International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I6P138

Volume 72 Issue 6, 432-441, June 2024

Original Article

Enhanced Rough K-Means and Bacterial Foraging Optimization Technique for Document Clustering

S. Periyasamy[1], R. Kaniezhil[2]

[1]*Department of Computer Science, Periyar University, Salem, Tamilnadu, India.*
[2]*Navarasam College of Arts and Science for Women, Erode, Tamilnadu, India.*

[1]*Corresponding Author : speriyasamyphd@yahoo.com*

Received: 15 February 2024 Revised: 14 May 2024 Accepted: 31 May 2024 Published: 29 June 2024

Abstract *- Document clustering is significant in Natural Language Processing (NLP) and Information Retrieval (IR) because it is widely applied in recommendation systems. Learning techniques perform the document clustering process, but it has varying challenges like high dimensionality, scalability, drifting, and large corpus data handling. The research issues are addressed with the help of Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO). The ERK-BFO was created by integrating the clustering and rough set approach to address the uncertainty and data imprecision issues. The clustering process analyses the document structures that group similar information and increases the clustering efficiency. During this process, the bacterial foraging optimization approach is utilized to predict the optimized cluster centre that improves the convergences and clustering quality. According to the cluster centre values, members are expected to minimize difficulties while exploring the high-dimensional data. Then, the system's effectiveness is evaluated using experimental results, and the ERK-BFO method ensures maximum convergence speed and robustness.*

Keywords *- Document clustering, Natural language processing, Information retrieval, Bacterial foraging optimization, Convergence speed, Cluster quality.*

1. Introduction

The propagation of textual data has significant challenges while storing, organizing, grouping, and extracting the information from unstructured documents[1,2]. Natural language processing and machine learning techniques are widely utilized in document clustering [3] to group similar documents depending on the content and context. During this process, descriptors are extracted and further analyzed to measure the content similarities [4] and a clustering process is performed. The clustering process divides the data into manageable formats that help analyze the unstructured data effectively. This clustering process is utilized in various applications such as text classification, information retrieval, content recommendation, text summarization, sentiment analysis and text classification [5], [6]. The main intention of this process is to analyze document structures and explore the document patterns according to the topics, themes, and characteristics. Regarding an unsupervised learning process, the clustering process does not require prior label knowledge [7]. As discussed, the clustering process works depending on the data similarities, patterns and relationships that minimize the difficulties in high-dimensional text analysis and data navigation issues. After analyzing the document's characteristics and patterns, the automatic system provides solutions by creating recommendation systems [8]. The developed system is used to understand the user demands, interests and requirements.

The clustering process examines every characteristic, unorganized textual data for deriving meaningful information, which leads to intrinsic difficulties. The main issue in document clustering is high-dimensionality [9] difficulties because the document consists of different information and high-dimension feature vectors. In addition, the overlapping and ambiguity of the natural language process cause uncertainty [10] issues. The incorporation of data noise causes inconsistencies and irrelevant matters, affecting the clustering quality. Another difficulty is scalability because the documents are collected from various locations, time and resources. Therefore, the system is confused while determining the number of clusters and cluster centroid identification [12], [13]. In the dynamic environment, cluster stability is affected due to the concept drift, subjects over time. Hence, allocating cluster members to a particular cluster is complex and requires human involvement to improve the clustering accuracy. The discussed issues create productivity, efficacy and scalability issues while analyzing the intricate and extensive textual data analysis [15], [16].

Then, the research issues are overcome by applying the Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-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 bacterial foraging optimization algorithm is incorporated with the clustering algorithm during this study, maximizing the overall clustering process. The system's efficiency is evaluated using experimental results and implemented using Python, and system efficiency is compared with neighbourhood 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 overall objective of the work is listed as follows.

- Analyzing the documents according to their structure and characteristics to improve the document clustering accuracy.
- To design the bacterial foraging optimization technique based on the 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 Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO) based document clustering and the system's excellence. Results and discussion 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, which 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. Then, the k-means clustering is applied to solve the joint clustering issues while clustering.

Yadav, N. (2021) [19] created Neighbourhood Rough set approach-based multi-document clustering systems (NR-MC) to group the similar documents. During the analysis, the rough set approach uses the lower and upper approximation process to identify the content similarity and group the similar documents with a minimum error rate. 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 introduced algorithm uses the global and local optimization functions to select the optimal solution for the population. Then, the particle swarm optimization process is applied to choose the cluster centre, allocating cluster members depending on the distance measure. The effective utilization of the optimization function reduces the error rate and improves the clustering accuracy.

Sangaiah, A. K. et al. 2019 [21] utilized dimensionality reduction and clustering algorithm for creating the arabic text clustering process. First, the arabic texts are gathered and processed using the k-means reduction technique. The dimensionality-reduced approach extracts the root word from the text. After that, stop words are eliminated from the text, minimizing the computation complexity. The extracted documents are further processed using the weighing method, which allocates the weight value for each document. The allocated weight value is used to identify the similarity between the documents. Then, similar documents are grouped using a support vector machine.

