Hybrid Data Clustering Technique: Review of Literature

*****Pradip Kumar Yadava and **Dr. Rajeev Kumar Yadav
[1]Research Scholar, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
[2]Professor, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
Corresponding author Email: - pradip.yadava@gmail.com

Abstract: Notably, real problems are increasingly complex and require sophisticated models and algorithms capable of quickly dealing with large data sets and finding optimal solutions. However, there is no perfect method or algorithm; all of them have some limitations that can be mitigated or eliminated by combining the skills of different methodologies. In this way, it is expected to develop hybrid algorithms that can take advantage of the potential and particularities of each method (optimization and machine learning) to integrate methodologies and make them more efficient. This paper presents an extensive systematic and bibliometric literature review on hybrid methods involving optimization and machine learning techniques for clustering and classification. It aims to identify the potential of methods and algorithms to overcome the difficulties of one or both methodologies when combined. After the description of optimization and machine learning methods, a numerical overview of the works published since 1970 is presented. Moreover, an in-depth state-of-art review over the last three years is presented. Furthermore, a SWOT analysis of the ten most cited algorithms of the collected database is performed, investigating the strengths and weaknesses of the pure algorithms and detaching the opportunities and threats that have been explored with hybrid methods. Thus, with this investigation, it was possible to highlight the most notable works and discoveries involving hybrid methods in terms of clustering and classification and also point out the difficulties of the pure methods and algorithms that can be strengthened through the inspirations of other methodologies; they are hybrid methods.

[Yadava, P.K. and Yadav, R.K. **Hybrid Data Clustering Technique: Review of Literature**. *The International Journal of Interpretation, Observation and Analysis*, 2024; Volume 1, Issue 1:19-23. (January-March). ISSN 2349-0713, Peer-reviewed (online/offline), Refereed, Indexed and International Journal (Since 2013), Global Impact Factor: 5.776]

Keywords: Data clustering, Literature, Clustering Techniques

Introduction: Data clustering is the process which divides a dataset into some groups or classes. It lets the data objects of the same group have high similarity, and the data objects of different groups have large differences. The similarity is often using the distance between the objects. The data clustering usually has two classes, namely the supervised clustering and the unsupervised clustering. Under the supervised clustering, learning algorithm has an external guidance signal, which offers the class marks for its data vectors. For the unsupervised clustering, there is not an external directive signal, and the algorithm groups the data vectors based on distance from each other.

Clustering is an unsupervised learning technology, and it groups information (observations or datasets) according to similarity measures. Developing clustering algorithms is a hot topic in recent years, and this area develops rapidly with the increasing complexity of data and the volume of datasets. In general, big data clustering methods can become categorized into two main groups: single-machine clustering techniques and multiple-machine clustering methods [12]. Lately multiple machine clustering techniques offers drawn more interest because they are even more versatile in scalability and provide faster response time to the users. Although the intricacy and velocity of clustering algorithms is definitely related to the quantity of situations in the dataset, but at the various other hands dimensionality of the dataset can be other important element [13]. In truth the more sizes data possess, the even more is complexity and it means the longer performance period. Sampling methods decreases the dataset size however they perform not really provide a answer for high dimensional datasets [14].

Although sampling and dimensions decrease strategies utilized in single-machine clustering algorithms enhances the scalability and velocity of the algorithms, but today the development of data size is usually method very much quicker than memory and processor developments, as a result one machine with a solitary processor and a memory cannot deal with terabytes of data and it underlines the want algorithms that can become operate on multiple machines [15].

De-duplication [16,17,18] can become divided into four measures: data chunking, chunk computation, chunk index search, and exclusive data shop. Resource de-duplication can be a well-known plan that works the 1st two guidelines of the de-duplication process at the customer aspect and chooses whether a chunk is certainly a duplicate before data transfer to conserve network bandwidth by staying away from the transfer of redundant

data, which varies from target de-duplication that performs all de-duplication techniques at the focus on side [19].

