Review Article

ACCENTS Transactions on Information Security, Vol 8(32)
ISSN (Online): 2455-7196
http://dx.doi.org/10.19101/TIS.2023.829003

Advancements in big data clustering: methods, applications, and insights

Chandan Kumar Soni* **and Mohan Kumar Patel**
School of Computer Science and Engineering, Madhyanchal Professional University, Bhopal, India

Received: 12-September-2023; Revised: 10-October-2023; Accepted: 15-October-2023

Abstract

The digital age has given rise to an unprecedented influx of data, marking the era of big data. In this landscape, clustering has emerged as a critical element of data analysis, enabling the discovery of latent patterns in vast datasets. This review paper explores the state-of-the-art in big data clustering, encompassing influential research, methodologies, advantages, and limitations. The paper highlights the significant advantages brought by different clustering algorithms, spanning domains from smart grids and education to e-commerce and different operations. However, it also acknowledges limitations such as scalability issues and generalization challenges, underlining the importance of addressing these constraints for future research.

Keywords

Big data clustering, Data mining, Unsupervised learning, Clustering algorithms.

1.Introduction

The ever-accelerating digital revolution has engendered a torrential deluge of data that has, in turn, ushered in the era of big data [1, 2]. This exponential growth in data generation, largely facilitated by advances in technology and the proliferation of digital platforms, has unleashed a cascade of opportunities and challenges [3]. Amidst this torrent of data, the process of extracting meaningful insights and valuable information has become a critical undertaking. At the forefront of this endeavor is the field of clustering, a pivotal element of data analysis that enables the discovery of underlying structures, patterns, and relationships within vast datasets [4–7].

Clustering, as a subdomain of unsupervised machine learning, entails the grouping of data points into clusters or categories based on shared characteristics, thereby allowing for the exploration of latent patterns and the identification of homogeneous subsets within the data [8–10]. This fundamental process has found applications across diverse domains, including but not limited to marketing, healthcare, finance, and scientific research [11, 12]. Its importance has been further accentuated with the advent of big data, where traditional data processing techniques prove inadequate in the face of sheer volume, velocity, and variety [13–17].

In recent years, the field of big data clustering has witnessed a surge in research and innovation [18–20]. Researchers and data scientists have been diligently developing novel clustering algorithms and approaches tailored to the unique challenges posed by vast and complex datasets. The motivation behind this review is rooted in the intrinsic value of clustering within the big data landscape. Clustering not only enables data reduction and pattern discovery but also underpins essential tasks such as anomaly detection, recommendation systems, and more [19–21].

The motivation for this comprehensive review of big data clustering stems from three key factors: the burgeoning volume and diversity of data, the wide-ranging applications across industries, and the ongoing evolution of clustering algorithms. As data-driven strategies become essential, understanding the latest clustering developments is crucial. The objective of this review paper is to comprehensively examine and synthesize the advancements, methodologies, contributions, and limitations in the field of big data clustering. By analyzing the notable contributions of various research endeavors, we aim to shed light on the state of the art, emerging trends, and challenges within this dynamic domain. The primary objective of this review paper is threefold: to explore, assess, and synthesize the state of the art in big data clustering. Our review entails exploring influential big data clustering research, critically

*Author for correspondence

assessing strengths and limitations, and synthesizing findings to identify trends, themes, and practical applications across domains. *Figure 1* explores the big data clustering valuable utility.

This paper is structured as follows: In Section 2, we provide a summary of key clustering algorithms and their applications across various domains. Section 3 covers the discussion of advantages and limitations found in the reviewed papers. Finally, in Section 4, we emphasize the significance of addressing these limitations in the context of big data clustering and discuss the paper's contribution and concluded it.