Abasi et al. (2020) [22] introduced a based Multi-Verse Optimization (LMVO) approach to solve the difficulties in document clustering. The main intention of this process is to reduce the dimensionality-related challenges while clustering the documents. The optimization algorithm uses the fitness function, which measures the inter and intra-cluster distance to improve the overall clustering accuracy. The effective utilization of the optimization algorithm ensures applicability generalizability and minimizes the computation complexity.

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 will examine existing methods and their extensive usage in text clustering. Based on the analysis, traditional clustering methods ensure

computation complexity while optimizing sensitive parameters. Therefore, this research provides a few meta-heuristic optimization methods to improve text clustering accuracy. The study by Alami et al. (2021) [24] recommended topic modelling with a document clustering approach to perform the arabic text summarization. Initially, arabic texts are gathered and analyzed using NLP techniques to generate logically interrelated descriptions. The NLP process simplifies the computation difficulties by removing the stop words and irrelevant text. Then, features are derived and fed into the modelling approach to cluster similar information with a minimum error rate. Mohammed, S. M. et al. (2021) [25] recommended GloVe embedding with a Density Clustering Approach (GloVe-DCT) to perform the clustering process. The system uses semantic similarity measures to predict the distance between the documents. Word embeddings are utilized during the analysis to predict the root word, simplifying the difficulties in unstructured data analysis. Finally, the density clustering approach forms the cluster according to the similarity measures.

The study by Guan et al. (2020) [26] introduced deep learning techniques to perform text clustering. The text information is collected and processed using deep learning techniques to derive the deep features. Then, NLP techniques are applied to reduce the involvement of irrelevant features. The extracted features are processed using a clustering technique that improves the overall clustering efficiency. According to various research studies, the document clustering procedure utilizes NLP, feature extraction, and clustering techniques to perform the document cluster.

However, the existing methods face difficulties when choosing cluster centres, and scalability and robustness are present in noisy data analysis. These problems maximize the computation complexity and reduce the clustering accuracy. Therefore, this study uses the Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO) to improve document clustering. Then, the detailed working procedure for ERK-BFO is illustrated in the section below.

3. Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO) Based Document Clustering

This ERK-BFO approach aims to maximize the clustering accuracy and robustness and preserve the convergence speed during high-dimensional data clustering. The research uses a combination of optimization algorithms, such as bacterial foraging optimization with rough k-means clustering, to improve the overall clustering efficiency. Initially, data was collected from various resources and processed using data preprocessing techniques to eliminate irrelevant and inconsistent information. The rough set theory feature reduction method minimizes the irrelevant features. The selected features are processed using the k-means clustering approach to group similar features.

The clustering approach predicts the number of clusters and centroid value to allocate the members into specific clusters according to the distance measure. A bacterial foraging optimization algorithm is applied to update the clustering process during the clustering process. The method uses the fitness function to predict the optimal cluster centre, which will continue to perform until convergence is achieved. Then, the overall structure of the ERK-BFO framework is illustrated in Figure 1. Figure 1 illustrates the working process of Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO) based document clustering. Here, documents are collected and processed with the help of NLP techniques that clean the documents and perform the steaming process. The features are extracted from the documents used to group similar clusters effectively. After removing the features, the similarity between the documents is computed to maximize the clustering accuracy. Finally, the enhanced rough set approach and optimization algorithm are utilized to group similar documents with a minimum error rate.

3.1. Noise Removal Process

Initially, the data was collected from the BBC datasets [14] derived from the BBC news. The primary justification for selecting this dataset is to conduct research and perform non-commercial analysis. The dataset comprises 2225 items sourced from BBC news websites between 2004 and 2005. The dataset encompasses five thematic domains: entertainment, business, technology, sports, and politics. The sports dataset comprises 737 records encompassing five games: tennis, rugby, football, cricket, and athletics. The collected data may include noisy information that might impact the entire document clustering process. Hence, it is imperative to eliminate the noise data from the documents before clustering. The noise removal process consists of several steps, such as cleaning, tokenization, stop word removal, and stemming, which fine-tune the documents. Then, the steps involved in the noise removal process are illustrated in Figure 2.

The irrelevant information, such as non-textual elements, HTML tags, unwanted data, and special characters, is analyzed and removed. This process is done with the help of special character removal, HTML tag removal, and regular expression analysis. This step eliminates punctuations, numbers, non-alphabetic characters, and irrelevant information from the document. The described procedures are performed with the help of a regular expression exploration process. Then, a natural language toolkit is utilized to divide the text into words or tokens. The derived tokens are further analyzed using a stopword removal procedure that reduces the dimensionality issues. Generally, "in, is, the and" are considered stop words. Then, the stemming procedure is applied to predict the root word; here, the Porter Stemmer tool kit is utilized to derive the root words, minimizing the computation complexity issues.

