IJCSMA

Jinan Redha Mutar, International Journal of Computer Science and Mobile Applications,
Vol.10 Issue. 10, October- 2022, pg. 44-50 **ISSN: 2321-8363**
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

A Review of Clustering Algorithms

Jinan Redha Mutar
Department of Computer Science, Collage of Education, Mustansiriyah University, Baghdad, Iraq
jinan_redha@uomustansiriyah.edu.iq

DOI: 10.5281/zenodo.7243829

Abstract:

Clustering is an unsupervised artificial intelligence methodology that has emerged as a good learning tool for evaluating the massive amounts of datasets made available by today's applications. There is an affluence of information available in the field of clustering, and many endeavors have been made to identify and evaluate it for a spectrum of uses; however, the major disadvantages of someone using a classical classification algorithm for big data analysis are their elevated complicity, humongous volume, variety, and generation rate. As a result, typical clustering algorithms for processing such data are rapidly becoming obsolete. This presents researchers with exciting problems in inventing new scalable and cost-effective clustering algorithms capable of extracting relevant information from huge volumes of data collected in numerous aspects of life. In this review, we categorize the review of big data with techniques clustering by identifying the main research concerns. Then, for each subject, we therefore provide an up-to-date review of research papers.

Keywords: clustering, Data, Unsupervised, grouping, Data collected.

1. INTRODUCTION

Clustering is classificationerized as that of the grouping of related goods into different categories classifications known as onliters [1]–[2], or more precisely as an unsupervised-supervised teaching method for categorizing Structures are classified into groups (clusters) based on their similarity, where a pattern is classified as an illustration of an object's features or attributes [3]. One of the most researched data mining tasks is data clustering. It seeks to find previously unknown groups within data sets using various methods [4]–[5]. Unsupervised clustering methodologies exist since they wouldn't know the categorization characteristics [6]–[7], data qualities, or even the amount of cluster communication between participants' classification techniques. As a result, clustering-based approaches attempt to estimate and learn these properties from available data. In general, it appears that there are two techniques for carrying out this procedure: offline for a specific preserved collection of information and online for coming data sequentially. Although offline methods are more accurate, they are ineffective for huge or real-time data sets [8]–[9]. Clustering is now widely acknowledged as a significant data-mining tool for data analysis. Categorization is useful in a variety of domains, including the ones listed below:

- Examining social networks.
- Filtering in collaboration.
- Data compilation.
- Analysis of multimedia data.
- Segmentation of custom.
- Analyzing biological data.

These several programs generate a variety of data types with varying functionalities. Quantitative methods, data categorical, textual, audiovisual, time - series data, intermittent patterns network data, and uncertain data are the most typical types of this data (Figure 1). Each of these data categories necessitates its own pre-processing or processing before using any data mining technology.

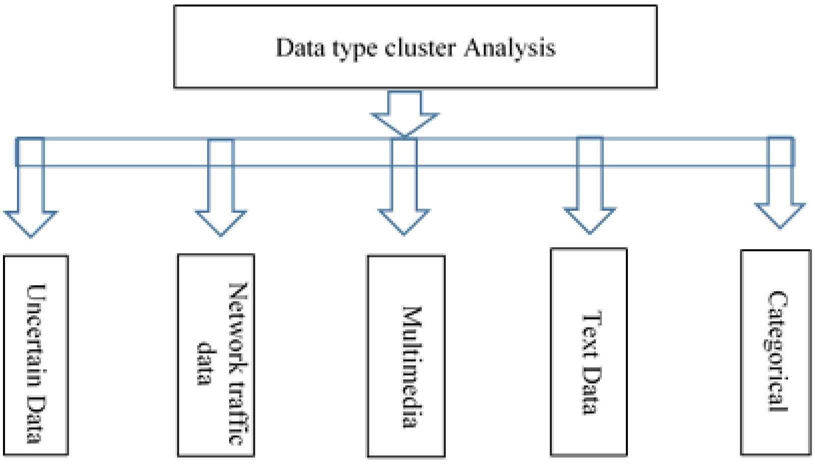

Figure (1) Cluster analysis using several data types

Dealing with enormous volumes of data has become unavoidable as data quantity and rate have increased. One definition of Big Data is the amount of information that just exceeds the capacity of technologies to store, manage, and program [10]. Big data is assumed to be enormous, complex, and expanding from separate or unconnected sources [11]–[12]. Big data has quickly emerged in many sectors and disciplines because of significant advancements in communication and data storage technologies, including scientific, architectural, physiological, pharmacological, and biomedical sciences [13]. Furthermore, Numerous potential apps can provide huge quantities of data in a short period of time; for example, social networks offer both huge opportunities for social relationships and immense troves of information [14]. Data streams, in the same vein, refer to huge amounts high-speed data streams, such as network traffic, site click streams, and sensor networks [15]. The continual emergence of new technologies and applications that generate data streams has boosted the attraction of streaming data mining.

