
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2113–2121, DOI 10.12694/scpe.v25i4.2091

# RESEARCH AND APPLICATION OF BIG DATA CLUSTERING ALGORITHM BASED ON AI TECHNOLOGY IN CLOUD ENVIRONMENT

DAN HUANG *AND DAWEI ZHANG †

**Abstract.** Traditional big data filling algorithms often give inaccurate results due to vulnerabilities to different data types. To solve this problem, this study presents a new big data clustering algorithm powered by AI technology in a cloud environment. The study proposes an advanced Big Data clustering algorithm that leverages AI technology in a cloud environment. It optimizes clustering based on predicted strength using parallel processing. The research focuses on optimizing the clustering algorithm based on the predicted intensity through parallel processing. Experimental results demonstrate that image clustering stability is achieved when the number of clusters exceeds 4, indicating reduced sensitivity to random factors. Although it was not possible to precisely determine the optimal number of clusters, the use of an optimization algorithm showed that at four clusters the prediction intensity reached its peak, ensuring more accurate cluster identification. Through rigorous testing, the optimal number of clusters was determined to be 4. Clustering results show that visitors characterized by certain attributes show higher interest in most columns. This algorithm makes it easier to cluster incomplete large data, improves clustering speed, and improves the accuracy of filling in missing data. Compared to existing methods, this algorithm leverages AI technology in the cloud environment to optimize clustering based on prediction intensity, providing improved accuracy and efficiency during processing in big data management.

**Key words:** Distributed Computing; Prediction Strength; Algorithm Research; Cloud Environment; Clustering Algorithm.

**1. Introduction.** In today's data-driven landscape, the growing amount of information has brought data to the forefront as a key strategic asset alongside natural resources and human capital. This paradigm shift, driven by the rapid development and widespread adoption of computer technology, internet connectivity, mobile devices, and cloud computing, has ushered in an era of data ubiquity across many different industries. Moreover, the challenge is not only to manage its size, but also its complexity, which is characterized by different data types, pervasive noise, and high-dimensional structures. Effectively harnessing the hidden potential of this vast amount of data requires advanced analytical techniques. Clustering algorithms, in particular, are essential tools for identifying meaningful patterns and deriving actionable insights from large datasets. However, given the complexity of big data, traditional clustering techniques often fail and yield inaccurate results, especially in situations where data integrity is compromised. To address these challenges, this study attempts to present an innovative solution in the form of an improved big data clustering algorithm based on artificial intelligence (AI) technology in electronic environments. By leveraging the power of cloud infrastructure and AI, the algorithm aims to overcome the limitations of traditional methods by optimizing the clustering process, improving prediction accuracy, and increasing computational speed. Through a comprehensive study on the use of advanced clustering techniques and parallel processing techniques, this research aims to not only improve the accuracy of clustering results but also streamline the processing of incomplete and noisy datasets. This section highlights the importance of addressing the challenges posed by big data clustering, highlighting the integration of AI and cloud computing as essential elements for achieving advanced clustering capabilities, and the proposed It paves the way to highlighting the rationale of the approach.

Data, as a quantitative representation of information, constitutes a strategic asset like natural and human resources, representing important economic and scientific value. The advent of computing and information technology, especially the widespread adoption of internet technology, digital platforms, mobile devices and cloud computing, has fueled explosive growth in Generate data in many different fields. This increase in data production manifests itself in many different types, characterized by extensive noise and high complexity [1].

*Public Basic Education Department,Beihai University of art and design Beihai Guangxi, 536000 China (danhuang132@yahoo.com).
†Telecom Department Beihai Vocational College, Beihai, Guangxi, 536000 China

For example, Baidu processes 10 to 100 PB of website data daily, while Taobao accumulates transaction data volumes of up to 100 PB. Meanwhile, Sina Weibo generates 80 million messages per day, and a provincial branch of China Mobile registers phone communications at prices ranging from 0.5 PB to 1 PB per month. In addition, a provincial police office accumulated 20 billion pieces of road vehicle monitoring data over three years, a total of 120 TB. Forecasts from IDC, a leading computer information analysis and consulting company, predict that the global annual volume of data will reach 35 ZB by 2020 [1]. The term "big data" has emerged to summarize the nature of such large, unstructured, digitized data sets. In 2008, recognizing the emerging challenges and opportunities inherent in handling big data, the journal Nature devoted a special issue to technical obstacles and officially introduced the concept of "Big Data" [2,3].Therefore, these large unstructured data sets, characterized by their digital and high-dimensional nature, are now often referred to as big data.

