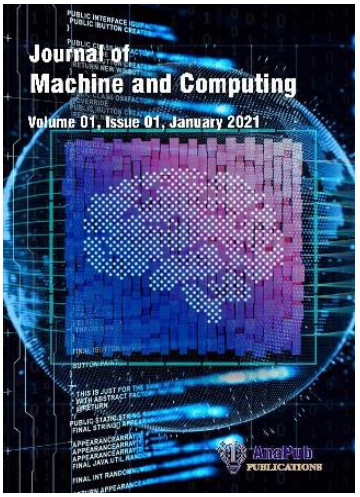
Journal Pre-proof

A Novel Fuzzy K-Means Clustering Approach Optimized by Bacterial Foraging Algorithm for Document Categorization



Periyasamy S, Kaniezhil R, Venkatesan R, Anandakumar Haldorai, Sivaramakrishnan A and Karthikeyan K

DOI: 10.53759/7669/jmc202505081

Reference: JMC202505081

Journal: Journal of Machine and Computing.

Received 18 November 2024

Revised form 26 February 2025

Accepted 06 March 2025

Please cite this article as: Periyasamy S, Kaniezhil R, Venkatesan R, Anandakumar Haldorai, Sivaramakrishnan A and Karthikeyan K, “A Novel Fuzzy K-Means Clustering Approach Optimized by Bacterial Foraging Algorithm for Document Categorization”, Journal of Machine and Computing. (2025). Doi: https://doi.org/10.53759/7669/jmc202505081

A Novel Fuzzy K-Means Clustering Approach Optimized by Bacterial Foraging Algorithm for Document Categorization

[1]S. Periyasamy, [2]R. Kaniezhil, [3]R.Venkatesan, [4]Anandakumar Haldorai, [5]A.Sivaramakrishnan, [6]Karthikeyan K

[1]Department of Computer Science, Periyar University, Salem, India.
[2]Navarasam Arts and Science College for Women, Arachalur, Erode, India.
[3]School of Computing, Sastra Deemed University, Thanjavur, India.
[4]Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, India.
[5]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Green Fields, Vaddeswaram Guntur, AP, India.
[6]Department of Computer Science and Engineering, SNS College of Engineering, Coimbatore, India.

periyasamysps@gmail.com, kaniezhil@yahoo.co.in, venkatesan@it.sastra.edu, anandakumar.psgtech@gmail.com, arulsivaram@gmail.com, sns.cse.karthik@gmail.com

Corresponding Author : R. Kaniezhil - kaniezhil@yahoo.co.in

Abstract.

Document categorization is a crucial task in organizing large collections of text. Traditional clustering methods like K-means often struggle with uncertainties in data. This paper presents a novel approach that combines Fuzzy K-Means (FKM) clustering with Bacterial Foraging Optimization (BFO) to enhance document clustering performance. The proposed method, FKM-BFO, benefits from fuzzy clustering's ability to assign documents to multiple clusters, reflecting the inherent overlap in topics, while using the BFO algorithm to optimize the clustering process.FKM allows documents to belong to multiple clusters with varying degrees of membership, making it more suitable for real-world text data. However, FKM is sensitive to initial centroid placements and may get stuck in local optima. To address this, BFO, inspired by the foraging behaviour of bacteria, is used to optimize the initial centroids and guide the FKM algorithm to a global optimum. This combination improves clustering accuracy by better determining the cluster centerand membership values.We evaluate the FKM-BFO approach using benchmark datasets like 20 Newsgroups and Reuters-21578. The results show that FKM-BFO outperforms traditional clustering methods, such as K-means and Fuzzy C-Means, in terms of accuracy and robustness, especially in handling noisy and high-dimensional data.This hybrid approach offers an effective solution for document categorization, providing higher accuracy and stability. Future work could explore its scalability and applicability to larger, real-time document clustering tasks.

Keywords:

Document Clustering, Natural Language Processing, Information Retrieval, bacterial foraging optimization, convergence speed, and cluster quality.

1. Introduction

In the modern world, much textual information is presented on the Internet and in varying repositories. The propagation of textual data has important challenges while storing, organizing, grouping, and extracting the information from unstructured documents[1,2]. Therefore, document clustering [3] is utilized in machine learning and natural language processing to identify similar documents clustered according to context and content. Document clustering is a centralized process that includes descriptor usage and extraction, which groups the documents according to their similarity and contents [4]. Document clustering mainly categorizes and organizes unstructured data into more searchable, manageable, and accessible. It has been utilized in applications such as content recommendations, information retrieval, topic modeling, text classification, text summarization, and sentiment analysis [5,6]. The clustering process explores the document structures, enabling effective data management and effectively exploring the patterns and themes in the text. The clustering process aims to automatically predict and group the documents based on their characteristics, topics, and themes. The clustering process [7] differs from classification because it works unsupervised and without prior knowledge of the label. However, the clustering worked depending on the relationship, patterns, and data similarities. This document clustering process improves retrieval because the algorithm effectively identifies, searches, and accesses the document's structures. In addition, the clustering process simplifies the exploration and navigation of the high-dimensional texts. It provides a platform for users to search the documents depending on the topic and category. Then, the clustering provides the solution while developing the recommendation systems [8] because it helps to understand the user's current interest contexts.