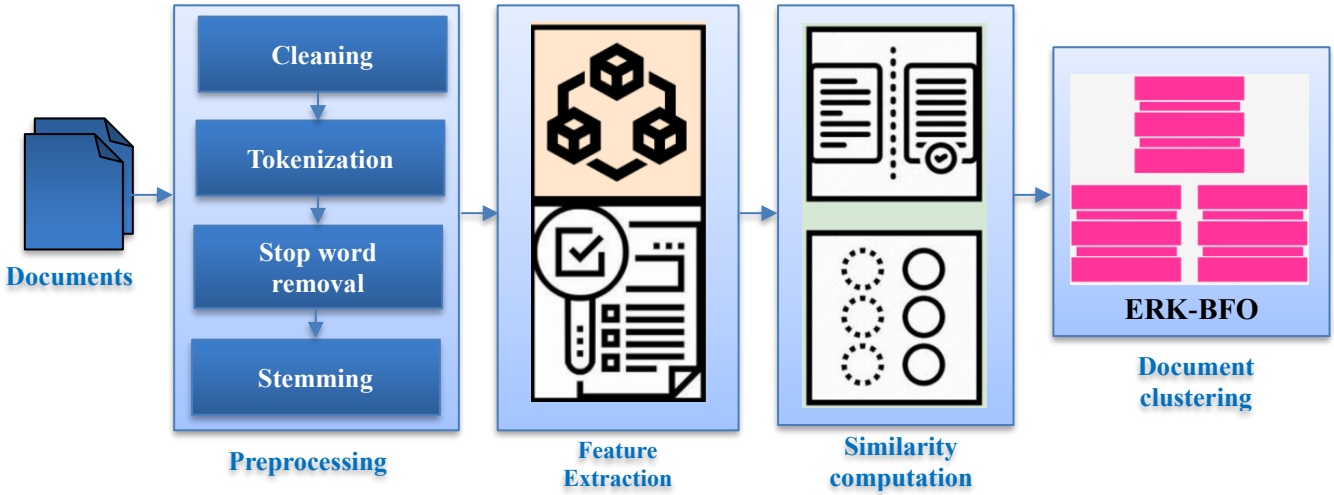


Fig. 1 Structure of ERK-BFO-based document clustering

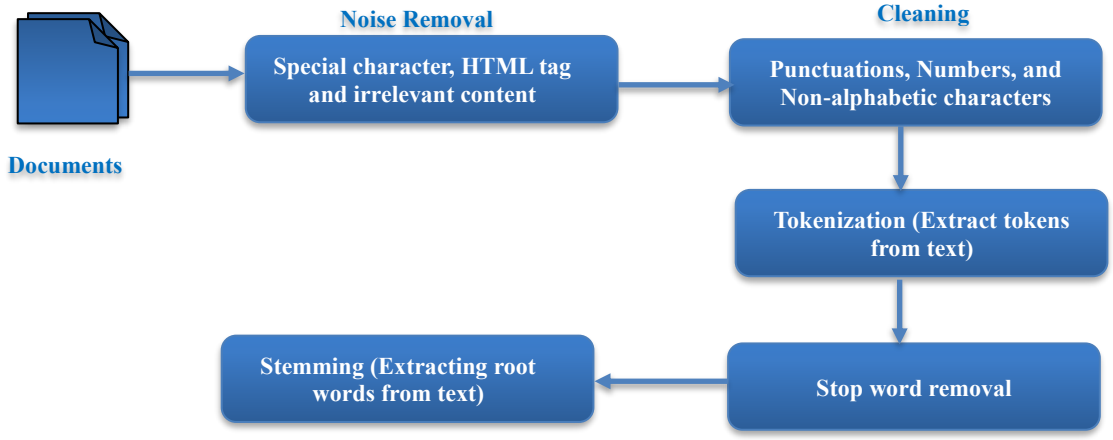


Fig. 2 Steps involved in the noise removal process

3.2. Feature Extraction

The next important step is feature extraction, in which raw text information is converted into numerical illustrations. The feature derivation reduces the computation difficulties and dimensionality issues and improves the clustering accuracy. This study extracts features from the Term Frequency and Inverse Document Frequency (TF-IDF). The TF-IDF process observes text documents and identifies the most relevant, significant features to get the exact text features. The feature extraction process generates the document term matrix in which the row is denoted as the papers and the column is denoted as the unique phrases. After that, the normalization process is performed to fine-tune the matrix score according to the document length. Initially, the term frequency is estimated according to the occurrence of words in the document. The TF value is calculated 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. The document will have a high score if the term frequently appears.

Then, the inverse document frequency is estimated to identify the document's unique terms, which is computed using equation (2).

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

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).

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

Based on equation (3), inverse document frequency is estimated if the document contains maximum terms in the entire corpus. The extracted features are fed into the clustering process to group similar clusters. The detailed clustering process is explained in the section below.

3.3. 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 clustering process integrates the meta-heuristic

optimization algorithm with the rough k-means algorithm. The bacterial optimization approach uses the exploration and optimization procedure to improve the convergence speed and provide effective solutions. In addition, the method works on noisy data, which can handle large dataset volumes with minimum computation difficulties. Moreover, the technique can adapt the clustering settings and provide effective solutions. 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 to improve further research analysis. Then, the working process of the Enhanced Rough K-Means clustering is shown in Figure 3. The Enhanced Rough 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 centres, which is done by using the K-Means algorithm. Subsequently, each document is assigned to the cluster that has the nearest centroid, based on a similarity metric like Euclidean distance. This study uses the robust k-means method to address concerns about ambiguity and uncertainty. During the process of updating the centroid, the recalibration of cluster centroids is performed by considering the documents that have been assigned to each cluster. The membership refinement step is a crucial element where rough set-based approaches are utilized to improve the precision of document membership inside clusters, mainly focusing on concerns with overlap and ambiguity. The process is repeated for a pre-established number of iterations or until convergence is attained, as evidenced by consistent cluster allocations and centroids. The initial cluster centre is determined using the K-means technique, which aids in predicting the appropriate members of the cluster. The initial cluster centre k is selected randomly in the search space. Here, k is the pre-defined number of clusters. For every data point in the document, compute the distance between each cluster centre. The distance is estimated using equation (4)