**1.1. Article Contribution.** This research contributes to the field of big data management by presenting a new big data clustering algorithm supported by AI technology in a cloud computing environment. Leveraging advanced clustering techniques and parallel processing, the algorithm optimizes clustering based on predicted strength. Experimental results demonstrate that clustering stability is achieved when the number of clusters exceeds four, indicating reduced sensitivity to random factors. Using an optimization algorithm, the study identifies four clusters as optimal, thereby improving the accuracy of cluster identification. The algorithm significantly improves clustering speed, missing data filling accuracy, and allows clustering of large incomplete data sets. Compared to existing methods, this algorithm delivers higher accuracy and efficiency by leveraging AI technology in the cloud environment. The results highlight the practical benefits of optimized clustering algorithms, showing a significant reduction in the influence of random factors on clustering results. Additionally, the study highlights distinct visitor behavior patterns, which have implications for improving user engagement. Overall, the proposed optimization algorithm shows significant application value across industries, promising to reduce the time complexity and economic costs associated with data clustering analysis. Further research efforts are needed to refine and apply advanced clustering algorithms to effectively meet changing business needs and user preferences.

**2. Literature Review.** In order to optimize network resources and improve user experience, Paknejad, P proposed a large-scale network traffic monitoring and analysis system based on Hadoop, which is an open source distributed computing platform established to process large amounts of data on hard disks. The system has been deployed to run on the core network of large cellular networks, and the system works well and has been widely praised [4]. Banerjee, A In order to solve the problem that the processing efficiency of frequent subgraph mining decreases when the amount of data increases, a frequent subgraph mining algorithm FSM-Ho FSM-H based on the MapReduce framework is proposed, which is suitable for all the latest FSM algorithms [5]. Through experiments, we verify that the parallel frequent subgraph FSM-H is effective with a large comprehensive dataset. Qin, X conducted in-depth research on the development status of MapReduce at home and abroad, pointed out the advantages and disadvantages of domestic and foreign MapReduce research results, and at the same time, deeply analyzed the development status and trends of key MapReduce technologies [6]. Finally, he expressed his views on the future development direction of the MapReduce distributed framework. Li, C decomposes noisy data into clean data, Gaussian noise and sparse error matrix, and then performs low-rank subspace clustering, which can improve the robustness of the model [7]. Wen, L. H combined sparsity and low rank to give a multi-subspace representation model; In the LRR model, the column representation coefficient matrix and the row representation coefficient matrix are simultaneously low-rank constraints, so that the row and column information can complement each other and de-noise each other, and a hidden low-rank representation model is proposed [8].

The applications of big data clustering algorithms are mainly concentrated in the fields of graph processing, pattern matching, and market analysis. There are various difficulties in cluster analysis research of big data, and these difficulties are determined by the characteristics of big data itself [9]. After entering the era of big data, the amount of data that needs to be processed has increased dramatically, and the traditional clustering method using serial data analysis has been difficult to adapt to the data processing requirements in the current cloud computing network environment, the author adopts a parallel method to study and optimize the big data clustering algorithm with the prediction strength as the starting point. Starting from one or several attributes of a specified data set, this process of classifying it is called clustering, and the process of clustering does not

require a full knowledge of all the properties of the dataset. Generally, each of the divided categories is called a cluster, and the similarity of the data on one or several specific attributes, as a standard for the division between different clusters. Therefore, in the process of clustering, it is not necessary to set classification criteria in advance, but the classification is performed automatically based on the specific attributes of the data itself.

**2.1. Research gaps in existing literature.** Despite significant progress in big data clustering algorithms and their applications in various fields such as graph processing, pattern matching, and market analytics, several research gaps still exist, requiring deeper inquiry and innovation.