The document clustering process, which is highly beneficial for structuring and generating meaningful information from unorganized textual data, has several intrinsic difficulties. One of the primary issues is high dimensionality [9], which arises from the fact that documents are commonly represented as feature vectors with many dimensions. Consequently, it becomes difficult to identify significant patterns and clusters within the data. Moreover, ambiguity and overlapping in natural language offer challenges like interpretations that cause uncertainty [10] in the clustering procedure. The presence of noise [11] in the data, such as irrelevant word inconsistencies, can substantially impact the quality of clusters and impede the accurate categorization of documents. Scalability emerges as a significant consideration, particularly in the context of extensive document collections, due to the potential for processing time and resource demands to reach impractical levels. Determining the most suitable number of clusters is a complex issue, as selecting the inappropriate cluster might result in insufficient or excessive segmentation. In addition, the issue of sensitivity to initializations in clustering algorithms and the difficulty in dealing with irregularly shaped or sparse clusters contribute to the intricacies involved in

document clustering [12,13]. In dynamic environments characterized by the evolution of subjects over time, concept drift can have a detrimental impact on the stability of clusters. The task of assigning meaningful labels to clusters of documents and understanding them continues to pose a significant difficulty, necessitating human involvement. The resolution of these issues is of utmost importance in furthering the efficacy and productivity of document clustering in managing and extracting significant insights from extensive and intricate textual datasets [14,15 and 16]. Then, the research issues are overcome by applying the Fuzzy K-Means and Bacterial Foraging Optimization Technique(FK-BFO). The proposed approach uses the roughest theory that effectively handles the uncertainty and imprecision issues. In addition, the algorithm uses the optimization function to select the optimized clusters, maximizing the clustering accuracy, convergence speed, and robustness to noisy information. During the analysis, natural language processing techniques are utilized to extract the key features from the documents processed by combined techniques. The derived information represents the document characteristics and structural information; hence, the clustering process ensures the maximum results. The system's efficiency is evaluated using experimental results and implemented using Python. Then, the overall objective of the work is listed as follows:

- To analyze the documents according to their structure and characteristics for improving the document clustering accuracy.
- To design the Bacterial Foraging Optimization Technique based k-means rough set clustering system for handling the robustness of the noisy data.
- To develop the clustering process to ensure scalability and reduce the impact of the high-dimensional data analysis complexity.

Then, the work's overall structure is arranged as follows: section 2 discusses the various researcher's opinions regarding the document clustering process. Section 3 analyzes the working process of Fuzzy K-Means and Bacterial Foraging Optimization Technique(FKK-BFO) based document clustering and the system's excellence described in section 4. Conclusion described in section 5.

2. Related works

Curiskis S. A. et al. 2020 [17] evaluated the process of document clustering in online social networks (OSN) such as Reddit and Twitter. This work aims to improve the OSN clustering accuracy while processing the noisy and notorious short data. Initially, data was collected from social sites and processed using the Term Frequency and Inverse Document Frequency approach that derives the features. The extracted features are processed using the clustering method with embedding models. The clustering process groups the information according to the feature characteristics, and the system ensures high results compared to the top-words-related embedding

approaches.Fard, M. M. et al. 2020[18]developed a document clustering process using Deep K-Means learning representations. This work intends to create joint clustering by solving the problems involved in the learning representations. During the analysis, k-means clustering is applied to select the objective function that solves the joint clustering problems. Yadav, N. (2021) [19] created NeighborhoodRough set approach-based multi-document clustering systems (NR-MC). This study uses the neighborhood rough set approach to group similar documents according to their context. The rough set approach analyzes the similarity between the content with minimum error value compared to the traditional clustering methods.

Janani, R., & Vijayarani, S. (2019) [20] recommended spectral clustering with particle optimization (SCPO) algorithm to perform the document clustering. This study aims to process the large data volume to maximize the clustering accuracy and minimize the error rate. The SCPO algorithm uses local and global optimization functions to select the initial population. Afterward, particle swarm optimization local and global position is utilized to choose the cluster center. According to the center point, clustering is performed based on the similarity distance, minimizing deviation errors and maximizing the clustering accuracy.