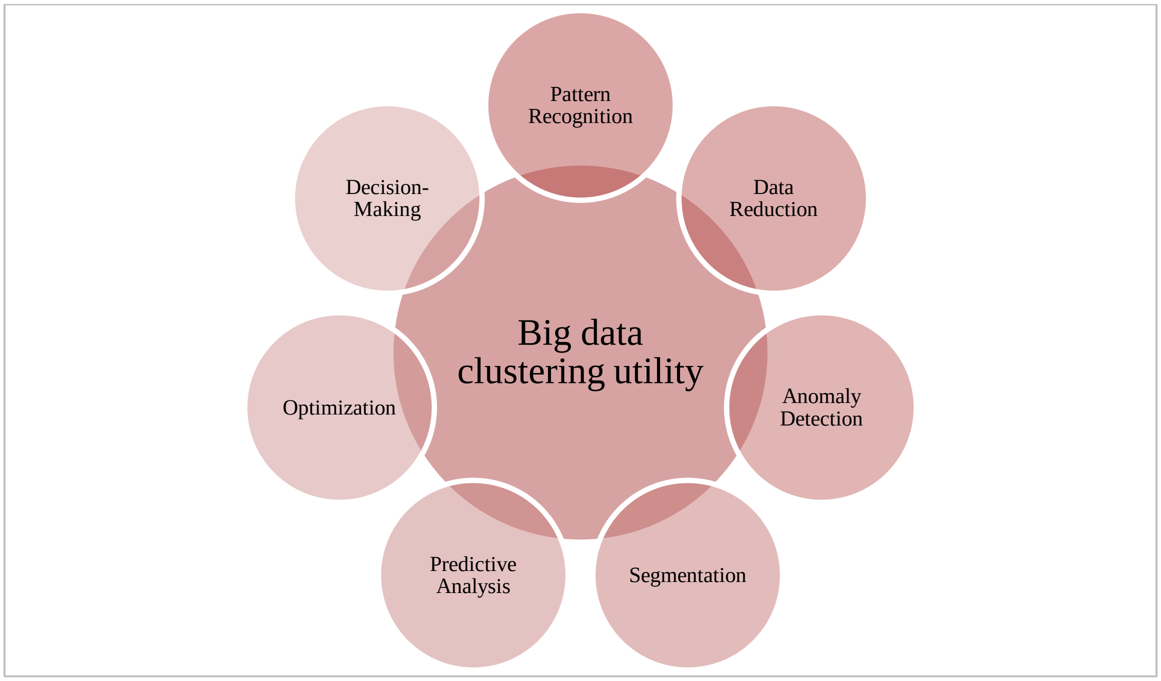

Figure 1 Big data clustering utility

2.Literature review

In this section related work has been discussed considering the method, results, advantages, and limitations.

In 2021, Du et al. [22] introduced the random sample partition-based clustering ensemble (RSP-CE) algorithm for big data clustering. It involves generating base clustering results on data blocks, harmonizing results with MMD criterion, and refining the clustering outcome. RSP data blocks ensure consistent distributions, yielding superior results and faster training times.

In 2021, Li et al. [23] addressed the need for advanced power applications in smart grids by introducing a curve-mean clustering algorithm for load big data. This algorithm improves data analysis accuracy and cluster selection by leveraging the low-rank property of load data, singular value calculation, and experimentation, enhancing smart grid data quality and analysis capabilities.

In 2021, Wang [24] focused on the application of data mining in education big data. The study highlighted the growing importance of data mining technology in education but noted existing deficiencies. Wang proposed an optimization scheme that enhanced data normalization, clustering, and prediction accuracy, achieving a 99.2% improvement in prediction accuracy compared to traditional methods. The research emphasizes the future role of data mining in education management.

In 2021, Shanshan and Zhiqiang [25] studied complex network-based statistical analysis in the context of big data and artificial intelligence. They developed clustering algorithms to reveal real network cluster structures, aiding in various applications.

In 2022, Shi et al. [26] introduced a big data classification algorithm for e-commerce using artificial intelligence. It enhances efficiency and reduces redundancy, utilizing the fast Spark architecture and vertical sequence control based on data jurisdiction. The algorithm demonstrated high accuracy and efficiency in classifying tourism e-commerce data.

In 2022, Deng and Hu [27] addressed data anomalies in emergency rescue operations. They introduced a new methodology, De-duplicated Record for Similarity (DDRfS), which combines Fuzzy C-means Clustering Algorithm and Levenshtein Distance Method to enhance fuzzy search accuracy. Their tests showed improved matching accuracy in data partitioning.

