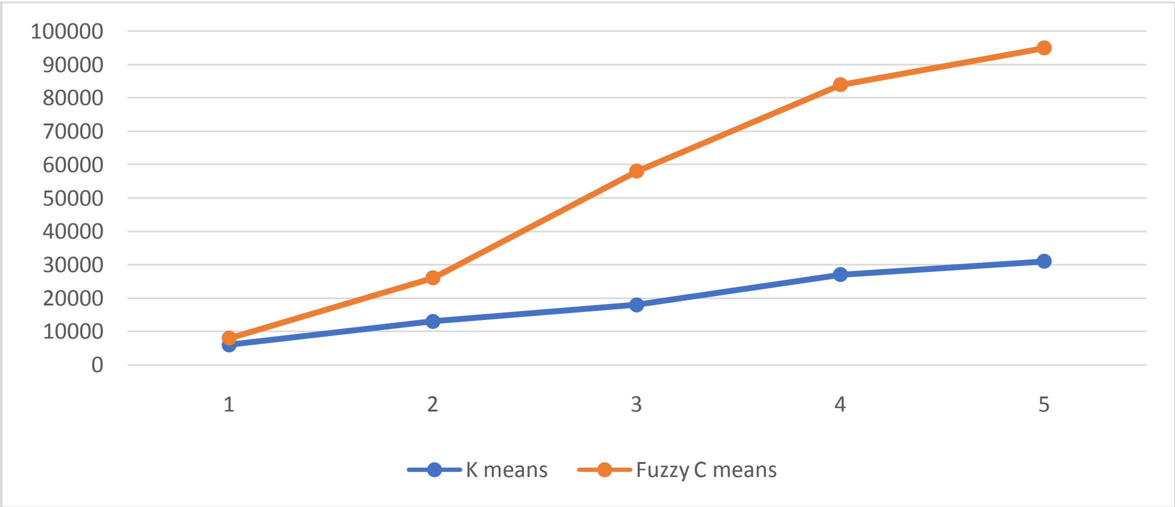

Figure 8: Graphical representation of time complexity of fuzzy C means and K means clustering with respect to cluster count

The K means and the fuzzy C means has a time complexity of 18000 and 58000 when the count of the cluster is 3. When the cluster count is 4, the time complexity of K means and C means are 27000 and 84000 respectively. The K means and the fuzzy C means has a time complexity of 31000 and 95000 when the count of the cluster is 5.

5 Discussion

The hybrid cat swarm optimization and the fuzzy C means clustering techniques have been proposed in this paper. The results have been compared with the existing techniques and with different attributes to prove the effectiveness of the proposed method. The attributes taken for the experimentation are average, standard deviation and time taken for clustering. The results obtained through the fuzzy C means clustering have been compared with the C means clustering with reference to the number of iterations and the number of clusters. The results prove that the proposed techniques are much more effective than the other existing methods.

Refrences

[1] Yun, Chidi, Miki Shun, Utian Junta, and IbrinaBrowndi. "Predictive Analytics: A Survey, Trends, Applications, Opportunities' and Challenges for Smart City Planning." International Journal of Computer Science and Information Technology 23, no. 56 (2022): 226-231.