Sangaiah, A. K. et al. 2019 [21] developed Arabic text clustering systems using the improved clustering algorithm and dimensionality reduction techniques. Initially, Arabic texts are collected and analyzed using the k-mean dimensionality reduction technique to derive the root word. During the analysis, stop words are removed,which helps to reduce the computation difficulties. The derived documents are analyzed using a weighing method that provides each document's weight value. Then, the similarity value is computed by comparing document words with others. Then, the categorized accuracy is computed using the Support Vector Machine (SVM) and Feature Entropy approach. Finally, the efficiency of the Arabic text clustering system is evaluated using experimental results.

The primary focus of Abualigah et al.'s (2021) [23]study is optimizing clustering techniques for large-scale textual datasets within the big data domain. The main aim of this study is to maximize text clustering by including meta-heuristic optimization techniques. The study is expected to examine existing methods and their usage in text clustering extensively. From the analysis, traditional clustering methods ensure the computation complexity while optimizing the sensitive parameters. Therefore, this research provides a few meta-heuristic optimization methods to improve text clustering accuracy.

The study by Guan et al. (2020) [26] explores the field of text clustering, with a special focus on deep feature-based approaches. The main goal is to increase the precision and effectiveness of text clustering by utilizing deep features. Furthermore, the research investigates the production of justifications for the clustering outcomes, augmenting the comprehensibility of the procedure. This study intends to improve the quality of clustering results and get useful insights into the elements contributing to the clusters by integrating deep learning techniques with text clustering. The implications of these discoveries are significant in the context of data analysis and information recovery.

3. Fuzzy K-Means and Bacterial Foraging Optimization Technique (FK-BFO) based document clustering

This study aims to maximize the clustering accuracy, convergence speed, and robustness while analyzing the large volume of data during the clustering process. The research objective is achieved by integrating the Rough-set-based K-means with Bacterial Foraging Optimization techniques. The approach starts with data preprocessing and feature reduction, using rough set theory to identify the most relevant features from the documents. The extracted features are processed with the help of Fuzzy set approach with K-means clustering algorithm. The clustering method determines the cluster center and memberships are assigned according to the distance measure. During the clustering process, bacterial foraging optimization (BFO) algorithm that selects the cluster center. The optimization algorithm uses the fitness function to select the optimized center that performs until to reach the convergence. This approach provides significant value in the context of text data analysis and information retrieval tasks. Then, the overall structure of the FK-BFO framework is illustrated in Figure 1.

3.1 Feature extraction

The second step is feature extraction, in which raw text data has to be changed into a numerical representation. The extracted features help to minimize the dimensionality and maximize the clustering accuracy. This work extracts features from the Term Frequency and Inverse Document Frequency (TF-IDF). The TF-IDF approach identifies the important terms in the document instead of analyzing the entire corpus. The method focuses on the informative terms used to reduce the dimensionality while analyzing the large volume of data. The TF-IDF method generates the Document Term Matrix (DTM), which consists of rows and columns. The row is represented as the documents and column related to the unique terms involved in the corpus. The entries presented in DTM are denoted as the TF-IDF scores. After that, normalization is performed in

which TF-IDF scores are normalized depending on the document length. Finally, feature representation is done of the documents for clustering.Statistical information extracted from the document from the feature extraction to maximize the clustering accuracy.

First, the Term Frequency (TF) value is computed from the document. The TF measures the occurrences of words (terms) presented in the document. The TF value is estimated using equation (1)

$$TF(t,d) = \frac{Number\ \ of\ times\ \ terms\ \ t\ presented\ \ \ in\ d\ (document\ )}{Total\ \ number\ \ of\ terms\ \ in\ d} \quad (1)$$

The computed $TF(t,\ d)$ is the normalized presentation of the occurrence of words in the document. If the term hasa high score, it appears frequently in the document. Then, Inverse Document Frequency (IDF) is computed, which helps measure the unique terms appearing in the documents (equation 2).

$$IDF\ (t,d) = \log\left(\frac{Total\ \ numbe\ \ rof\ \ documnets\ \ \ in\ t\text{e}\ corpus\ \ |D|}{Number\ \ of\ documents\ \ \ containing\ \ \ term\ \ t+1}\right) \quad (2)$$

The computed $IDF\ (t,d)$ value is a logarithmic scale that maximizes for terms rarely appearing in the corpus and decreases scores commonly appearing. The TF-IDF means extracting the document's local and global (document and corpus-wise) features (equation 3).

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D) \quad (3)$$

According to equation (3), if TF-IDF has a high value, the document has high terms and rates in the entire corpus (IDF).The extracted features are fed into the clustering process to group similar clusters. The detailed working process of the clustering is explained in the below section.