In parallel clustering [20], designers are included with not only parallel clustering difficulties, but also with information in data distribution procedure between different machines obtainable in the network as well, which makes it extremely difficult and time consuming. Difference between parallel algorithms and the MapReduce [21,22] framework is normally in the comfortless that MapReduce provides for developers and discloses them type unneeded networking complications and ideas such as weight handling, data distribution, fault tolerance and etc. by managing them instantly. This feature enables huge parallelism and less difficult and faster scalability of the parallel program. The field of data analysis and big data processing has seen a significant increase in the amount of huge data being generated and stored in recent years. Some studies argue that handling and using this huge data could become a new pillar of economics, scientific research, experimentation, and simulation. Indeed numerous chances of big data appearing in different areas similar to health (Enhancing the effectiveness of some treatments), transportation (reducing costs), finance (minimizing pitfalls), administration (decision stuff with high effectiveness and speed), social media, and government services. However, in today's era, big data is also fraught with problems and has some quality issues like issues of scale, heterogeneity, privacy, timeliness, and visualization, at all stages of the analysis pipeline from data acquisition to result in interpretation. To improve data processing's effectiveness and usefulness, the most recent techniques and technologies are used to deal with this large data [1]. Another crucial data analysis technique is cluster analysis, which aims to categorize physical or abstract sets into related object classes so that items within the same group share a high degree of similarity and differ significantly from one another. There are different clustering algorithms used to manage large sets of data. But no clustering algorithm can solve all the Big Data issues [2]. Among them, the K-means algorithm is widely used because of its simplicity, but how to make it more compatible with the development of the era of big data still faces very big challenges like how to reduce the time complexity of the K-means algorithm and improve our clustering effect still needs further optimization [3]. In this research work, we propose K-Means Clustering Algorithm with Artificial Bee Colony (ABC) algorithm and MapReduce Framework. It is a powerful approach for solving large-scale clustering problems.

Cluster analysis is a group objects like observations, events etc based on the information that are found in the data describing the objects or their relations. The main goal of the clustering is that the objects in a group will be similar or related to one other and dif erent from (or unrelated to) the objects in other groups. Extracting relevant information from large database is attaining huge significance. Clustering of relevant information from large database becomes dif icult. The major objective of this work is to proposed novel clustering methods for solving clustering problem. Data Mining is too possible to chunk away, concealed helpful acquaintance and data from profuse, imperfect, noisy, fuzzy and random realistic data. In data mining, the clustering method is one of the popular methods to be used. It is used to separate the data set into a significant set of reciprocally limited clusters with respect to relationship of data and it is used to create the more number of data in the same manner surrounded by a group and extra various among groups. Data clustering is a vital concept of mining as it partitions the given dataset into meaningful set of clusters based on data similarity. This concept enhances the computation ef iciency in the data analysis processes.

Literature Survey

In Cluster is a collection of data objects which are similar to one another within the same cluster but dissimilar to the objects in other clusters. The problem is to group N patterns into c possible clusters with high intra-class similarity and low interclass similarity by optimizing an objective function. In objective function-based clustering algorithms, the goal is to find a partition for a given value of c. The c-means algorithm represents each cluster by its center of gravity [1]. The aim of collaborative clustering is to make different clustering methods collaborate, in order to reach at an agreement on the partitioning of a common dataset. As different clustering methods can produce different partitioning of the same dataset, finding a consensual clustering from these results is often a hard task. The collaboration aims to make the methods agree on the partitioning through a refinement of their results. This process tends to make the results more similar. In this paper, after the introduction of the collaboration process, we present different ways to integrate collaboration into already existing methodologies. The implementation of fuzzy clustering has to be dealt with imprecise data that takes into consideration soft computing algorithms like c-means clustering. The fuzzy data is specifically used to deal with overlapping of data points. Whereas, the rough c-means incorporates the idea of vagueness and it is used cluster imprecise data. Rough sets are