$$dist\ (X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \qquad (4)$$

In equation (4), X and Y are the data and cluster centre vectors. According to the distance measure, the data point is allocated to the particular cluster. For the data point i, it is assigned to the cluster C_i which is described using equation (5)

$$i = argmin_k\ dist\ (D_j, C_k) \qquad (5)$$

After that, the cluster centres are updated frequently for every iteration, which enhances the document clustering process. During the clustering process, uncertainty is handled with the help of the roughest theory. For every document D_j, the cluster C_i membership is computed depending on the membership function μ_{ij}. Then, the membership value is defined in equation (6)

$$\mu_{ij} = \frac{1}{1+\frac{l_a\ of\ C_i}{U_a\ of\ C_i}} \qquad (6)$$

In equation (6), $l_a\ of\ C_i$ is defined as the lower approximation of the cluster C_i and $U_a\ of\ C_i$ is defined as the upper approximation of the cluster C_i. These approximation values are estimated depending on the rough set theory that helps to address the data vagueness and uncertainty. After computing the membership value, the centroid is updated to make the clusters effective. The centre updating process defined in equation (7)

$$C_i' = \frac{\sum_{j=1}^{N}\mu_{ij}\cdot D_j}{\sum_{j=1}^{N}\mu_{ij}} \qquad (7)$$

In equation (7), N is denoted as the number of documents, membership value of D_j is represented as μ_{ij} in clusterC_i. Next, utilize the most rudimentary approximation method to enhance the document memberships inside each cluster. The refining procedure efficiently addresses the concerns of document ambiguity. Upgrading cluster centres and refining membership is conducted regularly in each cycle. The convergence of the method is achieved when the cluster assignments and centroids reach a condition of stability.

Convergence is often established by a fixed number of iterations or by using a criterion based on changes in centroids. The resulting clusters represent groups of papers that have similar content, determined by the improved clustering technique based on rough set theory. Next, the method outlines the specific procedures for the improved roughest-based K-means clustering.

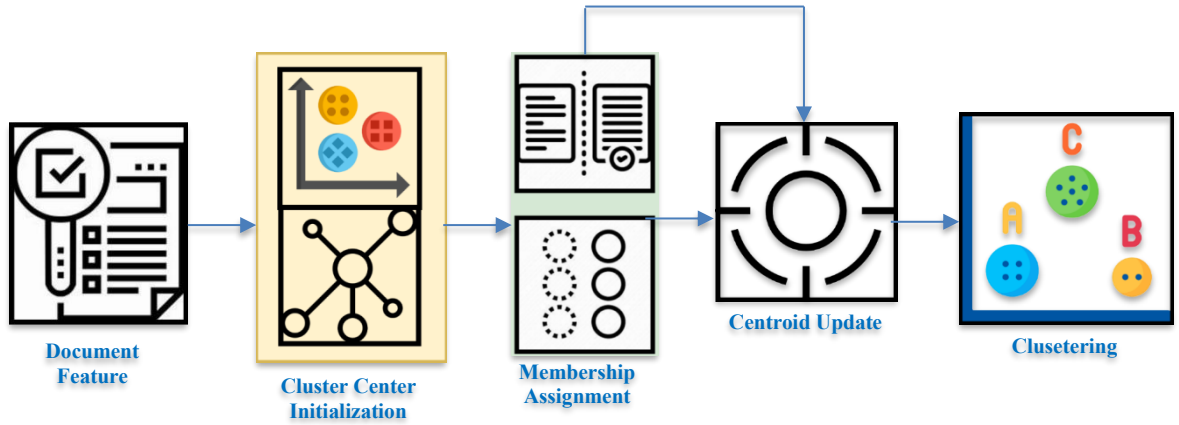


Fig. 3 Structure of enhanced rough K-means clustering

Algorithm for enhanced roughest k-means clustering

Input: k, $D = \{D_1, D_2 \ldots\ldots D_n\}$
Initialize: cluster center C= $\{C_1, C_2 \ldots\ldots C_n\}$ for randomly selected k points
Repeat the process to meet the convergence
For D_i computeμ_{ij} for C_j using equation (6)
Assign D_i to C_j if μ_{ij} is maximized.
Update cluster center C_i' using equation (7)
Check convergence by using the threshold value (no significant changes in cluster center)
Output: cluster C with respective data points

Selecting an optimized cluster centre is crucial because it affects the clustering accuracy. Furthermore, the analysis of data with large dimensionality has presented optimization challenges. Hence, the bacterial foraging optimization technique is included in this clustering procedure to reduce the variances between and within classes.

The utilization of BFO in clustering tasks, such as document clustering, is driven by many variables. The BFO framework achieves a harmonious balance between exploration and exploitation processes.

