Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 33 No. 3 (2023)

Enhancing Big Data Clustering: The Improved K-Means - Artificial Bee Colony Algorithm with MapReduce

Dr. Satish S. Banait[1], Prof. M. E. Maniyar[2], Prof. Archana L. Rane[3], Dr. Rajani P. K.[4], Dr. Jyoti S.Kulkarni[5], Prof. Swapnil S. Ayane[6]

[1] Department of Computer Engineering, K. K. Wagh Institute of Engineering Education Research, Nashik (SPPU), Maharashtra, India. ssbanait@kkwagh.edu.in

[2,3] Department of Master in Computer Application, K. K. Wagh Institute of Engineering Education Research, Nashik (SPPU), Maharashtra, India. [2]memaniyar@kkwagh.edu.in, [3]alrane@kkwagh.edu.in

[4,5,6] Electronics and Telecommunication Engineering, Pimpri Chinchwad College of Engineering, Nigdi,Pune. [4]rajani.pk@pccoepune.org, [5]jyoti.kulkarni@pccoepune.org, [6]swapnil.ayane@pccoepune.org

Article History:

Received: *10-06-2023*

Revised: *13-08-2023*

Accepted: *15-09-2023*

Abstract:

The volume and diversity of data produced by scientific applications and the corporate world have increased drastically in the modern era. Large data is challenging to collect, store, transform, and analyse. Processing vast amounts of data is a challenging task, and one major problem with big data is that it takes longer to run typical algorithms. One of the most common data mining jobs is clustering. It is applied in several fields. The K-Means clustering algorithm is widely recognized as one of the prominent unsupervised learning techniques in machine learning. The advantages involve fundamental clarity, good effect, and ease of execution. As the online world grew quickly, the rise in data collection locations occurred simultaneously, marking the era of big data and an explosion of information. This research introduces the IK-ABC Algorithm (Improved K Means - Artificial Bee Colony) to tackle various challenges encountered in k- means clustering algorithms. These issues encompass limitations in global search capabilities, the sensitivity of cluster center selection, randomness in initialization, early-stage development, and sluggish convergence observed in the original artificial bee colony algorithm. To expedite computation and enhance the effectiveness of the iterative optimization process, a custom fitness function tailored for the K-means clustering technique and a position update formula relying on global guidance was created through the utilization of MapReduce.

Keywords: Clustering Algorithms, Big Data, Swarm Optimization Techniques, MapReduce.

1. Introduction

In recent years, there has been a substantial increase in the volume of extensive data produced and stored within the realm of data analysis and big data processing. Accord- ing to some research, managing and utilizing this massive amount of data may establish a new foundation for experimentation, simulation, scientific research, and economics. Indeed, numerous prospects exist for the application of big data across diverse sectors including healthcare (enhancing treatment efficiency), transportation (cost reduction), finance (risk mitigation), administration (swift and

efficient decision-making), social media, and government services.

But in today's era, big data also comes with a lot of challenges and quality difficulties, including problems with scale, heterogeneity, privacy, timeliness, and visu- alization at every step of the analysis pipeline, from data collection to interpretation. The most modern methods and tools are employed to handle this massive amount of data to increase the efficiency and utility of data processing [1]. Cluster analysis is an essential data analysis method that aims to organize physical or abstract sets into related groups, ensuring that items in the same cluster exhibit strong similarities while differing significantly from one another. To handle big data sets, various clus- tering algorithms are employed. However, no clustering approach can address every Big Data problem [2]. Because of its simplicity, the K-means algorithm is one of the most popular among them; nonetheless, there are still significant challenges in adapt- ing it to the evolving big data era. There is still room for improvement in areas such as enhancing their clustering effect and decreasing the temporal complexity of the K-means algorithm [3].

2. Related Work

Big data is generated from various sources, necessitating high-performance, scalable systems for processing. Frequent updates are needed to keep up with the growing data volume [4]. Big data analytics involves the analysis of large datasets to uncover hidden patterns, and the complexity varies from traditional analysis to modern big data techniques [5]. The Knowledge Discovery in Databases (KDD) process serves as a framework, highlighting quality, security, and privacy concerns in computing [6].

