Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 6s (2024)

Revolutionary Data Deduplication with Fuzzy C-Means: Advancing Data Quality Management

[1*] Dr. P. Selvi

[1*] Assistant Professor, Department of Computer Science, KG College of Arts and Science Coimbatore, India Coimbatore-641 105. [1*]Corresponding author email id: selviragu19@gmail.com

Article History:

***Received:** 15-06-2024*

***Revised:** 26-07-2024*

***Accepted:** 06-08-2024*

Abstract:

Maintaining the integrity and precision of data depends on the crucial process of data deduplication, the search and elimination of duplicate data from a database. Conventional deduplication methods may not be useful when dealing with data with variances and uncertainty as from time to time depend on spotting the closest matches. In this point, another procedure frequently applied in data clustering but especially for data deduplication is employed in this work to propose a new approach to data deduplication involving Fuzzy C-use (FCM) clustering. FCM allows to set as many data points as desired to peculiar clusters in components, in addition to the account for variation and error. Therefore, it is possible to improve the further usage of fuzzy C-Means clustering in the data deduplication field. It provides significant practical support in the areas of statistical preprocessing and information quality management among the disciplines. It is now possible to assert that in the context of global development and growing importance of big data, the improvements in this sphere are expanding the potential for creating more efficient and powerful analytic tools. The proposed research focuses on establishing whether the FCM-based deduplication strategy will enhance the false positive and false negative ratios. Due to this, it is a reliable solution in cases where data duplication can be expected to occur.

Keywords: Fuzzy C-Means, Deduplication, Decision Making, Accuracy, Integrity,

1. Introduction

Data deduplication is one of the processes which can assist in making selections, predictions and obtaining improved decision via means of Data deduplication is the identification and organization of redundant data in a dataset. The elaborate strategy called data-driven decision-making or DD-DM, is based on the systematic analysis of integrated data to support and guide the business decisions. This process ensures that decisions made are in alignment with the organizational values, strategies as well as meeting organizational commitment with data supported by the help of market research. The used data has to be accurate and as reliable as possible, which is very crucial for this tool. To enhance the quality of the datasets that have been collected, business organizations may ensure the every document is accurate. Lack of reliable or fake information may cause wrong opinions and choices. Duplicates are removed by statistical dedupilation that are not based on unreliable information. The kind of data contained in repeat cycles can alter ways in which trends and patterns are addressed and the outcomes of analysis. DD-DM is generally an information extraction technique which mostly relies on the application of statistical analysis techniques to arrive at the most appropriate information. Firms may ensure that the data used for analysis is actually refined and free from multiple entries by using data deduplication, thus getting more specific results. One of the critical stages carried out in the DD-DM process is data deduplication. Companies can rightfully assume that knowledge that they obtain from their research will remain intact and correct, which will enhance the efficiency of their records and subsequently the effectiveness of their strategic choices.

Data deduplication is the process of identifying and subsequently removing all the data that is repetitive or is a duplicate of the others in a given dataset. As it will be shown, there are pros in deduplication, While these may seem like small issues, they point to serious problems in the integrity and quality of the data. In other words, when addressing this sophisticated problem, as data deduplication is, data reliability is of paramount importance. Accordingly, in order to accomplish the advantages of deduplication but to limit data loss or mistakes in their organization, companies have to scrupulously design and devise their deduplication procedures [15].

2. Literature review:

A method for cloud storage deduplication is described in Reference [8] to help execute duplicate checks on critical servers. Generating a fixed size ciphertext that is independent of the number of key servers, the work establishes better performance in plain theory and testing. Furthermore, the study shows an increased prevalence of the proposed technique and disputes some issues concerning document confidentiality in techniques of deduplication.

The problem of duplicating records in cloud-fog storage systems is talked address in [9] and MECC and Convergent Encryption is proffer as a probable solution to fixing this problem. The recommended method is just to avoid a huge amount of duplicated information, preserving good encryption at the same time. The constant outcomes also reveal how the novel approaches' computational efficiency and security surpass the one offered by the previous methods.