In the algorithm exploration phase, evaluate the entire solution and exploitation phase, focusing on the high accomplishment probabilities. These two phases used to choose the effective cluster centroid also help to maintain the equilibrium. The algorithm has adaptability characteristics that help fine-tune the cluster centre, maximizing the overall clustering efficiency.

The adaptive characteristics improve the overall clustering efficiency because it can explore the various document patterns. In addition, the optimization method handles the large volume of data with minimum computation difficulties and effective computation of global solutions.

In addition, the process can analyze the noise data with the help of exploitation and exploration characteristics, improving the overall efficiency, flexibility and reliability. Parallelizing BFO enables its application in large datasets, improving computational efficiency. The parallelism approach of BFO is highly successful in enhancing clustering outcomes. The gradient value is linked with fitness. The chemotactic process used to update the bacterial position updating process that is done by using equation (8)

$$X_{ij}^{new} = X_{ij}^{old} + stepsize.random().\left(atractant_{ij} - repellent_{ij}\right) \quad (8)$$

In equation (8), the updating position X_{ij}^{new} of bacteria, B is computed to select the optimized cluster centre. The random function calculates the values between 0 and 1 during this process. The concentration of the chemoattractant is represented as $atractant_{ij}$ and chemorepellent is denoted as $repellent_{ij}$.

After that, the reproduction step is carried out for new individuals in which higher fitness bacteria are selected. The position of reproduced bacteria has to be updated frequently to improve the search process. The reproduced offspring position is updated using equation (9)

$$X_{ij}^{offspring} = X_{ij}^{parent} + mutation.random() \quad (9)$$

Bacteria exhibiting reduced fitness are eradicated and substituted by freshly produced bacteria possessing randomly assigned locations.

This stage aims to facilitate the integration of variety within the population. The elimination-dispersal process entails the adjustment of the position of the deleted bacteria B_i as outlined below:

$$X_{ij}^{new} = X_{ij}^{old} + random().(X_{max} - X_{min}) \quad (10)$$

In equation (10), minimum and maximum values in solution space are represented as $X_{min}\, and\, X_{max}$. The chemotactic parameters, including the step size and the number of chemotactic steps, should be modified under the bacteria's performance in locating improved solutions.

Conduct a convergence assessment using a predetermined criterion, such as the number of iterations or a threshold alteration in the fitness values.

The complete sequence of chemotaxis, reproduction, and elimination-dispersal is iterated until convergence. Then, the steps involved in the Enhanced Rough K-means with the BFO clustering process are described as follows.

Algorithm: ERK-BFO-based document clustering

Step 1: Gather the cluster inputs k, D, and other parameters
Step 2: Initialize the bacterial colonies k and cluster center 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 position
Check convergence
Step 4: Get the final cluster C

According to the above steps, the documents are clustered with maximum clustering accuracy. During clustering, noise removal techniques are incorporated to minimize overfitting issues and computation difficulties.

In addition, the extracted features simplify the similarity computations. Then, the convergence and uncertainty issues are handled by updating the cluster centers using optimization techniques. The system's excellence is evaluated using performance metrics.