purposed at defining clusters in terms of upper and lower approximations, which are identified by a pair of parameters while computing cluster prototypes. It is to be noted that RCM assigns objects into two distinct regions, viz., lower and upper approximations, such that objects in lower approximation ensures that the object is absolutely in the cluster while those in the upper approximation indicate possible inclusion in it. Since there is no concept of membership involved, therefore any measure of closeness of patterns to the clusters cannot be determined. The paper [2] deals with a comparative study using RIFCM [3] with other related algorithms from their suitability in analysis of satellite images with other supporting techniques which deals with proving the superiority of RIFCM with RBP in clustering with other clustering methods and other supporting metrics with and without refined which integrates judiciously RIFCM with RBP. Finally, the superiority of the RIFCM using RBP is demonstrated, along with a comparison with other related algorithms, on satellite images with NASA.org images(Hills, Drought) and national geographic photographic images(Freshwater, Freshwater valley). Several papers have used image segmentation through clustering with various applications in view [4], [5], [6], [7], [28]. A family of clustering algorithms has been established with the use of the kernel function instead of the Euclidean distance [8], [9], [10], [11], [12], [13]. Algorithms have been devised to use mode as the measure of central tendency instead of mean some more clustering algorithms have been devised [14]. Using the possibilistic approach to clustering some algorithms have been proposed [15], [16], [17], [18]. Using covering based rough sets instead of basic rough sets some algorithms have been devised [19]. Some efforts have been done to improve the speed of existing algorithms like in [20]. Clustering of time series data is done in [21]. The initial assignment of input is done arbitrarily in almost all the above algorithms. But using genetic algorithms like the firefly algorithm an algorithm is proposed in [22]. This quick growth is certainly sped up by the dramatic boost in approval of social networking applications, such as Facebook, Twitter, etc., that enable users to produce material openly and enhance the currently huge Web volume [9].

Working with Big Data, the amount of space required to shop it is normally extremely relevant. There are two primary methods: compression where we do not lose anything or sampling where we select what is the data that is usually more relations [10].

Using compression, we may consider even more time and much less space, so we can consider it as a change from period to space. Using sampling, we are dropping info, but the benefits in space may end up being in orders of degree. Using merge-reduce the little units can after that be utilized for resolving hard machine learning complications in parallel processing [11]. Despite that the info found out by data mining can become extremely useful to many applications; people possess demonstrated raising concern about the additional part of the gold coin, specifically the privacy risks presented by data mining.

This section covers a broad review of big data difficulties, clustering algorithms, in particular, the KMeans Clustering Algorithm, the Artificial Bee Colony Algorithm, and the MapReduce Framework and big data applications. The development of big data has led to the analysis of a wide variety of data formats, most of which are streaming in nature. As a result, conventional techniques have a difficult time meeting Big Data needs.

Big data are generated through internal and external sources of data; thus, existing systems fail to handle the unprecedented data. High-performance, highly scalable systems with advanced techniques are required to process valuable information. The study shows that the current tool and technology must be updated with time as the data is continuously growing [4]. The term "big data" describes a collection of numerical data generated by applying new technologies for either personal or professional usage. Big data analytics is used to analyze large amounts of data to find hidden patterns. The complexity of the analysis of this data, however, varied depending on the process that was needed [5], from traditional data analysis to the more current big data analysis and data analytics. The KDD process serves as the study's framework from a systems perspective. The unresolved problems with computing are discussed, resulting in quality, security, and privacy [6]. By grouping data using a variety of clustering algorithms, we set out to identify the day of the year with the greatest heart rate. A more effective clustering technique with improved accuracy, recall, and F-measure is produced via hybrid methodology. The hybrid technique produces the most clusters and includes each data point in each cluster [32]. EM and FCM clustering algorithms exhibit good performance in terms of the quality of the clustering outputs. Future research should address each clustering algorithm's shortcomings because none performs well for all evaluation criteria [8]. K-means clustering is a highly traditional clustering algorithm, and its use will increase over time. Future research may enhance the capability to handle large or multidimensional data sets. An area of study is the clustering of exponential data using K-Means [9]. A popular