A hybrid and dependable deduplication technique for area computing is presented in Reference [10], which mixes client-side and server-side deduplication algorithms. While maximizing network throughput between edge nodes and the cloud, the system ensures the privacy of data between clients and side nodes. Furthermore, a novel encryption technique has been devised that enables quick deduplication operations on nodes with limited resources. This is an additively homomorphic encryption technique.

Reference [11] also develops a multi-stage Stackelberg game with data holders, owners, and CSPs. Hybrid Encrypted Cloud Data Deduplication (H-DEDU) is financially represented. The suggested model examines the Nash Equilibrium requirements and offers stakeholders a thorough set of gradient-based guidelines for selecting approaches that are almost ideal. This indicates that implementing H-DEDU is financially feasible.

The MTHDedup deduplication algorithm, which makes use of Merkle hash trees to mitigate convergent encryption problems, is introduced in Reference [12]. By using Merkle hash trees and other encryption methods during file and block-level deduplication, the method improves data security. In addition to reducing computational overhead, the method protects the important parking area against brute-force attacks from the outside as well as from inside.

A hybrid statistical deduplication technique that incorporates the chromatic correlation clustering algorithm and Euclidean distance is presented in Reference [13]. The goal of this method is to get over the limitations of both crowdsourced and machine-based techniques. The method seeks to improve accuracy, reduce deduplication time, and lower crowdsourcing expenses. According to empirical comparisons, a lower crowdsourcing value results in a higher deduplication accuracy performance.

A variety of methods for safely managing storage and getting rid of redundant data in cloud and edge computing environments are examined in these papers. The review examines a server-helped deduplication strategy for distributed storage with the objectives of delivering ciphertext of a steady size while at the same time further developing pace. It deals with data privacy concerns as well. Secure deduplication in cloud-fog storage is the subject of [9], which offers a solution to reduce data redundancy by combining MERCC with convergent encryption. The study presents the server-side-client-side algorithm hybrid stable deduplication scheme for edge computing that the research suggests could enhance the network performance while preserving data privacy. In the study, both hybrid encrypted cloud data deduplication is provided along with an economical analysis and a multi-level sport. It looks at Nash Equilibrium and then, offers an algorithm by gradients for deciding on a good strategy. Merkle hash trees are applied in MTHDedup, a process that strengthens security throughout hybrid cloud storage. Lastly, [13] suggests a half breed deduplication technique that uses Euclidean distance and support from the public. This method is attributed to the provision of cost effectiveness, less time for deduplication and increased accuracy. Several researches also show it as more effective than the existing procedures. Altogether, these publications give thoughtful information on how information deduplication is developing together with the issues of security and performance.

Categorized prior artworks may contain problems of data loss, deduplication, or have data security problems due to such factors. has been deduplicated. There are cases where the structure of information is complex, data settings can change from time to time or strict referential integrity should be maintained, and the existing methods of deduplication become unsuitable. However, accomplishing data security and secrecy especially through the entire deduplication process might be a daunting process when working with encrypted data.

For this reason, the artworks in this proposal employ fuzzy C-Means (FCM) clustering in order to offer an alternative approach to data deduplication which can possibly address all of these issues. In this way, FCM clustering differs from the conventional methods which are based on the best fit between elements and clusters since it enables record factors to be partially ‘belonging’ to specific clusters thus is able to handle fluctuations and uncertainty inherent in records. In particular, this version can boast satisfactory results on real datasets with mistakes, if any [16]. The recommended materials will contribute to enhancement of both the resistance to disturbances and such accuracy improvement techniques as FCM clustering to identify similar information. To some extent, one gets the impression that, given the objective of improving data quality and its subsequent integrity, the concentration on minimizing false positives as well as false negatives is the motivating factor of this approach. Regarding the abstract of the paper: FCM-based deduplication is advanced compared to previous approaches as it enhances the data processing accuracy and credibility in numerous cases. As to most solutions presented, the recommended artworks offer a more versatile and flexible strategy to the conventional deduplication approaches and consequently eliminate the problems [17].