2. RELATED WORK

The contemporary study's methodological and data statistics are totally based upon bibliometric data, which encompasses all characteristics of a publication, including authors, journal, keywords, territories, publication date, and so on. Web of Sciences (WoS), Scopus, Google Scholar, and other comparable services generally index this bibliometric data. In this effort, we used only the WoS database because it indexes only high-quality publications and recognized international conferences, ensuring that the articles are of excellent quality. This research relies on a selection of the most current recent automatic clustering methods.

Ezugwu [16] This study presented a current assessment of the most important to deal with automatic clustering challenges, meta-heuristic algorithms based on the environment have been used. To overcome concerns with automatic clustering, a comparison search of different customized well-known global metaheuristic approaches is performed. Three swarm intelligence and evolutionary algorithms are created to address the difficulty of automatic data grouping: particle swarm evolutionary algorithm, The invasive weed optimization evolutionary algorithm, and the firefly evolutionary algorithm technique The clustering methods introduced in this paper, Unlike many other classical and Darwinian computational clustering algorithms, classification somehow doesn't necessitate any prior knowledge or knowledge of the dataset Rather, algorithms would determine the optimal amount of partitions for the chunks of information even during the partitioning process. program execution. Forty-one performance test datasets are gathered and used to assess the effectiveness of common nature-inspired clustering algorithms, eleven synthetic datasets, and thirty real-world datasets Based on extensive experimental data, the firefly technique was discovered to be more suitable than other cutting-edge methodologies, comparisons, and statistical significance for improved cluster formation of both low-dimensional and high-dimensional datasets Experiments findings also reveal that the three suggested composite techniques outperform conventional government approaches when it comes to discovering meaningful grouping solutions for the problem at hand.

Hatamlou [17] This study presents a novel heuristic algorithm based on the phenomenon of black holes. The black hole algorithm (BH), like other community algorithms, begins considering an original population of plausible improvement strategies and a nonlinear objective that is determined for individuals Each time the black hole procedure is completed, the strongest applicant is selected to fill the black hole, which attracts further individuals known as starlight, around this one. Star that becomes too close to a black hole will indeed be swallowed by it and eventually vanish permanently. In this case, at different intervals, a single star (possible solution) is manufactured and placed in the search region, resulting in the beginning of a fresh search. The NP-hard clustering challenge is used to assess the performance of approaches such as the black hole technique. On a range of benchmark datasets, the experimental results show that even the theorized black hole technique outperforms existing traditional heuristic algorithms.

Agbaje et al [18] This research looked at the search algorithm is a based on genetic algorithm improvement strategy inspired by nature that has emerged as a significant method for tackling the most complex multi-objective classification problem in practically all efficiency approaches including agile methodologies. The performance of the firefly technique, like that of other metaheuristic algorithms, is dependent on excellent parameter selection. Furthermore, when used to handle issues with a high diversity of variables, such as statistical data analysis, its multiplicity as a global metaheuristic translates into slower performance and a lower degree of convergence. Clustering is an unsupervised data analysis technique that uses input parameters to identify homogeneous object groupings. To address the limitations noted above, a modified random forest algorithm is integrated with the well-known particle swarm optimization approach to address the problems of automated or semi-automated- automated clustering. Using twelve traditional samples from the UCI Machine Learning Repository and the twin moons dataset, the effectiveness of the new hybrid technique is compared to four common metaheuristics from the literature. Higher processing trials and data analysis revealed that the suggested technique not only outperforms conventional dragonfly and particle-swarm optimization procedures, but it also has a high level of predictability and may be apply other high-dimensional grouping situations.