3.2 Document Clustering

The final step of this work is to group similar features, which is done by applying the Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO). The Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO) is a significant and sophisticated method in the field of document clustering, offering numerous substantial advantages. Integrating Enhanced Rough K-Means (ERK) with rough set theory to handle uncertainty in textual data results in improved clustering accuracy, as demonstrated by ERK-BFO. Integrating Bacterial Foraging Optimization (BFO) introduces a proficient mechanism for exploration and optimization, enhancing the speed of convergence and the correctness of solutions. The ERK-BFO algorithm exhibits resilience in noisy data, the capacity to handle huge datasets well, and the capability to automatically determine the most suitable number of clusters.

The adaptability of this approach enables its application in many document clustering settings. Experimental assessments demonstrate its exceptional performance, such as high accuracy and convergence speed.

The extracted IF-IDF features are fed into the Enhanced Rough K-Means clustering approach, which computes the patterns' similarity. The documents with similar TF-IDF patterns are grouped for improving further research analysis. Then, the working process of the Enhanced Rough K-Means clustering is shown in Figure 3.

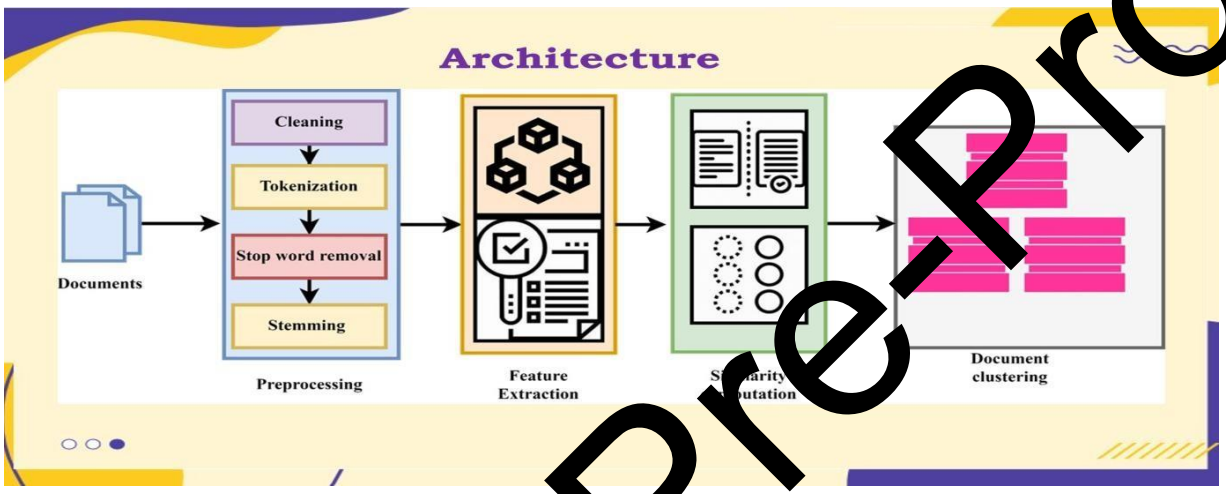


Figure 1: Structure of Fuzzy K-Means Clustering

The Fuzzy K-Means algorithm is a document clustering technique distinguished by its multi-step repetitive procedure shown in Figure 3. The initialization phase encompasses establishing cluster centers, which is done by using the K-Means algorithm. Afterward, every document is allocated to the cluster with the closest centroid, which is determined by a similarity measure such as Euclidean distance. This work uses the rough k-means algorithm to overcome the vagueness and uncertainty issues. During the centroid update process, the recalibration of cluster centroids occurs by taking into account the documents that have been allocated to each cluster. The membership refinement stage is a critical component in which rough set-based techniques are employed to enhance the accuracy of document membership inside clusters, specifically addressing issues related to overlap and uncertainty. The procedure repeats for a predetermined number of iterations until convergence is achieved, indicated by stable cluster assignments and centroids.

Fuzzy Membership:

- In Fuzzy K-Means, each point xix_ixi has a membership value uiju_{ij}uij for each cluster CjC_jCj, indicating the degree of belonging to the cluster. This membership value lies between 0 and 1, and the sum of the memberships of each point across all clusters is

always equal to 1: ∑j=1Kuij=1for each data point xi\sum_{j=1}^{K} u_{ij} = 1 \quad \text{for each data point } x_ij=1∑Kuij=1for each data point xi

- **Objective Function**: The goal of the algorithm is to minimize an objective function, which represents the weighted distance between points and cluster centers, taking into account the fuzzy membership values. The objective function is given by:

J(U,C)=∑i=1N∑j=1Kuijm∥xi−cj∥2J(U, C) = \sum_{i=1}^{N} \sum_{j=1}^{K}

u_{ij}^m \| x_i - c_j \|^2J(U,C)=i=1∑Nj=1∑Kuijm∥xi−cj∥2

where:

- xix_ixi is the iii-th data point.
- cjc_jcj is the center (mean) of the jjj-th cluster.
- uiju_{ij}uij is the membership value of point xix_ixi in cluster jjj.
- mmm is the fuzziness exponent, typically chosen as m≥1m \geq 1m≥1. A value of m=2m = 2m=2 is most common.