3. Clustering Fuzzy C-Means (FCM) for Data Duplication

3.1 Transforming FCM for Data Duplication:

Instead of using typical methods that rely on precise matching for facts deduplication, fuzzy C-Means (FCM) clustering is employed [18]. One clustering method that finds frequent application in data division is Fuzzy C-Means, or FCM. However, its particular use here is to eliminate duplicates from a dataset. When dealing with data that displays unpredictability and uncertainty, this version aims to address problems associated with optimum suit-based deduplication algorithms. To better deal with the complexity of real-world datasets during deduplication, FCM permits information components to partly belong to other clusters, offering a more versatile and advanced approach [19].

Fuzzy C-means clustering is the primary method used in this strategy for deduplicating data. The FCM approach is utilized strategically because to its capacity to handle ambiguity and uncertainty; it partially considers memberships while clustering data points. Thus, it allows for the recognition of duplicates and their elimination while at the same acknowledging data differences. Unlike the use of strict guidelines in traditional deduplication approach, the methodological overview present a complex and flexible method, which specifically uses FCM in dealing with certain issues in the data deduplication process [20].

From the above, data deduplication contributes significantly towards an increase in the workload in creating a firm strategy on FCM clustering.

In the first algorithm, known as Fuzzy C-Means clustering to demonstrate redundancy of data is used. Towards addressing the problem of data duplication, it aims at achieving its goal with the help of Fuzzy C-Means (FCM) clustering in order to detect and, further, remove the duplicate data from the given set containing such data. For the ruleset to work effectively, the dataset should include n facts that can be effective for the play-field utility. an exponent for the fuzzifier, a desired number of clusters, a convergence threshold for the stopping criterion, and the maximum number of iterations, which is denoted by max_i. As soon as the cluster centroids have been initialized, the rule set constructs a membership matrix U at random, where each row adds up to one. It then updates the membership matrix and repeatedly refines the cluster centroids depending on the spacing between the data points and the cluster centers. Once convergence is proven, the set of criteria is repeated until either convergence is achieved or the maximum fresh release limit is met.

The end outcome is a fuzzy partition matrix in the form of an o, which displays the degree of relationship between each data point and each cluster. Since the strategy makes versatile acclimations to the group tasks utilizing a fluffy rationale approach, obliging information changes and vulnerabilities, it is ideally suited for the deduplication setting. A fuzzy partition matrix that assists in the identification and removal of duplicate data is produced by providing a comprehensive picture of the structural properties of the dataset.

Algorithm 1: Data deduplication with fuzzy C-Means clustering:

```
Input:
 A dataset Đɘ containing n entries.
A value representing the exponent of a fuzzy function.
Cluster count.
```

The limiting condition for convergence is denoted as ṰḎ.
Upper limit on the maximum number of iterations.

Output:

Fuzzy partition matrix O.

Step 1: To begin, the membership matrix U O should be initialized at random, and it should be ensured that the sum of all the rows is equal to 1.

Cluster centroids C= {C1, C2,..., C} should be randomly initialized.

Keep doing this until you achieve convergence or max_i.

Step 2 Determine where the clusters are located:$C_j = \frac{\sum_{i=1}^{n} o_{ij}^{\dot{m}}.dp.}{\sum_{i=1}^{n} o_{ij}^{\dot{m}}}$j=1,2,3,,,……. Additionally, dp is the location-based data point.

step 3 Revise the array of members: $O_{ij} = \frac{1}{\sum_{k=1}^{ç}\left(\frac{\|dp-c_j\|}{\|dp-c_k\|}\right)^{\frac{2}{\dot{m}-1}}}$

Step 4 Verify whether there is convergence

Step 5 Produce a fuzzy partition diagram O.

Subsequently, you might group by using the Fuzzy C-Means systematically with the help of an iterative procedure. Using the membership matrix O it is possible to determine the degree of relatedness between each of the data point and each of the clusters to which they belong The clusters' as well as the centroids' memberships are iteratively altered by the approach until it is possible to either attain the required convergence or the stipulated maximal number of iterations. This is the exponent. that considers incomplete memberships and according to which the degree of the fuzziness in the clustering is specified.