ana et al [19] The K-Means technique is the most usually applied split classification method in this study simply because it is simple to construct and boasts the minimum execution time. It can, however, merely produce a partial optimization algorithm considering that it is dependent upon the initial partitioning. Although providing a broad search procedure, the Particles Swarm Optimization (PSO) algorithm suffers from prolonged completion approaching the optimum alternative. In this research, they suggest a novel Hybrid Sequencing classification method involving text categorization that combines PSO with an algorithm of K-Means algorithm. The suggested technique solves the flaws of both methods, promotes aggregation, and precludes remaining caught in a locally effective solution. Four distinct types of information sets were used in the investigations. In comparing to the conclusions of the K-Means, PSO, Hybrid, and K-Means+Genetic Approach, the proposed technique provides more accurate, resilient, and superior grouping results.

Aliniya et al [20] In this study, called "the automated clustering utilizing ICA (AC-ICA)," the Imperialist Competitive Algorithm (ICA) is utilized for the first time to deal with automatic clustering issues. To improve exploring capacities, the proposed methodology altered the movement of colonies toward imperialism during the assimilation phase. When joining a randomized as well as cultural integration intermix strategy, a new mechanism for modifying the number of clusters has been devised. A density-based strategy for continuation plans for unoccupied cluster centers has also been developed. The setup and competitiveness stages were changed to allow AC-ICA to be deployed independently. Based on changes within those two techniques, a framework for modifying various types of ICA to address automatic clustering difficulties was proposed. The basic ICA and its three newly discovered kinds were then changed, and its automatic grouping capabilities were compared to AC-ICA. The experiments used six 10 real-world data sets and 10 synthesized data repositories. The proposed technique outperforms basic ICA, its three presumably developed variations, and several state-of-the-art automatic classification methods in terms of speed of resolution to the optimal solution and quality of the manufactured solution. Researchers have applied our technique to a practical application (facial recognition), and the results appeared to be satisfactory.

Senthilnath et al [21] In this research This study investigates a grouping task utilizing three technology learning models are influenced by nature: Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Cuckoo Search (CS). In conjunction with the training method, Levy soaring is used. The levies flight's heavy-tail feature is used here. Three standard data sets and one legitimate multi-spectral satellite dataset are used to test these approaches. Many approaches are used to tabulate and analyze the results. Finally, we show that for the great majority of the dataset, the cuckoo search operates efficiently under the specified defined set of parameters, with levy flying playing a crucial role.

Sharma et al [22] This paper introduces the first sustainability grouping HPSOM for clustering data generated by multiple networks by integrating PSO with an evolutionary algorithm. The information recorded by such networks is often dynamic and unstructured, with no resolved number of clusters. As a result, the proposed approach will be developed further It will be used as an AHPSOM for autonomously manufacturing and adjusting clusters among mobile network devices, as well as to aid in the building of long-term clusters. First, the core HPSOM performance is tested on six real-world sets of data and compared when it comes of to CD and convergence speed to other well-known hierarchical clustering protocols. The effectiveness of AHPSOM is then measured against some of the most notable cutting-edge artificial clustering approaches, which are tested on a variety of simulated and real-world datasets (Group dimensions, interpersonal and inter distance, inter-cluster length, ARI, and F-measure are all variables involved.). The results show that the proposed technique produces clusters that are generally distinct, concise, attractive long-lasting.

Bouyer et al [23] This paper describe data a clustering hybrid strategy that utilizes algorithms for cuckoo search and differential evolution Locust Search (CS) is a recently proposed novel swarm-intelligence-based technique. This approach has fewer control settings and can handle a wide range of situations; nonetheless, it has substantial limitations, such as a large number of functionalizations and a proclivity to become caught in regional minimums The Differential Evolution Method (DE) and the Commonly found levy distribution are used in the suggested technique to assist the CS algorithm in completing a significant number of operational evaluations while also attaining fast convergence and precision in a short period of time. Six real standard benchmark datasets from the UCI machine learning collections were used to demonstrate the proposed technique. According to the results of the simulation, the suggested technique may surpass existing methods in terms of cluster formation, accuracy, and frequency of function evaluations.

Rajah et al [24] As a consequence, this paper offers five novel hybrid symbiotic creature search methods for autonomously separating datasets when the quantity of clusters present is unknown. The Davies-

Bouldin clustering validity index will be used to evaluate the solution quality of the methodologies. The simulation results suggest that the suggested symbiotic microorganisms search optimization by using the particle swarm's method outperforms earlier hybrid methodologies based on particle swarms in regards to performance.