In 2022, Xing et al. [28] addressed error checking in large-volume heterogeneous power big data. They introduced a K-Means clustering-based error calibration method, enhanced by the particle swarm optimization (PSO) algorithm. This approach effectively reduced errors, improving data credibility in power big data management.

In 2022, Gupta and Jain [29] discussed the growth of big data and the use of distributed analytics platforms. They addressed real-time streaming analytics and the challenges of accessing interim results in external memory within the context of Apache Spark. Their paper proposed an optimized data storage and retrieval pattern for external storage using standard tools, with an emphasis on practical production deployment.

In 2023, Mahmud et al. [30] tackled the challenge of clustering large distributed datasets. They introduced a distributed computing framework using multiple random samples to compute an ensemble result, managed in the data model. Component clustering results were integrated using two novel methods. Experimental results demonstrated the superior performance of their ensemble clustering methods in terms of efficiency and scalability.

In 2023, Wei [31] focused on spectral clustering's significance in data mining, particularly for big data. He emphasized the need to enhance spectral clustering's computational efficiency for large-scale datasets, with the growing challenge of handling terabytes or petabytes of data. Wei explored optimization through sampling and distributed parallelization using frameworks like MapReduce and Spark to achieve improved clustering outcomes and greater efficiency in big data analysis.

In 2023, Wang [32] focused on using big data technology to address elevator safety concerns. They studied data mining techniques to analyze elevator-related feature information, aiming to predict typical elevator faults and enhance elevator fault prediction services in China.

The reviewed papers cover diverse aspects of big data clustering, addressing its applications in various domains and proposing innovative algorithms and methodologies. These contributions significantly enhance data analysis, efficiency, and the quality of insights, supporting data-driven decision-making across industries.

3.Discussion and analysis

Several papers presented various advantages in the field of big data clustering. The RSP-CE Algorithm (Du et al. [22]) introduces the benefits of consistent distributions achieved through RSP data blocks, leading to superior clustering outcomes. The Curve-Mean Clustering Algorithm (Li et al. [23]) enhances data analysis accuracy, making it crucial for advanced power applications in smart grids. Additionally, it improves cluster selection by leveraging the low-rank property of load data and singular value calculations.

In the realm of education data mining, Wang [24] brings substantial advantages, achieving exceptional prediction accuracy improvements of up to 99.2% compared to traditional methods. This not only emphasizes the significance of data mining in education but also addresses existing deficiencies, highlighting its potential for enhancing education management.

Shanshan and Zhiqiang [25] delve into complex network-based statistical analysis, revealing real network cluster structures. This valuable insight has applications across various domains in big data and artificial intelligence. Furthermore, Shi et al. [26] introduce a big data classification algorithm for e-commerce that offers high accuracy and efficiency, contributing to effective data classification in this domain.

Deng and Hu [27] bring an advantage to emergency rescue operations with the DDRfS methodology, which enhances fuzzy search accuracy, leading to improved matching accuracy in data partitioning. This is particularly valuable in time-sensitive and critical situations. On the other hand, Xing et al. [28] address error checking in power big data with a K-

Means clustering-based error calibration method. Supported by the PSO algorithm, this method effectively reduces errors, enhancing data credibility in this important domain.

Gupta and Jain [29] focus on distributed analytics platforms, optimizing data storage and retrieval for real-time streaming analytics. Their practical production deployment emphasis ensures that the proposed solutions can be applied effectively in real-world scenarios. Mahmud et al. [30] introduce ensemble clustering methods that demonstrate superior efficiency and scalability, especially beneficial when clustering large distributed datasets.

In the context of spectral clustering, Wei [31] addresses the importance of improving computational efficiency for large-scale datasets. This enhancement in spectral clustering's computational efficiency offers improved clustering outcomes, which are essential for efficient big data analysis. Finally, Wang [32] employs data mining techniques to enhance elevator safety by predicting typical elevator faults. This contribution is significant in improving elevator safety and fault prediction services in China.