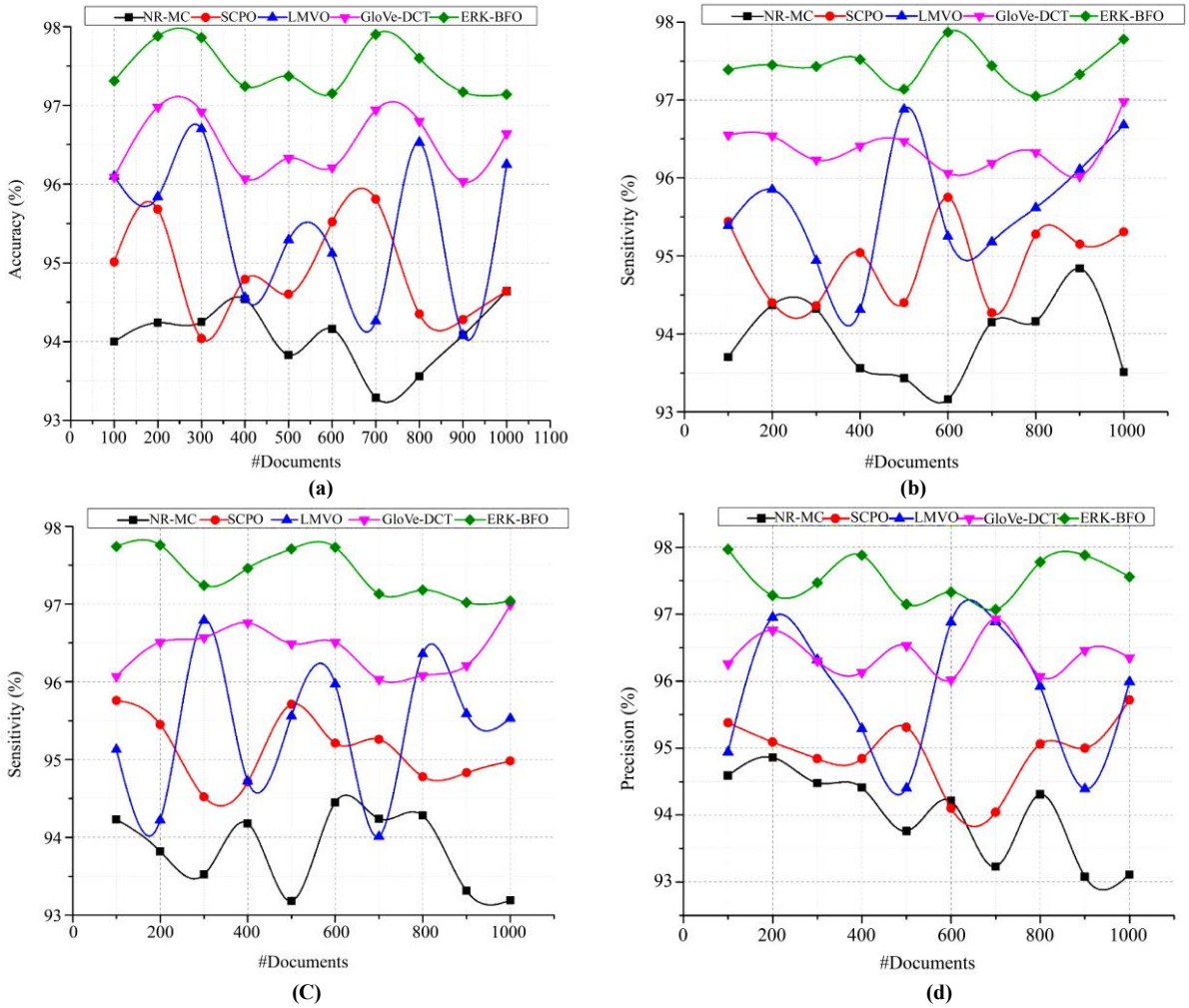


Fig. 4 Clustering efficiency analysis (a) Accuracy, (b) Sensitivity, (c) Specificity and (d) Precision

4. Results and Discussions

This section analyses the efficiency of the Enhanced Rough K-Means (ERK) 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 using the TF-IDF feature extraction technique, which extracts the meaningful key features.

The features are processed using the ERK-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 4 illustrates the clustering efficiency analysis graphical representation. The excellence of ERK-BFO is evaluated using accuracy, specificity, sensitivity, and precision compared with existing studies. The ERK-BFO algorithm's accuracy (Figure 4a) indicates its ability to capture the inherent patterns within the document data effectively. 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) indicates the algorithm's proficiency in accurately detecting and incorporating relevant documents within a cluster, leading to a reduction in false negatives and ensuring a comprehensive representation of the content. Figure 4c illustrates the specificity analysis of ERK-BFO-based clustering efficiency, and Figure 4d demonstrates the precision analysis of the introduced clustering systems. The precision and specificity analysis indicates how effectively the clustering methods identify the cluster members while grouping similar documents. Effectively selecting features improves the overall clustering efficiency and minimizes computation difficulties. In addition, the system's excellence was enhanced due to the rough set-based chosen parameters, which helped to handle the uncertainty issues and improve the system's reliability and efficiency. The chemotactic steps in the optimization algorithm balance the

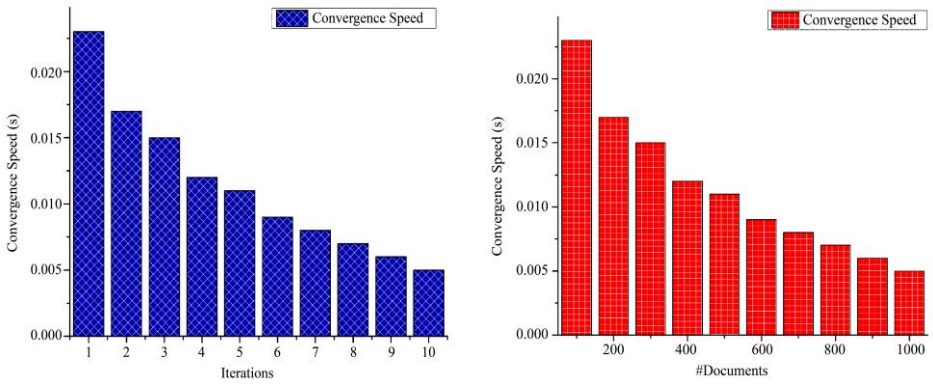
exploitation and exploration process, maximizing the convergence speed and clustering accuracy. The optimal cluster heads impact clustering efficiency and the centroid updating process. Table 1 demonstrates the clustering efficiency of the ERK-BFO method, which is analyzed using different metrics such as sensitivity, specificity, precision and accuracy. The ERK-BFO method attains 98% of the above sensitivity values, which means the introduced approach successfully recognizes the relevant documents. In addition, the technique attains a 97% precision value, indicating that the ERK-BFO approach effectively filters the unnecessary documents, directly improving the overall clustering quality. The high precision value is directly related to the clustering accuracy, in which the system ensures 97.69% to 98.28% accuracy. Therefore, the system is highly trustworthy and reliable while analyzing a large volume of data. The algorithm uses the exploitation and exploration phases that select the global and optimal solutions, ensuring reliability during the cluster head selection. Further, the algorithm manages the convergence speed by computing the distance similarities. Then, the convergence speed analysis is illustrated in Figure 5. Figure 5 illustrates the convergence speed analysis of the ERK-BFO-based clustering system, which ensures the effective value of various iterations and documents. The introduced approach can handle ambiguity issues, enabling smoother convergences. In addition, the optimization algorithm balances the exploitation and exploration process, which dynamically adapts the system for analyzing the large volume of data. The effective utilization of the rough set theory and relevant parameters maximizes the overall clustering efficiency and can handle high-dimensionality issues. The effective convergence speed provides the optimal solutions, and the respective convergence speed analysis is illustrated in Table 2.