1. Scale network traffic analysis: Although Paknejad's Hadoop-based system proves its effectiveness in monitoring large-scale network traffic, there is still a need to optimize scalability, especially in The mobile network landscape is expanding rapidly. Further research could focus on improving the system's ability to efficiently process and analyze network traffic data while accommodating growing data volumes.
2. Improving the efficiency of frequent subgraph mining: Banerjee's FSM-Ho algorithm offers a promising solution to improve the efficiency of frequent subgraph mining using Using the Map-reduce framework. However, there is a need to explore additional techniques to further improve the scalability and performance of the algorithm, especially with regard to handling larger data sets and complex graph structures.
3. Meeting the challenges of big data clustering research: Qin's comprehensive analysis of Map-reduce technology highlights both its advantages and limitations in meeting the challenges of data processing big material. Future research efforts could delve deeper into addressing specific barriers encountered in big data clustering, such as handling different data types, scalability issues Scaling and optimizing algorithm performance in distributed computing environments.
4. Improving robustness and handling noise in data clustering: Li and Wen's study of low-level subspace clustering models and multi-subspace representations provides valuable insights value in improving the robustness of clustering algorithms. However, further research is needed to develop techniques that can effectively handle noisy data and improve the accuracy and reliability of clustering results, especially in situations where large data are available data is heterogeneous and has many dimensions.
5. Automating cluster analysis in cloud computing environments: Despite advances in parallel big data clustering methods, there are still gaps in automating the cluster analysis process in networks cloud computing.

Future research could focus on developing intelligent clustering algorithms that are able to dynamically adapt to changing data patterns and network conditions, thereby optimizing resource utilization and improving performance. Improve user experience in cloud-based applications. Addressing these research gaps will help advance modern big data clustering algorithms, allowing more efficient analysis of large-scale datasets in various application domains.

## 3. Research Methods.

**3.1. Data fusion in cloud environment.** With the rapid development of network technology, especially the Internet, the amount of unstructured or semi-structured data is increasing day by day. Among them, the IDC survey report shows that: The amount of unstructured data in the enterprise accounts for 80 percent and is growing exponentially by 60 percent every year [10]. If the structured data in the enterprise records the development and transaction activities of the enterprise meticulously and intuitively, then the unstructured data is the key lifeline of the enterprise development and the method to improve the competitiveness of the enterprise. Therefore, in order to make the enterprise develop rapidly and steadily and improve the core competitiveness of the enterprise, the research on unstructured data is imminent.

Therefore, in order to make full use of unstructured data, grasp the lifeline of enterprise development, and improve the core competitiveness of enterprises, enterprises must process unstructured data to realize data fusion and provide high-quality data guidance for future enterprise decision-making. However, due to the huge amount of unstructured data, it is difficult to quantitatively analyze it, its manifestations are diverse, which makes the integration results inefficient and inaccurate, which has a great impact on the results of future data analysis and may bring unpredictable losses to enterprises [11]. Therefore, how to carry out data fusion

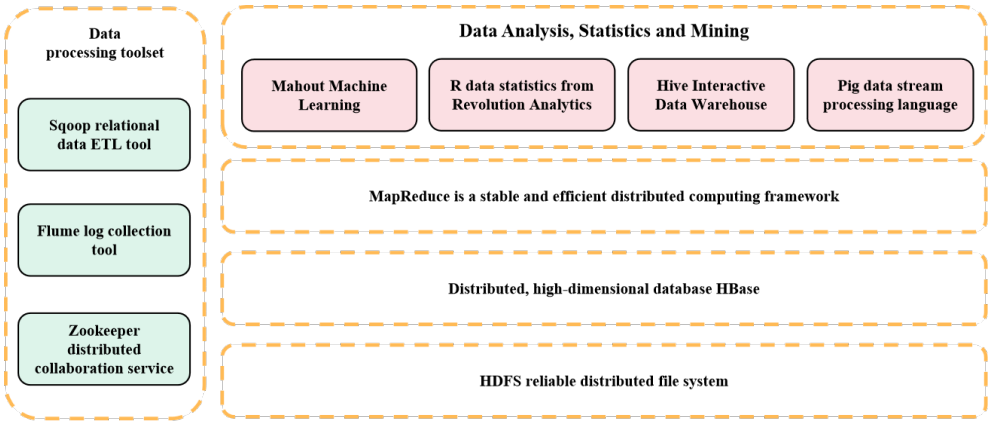


Fig. 3.1: Hadoop ecosystem structure diagram

efficiently and quickly, and provide reliable data guarantee for later data analysis and data mining, is a very challenging task.