Clustering algorithms are used to identify peak heart rates throughout the year. A hybrid methodology improves accuracy and clustering performance [7]. Clustering algorithms, such as EM and FCM, are effective but have shortcomings that need addressing [8]. K-means clustering is widely used, and research focuses on handling large or multidimensional datasets [9]. K-means is a popular method for clustering large datasets. This research presents an efficient approach that guarantees O(nk) time complexity for clustering [10]. However, K-means requires careful initial data point selection and cluster assignment. The study outlines improved techniques for more accurate data point assignment and initial centroid determination [11].

The ABC algorithm offers an adaptable and efficient solution for optimization problems [12]. ABC enhances its iterative optimization by utilizing a position update formula that combines local and global information to achieve greater effectiveness [13]. The ILABC variant of the ABC algorithm can be useful for data structuring and challenging optimization problems [14]. ABC-based systems offer efficient clustering, particularly in handling large and growing datasets [15].

IABC algorithm is proposed to address issues with the K-means algorithm, improving initial center selection and global search capabilities [16]. An ABC variant with variable-length food sources (ABCVL) improves clustering quality [17]. ABC-based clustering improves initial centroid values, resulting in increased inter-group variation and similarity in clustering outcomes [18]. A hybrid clustering algorithm combining modified ABC and K-Means is proposed [19].

MapReduce is a powerful tool for analysing large, unstructured data, and it was developed by Google [20]. Hadoop simplifies MapReduce programming, making it accessible to programmers without prior knowledge [21]. MapReduce framework is accessible to non-programmers and eliminates the need for load balancing, fault tolerance, serialization, and parallelization [22]. The K-Means algorithm in the Hadoop cluster aids in identifying valuable consumer segments [23]. K-Means Clustering is a reliable and cost-effective method for categorizing similar data, particularly when dealing with massive volumes on distributed networks. It also reduces the number of iterations needed for task completion, improving efficiency [24]. Parallel K-means methods based on Hadoop demonstrate efficient data processing, especially with larger datasets [25]. Improved K-means algorithms address initial center point sensitivity, enabling better parallelization and performance [26].

A parallelization technique for K-means in MapReduce reduces iteration time while maintaining accuracy [27]. ABC algorithms are applied in MapReduce for large-scale data clustering, offering efficiency and performance [28]. MapReduce-based optimization methods improve search quality and local search time, combining features of both MapReduce and specific methods [29]. Parallel ABC models in MapReduce leverage the parallelization capabilities of MapReduce and can handle large populations [30]. The Modified Artificial Bee Colony Algorithm (MABC) is used for optimization in cloud resource management, resulting in improved resource utilization [31].

3. Methodology

This section provides illustrations of the particular features of the suggested approach, as shown in Figure 1. To obtain the optimum model as a consequence, the suggested method made use of three essential components. Two effective tools for evaluating and comprehending big datasets are the swarm optimization algorithm, Artificial Bee Colony (ABC) and the clustering algorithm K-Means. However, the computing difficulties associated with handling massive volumes of data may make them less useful. By dividing big datasets into smaller, more manageable parts, the distributed computing framework MapReduce makes it possible to process enormous datasets efficiently.

