Received: February 27, 2024. Revised: April 17, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Text Document Clustering Using Chaotic Northern Goshawk Optimization with K-means Algorithm

Ratnam Dodda[1]* **A Suresh Babu**[1]

[1] *Department of Computer Science and Engineering,*
Jawaharlal Nehru Technological University, Ananthapur, India
* Corresponding author's Email: ratnam.dodda@gmail.com

Abstract: Text document clustering (TDC) is becoming the most famous technique in various real-world applications like information retrieval, data mining, and so on. The TDC poses a significant challenge in these applications as the volume of data collection continues to grow daily. In clustering, the metaheuristics algorithms are used for solving the NP-hard problem and to enhance the model performance. Once the solution space is maximum, traditional approaches are incapable of identifying the solution in minimal time. In this research, the Chaotic Northern Goshawk Optimization (CNGO) with K-means Clustering algorithm is proposed for TDC. Initially, three benchmark datasets such as Reuters-21578, 20-Newsgroup and BBC Sport are used to estimate the performance. Then, the pre-processing techniques such as tokenization, stop word removal, stemming and lemmatization are used for data denoising. The word2vec feature extraction technique is used for converting the document into numerical vectors and finally, the CNGO with K-means is used for TDC. The effectiveness of the proposed method is estimated by utilizing various metrics and it attains the accuracy of 98.36%, 96.26% and 97.36% by using 20-Newsgroup, Reuters-21578 and BBC datasets when compared to competitor approaches like Salp Swarm Algorithm (SSA) and Rider Optimization-based Moth Search Algorithm (Rn-MSA).

Keywords: Chaotic northern goshawk optimization, K-means clustering, Text document clustering, Tokenization, Word2vec.

1. Introduction

The clustering is a well-known unsupervised Machine Learning (ML) approach that has been majorly examined in text context, as well as it is utilized for arranging the maximum number of text documents into various clusters or groups [1]. Text Documents clustering (TDC) is significant for data retrieval, managing as well as removal of large amounts of text data. TDC is an automatic management of learning task, which groups large correlation documents into similar categories and at the same time splits the dissimilar documents into various categories [2, 3]. The clustering has a larger application range, which consists of the organization of text documents, conception as well as classification. Consequently, though each cluster involves unstructured data, relevant and irrelevant documents [4]. Furthermore, clustering is abundantly predicted as a significant characteristic for pattern determination, data mining, medical diagnosis, computer vision and so on. Recently, subject replicas have been significantly used in different applications such as document clustering, recapitulation, retrieval as well as classification for numerous linguistics [5, 6]. Natural Language Processing (NLP) is absolutely a substitute in some cases and requires unique computational difficulties as well as still not completely proven [7].

The traditional TDC approaches illustrate the documents as high-dimensional numeric vectors by utilizing Bag of Words (BOW) or Term Frequency Inverse Document Frequency (TF-IDF) [8, 9]. Then, these high-dimensional document feature vectors are clustered through popular partition or hierarchical clustering approaches. The metaheuristic algorithms have become the significant solution for the problem

DOI: 10.22266/ijies2024.0630.56

of TDC [10]. Based on the various solutions controlled in every iteration, these algorithms are classified as local search or population-based algorithms. The density-based clustering approaches are clustering groups that can group the text with random distribution. However, the traditional clustering approach's performance was poor because of its high-dimensional feature vectors [11, 12]. The general clustering approach with corpus minimizes the feature dimensionality as well as targets to eliminate extended data [13, 14]. In this research, the Chaotic Northern Goshawk Optimization (CNGO) algorithm with K-means clustering algorithm is introduced for TDC. The K-means clustering approach is integrated with the CNGO to prevent premature convergence in local optima. The outcome form CNGO is utilized as an initial seed of K-means clustering approach, which is applied for refining as well as developing the final outcome. This research has used the K-means clustering approach to acquire the local optimum or local search solution. The outcomes of K-means clustering approach supports in exploration and exploitation phase through CNGO. This approach will enable the optimal search solution and it enhances the global solution. The proposed TDC approach effectively performs the clustering process according to the relevant data of the documents. The proposed CNGO with K-means clustering approach achieves the advantages such as robustness, and less computational complexity when compared to the existing methods. The main contributions of this work are as follows:

- The SMOTE sampling technique is utilized for handling the data imbalance problem and then the word2vec feature extraction technique is performed for identifying the keywords for the documents.
- This research utilized the CNGO with K-means clustering algorithm for TDC on various benchmark datasets for enhancing the accuracy as well as robustness of the model.
- The proposed method's effectiveness is calculated by using various assessment metrices such as accuracy, precision, recall, F1-score, AMI, NMI and so on.

An outline of this research work is given as follows: Section 2 discusses the related works. Section 3 provides the materials as well as methods used in the proposed method. Section 4 gives the results and discussion and Section 5 presents the conclusion.

2. Literature survey

In this section, recent research on text document clustering with various approaches is discussed.