clustering method that is frequently used for clustering massive amounts of data is K-means. An effective method for clustering data points is presented in this research. The suggested approach guarantees that clustering is completed in O(nk) time [10]. However, Kmeans requires initial data point selection and nearest cluster assignment. This study explains how to more accurately assign data points to their nearest clusters and determine initial centroids using improved methodologies [11]. An analysis of previous work on artificial bee colony algorithm (ABC), ABC variations, and data clustering applications. ABC is a straightforward and adaptable method that requires less parameter tuning than other algorithms. The efficiency, precision, and usefulness of ABC in solving various optimization issues are demonstrated by numerous tests conducted in the pertinent literature [12]. ABC works on position updating formula and objective function. The iterative optimization procedure is more effective by using a position update formula based on local better and global best [13]. An artificial bee colony algorithm based on information learning (ILABC) could be useful for data structuring and data probation. The design of wireless telecommunications networks and the flow scheduling problem illustrate difficult optimization problems that can be solved with ILABC. Applying ILABC to more difficult issues may be worthwhile [14]. Our dataset's size has constantly been growing, making it challenging to cluster the data using conventional clustering algorithms. The fastest execution time is provided by the ABC system, which is also more effective for all sorts of data. To discover the optimal fitness value, the mapper phase simulates the behavior of an employed bee. In the reducer mode, the behavior of an observer bee is simulated to optimize the clusters [15].

The ease of use and quick convergence, the clustering algorithm has become a popular technique for cluster analysis. The IABC algorithm is suggested to solve the issues with the K-means clustering algorithm's randomly chosen initial centre points and poor global search capability [16]. The k-means algorithm challenges selecting an appropriate set of parameters, such as the number of clusters k and initial centroids. For the ABC algorithm, they have not discovered any attempts to date. A novel method to generate variable-length food sources for the ABC algorithm with a variable length (ABCVL) to supply the system with an appropriate level of diversity [17]. The ABC-based cluster has improved the influence of the initial center value and increased inter-group variation and similarity in the clustering [18]. A hybrid clustering algorithm based on modified ABC and K-Means algorithms. The relative fitness of each person - the ratio between their individual and overall fitness is used to create a roulette wheel. In the onlooker bee phase, variable tournament selection is used instead of roulette wheel selection [33]. This study aimed to provide an overview of the MapReduce ideas used in big data analytics. To analyze large data, which is unstructured data like web data, Google developed Map Reduce [20]. Big data and related technologies can positively impact the company's operations. A few guidelines must be followed to acquire fast and beneficial results from big data. Programming MapReduce using the Hadoop framework, which is an open-source system, accelerates the processing of massive amounts of data [21]. Without any prior programming knowledge, programmers can simply grasp the MapReduce framework. Load balancing, fault tolerance, serialization, and parallelization are no longer required [22]. The data mining environment of the Hadoop cluster is used to study the K-means method. With the help of the improved algorithm, catering decision-makers may identify highvalue consumer segments and provide superior service. The k-Means algorithm for processing data mining has superior expansion performance and mining efficiency in a cluster of cloud computing platforms, which has been demonstrated [23].

The K-Means Clustering Algorithm offers a reliable and effective method for classifying data that have similar features. It lowers the implementation costs associated with handling such massive data volumes via a distributed network. Reducing the number of iterations needed to finish a task allows for improvements [2]. A parallel Kmeans method based on Hadoop is given in work with quite good findings for data processing effectiveness and convergence. As the amount of data increases, the acceleration effect is better for processing huge amounts of data, especially in the MapReduce architecture [25]. The standard K-means method has been enhanced. The problem of the K-Means initial center point sensitivity was resolved by the modified approach, which successfully identified the initial clustering centers. Large data processing was made possible by better algorithm parallelization. The performance of the K-means algorithm has been increased, and both techniques significantly improve results [26]. K-means algorithm improves MapReduce design using an iteration-saving technique. They illustrate that this keeps 80% of the clustering accuracy while reducing the number of iterations and execution time in clustering techniques [2]. An effective artificial bee colony for MapReducebased large-scale data clustering is developed. In the Hadoop system, the ABC could be used to streamline the clustering of enormous amounts of