Table 1. Clustering efficiency of ERK-BFO

Documents	Sensitivity	Specificity	Precision	Accuracy
100	98.92	97.86	97.24	98.01
200	98.04	97.11	98.51	97.89
300	98.74	97.6	97.46	97.93
400	98.53	97.51	97.48	97.84
500	98.81	97.3	97.52	97.88
600	98.14	97.15	98.64	97.98
700	98.81	97.4	98.36	98.19
800	98.14	97.82	98.64	98.2
900	98.01	97.15	97.92	97.69
1000	98.82	97.43	98.58	98.28

Table 2. Convergence speed analysis (s)

Documents	NR-MC	SCPO	LMVO	GloVe-DCT	ERK-BFO
100	0.34	0.27	0.23	0.21	0.023
200	0.32	0.25	0.20	0.18	0.017
300	0.29	0.22	0.19	0.16	0.015
400	0.28	0.19	0.17	0.15	0.012
500	0.26	0.18	0.16	0.13	0.011
600	0.25	0.17	0.15	0.11	0.009
700	0.20	0.15	0.13	0.10	0.008
800	0.19	0.13	0.11	0.09	0.007
900	0.18	0.11	0.09	0.08	0.006
1000	0.16	0.10	0.08	0.07	0.005



(a) #Iterations **(b) #Documents**

Fig. 5 Convergence speed analysis of ERK-BFO

Several document clustering techniques are compared in Table 2, which provides a comparison of their convergence speeds. It emphasizes the efficiency of the Enhanced Rough K-Means and Bacterial Foraging Optimization Technique (ERK-BFO) algorithms when applied to various document quantities. The convergence speeds of NR-MC, SCPO, LMVO, and GloVe-DCT reduce as the databases' sizes increase, indicating possible scalability concerns. The ERK-BFO algorithm, on the other hand, regularly exhibits speedy convergence, demonstrating its adaptability and efficiency in achieving stable clustering configurations. It is possible to trace the rapid convergence of ERK-BFO to its technical components, including managing uncertainty in Enhanced Rough K-Means and achieving an ideal balance between exploration and exploitation in Bacterial Foraging Optimization. ERK-BFO algorithm shows potential as an efficient and scalable method for document clustering, according to the outcomes of this study, which suggest that the algorithm holds promise.

4.1. Discussions

In order to properly evaluate Enhanced Rough K-Means (ERK) and Bacterial Foraging Optimization Technique (BFO), it is necessary to conduct a comprehensive analysis of their technological capabilities in various fields. The idea of robustness to noise refers to the process of purposefully introducing disruptions and irrelevant features into the dataset to evaluate the algorithms' capacity to retain correct clusters despite interruptions. This study aims to investigate the technological complexities involved with the adaptive cluster assignments and centroid updates of the ERK and BFO algorithms. To demonstrate that the algorithms can effectively handle imperfect real-world data, the purpose of these adjustments is to reduce the impact of noisy aspects as much as possible. Regarding scalability analysis, the primary focus is on the technical aspects of resource allocation, which are becoming increasingly important as the volume of datasets increases. The effective exploitation of the bacterial foraging optimization algorithm exploration process can mitigate the issues associated with managing enormous volumes of data. The scaling strategies utilized in ERK and BFO are revealed through a study of the technological aspects, which sheds light on their capacity to manage computer resources and handle massive datasets effectively.

The employment of sophisticated approaches like ERK and BFO, which can independently determine the ideal number of clusters, is required to accomplish the automatic selection of clusters. The flexibility of these algorithms to a wide range of data scenarios is demonstrated by their use of sophisticated approaches to examine the structures and complexities of datasets.

The complex structure of this technique highlights the cognitive skills and autonomy of the algorithms, reducing the reliance on user-defined parameters and enhancing the algorithms' applicability to a wide variety of document clustering tasks. Compared to more conventional approaches, the effectiveness of ERK and BFO is evaluated in this study by examining benchmark datasets and parameters that have been determined. This work aims to investigate the complex technical elements of ERK and BFO, their unique properties, and whether or not they are compatible with existing clustering methods. According to the findings of the inquiry, the ERK-BFO strategy obtains an impressive accuracy of 97.50%, which is higher than the accuracy achieved by other methods such as NR-MC (94.18%), SCPO (94.96%), LMVO (95.78%), and GloVe-DCT (96.58%).