Muruganantham Ponnusamy [15] presented the Salp Swarm Algorithm (SSA) for TDC. The presented approach was enhanced with similarity as well as distance-based measurements aimed at clustering domains. The presented approach aimed to develop the exploration as well as the exploitation of search space by the utilization of a Support Vector Machine (SVM) with SSA. In this approach, the former identified the local solution and later identified the global solution from the local solution. The experimental analysis was performed to illustrate the SSA-based efficiency similarity distance measurement, which enhanced the clustering quality. However, the maximal combination had an adversarial effect on productivity as well as effectiveness.

Madhulika Yarlagadda [16] developed the Modsup-based continuous substances as well as Rider Optimization (ROA) -based Moth Search Algorithm called Rn-MSA for TDC. In this developed approach, initially, input documents were provided for the pre-processing and it was extracted based on the TF-IDF approach as well as Wordnet features. After that, the extracted features were taken out according to the frequent itemset for the feature understanding formation. Finally, the Rn-MSA was utilized for the clustering, which was intended to integrate the ROA as well as the Moth Search Algorithm (MSA). However, the developed approach had frequently inclined to local optima by the minimal number of iterations.

Nikhil V. Chandran [17] introduced the Topic Stricker, the approach that integrated the benefits of unsupervised topic modeling by supervised string kernels for the classification of the tasks. The acquired coincident words and proportions of the topic were utilized for the corpus minimization for the sequence of topic words. The minimized word was forwarded to the text classification with the string kernels aid, which crucially enhanced the accuracy as well as minimized the training time. The suggested approach outperformed the text classification by utilizing bag-of-words with kernel-based string embedding models. However, the suggested approach did not consider the long-term dependencies.

Kamal Berahmand [18] developed the Deep Text Clustering approach by local manifold in the Autoencoder layer (DCTMA) that performed various similarity metrics to acquired manifold data, such that the last similarity matrix was acquired from an average of those matrices. The acquired matrix was included in the bottleneck symbol layer in an autoencoder. The suggested approach's significant aim is to develop similar representations for samples belonging to a similar cluster. However, the suggested approach does not estimate the clustering performance even for a minimum number of clusters. However, the suggested approach exclusively utilized the various datasets, thus, this may not be comprehensive for other operations.

Sungwon Jung and Sangmin Ka [19] implemented the document embedding approach of Graph autoencoder for clustering. In that implemented approach, an undirected as well as weighted sparse graph from the document set was developed, wherein every document was decocted through the node, and weighted edges designed in the graph had greater cosine similarities among the end of the two nodes. Finally, an implemented approach was applied to the graph to estimate the node embedding vectors and every node in the graph was utilized as a document embedding vector. However, the suggested approach does not modify the dynamic system modification characteristics.

Purba Daru Kusuma and Ashri Dinimaharawati [20] presented the metaheuristic approach named Extended Stochastic Coati Optimizer (ESCO). The presented ESCO expanded the number of searches as well as references utilized in COA. The ESCO implemented the stochastic procedure for every unit to select the searches to perform and it varies from COA, which divided the population into two-fixed groups and performed based on its strategy. The ESCO had implemented multiple sequential phases in each iteration and two options could be selected in every phase. However, the strategy executed in initial group had more constraint compared to the metaheuristics in next group.

Purba Daru Kusuma and Anggunmeka Luhur Prasasti [21] introduced the stochastic optimization approach which integrated the direction-based search as well as neighbourhood search named Walk-Spread Algorithm (WSA). Two-direction-based searches were employed in each iteration where every search generated the individual child. In the meantime, there were two neighbourhood searchers were employed in each iteration, where every search generated number of children. The global best unit had the initial reference, while two shuffled units had the next reference during the direction-based search performance.

The limitations of the above-discussed existing TDC approaches considered that the long-term dependencies, modified individually for the dynamic system modification characteristics, infrequently disposed to local optima by the minimal number of iterations. To address these limitations, this research proposes the CNGO with a K-means clustering algorithm for clustering the text document. Thus, the proposed CNGO with K-means clustering provides better results for TDC.

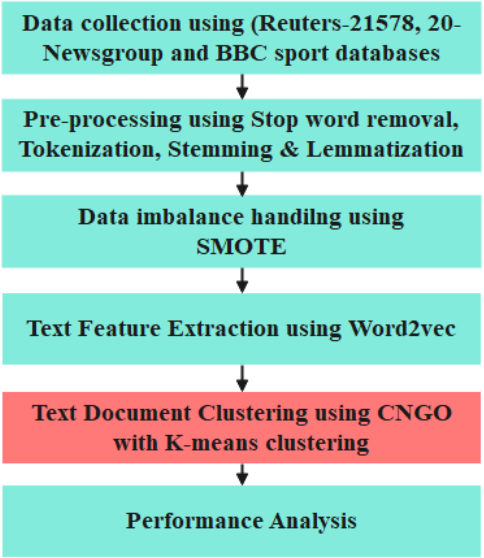
Data collection using (Reuters-21578, 20-Newsgroup and BBC sport databases
Pre-processing using Stop word removal, Tokenization, Stemming & Lemmatization
Data imbalance handilng using SMOTE
Text Feature Extraction using Word2vec
Text Document Clustering using CNGO with K-means clustering
Performance Analysis