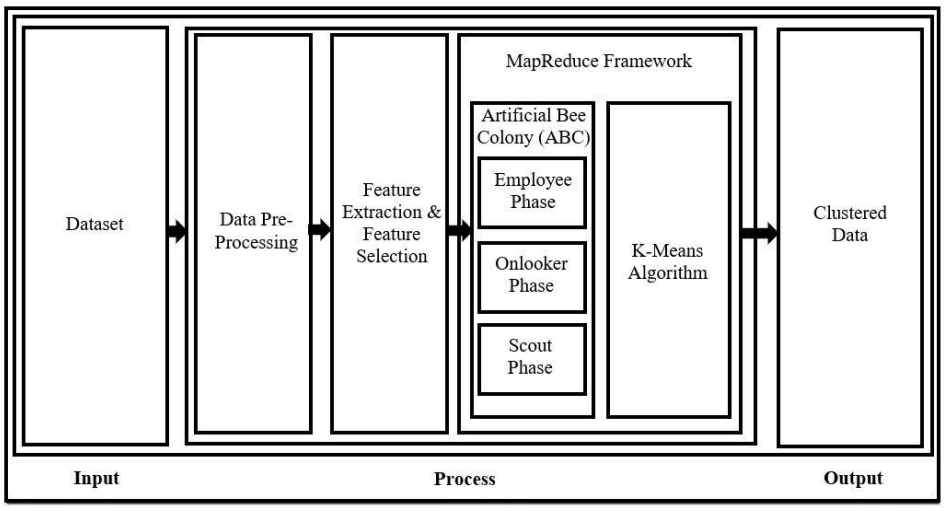

Figure 1. Block Diagram

The combination of K-Means, MapReduce, and ABC (Artificial Bee Colony) has been the subject of recent research due to its potential to improve the scalability and performance of clustering algorithms for large data processing. Apart from pinpointing the primary challenges and opportunities in this domain, this literature review article [32] aims to provide a summary of the current status of investigation on the application of MapReduce, ABC, and K-Means algorithms for clustering big data.

3.1 K-Means Clustering Algorithm

K-Means Clustering is an unsupervised learning method employed to categorize similar data points into groups. This technique separates a dataset into k predefined classes while aiming to minimize errors. Each cluster's center, represented as E_j (j = 1, 2,⋯, k) defines the cluster, and similarity is assessed using distance. $D(x_i, x_j)$ signifies the Euclidean distance between two data items, x_i and x_j, which is calculated as follows:

$$D(x_i, x_j) = \sqrt{(x_{i1} - x_{i2})^2 + \cdots (x_{iL} - x_{iL})} \qquad (1)$$

where L denotes the number of data object attributes. The primary goal of the algorithm is to minimize the within-cluster variance, which measures the similarity of data points within a cluster. Error square and SSE (Sum of Squared Errors) are utilized as objective functions to evaluate the quality of clustering and illustrate how closely the samples are grouped. A smaller SSE indicates higher similarity among the samples. Here's the formula for calculating SSE:

$$SSE = \sum_{j=1}^{k} \sum_{x \in E_j} D(x, e_j) \qquad (2)$$

$$e_j = \frac{1}{n_j} \sum_{x \in E_j} x \qquad (3)$$

where n_j denotes the number of sample data in the j^{th} cluster $E_{j..}$

3.2 Artificial Bee Colony Algorithm

Inspired by the way honeybees forage, artificial bee colonies use swarm intelligence in the form of algorithms. Using this technique, a colony of artificial bees is created, each of which represents a potential fix. These bees mimic the quest for nectar sources made by real bees, hoping to locate the finest answers within a specific problem domain.

The three primary components of the algorithm are a scouter, a follower, and a leader. The group leader finds a specific food source and informs everyone about it. In the dancing area of the hive, the follower bees first await information from the leader bees regarding the food supply. After obtaining this knowledge, they select one and start investigating it. Moreover it is the duty of the scouters to haphazardly look for fresh food sources.

There are four stages to the basic artificial bee colony algorithm.

3.2.1 Initialization Phase

The algorithm commences by initializing the bee population in a random manner. Each of the N food sources represents a valid solution, and these solutions are generated randomly within the solution space. The specific formula is outlined as follows:

$$x_{i,j} = x_j^{min} + random(0,1) \times (x_j^{max} - x_j^{min}) \quad (4)$$

where D is the dimension of the feasible solution and i=1, 2,⋯, N; j=1, 2,⋯, D. The j^{th} parameter's upper and lower bounds are represented by the variables x_j^{max} and x_i^{max}. Set a counter as well, with a value of 0, for each food source.

3.2.2 Leader Search Phase

The leader locates a new food source vi in the neighbourhood of the corresponding food source, using the following formula:

$$v_{i,j} = x_{i,j} + (-1 + 2 \times random) \times (x_{i,j} - x_{k,j}) \quad (5)$$

where k represents a collection of randomly selected food sources that are different from i, specifically k ≠ i. The "greedy selection" approach is applied to both the new and old food sources, x_i and v_i. In other words, the qualities of the new and old food sources are compared. If the quality of the new food source is superior, it is retained, and its counter is reset to zero. If the new source's quality is not better, it is discarded. If needed, the old food source is retained, but its counter is increased by one.