The limitations across these reviewed papers are diverse. Common challenges include potential scalability issues when dealing with extremely large datasets, generalization limitations of proposed algorithms to various domains, and the need for rigorous testing in real-world scenarios to validate their effectiveness. Some papers lack in-depth discussions of potential drawbacks, and others may face challenges related to computational resources required for their methods. Additionally, the transferability of findings between different data types and contexts may be a limitation. It's crucial to recognize these constraints to inform future research and refine the practical implementation of big data clustering techniques.

4.Conclusion

The evolving landscape of big data clustering presents a multitude of opportunities and challenges. This comprehensive review emphasizes the growing importance of clustering in the context of vast and complex datasets. It highlights the advantages introduced by innovative clustering algorithms across various domains and emphasizes the necessity of addressing potential limitations, such as scalability and generalization issues. As big data continues to shape diverse industries, a deeper understanding of the latest developments and their impact is crucial. This review serves as a valuable resource for researchers, data scientists, and professionals seeking to harness the potential of big data clustering in their respective domains.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Mussabayev R, Mladenovic N, Jarboui B, Mussabayev R. How to use K-means for big data clustering?. Pattern Recognition. 2023; 137:109269.

[2] Hu H, Liu J, Zhang X, Fang M. An effective and adaptable k-means algorithm for big data cluster analysis. Pattern Recognition. 2023; 139:109404.

[3] Pina AF, Meneses MJ, Sousa- Lima I, Henriques R, Raposo JF, Macedo MP. Big data and machine learning to tackle diabetes management. European Journal of Clinical Investigation. 2023; 53(1):e13890.

[4] Alghamdi A. A hybrid method for big data analysis using fuzzy clustering, feature selection and adaptive neuro-fuzzy inferences system techniques: case of Mecca and Medina hotels in Saudi Arabia. Arabian Journal for Science and Engineering. 2023; 48(2):1693-714.

[5] Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard DA, Najarian K. Big data analytics in healthcare. BioMed Research International. 2015; 2015.

[6] Dubey A, Gupta U, Jain S. Medical data clustering and classification using TLBO and machine learning algorithms. Computers, Materials and Continua. 2021; 70(3):4523-43.

[7] Jahani H, Jain R, Ivanov D. Data science and big data analytics: a systematic review of methodologies used in the supply chain and logistics research. Annals of Operations Research. 2023:1-58.

[8] Pandey KK, Shukla D. Min–max kurtosis mean distance based k-means initial centroid initialization method for big genomic data clustering. Evolutionary Intelligence. 2023; 16(3):1055-76.

[9] Li J, Herdem MS, Nathwani J, Wen JZ. Methods and applications for artificial intelligence, big data, internet of things, and blockchain in smart energy management. Energy and AI. 2023; 11:100208.

[10] Dubey AK, Kushwaha GR, Shrivastava N. Heterogeneous data mining environment based on dam for mobile computing environments. In international conference on advances in information technology and mobile communication 2011 (pp. 144-9). Berlin, Heidelberg: Springer Berlin Heidelberg.

[11] Hussin SK, Omar YM, Abdelmageid SM, Marie MI. Traditional machine learning and big data analytics in virtual screening: a comparative study. International Journal of Advanced Computer Research. 2020; 10(47):72-88.

[12] El Hilali W, El Manouar A, Idrissi MA. The mediating role of big data analytics in enhancing firms' commitment to sustainability. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(80):932-44.

[13] He W, Hung JL, Liu L. Impact of big data analytics on banking: a case study. Journal of Enterprise Information Management. 2023; 36(2):459-79.

[14] Izhar A, Rastogi A, Ali SS, Quadri SM, Rizvi SA. Feature-driven label generation for congestion detection in smart cities under big data. International Journal of Advanced Technology and Engineering Exploration. 2022; 9(86):94-110.

[15] Dubey AK, Shandilya SK. A comprehensive survey of grid computing mechanism in J2ME for effective mobile computing techniques. In 5th international conference on industrial and information systems 2010 (pp. 207-12). IEEE.

[16] Guan S, Zhang C, Wang Y, Liu W. Hadoop-based secure storage solution for big data in cloud computing environment. Digital Communications and Networks. 2023.