Figure. 1 Schematic diagram of the proposed method

Table 1. Characteristics of the TDC datasets

Datasets	Documents	Cluster	Words in document		
			Minimum	Average	Maximum
Reuters-21578	7674	8	5	102	964
20-Newsgroups	18846	20	20	374	23031
BBC sport	2225	5	86	372	4342

3. Proposed methodology

Fig. 1 illustrates the schematic diagram of the proposed method. This section shows the TDC utilizing Chaotic Northern Goshawk Optimization (CNGO). Initially, the data is collected from three benchmark datasets such as Reuters-21578, 20-Newsgroup and BBC-sport group. After that, pre-processing is performed to eliminate the unwanted words, punctuation and so on by the utilization of stop word removal, tokenization, stemming and lemmatization. Then, pre-processed data is provided by the SMOTE technique for solving the problem of data imbalance. Then, the textual feature extraction is performed to identify the keywords from the document by using the TF-IDF technique. Finally, the extracted features are provided to the TDC by using CNGO with K-means clustering algorithm.

3.1 Dataset

The initial stage of the proposed method is data collection. This research presented the results by utilizing the three benchmark datasets such as

Reuters-21578 [22], 20Newsgroups [23] as well as BBC-sports databases [24]. Table 1 illustrates a detailed description of the characteristics of these TDC datasets.

3.2 Pre-processing

In this pre-processing step, the collected dataset is utilized as input at the mapper level to eliminate unwanted words. The collected dataset contains stopping words, punctuations, lower cases, and so on, which causes the accuracy results. To solve these problems, this research utilizes various pre-processing techniques such as stop word removal, and lemmatization. The detailed description of these techniques is discussed below.

3.2.1. Stop word removal

The words are commonly utilized in a sentence in which articles, and prepositions are involved. This approach ignores the particular words that are presented repeatedly in the text. The lack of informative words is eliminated to minimize the noises presented in input data. Hence, it is utilized for stop word removal to minimize the greater accumulation as well as faster processing.

3.2.2. Tokenization

It is an approach for dividing the letter strings in text into tokens. Any characters crushed among two spaces like words, phrases, keywords, or symbols are examined as tokens [23].

3.2.3. Stemming

It is an approach for eliminating the prefixes as well as suffixes from original words to acquire general word vision. This procedure is taken out by utilizing the Porter stemmer, which eliminates prefixes and suffixes like “ly”, “ed”, “ing” and so on.

3.2.4. Lemmatization

It is the process of grouping various inflected forms of similar words. It is utilized in NLP, linguistics as well as chatbots. For illustration, the lemmatization approach has eliminated the better word to its root word [24].

This research analyzed that 20-Newsgroups has class balanced datasets as well and Reuters-21578 and BBC sport datasets are class imbalance datasets. To overcome the problem of class imbalance problem by using the SMOTE technique and a detailed description of this technique is discussed in the following.

3.3 Data imbalance handling

In this section, the pre-processed output is utilized as input to the data imbalance handling technique. The Synthetic Minority Over-Sampling Technique (SMOTE) is the most utilized oversampling model for solving the problem of class imbalance [25]. In SMOTE, the minority class is over samples through developing the new samples from every sample in the minority class as well as the nearest neighbors where every synthesized sample is expressed in Eq. (1) as follows:

$$Xsyn = Xi + rand(0,1) \times |Xi - Xneighbour| \quad (1)$$

Where, $Xsyn$ – synthesized sample; Xi – sample provided by the minority class; $Xneighbour$ – arbitrarily chosen sample from K-Nearest Neighbors (KNN) to sample Xi; $rand(0,1)$ – arbitrary number from 0 to 1. This technique extends the occurrence of synthetic, arbitrarily, along the line associated with two actual occurrences. The significant benefit of utilizing SMOTE for oversampling instead of utilizing other methods like Random Over Sampling (ROS) is overfitting avoidance. The outcomes can be acquired through synthesizing the new samples from the minor class rather than the replication of samples applied in ROS. Despite the acceptance as well as better effectiveness in various application areas, inappropriately no one is efficiently suitable, hence the SMOTE has various issues. The initial issue is that the high dimensionality of the dataset has been utilized which limits various classifiers from enhancing the performance despite minor class oversampling. The significant drawback of SMOTE nevertheless of the dimensionality of the dataset is the noisy samples developed as an outcome of its inherent randomness. This is by the noise that occurs in the actual dataset as well as utilized in new samples development tends to number of propagations as well as noise enhancement. Hence, it has a better attempt at the utilization of the SMOTE approach. This technique helps to minimize the noise-developed effect through SMOTE as well as enhance the performance. The obtained output is then forwarded to the feature extraction process.