3.2.3 *Follower Search Phase*

Follower bees select a food source from the options at hand using a roulette wheel method. The probability of selecting each food source is as described below:

$$P_i = \frac{fit_i}{\sum_{j=1}^{N} fit_j} \quad (6)$$

where fit_i is a measure of the quality of the food source and is determined using the formula below:

$$fit_i = \begin{cases} \frac{1}{1+f_i} & f_i \geq 0 \\ 1 + |f_i| & otherwise \end{cases} \quad (7)$$

where fi is the value of the objective function.

After selecting the food source, the follower then carried out operations related to the "greedy selection" method while exploring the field in line with phase 2.

3.2.4 Scouter Search Phase

To maintain population diversity during the evolutionary phase, a unique search mode for scout bees was incorporated into the bee colony algorithm. If the counter value of a food source exceeds a predefined threshold, indicating it's exhausted or abandoned, the leading bee responsible for that source transitions to the role of a scout bee. Subsequently, a new food source is randomly generated within the feasible solution space using the phase 1 approach.

3.3 Improved K-Means – Artificial Bee Colony (IK-ABC) Algorithm

For complex clustering problems, the Enhanced K-Means Clustering Algorithm, which combines MapReduce and the Artificial Bee Colony (ABC) algorithm, offers a reliable solution. One well-known technique for grouping comparable data points is K-Means. Still, there's a chance that the

conventional K-Means algorithm will run into prob- lems with slow convergence or getting stuck in local optima. These drawbacks of the K-Means method are addressed by the Artificial Bee Colony (ABC) algorithm, which relies on population-based optimization and is inspired by honeybee foraging behaviors.

Utilizing the ABC algorithm improves the centroids' initialization in the K-Means algorithm, which raises the final cluster quality. However, there are two significant drawbacks to the basic artificial bee colony algorithm: 1) inefficiency because of random initialization, and 2) sluggish convergence because of one-dimensional domain search. The randomness problem is mitigated in this study by initializing the sealed group using the maximum and minimum distance product method. Iterative optimization also makes use of an updated fitness function and a modified formula for the position change of the global guide factor.

3.3.1 Maximum and Minimum Distance Product

The ABC algorithm produces potential solutions by randomly choosing from the search space and evaluating each solution's fitness. It adjusts the search space based on the product of maximum and minimum distances to enhance solution quality. This approach reduces the k-means algorithm's reliance on the initial point and overcomes colony initialization randomness.

3.3.2. *Fitness Function*

The ABC algorithm evaluates the quality of possible solutions by using a fitness function, focusing its search on those with higher fitness values. This method offers an effective way to explore the solution space while ensuring quick convergence to the ideal solution. The following formula defines the fitness function:

$$fitness_i = \frac{CM_i}{Dist_i}\ i = 1,2,\cdots,N \qquad (8)$$

where CM_i is the number if points belonging to class i; $Dist_i$ is the sum of distances between all objects in class *i* and center

$$C_i, Dist_i = \sum_{x_j \in C_i} d(x_j, C_k), Dist = \sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, C_j) \qquad (9)$$

3.3.1 Position Updating Formula

The Position Updating Formula serves to modify the location of each food source, which signifies a potential solution to the optimization problem. This adjustment relies on the information collected by both the employed and onlooker bees. The formula considers the food source's quality, as indicated by its fitness function, and the direction suggested by the employed bee towards a potentially superior food source. Essentially, this formula strikes a balance between exploitation (seeking the best current solution) and exploration (searching for new and potentially improved solutions) in the quest for an optimal solution.

$$v_{ij} = x_{ij} + r_{ij}(x_{mj} - x_{kj}) + \mu\,(x_{best,j} - x_{i,j}) \qquad (10)$$

where v_{ij} is a new position generated near x_{ij}; k, m, and j are random numbers generated by random formulas, k, m $\in\{1,2, \cdots,N\}$, k and m are mutually exclusive and neither is equal to i; $r_{ij} \in$ [-1,1]; μ

$\in[0,1]$ is a random number; $x_{best,j}$ is the most abundant source of honey which is greatest in proportion.

Moreover, the MapReduce framework is employed for distributed computing on the dataset, enabling the streamlined handling of extensive data volumes. Within this system, the mapper function assigns each data point to its closest centroid, while the reducer function adjusts the centroids according to the assigned data points.