Slightly less essential, but still vital for the process, is the recognition of the fact that one has to address the issues of mismatched datasets and vagueness. Applying conventional deduplication methods may prove challenging when dealing with the data with hard-to-interpreted patterns or that has low match rates. The technique avoids these issues by employing FCM clustering, which is less likely to face issues regarding the categorization of data points. The approach also considers such factors such as uncertainty that is apparent in real-world data sets as well as factual deviations. The deduplication approach used here is more reliable due to the fact of flexibility; it is capable of weed out duplicity items irrespective the status of the data situation.

4. Result & Analysis

Therefore, it is essential to employ the fuzzy C-Means clustering approach because it is robust, elastic and apt to manage the non-certain environment. It also brings soft tasks. The flexibility of the software to various industries such as image processing, research, data discovery affirms how critical it is to respond to actual issues involving complex and ambiguous information. The figure 1 denotes the FCM Clustering Result' which graphically enhances the given dataset's readability through the outlining of inherent tendencies and enabling further data analysis. It is among the components of the information exploration system that provides valuable characteristics of the data and indicates further studies.

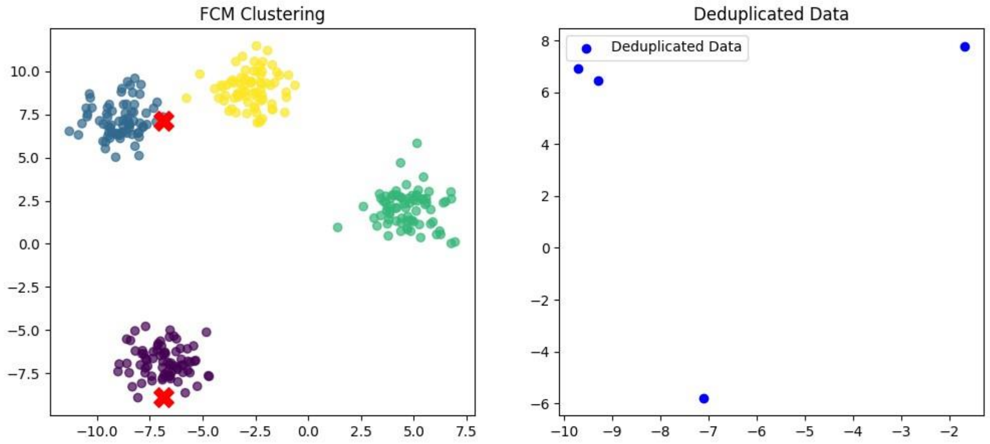

Figure 1 FCM clustering & Deduplicated data

The phases of data processing are presented in figure 2 in a temporal sequence; the "Original Data" phase stands out since the factors and values are displayed prominently. Consequently, The fuzzy C-Means clustering organizes the data points into cluster outcomes known as the "FCM Cluster Result" and uses a color code to depict the data points' membership in the clusters. It can be suggested that by view of the "Deduplication Data" picture, it may shrink the file size. regarding the data by identifying the items that are peculiar and emphasizing them during the process of deduplication, which, in its turn, improves the clusters. In biometric or when trying to distinguish, it's obvious that it becomes easy to distinguish the data and understand how the data is grouped, by these visuals highlighting the difference from the unstructured data to the grouped representation and then the improved collection of the unique records.