The concept of data fusion originated from sensor data fusion, which is a series of processing of data from multiple sensors, in order to obtain unknown, useful information. Sensor data fusion has existed for a long time and was proposed around the 1970s, at the same time, it is widely used in military, biomedical and transportation [12]. The concept of data fusion also exists in the field of information retrieval. Data fusion in information retrieval mainly refers to merging the retrieval results of each independent data set into a unified result, so that the combined effect is as close as possible to the retrieval effect on a centralized data set.

**3.1.1. Data Analysis System.** There are two main directions of the current big data analysis and processing system: Batch processing systems represented by Hadoop; Stream Processing systems developed for specific applications. The main difference is that batch processing systems need to store and then process, while stream processing systems process directly [13]. A single data analysis and processing system is difficult to adapt to the rapid increase in data volume in the current cloud computing network environment, a hybrid data analysis system that mixes application architecture with underlying design languages and high-level computing modes for big data processing is more suitable for current application needs.

The Hadoop ecosystem structure is shown in Figure 3.1.

The core of Hadoop is Distributed File System (HDFS) and MapReduce. HDFS has high fault tolerance and high throughput data access, suitable for applications deployed on cheap machines and large datasets [14]. MapReduce is a mature programming model for parallel computing of large data sets. HBase is a column-oriented data database, it runs on HDFS. The main goal of HBase is to quickly locate and access the required data for billions of rows of data that exist on the host. HBase can also be used in combination with Hive and Pig, with their high-level language support, HBase can easily and quickly perform statistics on data.

The core of the Hadoop framework is HDFS and MapReduce. Here is mainly to explain the composition of HDFS, as shown in Figure 3.2.

Next, we describe the relationship between NameNode, DataNode and Client from three operations: file writing, file reading, and file block copying [15].

(1) File writing: First, the Client makes a request to write a file to the NameNode, after the NameNode receives the request, according to the size and configuration of the file, it feeds back the DataNode information under its jurisdiction to the Client, after receiving the DataNode address information returned by the NameNode, the Client divides the file into blocks and writes them to the DataNode in sequence [16].

(2) File reading: First, the Client sends a request to the NameNode to read the file, the NameNode responds after receiving the request, and feeds back the DataNode information of the stored file to the Client, the Client receives the DataNode information sent by the NameNode and performs a read operation on the file.