The algorithm for combining the Artificial Bee Colony (ABC) optimization algorithm with the K-means clustering algorithm and MapReduce as follows:

(1) Initialize by setting up the bee population, where every individual bee serves as a potential solution for the initial centroids in the K-means clustering algorithm.
(2) Divide the dataset into more manageable segments and allocate them to various worker nodes through the utilization of the MapReduce framework.
(3) Every worker node executes the K-means clustering algorithm, with the initial centroids set as the solution provided by the bee, on the portion of data it has received. The fitness value is determined based on the quality of the clustering solution, which includes metrics like the sum of squared errors.
(4) Utilize the ABC algorithm to enhance the bee population's optimization by following these steps:
    a. Employed bees: The bees assigned to specific bees improve the solution of their associated bee through the application of a local search strategy.
    b. Onlooker bees: The onlooker bees choose a solution for an update by considering the likehood of each solutions fitness.
    c. c. Scout bees: The scout bees explore novel solution by randomly creating fresh alternatives and substituting them for the least favorable solutions within the population.
(5) The outcomes from the worker nodes are sent back to the master node, where they are amalgamated to create a unified bee population.
(6) Repeat steps 3 to 5 until the stopping criteria are met, such as a maximum number of iterations or a satisfactory level of convergence.
(7) The ultimate outcome corresponds to the finest bee, symbolizing the most favorable initial centroids for the K-means clustering algorithm.

4. Result

This section presents the data clustering performance of the proposed IK-ABC algorithm. The objective was to evaluate the outcome of the proposed approach in relation to the efficiency of the parallel algorithm and the quality of the clustering.

To assess the performance of the IK-ABC approach on a big dataset it is divided into two stages. The first stage involves the pre-processing and synthesizing of datasets. The datasets were taken from the UCI Machine Learning Repository, each of which has a different number of dimensions and clusters. These datasets are described in Table 1. The second stage involves the implementation of the

proposed IK-ABC algorithm with the MapReduce framework on the processed data from the previous stage.

Table 1. Dataset Description

Dataset	Dataset Name	Instances	Features	Clusters
1	Household Power Consumption	2075259	9	3
2	IRIS	102001	5	3

The performance metrics are assessed through an evaluation of clustering accuracy and execution time. Clustering quality is determined by employing metrics such as SSE (Sum of Squared Errors) and the Silhouette Score.

SSE (Sum Squared Error):

Sum Squared Error (SSE) is an accuracy measure. The sum of the squared differences between each observation and its group's mean. It is used as the objective function to measure the clustering quality and variation within a cluster. The lower the SSE have more accurate clustering. Calculate using the following formula:

$$SSE = \sum_{j=1}^{k} \sum_{x \in E_j} D(x, e_j) \quad (11)$$

Silhouette Score:

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Use to calculate the goodness of a clustering technique. The value of the silhouette coefficient is between [-1, 1].

$$S_i = \frac{b_i - a_i}{Max(a_i, b_i)} \quad (12)$$

Where, a = average intra-cluster distance i.e. the average distance between each point within a cluster and b = average inter-cluster distance i.e. the average distance between all clusters.

Table 2 illustrates the comprehensive performance of the IK-ABC algorithms when utilized on large datasets, consistently revealing positive results across all datasets.

Table. 2 Clustering Result with Performance Metrics

Dataset	SSE	Silhouette Score	Execution Time
Household Power Consumption	131753.25	0.57	20min 58sec
Iris	43406.47	0.61	4min 38sec

With an increase in the dataset size, the IK-ABC algorithm demonstrates improved performance. As the problem scale expands, the benefits of distributing computations across more nodes become more pronounced, resulting in decreased execution time without compromising accuracy.