[2] Mabu, Audu Musa, Rajesh Prasad, and RaghavYadav. "Mining gene expression data using data mining techniques: A critical review." Journal of Information and Optimization Sciences 41, no. 3 (2020): 723-742.
[3] Hamerly, Greg, and Jonathan Drake. "Accelerating Lloyd's algorithm for k-means clustering."Partitional clustering algorithms (2015): 41-78.
[4] Mustafa, Hossam MJ, MasriAyob, MohdZakree Ahmad Nazri, and Graham Kendall. "An improved adaptive memetic differential evolution optimization algorithm for data clustering problems." PloS one 14, no. 5(2019): e0216906.
[5] Douzas, Georgios, Fernando Bacao, and Felix Last. "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE." Information Sciences 465 (2018): 1-20.
[6] Xie, Hailun, Li Zhang, CheePeng Lim, Yonghong Yu, Chengyu Liu, Han Liu, and Julie Walters. "Improving K-means clustering with enhanced Firefly Algorithms." Applied Soft Computing 84 (2019): 105763.
[7] Al-Zoubi, Ala' M., Mohammad A. Hassonah, Ali AsgharHeidari, HossamFaris, MajdiMafarja, and IbrahimAljarah. "Evolutionary competitive swarm exploring optimal support vector machines and feature weighting." Soft Computing 25 (2021): 3335-3352.
[8] Hemati, Sobhan. "Learning Compact Representations for Efficient Whole Slide Image Search in Computational Pathology." (2022).
[9] Rana, Madhurima, PrachiVijayeeta, UtsavKar, Madhabananda Das, and B. S. P. Mishra. "Unsupervised machine learning approach for gene expression microarray data using soft computing technique." In Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics: ICACNI 2015, Volume 1, pp. 497-506. Springer India, 2016.
[10] Noor Basha, Ashok kumar P S "Early detection of heart disease using machine learning techniques" 2019 4th International Conference on Electrical Electronics Communication Computer Technologies and Optimization Techniques (ICEECCOT) 387-391.
[11] Noor Basha, Ashok kumar P S et al." Cat-Ant Swarm Optimization Based On Repetitive Deep Learning Neural Network For Big Data Processing" International journal of European chemical bulletin volume 12 issue 10 (2023): 9180-9195.
[12] Huang, Ruyi, Yixiao Liao, Shaohui Zhang, and Weihua Li. "Deep decoupling convolutional neural network for intelligent compound fault diagnosis." Ieee Access 7 (2018): 1848-1858.
[13] Ezugwu, Absalom E., Abiodun M. Ikotun, Olaide O. Oyelade, LaithAbualigah, Jeffery O. Agushaka, Christopher I. Eke, and Andronicus A. Akinyelu. "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects." Engineering Applications of Artificial Intelligence 110 (2022): 104743.
[14] Noor Basha, Ashok kumar P S " Distance-based K-Means Clustering Algorithm for Anomaly Detection in Categorical Datasets" International Journal of Computer Applications (0975 – 8887) Volume 183 – No. 11, June 2021.
[15] Bouguettaya, Athman, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. "Efficient agglomerative hierarchical clustering." Expert Systems with Applications 42, no. 5 (2015): 2785-2797.
[16] Weikl, Fabian, Christina Tischer, Alexander J. Probst, Joachim Heinrich, IanaMarkevych, Susanne Jochner, and Karin Pritsch. "Fungal and bacterial communities in indoor dust follow different environmental determinants." PloS one 11, no. 4 (2016): e0154131.
[17] Noor Basha, Ashok kumar P S " Reduction of Dimensionality in Structured Data Sets on Clustering Efficiency in Data Mining" 2017 IEEE International Conference on Computational Intelligence and Computing Research.
[18] Saggi, MandeepKaur, and Sushma Jain. "A survey towards an integration of big data analytics to big Insightsfor value-creation." Information Processing & Management 54, no. 5 (2018): 758-790.
[19] Lähnemann, David, Johannes Köster, EwaSzczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos et al. "Eleven grand challenges in single-cell data science." Genome biology 21, no. 1 (2020): 1-35.
[20] Yang, Aimin, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, and Limin Zhang. "Review on the

application of machine learning algorithms in the sequence data mining of DNA." Frontiers in Bioengineering and Biotechnology 8 (2020): 1032.

[21] Campello, Ricardo JGB, Peer Kröger, Jörg Sander, and Arthur Zimek. "Density- based clustering." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, no. 2 (2020): e1343.

[22] Sinky, Hassan, BassemKhalfi, BechirHamdaoui, and AmmarRayes. "Responsive content-centric delivery in large urban communication networks: A LinkNYC use-case." IEEE Transactions on Wireless Communications 17, no. 3 (2017): 1688-1699.

[23] Schaufeli, Wilmar B. "Engaging leadership in the job demands-resources model." Career Development International (2015).

[24] Reddy, Chandan K., and BhanukiranVinzamuri. "A survey of partitional and hierarchical clustering algorithms." In Data clustering, pp. 87-110.Chapman and Hall/CRC, 2018.