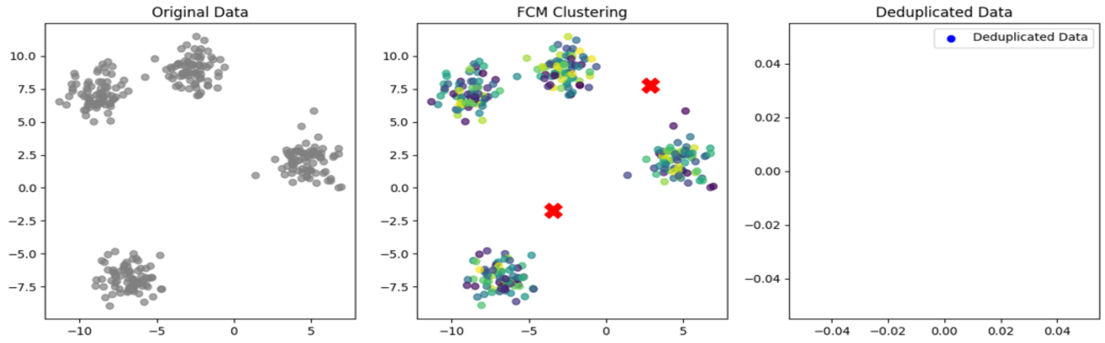

Figure 2 Data without duplication and FCM clustering

Figure 3 illustrates the three components focusing on the data processing procedure. The first data organization level seen is the "Original Data" as depicted on the screen below. The 'FCM Cluster' image which helps in identifying the patterns after applying the Fuzzy C-Means clustering sometime has clustered pattern in the different colors. The "Deduplication Data" image gives a better understanding of these clusters through differentiating the data that has been deduplicated. Together with these visuals, the histogram with the label "Distribution of Cluster Membership" provides a wider view of how the points are distributed within the obtained clusters The numeric evaluation of the deduplication efficiency and time-consuming is available under the titles “Number of Unique Records Retained” and “Execution Time” correspondingly. As a whole, such visualizations offer the entire story about the dataset including data processing, grouping, deletion of similar data, assessment of the performance, and providing the viewers with complete information about the properties of the dataset and patterns regarding its processing.