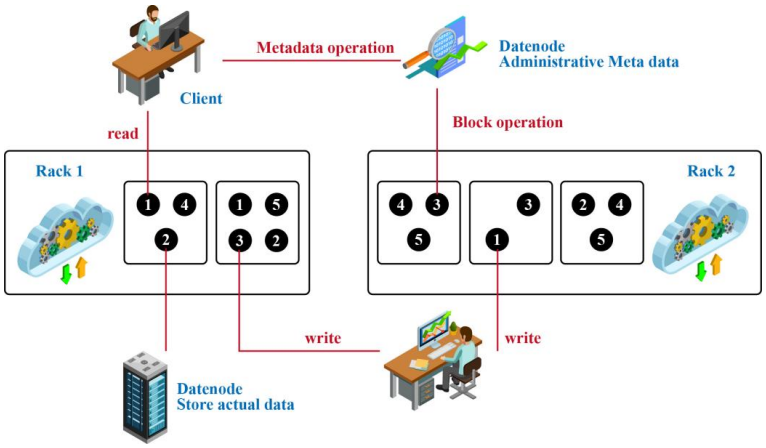


Fig. 3.2: HDFS structure diagram

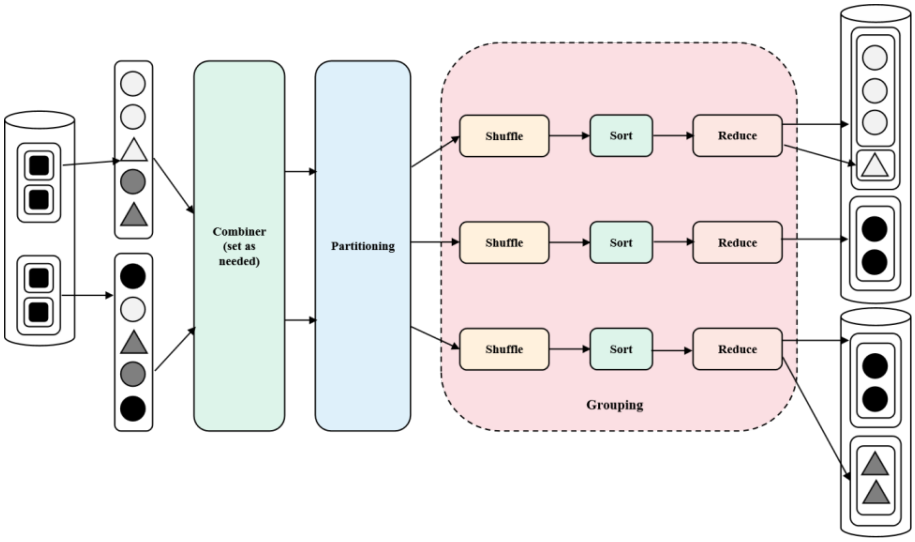


Fig. 3.3: MapReduce flowchart

(3) File block copy: When the NameNode finds that the number of file blocks has not reached the minimum number of replications, and detects that some DataNodes have failed, the NameNode will issue commands to the DataNodes under its jurisdiction to copy file blocks with each other, and these DataNodes will perform block copy operations with each other after receiving the NameNode instructions.

In addition to the map and reduce functions, a mature MR model also contains three functions: Part, comp, group, the following author will introduce the relationship between them [17]. First, the part function divides the data output by the map and distributes it to the available reduce tasks; Then all the keys will be sorted under the comparison function comp; Finally, the data is grouped by group for the convenience of reduce calls. It is important to note here that functional operations are performed on keys in key-value pairs without involving values, and the keys here are any kind of data that are comparable [18]. Proper selection of part, comp, and group functions can realize the division and grouping of complex tasks, which is particularly important when using mixed keys. The following is a detailed description through the flow chart of MapReduce, as shown in Figure 3.3.

As can be seen from Figure 3.3, MapReduce can be generally divided into two processes: Map and reduce. In Figure 3.3, it can be seen that the MapReduce instance in the figure contains 2 map tasks and 3 reduce tasks. The map function is called once for each block of input data (the four light gray squares in the Figure). When the map stage is completed, it generates 10 sets of intermediate key-value pairs and outputs them for further processing. Each set of intermediate keys corresponds to a shape (triangle and circle) and a color (light gray, dark gray, and black). These key-value pairs are assigned to 3 reduce tasks through the Partition function based on a part of the key (like color) (and of course also by shape). You can also formulate a Combiner class for the map function as needed to combine the output composite key-value pairs. Finally, the group function groups the entire key-value pair, so that the reduce function completes the classification of the five key-value pairs [19].

The actual execution of MR programs (that is, what we call jobs) is implemented through an MR architecture, such as Hadoop. An MR cluster consists of a series of nodes running on a fixed number of map and reduce processes. It's worth noting that the partition function is partially dependent on the number of reduce tasks, because it sends key-value pairs to the available reduce tasks. If the MapReuduce instance program in Figure 3.3 is running in a Hadoop cluster, and it contains 1 map process and 2 reduce processes, that is to say, 1 map task and 2 reduce tasks can run at the same time; Therefore, this 1 map task will be processed in this 1 map process, and these 3 reduce tasks will need these 2 reduce tasks to be processed.

**3.2. Clustering Algorithm for Prediction Strength Optimization.** The number of clusters expected to be divided into clusters is an important parameter in the clustering process, and the authors used the prediction strength-based method proposed by Tibshirani in 2001 to calculate the number of clusters. The prediction strength is defined as formula (1):

$$pk(s) = \min_{1\le j\le k} 1/(n_{kj}(n_{kj}-1)) \sum_{i\neq i'@i,i^*\in A_{kj}} I(D[C(A,k),B]_{ii'}=1) \tag{3.1}$$

The specific calculation steps are:

1. Divide the current data set into test set A and test set B by random division;
2. Using k as the current number of clusters, cluster the two subsets and record the results;
3. Distinguish the clustering results of the two subsets;
4. Count the classification errors of all samples in set A in set B, and calculate the correct rate of allocation;
5. The prediction strength with k as the number of clusters is the minimum value among all the correct rates.