Abdulwahab et al [25] This research looked at The Black Hole (BH) efficient algorithm created as a resolution to knowledge grouping problems; it is a demographically heuristic algorithm that simulates the preponderance of black holes in the data nature. Each solution in motion within the search window symbolizes a single star in this scenario. The classic BH outperforms it whenever implemented to a benchmark dataset, it performs well; however, it lacks exploratory skills in some populations. This work incorporates the levy flight into the BH algorithm to address the investigation problem, providing a customized data classification designated as that the "Levy Flight Black Hole (LBH)," which was later released. The step size of the Levy distribution affects the movement of each star in LBH. When the value steps are large, the star moves to a location far from the current black hole, as well as vice versa Using a large number of unimodal and multimodal numerical optimization problems, the performance of LBH in terms of identifying the best solutions, avoiding becoming stuck in local optimums, and the convergence rate has been explored. LBH is then tested using six real datasets from the UCI machine learning lab. Experimental findings revealed that the suggested technique for data clustering was successful and robust. The type algorithms and details of each algorithm will be shown in the following table (1).

Table (1) The type Algorithms of clustering

Nm	Clustering Algorithm	area of Application	Type of Clustering	Cluster validity index (CVI)
1	FA, DE, PSO, IWO	Cluster analysis	Automatic	DB and CS index
2	BH	Cluster analysis	Non-automatic	Error rate Inter-cluster distance Intra-cluster distance
3	FAPSO	ClFASO analysis	Automatic	DB and CS index
4	Hybrid sequential	Cluster analysis	Non-automatic	Intra-cluster distance Inter-cluster distance Quantization error
5	AC-ICA	Cluster analysis Face recognition	Automatic	Purity entropy Rand index (PERI) ARI F-measure
6	PSO, GA, CS	Cluster analysis	Non-automatic	CEP Statistical significance test
7	HPSOM AHPSOM	Cluster analysis Mobile network data	Automatic	Inter-cluster distance Intra-cluster distance Adjusted Rand index (ARI) F-measure
8	HCSDE	Cluster analysis	Non-automatic	Inter-cluster distance Intra-cluster distance Error rate
9	SOS, SOSFA, SOSDE, SOSTLBO, SOSPSO	Cluster analysis	Automatic	DB and CS index
10	LBH	Cluster analysis	Non-automatic	Euclidean distance

3. CONCLUSION

The value of clustering analysis in practice cannot be emphasized, especially in applications requiring selection and experimental pattern analysis utilizing large-scale datasets important information extraction Gathering information from a large number of data samples with tens of thousands of measurements is a difficult task in most cases. As a result, unsupervised learning algorithms endowed Only with strategies that have shown to be beneficial in large-scale data analysis have techniques that can substantially assist computing capacity in researching and comprehending the complex structural component of almost any metadata. Recent breakthroughs. Academics have invented a variety of the simplest yet robust database abstractions and analysis techniques, which do not demand any preexisting information about the information to be processed, thanks to the use of nature-inspired meta-heuristic optimization techniques. This review provided a comprehensive overview examination of a really well clustering algorithm in the format of a detailed, cutting-edge database. A classification A set of clustering techniques is described, as well as studied. The emphasis seems to be on grouping approaches and the investigation of autonomous clustering algorithms. The publication and citation structures are investigated from the early 1990s until 2022. Articles comprised 98.25 percent of the retrieved publications. Every year, the number of publications and citations has increased considerably, which can in nearly all fields, this can be related to the contemporary era of big databases.

References

[1] Kaufman L, Rousseeuw PJ. “Finding groups in data: an introduction to cluster analysis”. *John Wiley & Sons*; 2009 Sep 25.

[2] Hastie T, Tibshirani R, Friedman J. “Prototype methods and nearest-neighbors”. *The elements of statistical learning* 2009 (pp. 459-483). *Springer, New York, NY.*

[3] Berry MW, Browne M. “Lecture notes in data mining”. World Scientific; 2006.

[4] Kaisler SH, Armour FJ, Espinosa A, Money WH. “Big Data and Analytics Challenges and Issues” 2014.