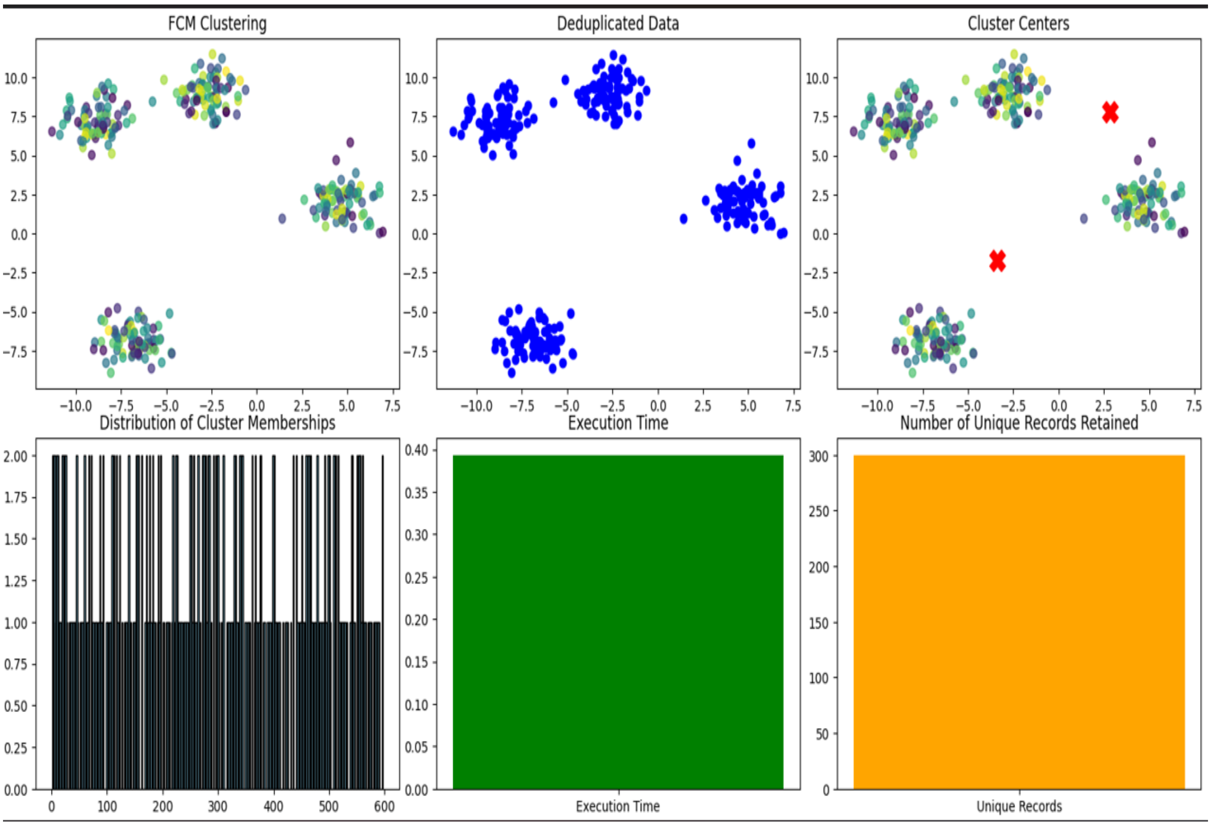

Figure 3 FCM with Cluster Membership Distribution, Execution Time, and Unique Record Retention

About the influence of the fuzzifier exponent, figure 4 depicts the execution time regarding the number of records that are kept in their original state. The value on the x-axis concerning the fuzzifier setting, which measures the degree of freedom in the allocation of the clusters. The total number of unique records and execution time of the process is represented on the y-coordinate axis. As the extent to which the fuzzifier exponent is adjusted, it is demonstrated that this has an impact on the relationship between the ratio of computational efficiency that is expressed in terms of execution time and the database distinctiveness where the proportion of records that are made unique within a specific database are considered. This figure also shows a correlation between these two factors. By observing this picture, you can decide the proper worth of the fuzzifier so that not that much amount of computation is needed for deduplication but at the same time, a huge amount of data that is entirely rid of duplication can be stored. In this study, significant insights on how the fuzzifier setting affects the algorithm response is contained in this paper.

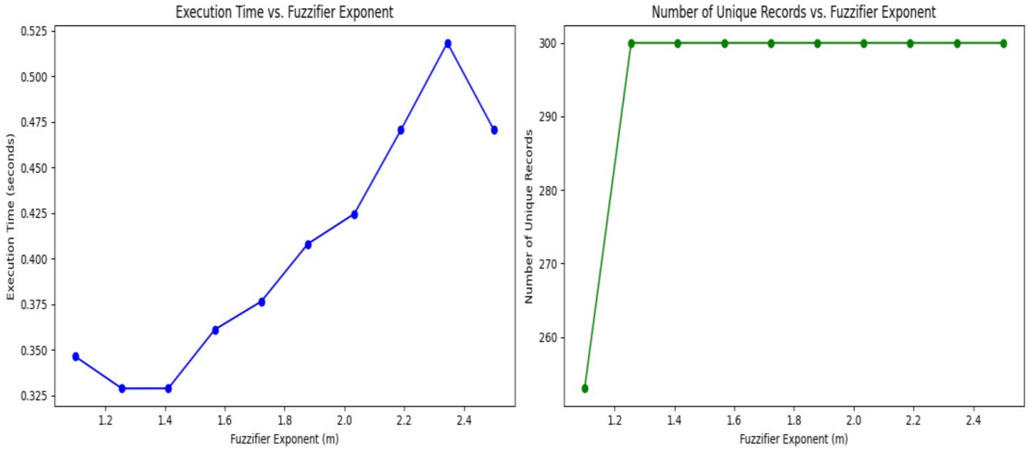

Figure 4 Factors Affecting Fuzzifier Performance vs. Runtime and Individual Records