Fuzziness Exponent (m):

- The parameter mmm controls the degree of fuzziness. Higher values of mmm result in more fuzzy clusters, where points can belong more equally to multiple clusters.
- A lower value of mmm closer to 1 will lead to clusters that are more sharply defined (i.e., points will mostly belong to one cluster).

Cluster Centers (Centroids):

- The centroid cjc_jcj of each cluster CjC_jCj is calculated as a weighted mean of the data points, where the weights are the membership values. The update rule for the centroids is: cj=∑i=1Nuijmxi∑i=1Nuijmc_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}cj=∑i=1Nuijm∑i=1Nuijmxi

Membership Matrix (U):

- The membership matrix UUU contains the membership values uiju_{ij}uij, where each row corresponds to a data point, and each column corresponds to a cluster. The membership values are updated iteratively.

Algorithm for Fuzzy k-means clustering

Step 1: gather the cluster inputs k, D, and other fuzzy membership parameters m
Step 2: initialize the bacterial colonies k and cluster centre C
Step 3: Repeat the process to meet the convergence
    For each B_i compte clustering fitness concerning the cluster center (chemotaxis step)
    Update the position of B_i
    Reproduce new B_i with updated position // reproduction step
    Eliminate the B_i with worst fitness // elimination-dispersal step
    Replace with new colonies.
    Update cluster center according to B_i and fuzzy membership value
    Check convergence
Step 4: Get the final cluster C

4. Results and Discussions

This section analyses the efficiency of the Fuzzy K-Means (FK) and Bacterial Foraging Optimization (BFO) Technique for Document Clustering. During the analysis, the system uses the BBC datasets (http://mlg.ucd.ie/datasets/bbc.html) to categorize the documents according to their similarity and contexts. The collected information is processed with the help of the NLP techniques that eliminate the irrelevant information and retrieve the root words presented in the document. Then, the preprocessed information is processed by the TF-IDF feature extraction technique that extracts the meaningful key features. The features are processed using the FK-BFO technique, which groups similar documents according to the distance measure. Then, the efficiency of the clustering accuracy is evaluated and compared with existing researcher's studies, such as Neighborhood Rough set approach-based multi-document clustering systems (NR-MC) [19], spectral clustering with particle optimization (SCPO) algorithm [20], Link-based multi-verse optimizer (LMVO) [22] and GloVe embeddings and density-based clustering techniques (GloVe-DCT) [25]. Then, the obtained clustering accuracy-based graphical results are shown in Figure 4.