In the prediction strength definition, $C(A,k)$ indicates that the set A is clustered into k categories, $A_{kj}$ indicates the $j$th category that the set B is clustered into, $n_{kj}$ indicates the number of elements in $A_{kj}$, $D[C(A,k),B]_{ii'}$ is the element value of the i row and i' column in the discriminant matrix of the clustering result. It can be seen that the predicted strength value $pk(s) \in [0,1]$ is affected by the number of clusters. The larger the prediction strength value, the stronger the prediction ability of the current clustering algorithm to classify new data elements into correct clusters. The random division of the data between the prediction set and the test set will cause the prediction strength to be seriously disturbed by accidental factors [20]. The author proposes to divide the data into multiple random classes first, and use them as the test set to calculate the prediction strength, and take the average of multiple prediction strengths as the final prediction strength under the current number of clusters, so as to reduce the interference of accidental factors on the prediction strength, function as an optimization algorithm.

The number of clusters k is determined by the prediction strength, and the corresponding algorithm for clustering is as follows:

Input: dataset D= $d_1, d_2, \ldots, d_i, \ldots, d_n$, the current optimal number of clusters k

1. Select k data points $d'_1, \ldots d'_k$belonging to D from the D set as the centroids of the clustered clusters;
2. For $\forall d_i \in D$, its corresponding cone should be $N_j$=arg $\min_j \|d_i - d'_j\|$;
3. Modify the centroid position $d'_i = (\sum_{i=1}^{n} \text{sign}(N_j = j)d_i)/(\sum_{i=1}^{n} \text{sign}(N_j = j))$ of each cluster
4. Take the sum of squares of errors between the centroid of each cluster and the data points in the cluster as the criterion function $G(N, d') = \sum_{i=1}^{n} \|d_i - d'_j\|$;

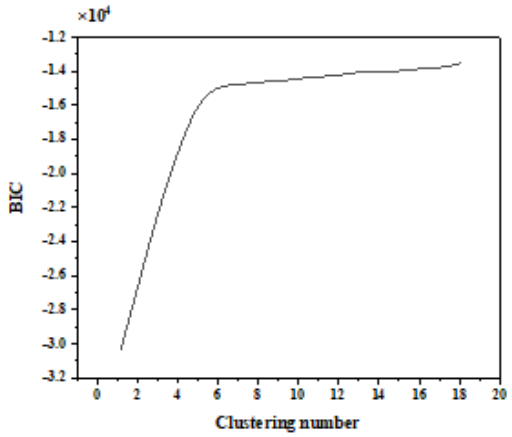


Fig. 4.1: BIC criterion model function curve

5. Repeat steps (2) and (3) until the value of (4) no longer changes, obviously, the value of the criterion function is shrinking.

Output: The cluster $N_1, \ldots, N_k$ where the centroid $d'_1, \ldots, d'_k$ is located

Among them, $N\_j$ should be the cluster where the centroid $d'_j$ closest to the data $d_i$ is located, sign$(N_j = j)$ indicates that its value is 1 when $N_j = j$, and its value is 0 in all other cases [21].

**4. Analysis of results.**

**4.1. System Composition.** To understand visitor behavior on live streaming platforms, visitor engagement duration data was analyzed in different columns. The analysis involved importing data from three columns and using a model built using the Bayesian Information Criterion (BIC). This criterion serves as the main measure for determining the optimal number of clusters and sets the number of anchor points for the analysis to her three variables.

**4.1.1. Model Optimization and Prediction Strength.** Figures 4.1 and 4.2 show function curves showing the relationship between the number of variables and the number of clusters. Note that the BIC criterion alone does not have a significant impact on determining the number of clusters when the number of variables is the same. As shown in Figures 4.1 and 4.2, when the number of clusters exceeds 4, the cluster images tend to become stable, indicating reduced sensitivity to random factors. Although it is clear that the optimal number of clusters should exceed 4, it is still difficult to determine the exact value [22].

**4.1.2. Optimized algorithm and improved cluster identification.** We find that using the optimized algorithm, the prediction strength reaches its maximum value at four clusters. This finding indicates that the number of clusters can be determined more accurately using an optimization algorithm. Based on the test strength, it was concluded that the optimal number of clusters is actually 4. The resulting curve after cluster analysis is shown in Figure 4.3.