The results provide a thorough depiction of the data processing procedure, starting with the figure titled "Original Data," going on to the figure titled "FCM Cluster," and concluding with the figure titled "Deduplication Data," which distinguishes between various entries. Histograms illustrate how data points are distributed inside clusters, and the "Distribution of Cluster Membership" histogram is an improvement over prior representations that showed how data points were dispersed. Two measures that have been suggested to be used in order to make the evaluation of the deduplication process with regards to computational efficiency and success are Time to Execution and Unique Records Retained. The final example shows how the structure of the fuzzifier exponents influences on the number of distinctive records and on the time employed, which is valuable information in order to understand how the method reacts to changes on the parameters. Significant information about the structure of the dataset, the success of the methods used for grouping, deduplication, and filtering steps, as well as the advantages and disadvantages of the selected parameters' cost, can be gathered when all proposed visuals are combined. This final and most encompassing figure parrots the need for more information search in a simple and effective manner by emphasizing that these two factors must be carefully balanced: computation on the one hand and the uniqueness of data on the other.

There is a significant amount of literature on the safe management of stored data and the process of deduplication in cloud and edge computing, which shows how the approaches to performance and security challenges are evolving. In the context of cloud-fog storage, the articles contribute with new approaches for diminishing the data redundancy, enhancing the processing time and strengthening the security of the data (Ref. [9]). As mentioned by [10], the hybrid secure deduplication system for edge computing is an all-around solution that improves communication between cloud nodes and edge nodes and improves data security at the same time. A short example of an approach that uses financial concepts to influence stakeholders' decisions is the Hybrid Encrypted Cloud Data Deduplication (H-DEDU) economic model that is discussed in [11]. In reference [12], the deduplication procedure applied by MTHDedup is illustrated; this technique successfully withstands internal and external attacks. This strategy uses Merkle hash trees. The inefficiencies of current approaches are managed by the proposed hybrid deduplication methodology, chromatic correlation clustering with Euclidean distance (Ref [13]) which is fast and inexpensive solution than earlier approaches.

Nevertheless, there are obstacles that come with these technological breakthroughs. Notable drawbacks highlighted in the aforementioned papers include data loss potential, deduplication process problems, and security concerns, particularly with encrypted data. Conventional methods of data deduplication face obstacles in the form of referential integrity and ever-changing data contexts. Despite the fact that the server-aided deduplication method (Ref [8]) takes data privacy concerns into account, problems may arise with dynamic data and complicated data structures. The data processing approach is shown in detail by the offered findings, which emphasize the requirement of maintaining a balance between computing efficiency and the preservation of unique and relevant information. Visualizations such as a histogram depicting the "Distribution of Cluster-Membership" and a look at how different fuzzifier exponents affect execution time and the number of unique entries might be helpful for making decisions. As we'll find in the last segment, the FCM-based deduplication technique offers a promising way forward for the business. In our data-centric society, it emphasizes the need for more trustworthy data processing.

5. Conclusion

Maintaining data quality should be the top priority in today’s context, as data forms the foundation for critical decision-making. It is an essential process to remove all the duplicate entries from the dataset in order to enhance correctness as well as integrity of data and for this specific process, there is a term called data deduplication. Because of their match approach, when it comes to handling data that has times and fluctuations and uncertainties, standard deduplication processes may at times not be sufficient. This study proposes a novel approach to data deduplication intgrates fuzzy C-Means (FCM) clustering. One of the most used methods is FCM clustering which originally has been adapted for the purpose of deduplication. The handling of inherent uncertainties and data changes is attained through the utilization of the C-means flexible clustering methods, which allow data to belong to more than one cluster. This study contributes a lot towards the area of data quality management and preprocessing section by proposing a new application of FCM clustering in the handler of data deduplication. Since the use of data, advancements in this sector are likely to enhance the reliability and efficiency of data handling.

References

[1] He, Q., Li, Z., & Zhang, X. (2010, October). Data deduplication techniques. In 2010 international conference on future information technology and management engineering (Vol. 1, pp. 430-433). IEEE.

[2] Mahesh, B., Pavan Kumar, K., Ramasubbareddy, S., & Swetha, E. (2020). A review of data deduplication techniques in the cloud. Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco, 825-833.

[3] Malhotra, J., & Bakal, J. (2015, January). A survey and comparative study of data deduplication techniques. In 2015 International Conference on Pervasive Computing (ICPC) (pp. 1-5). IEEE.