3.4 Feature extraction

The word2vec feature extraction technique is performed after the data preprocessing, and in training which converts the text data into vectors. The general aim of feature extraction is utilized for reducing the dimensionality as well as data compaction. The word2vec is the most significant approach in NLP for learning word embeddings, which are dense numerical word depictions in continuous vector space. This technique utilizes the advantage of a Neural Network to design “equivalence classes” of provided words. Every word in the text is depicted as a vector, permitting to estimate of the similarity degree among the words as the distance between two vectors. This technique utilized in this research was developed by utilizing the actual word2vec execution on news texts as well as Wikipedia3 acquired from the G14 portal from Sep 15, 2016. In this research, the avoidance parameters for word2vec training are accepted. In this step, the stop words are eliminated from every post that experienced lemmatization before the similarity analysis. After that, the matrix similarity approach is performed. The output of extracted features is then provided to the clustering process.

3.5 Text document clustering

This section utilizes the input as extracted textual features for clustering the text document. The text document can be clustered by utilizing the integration of the metaheuristic optimization algorithm of CNGO with the K-means clustering algorithm. A detailed description of how these algorithms integrated and clustered the text document is provided in the following section.

3.5.1. Chaotic northern goshawk optimization

The NGO is a population-based optimization approach in which the northern Goshawks are the search members in this approach. The NG belongs to Slip-up Genius, which hunts a variety of prey, particularly of both small and large birds, small mammals such as rabbits, rats, and squirrels and larger creatures such as raccoons and foxes. The two levels of the hunting process of NG involve fast approaching its prey in the initial phase, then broadly following the prey in another phase. The main inspiration source for NGOs is the mathematical modeling of the before-discussed approach.

3.5.1.1. Mathematical modelling

The population members are updated by utilizing the NG hunting approach simulation. In this approach, the two significant NG activities are regenerated in two phases such as exploration and exploitation.

- **Exploration or prey identification:**

In the first phase of hunting, the NG selects the target arbitrarily as well as attacks it randomly. Because the prey in the search area is randomly selected, this phase improves the exploring capability of the NGO. This phase tends to the global search space with respect to best position identification. The notations of this initial phase are mathematically expressed in Eqs. (2)-(4) as follows:

$$P_i = X_{num} \quad (2)$$

$$x_{(i,j)}^{new,p1} = \{x_{i,j} + r\left(P_{i,j} - I \times x_{i,j}\right), F_{pi} < F_i \; x_{i,j} + r\left(x_{i,j} - P_{i,j}\right), \; F_{pi} \geq F_i \quad (3)$$

$$X_i = \{x_i^{new,p1}, F_i^{new,p1} < F_i \; X_i, F_i^{new,p1} \geq F_i \quad (4)$$

Where, P_i - ith prey position; num – the random number in the range between 1 and N ; F_{pi} – objective function; $x_i^{new,p1}$ – new status for i th solution; $x_{(i,j)}^{new,p1}$ – j th dimension; $F_i^{new,p1}$ – the value of objective function based on the initial NGO phase; r random number among 0 and 1; I – random number 1 and 2 respectively.

- **Exploitation or chasing and escaping operation**

The prey tries to run just after the NG attacks. As an outcome, the NG endures to chase the prey in tail-and-chase custom. Simulating this inclination enhances the exploitation of NGO capability for local search of search space. This hunting is hypothetical to be near an attack point with radius R in the proposed NGO approach. The chasing ideas are formulated in this phase and which is expressed in Eqs. (5)-(7) as follows:

$$x_{i,j}^{new,p2} = x_{i,j} + R(re - 1)x_{i,j}, \quad (5)$$

$$R = 0.02\left(1 - \frac{t}{T}\right), \quad (6)$$

$$X_i = \{x_i^{new,p2}, F_i^{new,p2} < F_i \; X_i, \; F_i^{new,p2} \geq F_i \quad (7)$$

Where, t – iteration counter; T – maximum number of iterations; $x_i^{new,p2}$ – new status for ith solution; $x_{i,j}^{new,p2}$ – j th dimension; $F_i^{new,p2}$ – objective function value using NGO phase. The chaotic series is performed in location as well as identification phase, as well as enhancing the exploration capability of NGO. The chaotic map is one with the chaotic pattern as well as the capability to encourage the chaotic movement. This research performs the popular logistic map and identified by using Eq. (8) as:

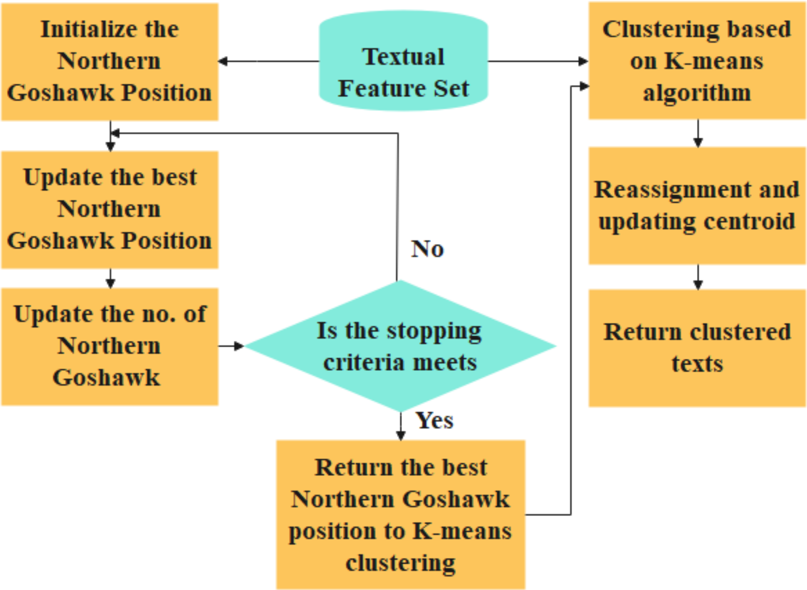
Initialize the Northern Goshawk Position
Textual Feature Set
Clustering based on K-means algorithm
Update the best Northern Goshawk Position
No
Reassignment and updating centroid
Update the no. of Northern Goshawk
Is the stopping criteria meets
Return clustered texts
Yes
Return the best Northern Goshawk position to K-means clustering