[17] Rani P, Lamba R, Sachdeva RK, Kumar R, Bathla P. Big data analytics: integrating machine learning with big data using hadoop and mahout. Intelligent Systems and Smart Infrastructure: Proceedings of ICISSI 2022. 2023:366.

[18] Al-Jumaili AH, Muniyandi RC, Hasan MK, Paw JK, Singh MJ. Big data analytics using cloud computing based frameworks for power management systems: status, constraints, and future recommendations. Sensors. 2023; 23(6):2952.

[19] Dubey AK, Shandilya SK. A novel J2ME service for mining incremental patterns in mobile computing. In information and communication technologies: international conference, ICT 2010, Kochi, Kerala, India, (pp. 157-64). Springer Berlin Heidelberg.

[20] Fan L. Research on precision marketing strategy of commercial consumer products based on big data mining of customer consumption. Journal of the Institution of Engineers (India): Series C. 2023; 104(1):163-8.

[21] Marichamy VS, Natarajan V. Blockchain based securing medical records in big data analytics. Data & Knowledge Engineering. 2023; 144:102122.

[22] Du X, He Y, Huang JZ. Random sample partition-based clustering ensemble algorithm for big data. In international conference on big data (Big Data) 2021 (pp. 5885-7). IEEE.

[23] Li C, Yang B, Chen X, Zhang E, Huang H, Li D. Research on smart grid big data's curve mean clustering algorithm for edge-cloud collaborative application. In international conference on wireless communications and smart grid (ICWCSG) 2021 (pp. 395-8). IEEE.

[24] Wang CL. Research on the core technology of education big data based on data mining. In 6th international conference on big data analytics (ICBDA) 2021 (pp. 5-8). IEEE.

[25] Shanshan F, Zhiqiang R. Analysis of big data complex network structure based on fuzzy clustering algorithm. In international conference on networking, communications and information technology (NetCIT) 2021 (pp. 348-52). IEEE.

[26] Shi Z, Zhang K, Liu B, Zhao Y, Zhang J, Li Z. Classification of e-commerce big data based on iterative fuzzy clustering algorithm. In international conference on intelligent transportation, big data & smart city (ICITBS) 2022 (pp. 78-81). IEEE.

[27] Deng J, Hu J. An investigation into big data of emergency rescue based on an improved DDRfs. In 4th international conference on machine learning, big data and business intelligence (MLBDBI) 2022 (pp. 52-6). IEEE.

[28] Xing W, Wu B, Liang M, Li Y, Cheng L. Research on error calibration method for power big data based on k-means clustering. In 9th international forum on electrical engineering and automation (IFEEA) 2022 (pp. 679-82). IEEE.

[29] Gupta A, Jain S. Optimizing performance of Real-Time Big Data stateful streaming applications on Cloud. In IEEE international conference on big data and smart computing (BigComp) 2022 (pp. 1-4). IEEE.

[30] Mahmud MS, Huang JZ, Ruby R, Ngueilbaye A, Wu K. Approximate clustering ensemble method for big data. IEEE Transactions on Big Data. 2023; 9(4): 1142-55.

[31] Wei C. Research on efficient parallelization of spectral clustering algorithm based on big data. In 2nd international conference on electrical engineering, big data and algorithms (EEBDA) 2023 (pp. 1912-6). IEEE.

[32] Wang C. Fault analysis and research on elevator clustering based on big data. In2023 4th international conference on big data, artificial intelligence and internet of things engineering (ICBAIE) 2023 (pp. 51-5). IEEE.

Chandan Kumar Soni is pursuing M.Tech in Computer Science & Engineering at Madhyanchal Professional University, Bhopal(MP), India, and hols a B.Tech degree in Computer Science & Engineering from Bhagalpur College of Engineering, Bhagalpur, Bihar, India. His primary area of interest is Network Security.
Email: chandansonics06@gmail.com

Mohan Kumar Patel is working as Assistant professor with the department of Computer Science and Engineering at Madhyanchal Proffessional University , Bhopal , India. He has completed his Bachelor of Engineering and Master of Technology in Computer Science Engineering from RGPV Technical University Bhopal (M.P). He has 1 publication and conferences. His reserch area in network and network security, cryptography etc.
Email: patel.mohan67@gmail.com