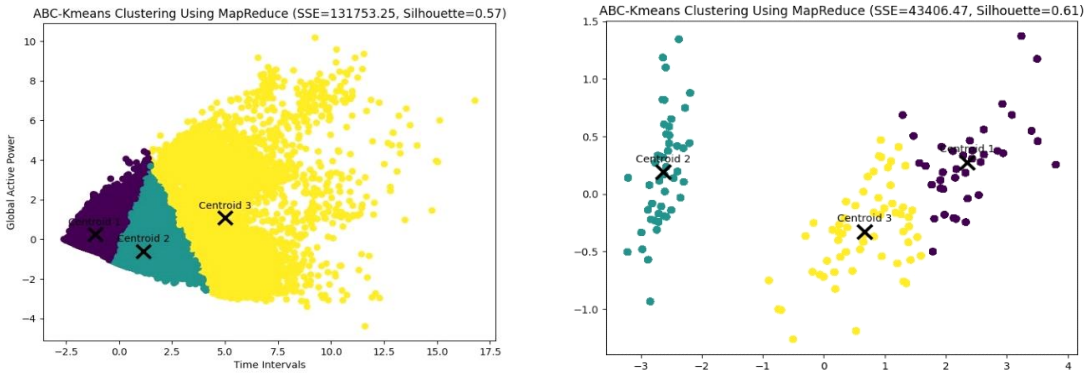

Figure. 2 Visualization of clustering results on each dataset

The K-Means method is generally quick but highly dependent on the initial cluster centroid's placement in the problem space, often converging to the local optimum nearest to the initial point. In contrast, the ABC method, when compared to other swarm optimization techniques, tends to yield superior results. The ABC approach combines both exploitation and exploration in its search process, with scout bees focusing on exploration and onlooker bees on exploitation. When some solutions get stuck in local optima, scout bees make random searches for alternatives. Moreover, our IK-ABC approach, when implemented with MapReduce, can efficiently reduce costs and processing time in large-scale data clustering. These findings suggest that our distributed method is well-suited for handling substantial input data volumes within a reasonable timeframe while maintaining high-quality solutions.

5. Conclusion

This research paper highlights the significant contribution of the IK-ABC Algorithm to the field of big data analytics. Through practical demonstrations, we have shown that our approach not only rectifies the shortcomings of traditional clustering methods but also substantially accelerates the clustering process. This work provides a powerful tool for efficiently handling large datasets, thereby enabling more effective and timely insights from the burgeoning volumes of data in today's data-driven world. The IK-ABC Algorithm promises to play a pivotal role in shaping the future of big data analytics, offering enhanced scalability and performance for diverse applications in science, industry, and beyond.

References

[1] G. Lakshen, S. Vranes, V. Janev, "Big Data & Quality- A Literature Review," *24th Telecommunications forum TELFOR*, (2016)
[2] Prajesh P. Anchalia, Anjan K. Koundinya, Srinath N. K, "MapReduce Design of K-means Clustering Algorithm," *IEEE Xplore*, (2013)
[3] Chen Jie, Zhang, Wu Junhui, Wu Yusheng, Si Huiping, Lin Kaiyan, "Review on the Research of K-means Clustering Algorithm in Big Data," *IEEE, International Conference on Electronics and Communication Engineering*, (2020)
[4] R Rawat and R Yadav, "Big Data: Big Data Analysis, Issues and Challenges and Technologies," *IOP Conf. Series-Materials Sci and Engineering*, (2021)
[5] Abdulbaset S. Albaour, Yousof A. Aburawe, "Big Data: Review Paper," *Intl Journal of Adv. Research And Innovative Ideas In Edu*, (2021)
[6] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos5, "Big data analytics: a survey," *Tsai et Journal of Big Data*, (2015)