[5] Wu X, Zhu X, Wu GQ, Ding W.” Data mining with big data”. *IEEE transactions on knowledge and data engineering*. 2013 Jun 26;26(1):97-107.

[6] Tsai CW, Lai CF, Chao HC, Vasilakos AV.” Big data analytics: a survey”. *Journal of Big data*. 2015 Dec;2(1):1-32.

[7] Abualigah LM, Khader AT.” Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering”. *The Journal of Supercomputing*. 2017 Nov;73(11):4773-95.

[8] Abualigah LM, Khader AT, Hanandeh ES. “A new feature selection method to improve the document clustering using particle swarm optimization algorithm”. *Journal of Computational Science*. 2018 Mar 1;25:456-66.

[9] Ezugwu AE, Ikotun AM, Oyelade OO, Abualigah L, Agushaka JO, Eke CI, Akinyelu AA. “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects”. *Engineering Applications of Artificial Intelligence*. 2022 Apr 1;110:104743.

[10] Oyelade J, Isewon I, Oladipupo O, Emebo O, Omogbadegun Z, Aromolaran O, Uwoghiren E, Olaniyan D, Olawole O. “Data clustering: Algorithms and its applications”. *In 2019 19th International Conference on Computational Science and Its Applications* (ICCSA) 2019 Jul 1 (pp. 71-81). *IEEE*.

[11] Agrawal D, Budak C, Abbadi AE, Georgiou T, Yan X. "Big data in online social networks: user interaction analysis to model user behavior in social networks". *In International Workshop on Databases in Networked Information Systems* 2014 Mar 24 (pp. 1-16). *Springer, Cham.*

[12] Ezugwu AE, Shukla AK, Agbaje MB, Oyelade ON, José-García A, Agushaka JO. "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature". *Neural Computing and Applications.* 2021 Jun;33(11):6247-306.

[13] Nguyen HL, Woon YK, Ng WK. "A survey on data stream clustering and classification". *Knowledge and information systems. 2015* Dec;45(3):535-69.

[14] Zubaroğlu A, Atalay V. "Data stream clustering: a review". *Artificial Intelligence Review.* 2021 Feb;54(2):1201-36

[15] Regin R, Rajest SS, Singh B. "Spatial data mining methods databases and statistics point of views". *Innovations in Information and Communication Technology Series. 2021* Feb 28:103-9.

[16] Ezugwu AE. "Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study". *SN Applied Sciences. 2020* Feb;2(2):1-57.

[17] Hatamlou A. Black hole: "A new heuristic optimization approach for data clustering". *Information sciences*. 2013 Feb 10;222:175-84.

[18] Agbaje MB, Ezugwu AE, Els R. "Automatic data clustering using hybrid firefly particle swarm optimization algorithm". *IEEE Access*. 2019 Dec 19;7:184963-84.

[19] Rana S, Jasola S, Kumar R. "A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm*". International Journal of Engineering, Science and Technology*. 2010;2(6).

[20] Aliniya Z, Mirroshandel SA. "A novel combinatorial merge-split approach for automatic clustering using imperialist competitive algorithm". *Expert Systems with Applications.* 2019 Mar 1;117:243-66.

[21] Senthilnath J, Das V, Omkar SN, Mani V. "Clustering using levy flight cuckoo search". *In Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications* (BIC-TA 2012) 2013 (pp. 65-75). *Springer, India*.

[22] Sharma M, Chhabra JK. "Sustainable automatic data clustering using hybrid PSO algorithm with mutation". *Sustainable Computing: Informatics and Systems.* 2019 Sep 1;23:144-57.

[23] Bouyer A, Ghafarzadeh H, Tarkhaneh O. "An efficient hybrid algorithm using cuckoo search and differential evolution for data clustering". *Indian Journal of Science and Technology.* 2015 Sep;8(24):1-2.

[24] Rajah V, Ezugwu AE. "Hybrid symbiotic organism search algorithms for automatic data clustering*". In 2020 Conference on Information Communications Technology and Society (ICTAS)* 2020 Mar 11 (pp. 1-9). *IEEE.*

[25] Abdulwahab HA, Noraziah A, Alsewari AA, Salih SQ. "An enhanced version of black hole algorithm via levy flight for optimization and data clustering problems". *IEEE Access*. 2019 Aug 22;7:142085-96.