[4] Maddodi, S., Attigeri, G. V., & Karunakar, A. K. (2010, November). Data deduplication techniques and analysis. In 2010 3rd International Conference on Emerging Trends in Engineering and Technology (pp. 664-668). IEEE.

[5] Mandagere, N., Zhou, P., Smith, M. A., & Uttamchandani, S. (2008, December). Demystifying data deduplication. In Proceedings of the ACM/IFIP/USENIX Middleware'08 Conference Companion (pp. 12-17).

[6] Dutch, M. (2008, June). Understanding data deduplication ratios. In SNIA Data Management Forum (Vol. 7).

[7] Zhang, X., & Deng, M. (2017). An overview of data deduplication techniques. In Information Technology and Intelligent Transportation Systems: Volume 2, Proceedings of the 2015 International Conference on Information Technology and Intelligent Transportation Systems ITITS 2015, held December 12-13, 2015, Xi’an China (pp. 359-369). Springer International Publishing.

[8] Nayak, S. K., & Tripathy, S. (2020). SEDS: secure and efficient server-aided data deduplication scheme for cloud storage. International Journal of Information Security, 19(2), 229-240.

[9] PG, S., RK, N., Menon, V. G., P, V., Abbasi, M., & Khosravi, M. R. (2020). A secure data deduplication system for integrated cloud-edge networks. Journal of Cloud Computing, 9, 1-12.

[10] Shin, H., Koo, D., & Hur, J. (2022). Secure and Efficient Hybrid Data Deduplication in Edge Computing. ACM Transactions on Internet Technology (TOIT), 22(3), 1-25.

[11] Liang, X., Yan, Z., Deng, R. H., & Zheng, Q. (2020). Investigating the adoption of hybrid encrypted cloud data deduplication with game theory. IEEE Transactions on Parallel and Distributed Systems, 32(3), 587-600.

[12] Gang, F., & Wei, D. (2022). Dynamic Deduplication Algorithm for Cross-User Duplicate Data in Hybrid Cloud Storage. Security and Communication Networks, 2022.

[13] Haruna, C. R., Hou, M., Eghan, M. J., Kpiebaareh, M. Y., & Tandoh, L. (2019). An effective and cost-based framework for a qualitative hybrid data deduplication. In Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018 (pp. 511-520). Springer Singapore.

[14] Anwarbasha, H., Sasi Kumar, S., & Dhanasekaran, D. (2021). An efficient and secure protocol for checking remote data integrity in multi-cloud environment. Scientific reports, 11(1), 13755.

[15] Grealy, A., Langmore, N. E., Joseph, L., & Holleley, C. E. (2021). Genetic barcoding of museum eggshell improves data integrity of avian biological collections. Scientific Reports, 11(1), 1605.

[16] Wang, X., Chen, Y., Jin, J., & Zhang, B. (2022). Fuzzy-clustering and fuzzy network based interpretable fuzzy model for prediction. Scientific Reports, 12(1), 16279.

[17] Guo, L., Wang, P., Sun, R., Yang, C., Zhang, N., Guo, Y., & Feng, Y. (2018). A fuzzy feature fusion method for auto-segmentation of gliomas with multi-modality diffusion and perfusion magnetic resonance images in radiotherapy. Scientific reports, 8(1), 3231.

[18] Ahmadianfar, I., Shirvani-Hosseini, S., He, J., Samadi-Koucheksaraee, A., & Yaseen, Z. M. (2022). An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction. Scientific Reports, 12(1), 4934.

[19] Pasin, O., & Gonenc, S. (2023). An investigation into epidemiological situations of COVID-19 with fuzzy K-means and K-prototype clustering methods. Scientific Reports, 13(1), 6255.

[20] Mosavi, A., Golshan, M., Choubin, B., Ziegler, A. D., Sigaroodi, S. K., Zhang, F., & Dineva, A. A. (2021). Fuzzy clustering and distributed model for streamflow estimation in ungauged watersheds. Scientific Reports, 11(1), 8243.