[7] Dr. Satish S. Banait, Dr. S. S. Sane and Dr. Sopan A.Talekar, "An Efficient Clustering for Big Data Mining", *International Journal of Next-Generation Computing*, (2022)
[8] Adil Fahad, Najlaa Alshatri , Zahir Tari , (Member, IEEE), Abdullah Alamri , Ibrahim Khalil, Albert Y. Zomaya, (Fellow, IEEE), Sebti Foufou, And Abdelaziz Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," *IEEE Trans. On Emerging Topics In Computing*, (2013)
[9] Bao Chong, "K-means clustering algorithm: a brief review," *Academic Journal of Computing & Information Science, ISSN 2616-5775 Vol. 4,* (2021)
[10] S. Na, L. Xumin, G. yong, "Research on k-means Clustering Algorithm", *3rd Intl Symposium on Intelligent Info. Tech. and Security Informatics*, (2010)
[11] U. Raval, C. Jani, "Implementing & Improvisation of K-means Clustering Algorithm," *International Journal of Comp. Science & Mobile Computing, Vol.5 Issue.5*, (2016), pg. 191-203
[12] Ajit Kumar, Dharmender Kumar, S. K. Jarial, "A Review on Artificial Bee Colony (ABC) and Their Applications to Data Clustering," *Cyber And Information Technologies, Volume 17, No 3, Sofia,* (2017)
[13] Yi Yang, Ke Luo, "An Artificial Bee Colony Algorithm Based on Improved Search Strategy," *2nd Intl Conf. on Artificial Intelligence and Info Sys,* (2021)
[14] Wei-Feng Gao, Ling-Ling Huang, San-Yang Liu, and Cai Dai, "Artificial Bee Colony Algorithm Based on Information Learning," *IEEE Transactions On Cybernetics*, (2015)
[15] S. Sudhakar Ilango, S. Vimal, M. Kaliappan, P. Subbulakshmi, "Optimization using Artificial Bee Colony based clustering approach for big data," *Springer Science, Business Media, LLC, part of Springer Nature*, (2018)
[16] Zhenrong Zhang, Jiayi Lan and Zhenrong Zhang, "K-means clustering algorithm based on bee colony strategy," *2nd Signal Processing and Computer Science*, (2021)
[17] S. Raheem, M. Alabbas, "Optimal k-means clustering using abc algorithm with variable food sources length," *Intl. Journal of Electrical and Comp Engg (IJECE),* (2022)
[18] Ting-En Lee, Jao-Hong Cheng2, Lai-Lin Jiang, "A new Artificial Bee Colony Based Clustering method & its Application to the Business Failure Prediction," *International Symposium on Computer, Consumer and Control*, (2022)
[19] A. Kumara, D. Kumarb and S. Jarial, "A novel hybrid K-means & Artificial Bee Colony Algorithm approach for data clustering," *Decision Science Letters 7, (2018),* pp. 65–76
[20] P. Sudha, Dr. R. Gunavathi, "A Survey Paper on Map Reduce in Big Data," *IJSR,* (2016)
[21] Seema Maitreya, C.K. Jha, "MapReduce: Simplified Data Analysis of Big Data," *3rd International Conf. on Recent Trends in Computing*, (2015)
[22] Muthu Dayalan, "MapReduce: Simplified Data Processing on Large Cluster," *Intl Journal of Research and Engineering, Vol. 5 No. 5, (2018)*, PP. 399-403
[23] Hongqin Wang , Hongxia Wang , Li Jiang and Zhengjun Pan, "Research & application of improved K-means based on MapReduce," *Journal of Physics: Conf. Series*, (2020)
[24] P. Anchalia, A. Koundinya, Srinath K, "MapReduce Design of K-means Clustering Algorithm," *IEEE*, (2013)
[25] Jiyang Jia, Hui Xie, Tao Xu, "Design and implementation of K-means parallel algorithm based on Hadoop," *2nd Intl. Conf. on Artificial Intelligence and Info. Systems*, (2021)
[26] Li Ma, Lei Gu, Bo Li, Yue Ma and Jin Wang, "An Improved K-means Algorithm based on Mapreduce and Grid," *Intl Journal of Grid Distribution Computing Vol.8, No.1*, (2015), pp.189-200
[27] O. Lachiheb, M. Salah Gouider, L. Ben, "An improved MapReduce design of Kmeans with iteration reducing for clustering stock exchange the very large Datasets," *11th International Conference on Semantics, Knowledge and Grids*, (2015)
[28] A. Banharnsakun, "A MapReduce-Based Artificial Bee Colony for Large Scale Data Clustering," *Pattern Recognition Letters*, (2016)
[29] Parikshit Patil, Hrishikesh Mhatre, Ruchita Patil, Apurva Shinde, Bharti Joshi, "Optimization of Data using Artificial Bee Colony Optimization with Map Reduce," *ITM Web of Conf,* (2020)
[30] N. Bansal, S. Kumar, A. Tripathi, "Application of Artificial Bee Colony Algorithm Using Hadoop," *IEEE Xplore*, (2016)
[31] S.A.Gowri Manohari, Mr.S.Jawahar "Large Biological Dataset Analysis Using Enhanced Map Reducing Method With Modified Artificial Bee Colony Optimization (Mabc)," *JETIR*, (2018)
[32] Satish S. Banait, Tanuja. B. Kaklij, Gauri K. Bankar, Srushti B. Hire, Digvijay B. Wagh, "Literature Review: An Efficient Clustering Approach to Big Data", International Journal of Computer Trends and Technology, (2023)