**4.1.3. Visitor Attribute Analysis.** Analysis Figure 4.3 shows the different behavior patterns of visitors based on the categories assigned to them. Visitors characterized by the first attribute type have increased interest in most columns, which is very much in line with the content focus of the platform. This idea suggests opportunities for customized services to improve user engagement. In contrast, the fourth category of users shows little interest in the platform's current content, and there is no potential for targeted or strategic content adjustments to re-target that segment of users.

This result highlights the importance of using advanced clustering algorithms, such as the BIC criterion and algorithms using predictive strength optimization, to derive valuable and actionable insights from big

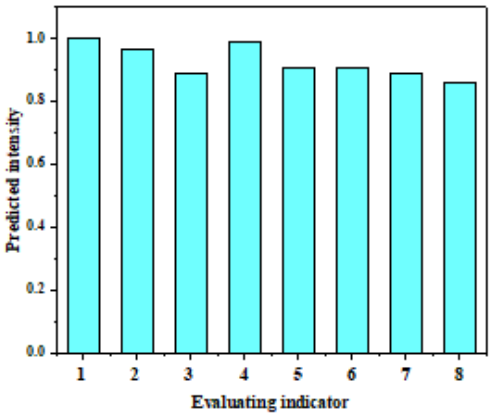


Fig. 4.2: The function curve of the optimized prediction strength model

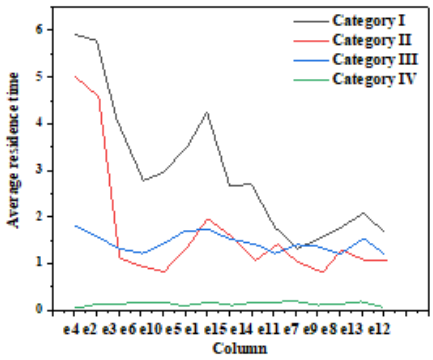


Fig. 4.3: The average stay time of different types of visitors on all columns

data. Understanding visitor behavior patterns allows platform operators to refine content strategies, improve user experience, and optimize resource allocation. Future research efforts may focus on improving clustering methods and exploring real-time adaptation strategies to respond to changes in user preferences.

**5. Conclusion.** The results of the performed experiments highlight the tangible benefits of using optimized clustering algorithms in real applications. By improving traditional clustering methods and optimizing the determination of specific cluster numbers, our study shows a significant reduction in the influence of random factors on clustering results. Our results reveal that when the number of clusters exceeds four, a stable trend emerges, suggesting a reduced susceptibility to random influences. Although the exact value of the optimal number of clusters remains elusive, our optimization algorithm facilitates a more precise determination, with maximum prediction strength observed for four clusters. Our analysis highlights distinct visitor behavior patterns, with a particular focus on increasing visitor interest with attributes that align with the platform's content focus. The use of this optimization algorithm promises significant practical benefits, including reduced time complexity and economic costs associated with large-scale data clustering analysis. By streamlining clustering processes and improving accuracy, our approach has significant application value in various industries. Further refinement and adoption of advanced clustering algorithms are warranted to continuously improve the efficiency and effectiveness of data analytics to meet growing business needs and user preferences.

REFERENCES


[1] Krammer, N. , *New challenges for distributed computing at the cms experiment.* , Journal of Instrumentation, 15(7), C07038-C07038, 2020.

[2] Seifollahzadeh, P., Alizadeh, M., Abbasi, M. R. , *Strength prediction of multi-layered copper-based composites fabricated by accumulative roll bonding - sciencedirect.* , Transactions of Nonferrous Metals Society of China, 31( 6), 1729-1739, 2021.

[3] [3] Rani, D. R., Geethakumari, G., *A framework for the identification of suspicious packets to detect anti-forensic attacks in the cloud environment*, Peer-to-Peer Networking and Applications, 14(3), 1-14, 2021.

[4] Paknejad, P., Khorsand, R., Ramezanpour, M. , *Chaotic improved picea-g-based multi-objective optimization for workflow scheduling in cloud environment*, Future Generation Computer Systems, 117(10), 12-28, 2021.

[5] Banerjee, A., De, S. K., Majumder, K., Dash, D., Chattopadhyay, S. , *[5] Banerjee, A., De, S. K., Majumder, K., Dash, D., Chattopadhyay, S. .* , The Journal of Supercomputing, 78(8), 11015-11050, 2022.