5. Conclusion

Thus, the work analyzes the working process of Enhanced Rough K-Means and Bacterial Foraging Optimization Technique for Document Clustering. The BBC dataset information is collected and processed frequently during the analysis, and irrelevant information is eliminated. Then, term frequencies are examined to identify the essential features in the documents. After that, extracted features are processed by K-means clustering, which clusters the similarities. During the clustering process, lower and upper approximation values are identified using a rough set algorithm to predict the global solution while clustering. Then, the convergence speed and optimization solutions are obtained by selecting the optimal cluster according to the BFO fitness function. The optimization process reduces the clustering process's difficulties and maximizes the clustering accuracy (97.50%) with minimum convergence speed compared to other methods. The effective utilization of NLP and term frequencies reduces the overfitting issues. However, the system requires training and learning systems to improve the overall clustering accuracy in the future.

References

[1] Chengke Wu et al., "Natural Language Processing for Smart Construction: Current Status and Future Directions," *Automation in Construction*, vol. 134, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Pranav Kumar et al., "Challenges to Opportunity: Getting Value Out of Unstructured Data Management," *SPE Gas & Oil Technology Showcase and Conference*, Dubai, UAE, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Yuzhuo Wang, and Chengzhi Zhang, "Using the Full-Text Content of Academic Articles to Identify and Evaluate Algorithm Entities in the Domain of Natural Language Processing," *Journal of Informetrics*, vol. 14, no. 4, pp. 1-21, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Guang Li et al., "Research on the Natural Language Recognition Method Based on Cluster Analysis Using Neural Network," *Mathematical Problems in Engineering*, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Laith Abualigah et al., "Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis," *Algorithms*, vol. 13, no. 12, pp. 1-32, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Majid Hameed Ahmed et al., "Short Text Clustering Algorithms, Application and Challenges: A Survey," *Applied Sciences*, vol. 13, no. 1, pp. 1-38, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Wonjik Kim, Asako Kanezaki, and Masayuki Tanaka, "Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering," *IEEE Transactions on Image Processing*, vol. 29, pp. 8055-8068, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Saurabh Kulkarni, and Sunil F. Rodd, "Context Aware Recommendation Systems: A Review of the State of the Art Techniques," *Computer Science Review*, vol. 37, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib, "Overview and Comparative Study of Dimensionality Reduction Techniques for High Dimensional Data," *Information Fusion*, vol. 59, pp. 44-58, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Feifei Shen et al., "Large-Scale Industrial Energy Systems Optimization under Uncertainty: A Data-Driven Robust Optimization Approach," *Applied Energy*, vol. 259, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Bassoma Diallo et al., "Multi-View Document Clustering Based on Geometrical Similarity Measurement," *International Journal of Machine Learning and Cybernetics*, vol. 13, pp. 663-675, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Bilal Bataineh, and Ahmad A. Alzahrani, "Fully Automated Density-Based Clustering Method," *Computers, Materials & Continua*, vol. 76, no. 2, pp. 1833-1851, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Seyed Mahdi Miraftabzadeh et al., "K-Means and Alternative Clustering Methods in Modern Power Systems," *IEEE Access*, vol. 11, pp. 119596-119633, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Derek Greene, and Pádraig Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania USA, pp. 377-384, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[15] Vivek Mehta, Seema Bawa, and Jasmeet Singh, "WEClustering: Word Embeddings Based Text Clustering Technique for Large Datasets," *Complex & Intelligent Systems*, vol. 7, no. 6, pp. 3211-3224, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ehsan Sherkat, Evangelos E. Milios, and Rosane Minghim, "A Visual Analytics Approach for Interactive Document Clustering," *ACM Transactions on Interactive Intelligent Systems*, vol. 10, no. 1, pp. 1-33, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Stephan A. Curiskis et al., "An Evaluation of Document Clustering and Topic Modelling in Two Online Social Networks: Twitter and Reddit," *Information Processing & Management*, vol. 57, no. 2, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier, "Deep K-Means: Jointly Clustering with K-Means and Learning Representations," *Pattern Recognition Letters*, vol. 138, pp. 185-192, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Nidhika Yadav, "Neighborhood Rough Set Based Multi-Document Summarization," *arXiv*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] R. Janani, and S. Vijayarani, "Text Document Clustering Using Spectral Clustering Algorithm with Particle Swarm Optimization," *Expert Systems with Applications*, vol. 134, pp. 192-200, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Arun Kumar Sangaiah et al., "Arabic Text Clustering Using Improved Clustering Algorithms with Dimensionality Reduction," *Cluster Computing*, vol. 22, pp. 4535-4549, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ammar Kamal Abasi et al., "Link-Based Multi-Verse Optimizer for Text Documents Clustering," *Applied Soft Computing*, vol. 87, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Laith Abualigah et al., "Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering," *Electronics*, vol. 10, no. 2, pp. 1-29, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Nabil Alami et al., "Unsupervised Neural Networks for Automatic Arabic Text Summarization Using Document Clustering and Topic Modeling," *Expert Systems with Applications*, vol. 172, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Shapol M. Mohammed, Karwan Jacksi, and Subhi R. M. Zeebaree, "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*, vol. 22, no. 1, pp. 552-562, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Renchu Guan et al., "Deep Feature-Based Text Clustering and its Explanation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3669-3680, 2020. [CrossRef] [Google Scholar] [Publisher Link]