Figure. 2 Flowchart of the proposed CNGO with K-means clustering

$$\mu(t+1) = a \times \mu(t) \times (1 - \mu(t)) \qquad (8)$$

Where, $\mu(t)$ – chaotic map; t – iteration number; a – constant value equal to 4. The chaotic NGO (CNGO) enhances a stochastic behavior, by eliminating a premature convergence through utilizing the chaotic map instead of facilitating arbitrary numbers in NGO to a prey. Hence, for supervising the NG throughout the prey at an exploration phase, which is expressed in Eq. (9) as:

$$x_{i,j}^{new,p_1} = \{x_{i,j} + \mu_t.(p_{i,j} - I.x_{i,j}), F_{pi} < F_i \; x_{i,j} + \mu_t.(x_{i,j} - p_{i,j}), \; else \qquad (9)$$

Furthermore, to enhance the exploration as well as capability of the searching phase, the humblest NGO's contribution of maximum fitness value will be varied at each iteration through new arbitrary NGOs. This process is expressed in Eq. (10) as:

$$X_{worst} = X_{i\,min} + rand \times (X_{i\,max} - X_{i\,min}) \; (10)$$

Where, X_{worst} – NGO with maximum fitness value.

3.5.2. K-means clustering

The K-means clustering is an ML approach that groups the data points surrounding each other according to their comparisons, whereas, the clustering approach is a non-supervised algorithm, in which an input is unlabeled as well as problem-solving.

The K-means clustering for provided sample data divides the sample set into k clusters based on the distance among the samples, which designs the points in the cluster as minimum as possible as well as distance among the clusters as maximum as possible. This approach involves two autonomous phases: Initially, the k clusters are arbitrarily chosen as well as earlier set the k values. Then, it aims to classify every data into the closest center. The process of iteration is continuous until the criterion function attains minimal value. It is hypothetical that the cluster is divided into $C_1, C_2, \dots C_k$ as well as aimed to reduce the square error E, which is expressed in Eqs. (11) and (12) as:

$$E = \sum_{i=1}^{k} \sum_{t \epsilon C_i} \|t - u_i\|_2^2 \qquad (11)$$

$$u_i = \frac{1}{|C_i|} \sum_{t \epsilon C_i} \mathrm{t} \qquad (12)$$

Where, u_i – mean vector of cluster C_i as well as known as centroid. A Euclidean distance is majorly utilized for the identification of distance among every data object as well as cluster center. A Euclidean distance $d(x_i, y_i)$ among two vectors $x = (x_1, x_2, \dots x_n)$ and $y = (y_1, y_2, \dots y_n)$, which is expressed in Eq. (13) as:

$$d(x,y) = [\sum_{i=1}^{n} (x_i - y_i)^2]^{1/2} \qquad (13)$$

The K-means clustering algorithm is simple in principle, has faster convergence as well as easy to implement, and is stronger as well as more robust.

Received: February 27, 2024. Revised: April 17, 2024.

3.5.3. CNGO-K-Means clustering algorithm

Fig. 2 shows the workflow of the proposed CNGO with K-means clustering. The CNGO algorithm is hypothetically capable of enhancing the primary arbitrary solutions as well as better convergence points in search space. Utilizing the CNGO, the number of clusters as well as their centroids is identified, and then it is provided as an initial kernel to the K-means algorithm. The steps need to be followed for combining the CNGO with K-means clustering.

Step 1: Goshawk matrix value is initialized through a textual feature vector. In the clustering context, the individual Goshawk position indicates the cluster centroid.

Step 2: For each Northern Goshawk

a) Estimate the Goshawk's fitness according to the cluster criteria. The best Goshawk position is updated.
b) Estimate the number of Goshawks
c) Repeat until the stopping criteria the satisfied.

Step 3: Apply the K-means clustering algorithm using the best Goshawks position and number of Goshawks acquired in CNGO.

Step 4: Return the clustering texts as well as centroids.

4. Results and discussion

This section illustrates the result and discussion of the proposed CGNO with K-means clustering for TDC. Furthermore, this section represented the experimental setup, evaluation metrics, performance analysis, and comparative analysis.