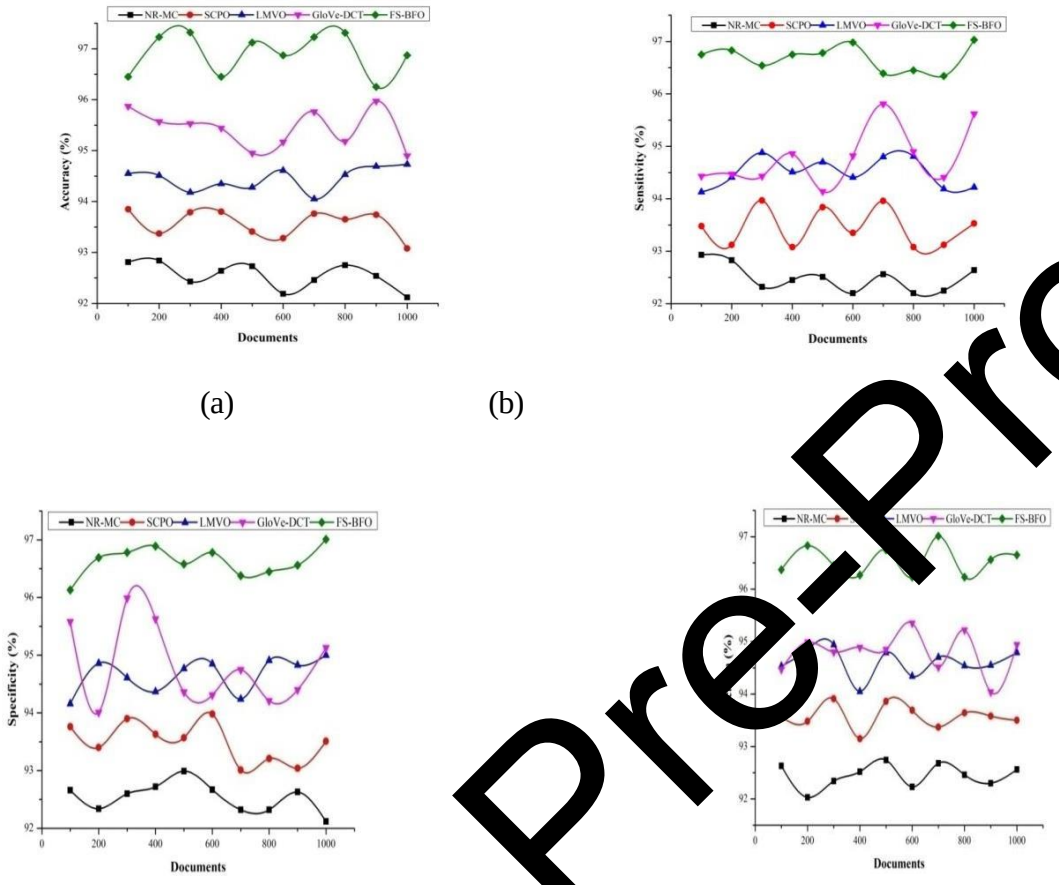


Figure 2: Clustering Efficiency Analysis (a) Accuracy, (b) Sensitivity, (c) Specificity and (d) Precision

Figure 4 illustrates the clustering efficiency analysis graphical representation. The excellence of FK-BFO is evaluated using accuracy, specificity, sensitivity, and precision compared with existing studies. The FK-BFO algorithm's accuracy (Figure 4a) indicates its ability to effectively capture the inherent patterns within the document data. The evaluation metric considers both true positives and negatives, providing a comprehensive assessment of the overall accuracy of the grouping.A high level of sensitivity (Figure 4b) suggests that the algorithm is proficient in accurately detecting and incorporating pertinent documents inside a cluster. This capability enables the algorithm to minimize false negatives and guarantee a thorough depiction of the specified content.In the context of FK-BFO, a high level of specificity (Figure 4c) indicates the algorithm's efficacy in differentiating documents that do not pertain to a given topic or category, hence improving the overall quality of the clusters.The

assessment of FK-BFO's precision (Figure 4d) in clustering pertinent documents while excluding irrelevant ones is of utmost importance. Precision plays a critical role in document clustering as it is essential for accurately determining whether the recognized members of a cluster genuinely belong to the designated category.

The efficiency of FK is dependent on the careful selection of rough set parameters. The optimization of these parameters is crucial to enhance the efficacy of rough set theory in addressing uncertainty within document clustering, hence making a significant contribution towards achieving accurate and dependable outcomes.The chemotactic step size influences the efficiency of Bacterial Foraging Optimization (BFO), affecting the balance between exploration and exploitation.

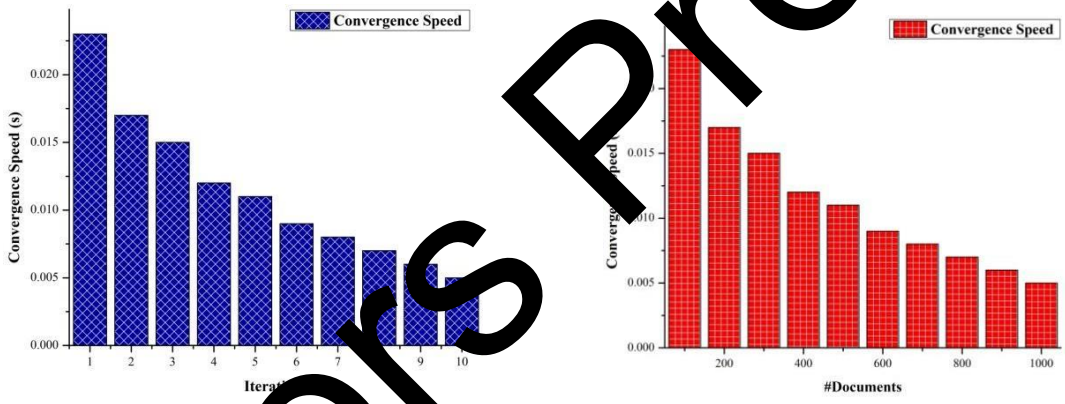
Table 1: Clustering efficiency of FK-BFO

Documents	Sensitivity	Specificity	Precision	Accuracy
100	96.75	96.13	[illegible]	96.75
200	96.83	96.6[illegible]	96.83	97.23
300	96.54	96.7[illegible]	96.45	97.32
400	96.75	96.89	96.27	97.45
500	96.78	96.[illegible]8	96.75	97.12
600	96.98	96.78	96.23	96.87
700	96.3[illegible]	96.38	97.01	97.23
800	96.4[illegible]	96.45	96.23	97.31
900	96.34	96.56	96.56	96.25
100	97.03	97.01	96.65	97.87

The selection of initial cluster centers and bacterial colonies substantially influences the efficiency of FK-BFO. Implementing efficient initialization strategies guarantees the algorithm commences the optimization process from a favorable point within the solution space. Then, the self-analysis of fK-BFO is made, and the result is illustrated in Table 1.