[6] Qin, X., Li, J., Hu, W., Yang, J. , *Machine learning k-means clustering algorithm for interpolative separable density fitting to accelerate hybrid functional calculations with numerical atomic orbitals.*, The Journal of Physical Chemistry A, 124(48), 10066-10074, 2020.

[7] Li, C., Bai, J., Zhao, W., Yang, X. , *Community detection using hierarchical clustering based on edge-weighted similarity in cloud environment*, Information Processing Management, 56(1), 91-109, 2019.

[8] Wen, L. H., Shi, Z. H., Liu, H. Y. , *Research on risk assessment of natural disaster based on cloud fuzzy clustering algorithm in taihang mountain*, Journal of Intelligent and Fuzzy Systems, 37(4), 1-9, 2019.

[9] Yang, Y., Yu, J., Fu, Y., Hu, J. *Research on geological hazard risk assessment based on the cloud fuzzy clustering algorithm* Journal of Intelligent and Fuzzy Systems, 37(2017), 1-8, 2019.

[10] Shi, T., Ma, H., Chen, G., Hartmann, S. *Location-aware and budget-constrained service deployment for composite applications in multi-cloud environment.* IEEE Transactions on Parallel and Distributed Systems, 31(8), 1954-1969, 2020.

[11] Ci, X., Wen, K., Sun, Y., Sun, W., Deng, W. *An energy efficient clustering algorithm in wireless sensor networks for internet of things applications* Journal of Physics: Conference Series, 1881(4), 042035, 2021.

[12] Dong, M., Fan, L., Jing, C. *Ecos: an efficient task-clustering based cost-effective aware scheduling algorithm for scientific workflows execution on heterogeneous cloud systems.* The Journal of Systems and Software, 158(Dec.), 110405.1-110405.1-11, 2019.

[13] Huang, C. *Data-parallel clustering algorithm based on mutual information mining of joint condition* IOP Conference Series: Materials Science and Engineering, 914(1), 012030 (8pp), 2020.

[14] Chaudhari, A. Y., Mulay, P. , *Cloud4nfica-nearness factor-based incremental clustering algorithm using microsoft azure for the analysis of intelligent meter data*, International Journal of Information Retrieval Research, 10(2), 21-39, 2020.

[15] Swagatika, S., Rath, A. K. , *Sla-aware task allocation with resource optimisation on cloud environment.*, International Journal of Communication Networks and Distributed Systems, 22(2), 150, 2019.

[16] Charles, P., Alagumalai, V. , *Load balancing in cloud computing using agglomerative hierarchical clustering approach*,Journal of Advanced Research in Dynamical and Control Systems, 4(11), 1720-1723, 2019.

[17] Wu, C., Yu, R., Yan, B., Huang, Z., Zhou, X. , *Data design and analysis based on cloud computing and improved k-means algorithm.*, Journal of Intelligent and Fuzzy Systems, 39(1), 1-8, 2020.

[18] Sharma, A., Rai, A. , *RAn approach of leading sequence clustering (lsc) algorithm based scheduling and agglomerative mean shift clustering for load balancing in cloud.*, Journal of Advanced Research in Dynamical and Control Systems, 11(10-SPECIAL ISSUE), 618-624, 2019.

[19] Sun, X., Ma, H., Sun, Y., Liu, M. , *A novel point cloud compression algorithm based on clustering*, IEEE Robotics and Automation Letters, 4(2), 2132-2139.

[20] Yu, Z. , *Big data clustering analysis algorithm for internet of things based on k-means*, International Journal of Distributed Systems and Technologies, 10(1), 1-12, 2019.

[21] Sharma, A., Kumar, R. , *Performance comparison and detailed study of AODV, DSDV, DSR, TORA and OLSR routing protocols in ad hoc networks*, 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2016.

[22] Sharma, K., Chaurasia, B. K. , *Trust Based Location Finding Mechanism in VANET Using DST*, Fifth International Conference on Communication Systems Network Technologies (pp.763-766), 2015.

[23] Ren, X., Li, C., Ma, X., Chen, F., Wang, H., Sharma, A. , *Design of multi-information fusion based intelligent electrical fire detection system for green buildings*, Sustainability, 13(6), 3405, 2021.


*Edited by:* Pradeep Kumar Singh
*Special issue on:* Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
*Received:* Jan 3, 2023
*Accepted:* Mar 21, 2024