4.1 Experimental results and evaluation metrics

The proposed CGNO with K-means clustering for TDC is executed by utilizing the platform of Python 3.9 with Windows 10 OS, 16GB RAM with intel-i7 processor. The proposed method is analyzed by utilizing various assessment metrics like accuracy, precision, recall, F1-score, Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). The mathematical expressions of these metrics are expressed in Eqs. (14)-(19) as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

$$F1-score = \frac{2TP}{2TP+FP+FN} \tag{17}$$

$$ARI = \frac{\sum_{i,j}\binom{n_{i,j}}{2}-\left[\sum_i(n_{i_2})\sum_j\binom{n_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{n_i}{2}\right]-\left[\sum_i\binom{n_i}{2}\sum_j\binom{n_j}{2}\right]/\binom{n}{2}} \tag{18}$$

$$NMI = \frac{2\times I(cl:k)}{[e(cl)+e(k)]} \tag{19}$$

Where, TP - True Positive; TN – True Negative; FP – False positive; FN – False Negative; n- number of documents, $n_{i,j}$ - documents that are present in both cluster and class, n_i and n_j – documents in cluster and class. k - set of clusters, e - entropy, cl - class label, and $I(cl: k)$ - mutual data among k and cl.

4.2 Performance analysis

In this section, the effectiveness of the proposed method for TDC is analyzed by using three standard datasets. Tables 2, 3 and 4 show the results on TDC using 20-Newsgroups, Reuters-21578 and BBC sports datasets.

As shown in Table 2 and Fig. 3, it is obvious that NCGO-K-means outclasses various optimization algorithms of K-means clustering for TDC by using 20-Newsgroups dataset. Moreover, the NCGO-K-means attains better performance as compared to the existing methods with various assessment metrics. The existing methods like Ant Colony Optimization (ACO), Cuckoo Search Optimization (CSO), Grey Wolf Optimization (GWO) and Particle Swarm Optimization (PSO) are analyzed. Furthermore, the metaheuristic algorithms of ESCO [20] and WSA [21] are analyzed with the K-means clustering approach with the proposed NCGO-K-means in the statistical measures. The proposed NCGO-K-means attains the accuracy, precision, recall, F1-score, ARI and NMI are 97.38%, 92.14%, 97.51%, 89.21%, 0.655 and 0.724 respectively. The proposed method attains effective outcomes when compared to other competitor approaches.

As shown in Table 3 and Fig. 4, it is obvious that NCGO-K-means outclasses various optimization algorithms of K-means clustering for TDC by using the Reuters-21578 dataset. Moreover, the NCGO-K-means attains better performance as compared to the existing methods with various assessment metrics. The existing methods such as ACO, CSO, GWO, and PSO are analyzed. Furthermore, the metaheuristic algorithms of ESCO [20] and WSA [21] are analyzed with the K-means clustering approach with the proposed NCGO-K-means in the statistical measures. The proposed NCGO-K-means attains the accuracy, precision, recall, F1-score, ARI and NMI are 96.26%, 96.11%, 98.45%, 97.98%, 0.495 and 0.634 respectively. The proposed method attains effective

Table 2. Results on TDC using 20-Newsgroups dataset

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ARI	NMI
ACO-K-means	92.18	90.35	91.32	84.33	0.489	0.490
CSO-K-means	94.26	92.19	92.36	85.58	0.503	0.576
GWO-K-means	96.01	94.27	94.12	87.28	0.587	0.590
PSO-K-means	96.39	95.98	95.19	88.26	0.612	0.687
ESCO-K-means	96.52	96.26	95.78	88.95	0.634	0.689
WSA-K-means	96.88	97.37	96.87	89.04	0.652	0.713
NCGO-K-means	97.38	98.14	97.51	89.21	0.655	0.724

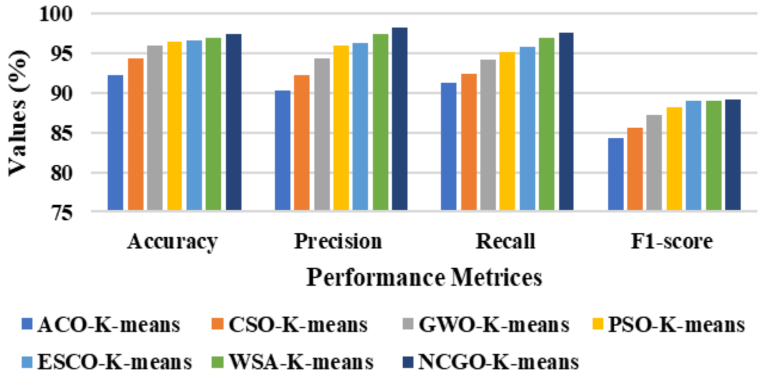

Figure. 3 Graphical representation of clustering results using 20-Newsgroups dataset

Table 3. Results on TDC using Reuters-21578 dataset

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ARI	NMI
ACO-K-means	90.45	91.67	92.88	91.46	0.298	0.308
CSO-K-means	92.18	92.67	93.29	92.57	0.313	0.350
GWO-K-means	93.29	93.28	95.38	93.57	0.329	0.480
PSO-K-means	95.35	95.13	97.46	95.39	0.467	0.612
ESCO-K-means	95.42	94.12	97.34	95.67	0.455	0.615
WSA-K-means	95.78	95.23	97.88	96.26	0.484	0.621
NCGO-K-means	96.26	96.11	98.45	97.98	0.495	0.634