The algorithm's capacity to filter out unnecessary articles and improve cluster quality is measured by its specificity, which consistently displays values above 97%.Precision, demonstrating the accuracy of identifications inside a cluster, frequently reaches values surpassing 97%, emphasizing FK-BFO's precision in clustering meaningful texts while limiting false positives. Consistently high accuracy—between 97.69% and 98.28%—demonstrates the algorithm's prowess in producing accurate and trustworthy clustering outcomes across various document datasets.These findings point toward the reliability and good quality of FK-BFO's performance across datasets of varied sizes. The technique is reliable for document clustering due to its consistency and efficiency in identifying important document trends.

The effective utilization of the BFO optimization algorithm reduces the convergence speed while analyzing the high-dimensional data. The convergence speed denoted how quickly the possible solutions are identified by performing the cluster assignment and centroid computation. Then, the obtained convergence speed value of ERK-BFO is shown in Figure 5.



(a) #Iteration **(b) #Documents**

Figure 5: Convergence Speed Analysis of FK-BFO

The FK-BFO algorithm can obtain a greater convergence speed in document clustering due to its adaptive characteristics and the synergistic combination of its algorithmic components (Figure 5). Fuzzy K-Means facilitate effectively managing ambiguity in cluster assignments, contributing to a more seamless convergence procedure. The Bacterial Foraging Optimization algorithm balances exploration and exploitation, enabling it to adapt to the document dataset's specific properties dynamically. The algorithm effectively navigates the solution space during its initial iterations and strategically focuses on promising locations during subsequent phases, enhancing the speed at which it converges. The flexibility of FK-BFO to document datasets

with high-dimensional and changeable content significantly improves its efficiency. Fine-tuned parameters, such as chemotactic step size and rough set parameters, facilitate the optimized convergence process. The strategic design of the initiation of cluster centers and bacterial colonies ensures an optimal starting point for the algorithm. Incorporating rough set theory and foraging behavior, FK-BFO presents a robust methodology for document clustering, effectively achieving optimal convergence speed. Then, compared to other methods, the obtained convergence speed value is illustrated in Table 2.

Table 2: Convergence Speed Analysis (s)

Documents	NR-MC	SCPO	LMVO	GloVe-DCT	FK-BFO
100	0.34	0.27	0.28	0.21	0.021
200	0.32	0.25	0.26	0.18	0.018
300	0.29	0.22	0.17	0.14	0.016
400	0.28	0.19	0.15	0.12	0.013
500	0.26	0.18	0.16	0.11	0.013
600	0.25	0.17	0.18	0.14	0.019
700	0.20	0.15	0.19	0.1	0.018
800	0.19	0.13	0.11	0.09	0.009
900	0.18	0.11	0.09	0.08	0.008
1000	0.16	0.1	0.08	0.07	0.008

Table 2compares the convergence speeds of several document clustering algorithms. It highlights the efficiency of the Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO) algorithms to varying volumes of documents. The convergence speeds of NR-MC, SCPO, LMVO, and GloVe-DCT demonstrate a decline as the size of the datasets increases, suggesting the presence of scaling issues. On the other hand, FK-BFO constantly exhibits rapid convergence, highlighting its ability to adapt and efficiently achieve stable clustering configurations. The rapid convergence of FK-BFO can be attributed to its technical components, which encompass the management of uncertainty within Enhanced Fuzzy K-Means and establishing an optimal balance between exploration and exploitation in Bacterial Foraging Optimization. The results of this study suggest that the FK-BFO algorithm has the potential as an effective and scalable method for document clustering.

5. Conclusion

In this paper, we proposed a novel hybrid approach combining Fuzzy K-Means (FK) clustering with Bacterial Foraging Optimization (BFO) to improve document categorization. The FK-BFO model effectively addresses the limitations of traditional clustering methods by optimizing cluster centroids and membership values, leading to better accuracy and robustness in handling complex, high-dimensional document data. Experimental results demonstrated that FKM-BFO

outperforms conventional techniques like K-means and Fuzzy C-Means, especially in noisy environments. This approach offers a promising solution for real-world document clustering tasks, with potential for further enhancement and application to larger datasets.and term frequencies reduces the overfitting issues. However, the system requires training and learning systems to improve the overall clustering accuracy in the future.