data. It provides an adequate level of grouping and performance in comparison to more current methods [28]. The novel optimization method has effective search capabilities in the solution space, and a pattern is applied to achieve the best outcomes with fewer iterations. Many methods enhance the search quality and fast local search time in global search by integrating and extracting the features of both MapReduce and a specific method [29]. MapReduce's parallelization capabilities make using the Artificial Bee Colony technique simple. Each member of the population just needs to look in a very small area, which allows them to find the answer more quickly. Because the particles continually update themselves after each iteration, the proposed model for parallel ABC can use a huge population but cannot be used with a large dataset [30]. The Modified Artificial Bee Colony Algorithm is the optimization algorithm we used (MABC). A method for utilizing the map-reducing algorithm to solve resource issues in clouds. With the aid of the optimization algorithm, the MapReduce algorithm creates a further improved solution. The suggested approach to resource problem reduction works better because it requires less space for data storage [31].

References:

[1] Hartigan, J.A.; Wong, M.A.; "A K-Means Clustering Algorithm" Journal of the Royal Statistical Society. Series C (AppliedStatistics), vol. 28, no. 1, pp 100-108, 1979

[2] Nagy, G.; "Feature Extraction on Binary Patterns" IEEE Trans. Systems Science and Cybernetics, vol. 5, issue 4, pp 273 – 278, 1969

[3] Coleman, G.B.; Andrews, H.C.; "Image segmentation by clustering" Proceedings of the IEEE,vol. 67, issue: 5, pp 773 –785, 1979

[4] Bezdek, J.C.; Dunn, J.C.; "Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions" IEEE Trans. Computers , vol. C-24, issue 8, pp 835 – 838, 1975

[5] Bezdek, James C.; "A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-2, issue 1, pp 1 – 8, 1980

[6] Eberhart, R.; Kennedy, J.; "A new optimizer using particle swarm theory", Sixth Int. Symp. Micro Machine and Human Science, pp 39 – 44, 1995

[7] Backer, Eric; Jain, Anil K.; **"**A Clustering Performance Measure Based on Fuzzy Set Decomposition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-3, issue 1, pp 66 – 75, 1981

[8] Selim, Shokri Z.; Ismail, M. A.; "K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-6, issue 1, pp 81 – 87, 1984

[9] Gath, I.; Geva, A.B.; **"**Unsupervised optimal fuzzy clustering**"** IEEE Trans. Pattern Analysis

[10] Guma Abdulkhader Lakshen, Sanja Vranes, and Valentina Janev, "Big Data & Quality- A Literature Review," 24th Telecommunications forum TELFOR, pp. 1-4, 2016.

[11] Prajesh P. Anchalia, Anjan K. Koundinya, and Srinath N. K, "Map Reduce Design of K-means Clustering Algorithm," IEEE International Conference on Information Science and Applications (ICISA), pp. 1-5, 2013. [12] Chen Jie et al., "Review on the Research of K-means Clustering Algorithm in Big Data," IEEE, International Conference on Electronics and Communication Engineering, pp. 107-111, 2020.

[13] R Rawat and R Yadav, "Big Data: Big Data Analysis, Issues and Challenges and Technologies," IOP Conference SeriesMaterials Science and Engineering, vol. 1022, 2021.

[14] Abdulbaset S. Albaour, and Yousof A. Aburawe, "Big Data: Review Paper," International Journal Of Advance Research And Innovative Ideas In Education, vol. 7, no. 1, 2021.

[15] Chun-Wei Tsai et al., "Big Data Analytics: A Survey," Journal of Big Data, vol. 2, no. 20, 2015.

[16] Fatema Jamnagarwala, and P.A.Tijare "Implementation of Data Mining With lustering of Big data for Shopping mall's data using SOM and K-means Algorithm," International Journal of Computer Trends and Technology, vol. 67, no. 12, pp. 3-7, 2019.

[17] Adil Fahad et al., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, 2013.

[18] Bao Chong, "K-Means Clustering Algorithm: A Brief Review," Academic Journal of Computing & Information Science, vol. 4, no. 5, 2021.

[19] Shi Na, Liu Xumin, and Guan Yong "Research on k-means Clustering Algorithm", 3 rd Intl Symposium on Intelligent Information Technology and Security Informatics, pp. 63-67, 2010.