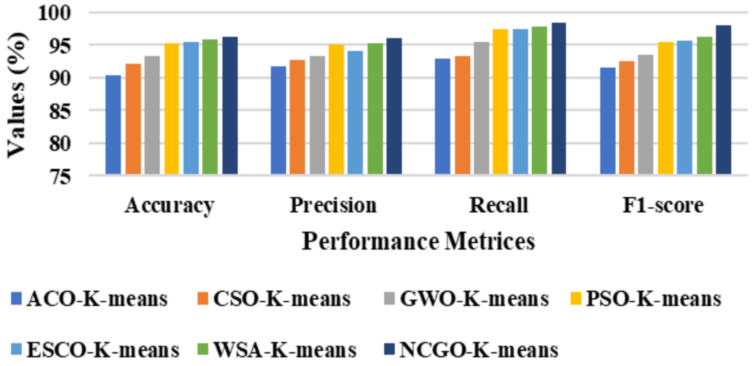

Figure. 4 Graphical representation of clustering results using Reuters-21578 dataset

outcomes when compared to other competitor approaches.

As shown in Table 4 and Fig. 5, it is obvious that NCGO-K-means outclasses with various optimization algorithms of K-means clustering for TDC by using the BBC sports dataset. Moreover, the NCGO-K-means attains better performance as compared to the existing methods with various assessment metrics. The existing methods such as ACO, CSO, GWO, and PSO are analyzed. Furthermore, the metaheuristic algorithms of ESCO [20] and WSA [21] are analyzed with the K-means

Table 4. Results on TDC using BBC sport dataset

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ARI	NMI
ACO-K-means	90.87	92.67	90.67	90.01	0.601	0.647
CSO-K-means	91.39	93.29	91.93	90.48	0.679	0.680
GWO-K-means	94.47	95.39	93.29	91.48	0.723	0.768
PSO-K-means	96.20	96.46	95.39	92.67	0.854	0.794
ESCO-K-means	96.78	96.68	95.65	92.55	0.867	0.802
WSA-K-means	97.23	97.02	95.89	92.79	0.901	0.899
NCGO-K-means	97.36	97.54	96.33	93.21	0.912	0.901

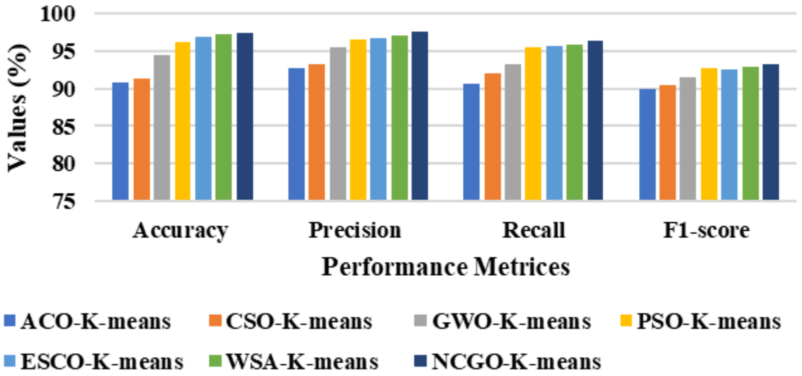

Figure. 5 Graphical representation of clustering results using BBC sports dataset

Table 5. Comparative Analysis

Datasets	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ARI	NMI
20-Newsgroup	SSA-SVM [15]	97.38	92.14	97.51	89.21	N/A	N/A
	Modsup + Rn-MSA [16]	94.38	95.90	94.37	95.57	N/A	N/A
	String Kernels [17]	85.72	N/A	N/A	N/A	N/A	N/A
	GDEM [19]	79.12	N/A	N/A	N/A	0.655	0.724
	Proposed CNGO - K-means	98.36	96.44	98.21	96.36	0.745	0.845
Reuters-21578	Modsup + Rn-MSA [16]	95.12	95.61	96.41	96.41	N/A	N/A
	String Kernels [17]	95.04	N/A	N/A	N/A	N/A	N/A
	GDEM [19]	39.24	N/A	N/A	N/A	0.308	0.537
	Proposed CNGO - K-means	96.26	96.11	98.45	97.98	0.495	0.634
BBC sport	String Kernels [17]	96.33	N/A	N/A	N/A	N/A	N/A
	GDEM [19]	96.02	N/A	N/A	N/A	0.907	0.878
	Proposed CNGO -K-means clustering	97.36	97.54	96.33	93.21	0.912	0.901

clustering approach with the proposed NCGO-K-means in the statistical measures. The proposed NCGO-K-means attains the accuracy, precision, recall, F1-score, ARI and NMI are 97.36%, 97.54%, 96.33%, 93.21%, 0.912 and 0.901 respectively. The proposed method attains effective outcomes when compared to other competitor approaches.

4.3 Comparative Analysis

Table 5 represents the comparative analysis of the proposed NCGO with K-means clustering by estimating the performance of various comparative techniques such as [15], [16], [17] and [19]. The comparative analysis is performed by the results are estimated by accuracy, precision, recall and F1-score.