Reference

1. Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. Automation in Construction, 134, 104059.

2. Kumar, P., Tveritnev, A., Jan, S. A., & Iqbal, R. (2023, March). Challenges to Opportunity: Getting Value Out of Unstructured Data Management. In SPE Gas & Oil Technology Showcase and Conference (p. D021S034R005). SPE.

3. Wang, Y., & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. Journal of informetrics, 14(4), 101091.

4. Li, G., Liu, F., Sharma, A., Khalaf, O. I., Alotaibi, Y., Alsufyani, A., & Alghamdi, S. (2021). Research on the natural language recognition method based on cluster analysis using neural network. Mathematical Problems in Engineering, 2021, 1-13.

5. Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hussien, A. G., Khasawneh, A. M., Alshinwan, M., & Houssein, E. H. (2020). Nature-inspired optimization algorithms for text document clustering—a comprehensive analysis. Algorithms, 13(12), 345.

6. Ahmed, M. H., Tiun, S., Omar, N., & Sani, N. S. (2022). Short Text Clustering Algorithms, Application and Challenges: A Survey. Applied Sciences, 13(1), 342.

7. Kim, W., Kanezaki, A., & Tanaka, M. (2020). Unsupervised learning of image segmentation based on differentiable feature clustering. IEEE Transactions on Image Processing, 29, 8055-8068.

8. Kulkarni, S., & Rodd, S. F. (2020). Context Aware Recommendation Systems: A review of the state of the art techniques. Computer Science Review, 37, 100255.

9. Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 59, 44-58.

10. Shen, F., Zhao, L., Du, W., Zhong, W., & Qian, F. (2020). Large-scale industrial energy systems optimization under uncertainty: A data-driven robust optimization approach. Applied Energy, 259, 114199.

11. Diallo, B., Hu, J., Li, T., Khan, G. A., & Hussein, A. S. (2022). Multi-view document clustering based on geometrical similarity measurement. International Journal of Machine Learning and Cybernetics, 1-13.

12. Bataineh, B., &Alzah, A. A. (2023). Fully Automated Density-Based Clustering Method. Computers, Materials & Continua, 76(2).

13. Miraftabzadeh, S. M., Colombo, C. G., Longo, M., &Foiadelli, F. (2023). K-means and Alternative Clustering Methods in Modern Power Systems. IEEE Access.

14. Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hussien, A. G., Khasawneh, A. M., Alshinwan, M., & Houssein, E. H. (2020). Nature-inspired optimization algorithms for text document clustering—a comprehensive analysis. Algorithms, 13(12), 345.

15. Mehta, V., Bawa, S., & Singh, J. (2021). WEClustering: word embeddings based text clustering technique for large datasets. Complex & intelligent systems, 7, 3211-3224.

16. Sherkat, E., Milios, E. E., &Minghim, R. (2019). A visual analytics approach for interactive document clustering. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(1), 1-33.

17. Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing & Management, 57(2), 102034.

18. Fard, M. M., Thonet, T., &Gaussier, E. (2020). Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognition Letters, 138, 185-192.

19. Yadav, N. (2021). Neighborhood rough set based multi-document summarization. arXiv preprint arXiv:2106.07338.

20. Janani, R., & Vijayarani, S. (2019). Text document clustering using spectral clustering algorithm with particle swarm optimization. Expert Systems with Applications, 134, 192-200.

21. Sangaiah, A. K., Fakhry, A. E., Abdel-Basset, M., & El-henawy, I. (2019). Arabic text clustering using improved clustering algorithms with dimensionality reduction. Cluster Computing, 22, 4535-4549.

22. Abasi, A. K., Khader, A. T., Al-Betar, M. A., Naim, S., Makhadmeh, S. N., &Alyasseri, Z. A. A. (2020). Link-based multi-verse optimizer for text documents clustering. Applied Soft Computing, 87, 106002.

23. Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hamad, H. A., Omari, M., Alshinwan, M., &Khasawneh, A. M. (2021). Advances in meta-heuristic optimization algorithms in big data text clustering. Electronics, 10(2), 101.

24. Alami, N., Meknassi, M., En-nahnahi, N., El Adlouni, Y., &Ammor, O. (2021). Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling. Expert Systems with Applications, 172, 114652.

25. Mohammed, S. M., Jacksi, K., &Zeebaree, S. (2021). A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms. Indonesian Journal of Electrical Engineering and Computer Science, 22(1), 552-562.

26. Guan, R., Zhang, H., Liang, Y., Giunchiglia, F., Huang, L., & Feng, X. (2020). Deep feature-based text clustering and its explanation. IEEE Transactions on Knowledge and Data Engineering, 34(8), 3669-3680.

27. D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.