4.4 Discussion

This research considers the detailed TDC by using the various assessment metrices with three benchmark datasets. The proposed CNGO with K-means clustering approach outclasses the baseline as well as competitor approaches are analyzed through experimentation. The hyperparameter selection values play a significant role in the effectiveness of Machine Learning (ML) techniques. By utilizing the utilization of the entire parameter search, the hyperparameter selection is computationally impracticable. Moreover, an initialization of arbitrary hyperparameter selection resulted in the usual effectiveness. Furthermore, the hyperparameter selection is dependent on the collected datasets. In comparison analysis, the proposed CNGO with K-means clustering approach is estimated using three benchmark datasets such as 20-Newsgroups, Reuters-21578 and BBC sport datasets to achieve better values for the metrices. The proposed method achieved an accuracy of 98.36%, 96. 26% and 97.36% respectively by using 20-Newsgroups, Reuters-21578 and BBC sports datasets.

5. Conclusion

In this research, the new metaheuristic algorithm named CNGO integrated with the K-means clustering algorithm is proposed for TDC. The CNGO efficiently clusters the text documents as well as eliminates the quality cluster degradation. The CNGO has minimized the local search because of the utilization of local exploitation. This research acquired the three benchmark datasets: 20-Newsgroups, Reuters-21578, and BBC Sport for estimating the proposed model’s effectiveness. In comparison results, the proposed CNGO with K-means clustering algorithm attains better accuracy results of 98.36%, 96. 26% and 97.36% respectively by using 20-Newsgroups, Reuters-21578, and BBC sports datasets as compared to the previous TDC methods such as SSA-SVM Modsup + Rn-MSA and GDEM. Furthermore, the proposed method shows better performance in overall metrics on all benchmark datasets. Moreover, the proposed model is computationally effective in terms of time, due to the optimal feature selection by using CNGO. In future work, the proposed work will extend to increase the global search ability and fast convergence speed in the optimization process.

Notation

Variables	Descriptions
$\hat{x}$	Normalized data
x	Actual data
x_{min} and x_{max}	Minimum and maximum of every feature.
χ^2	Chi-square score
s_{ij}	ith feature value along with instances.
μ_{ij}	Expected count
f	Features
c	Classes
s_{i*}	ith value of particular feature
s_{*j}	number of instances in class j
s	number of instances
$V_{i,G}$	Mutation vector
$X_{i,G}$	Target vector
$U_{i,G}$	Trial vector
$X_{a,G}$, $X_{b,G}$ and $X_{c,G}$	The vectors are arbitrarily chosen from the present population
$u^j_{i,G}$, $v^j_{i,G}$ and $x^j_{i,G}$	jth parameter in ith trial, mutant as well as target vector
$Rand_2$, $Rand_3$, $Rand_4$	independently developed arbitrary numbers in the range between 0 and 1 respectively
$CR_{i,G}$	Crossover mutation
$f(U_{i,G})$	Objective function
$X_{i,G+1}$	vector with minimal objective function survives into further generation
n	Number of layers in the coding network
h	Activation function
m_x and y_k	Input and output feature map
$w_{n,k}$	Kernel weight
$*$	Convolutional operation
b_k	Bias
$r_{i,E}$	Mapping feature
$e_{i:1+z-1}$	Sword embedding vector
$\sigma(.)$	Nonlinear activation function of convolutional operation
$\circ$	Hadamard product between two matrices
W	Weight
z	Softmax function input vector contains “n” features of “n” target values
zi	ith item of input vector
e^{zj}	Standard exponential function
$\sum e^{zj}$	Normalization term to acquire valid probability distribution
TP	True Positive
TN	True Negative
FP	False positive
FN	False Negative

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] P.S. More, and B.S. Saini, "KH-FC: krill herd-based fractional calculus algorithm for text document clustering using MapReduce structure", *International Journal of Computational Science and Engineering*, Vol. 25, No. 6, pp. 668-684, 2022.

[2] S. Selvaraj, and E. Choi, "Dynamic Sub-Swarm Approach of PSO Algorithms for Text Document Clustering", *Sensors*, Vol. 22, no. 24, p. 9653, 2022.

[3] M. Asif, A.A. Nagra, M.B. Ahmad, and K. Masood, "Feature selection empowered by self-inertia weight adaptive particle swarm optimization for text classification", *Applied Artificial Intelligence*, Vol. 36, No. 1, p. 2004345, 2022.

[4] S. Adinugroho, R.C. Wihandika, and P.P. Adikara, "Newsgroup topic extraction using term-cluster weighting and pillar K-means clustering", *International Journal of Computers and Applications*, Vol. 44, No. 4, pp. 357-364, 2022.

[5] K. Thirumoorthy, and J.J.J. Britto, "A feature selection model for document classification using Tom and Jerry Optimization algorithm", *Multimedia Tools and Applications*, Vol. 83, pp. 10273-10295, 2023.

[6] F. Malik, S. Khan, A. Rizwan, G. Atteia, and N.A. Samee, "A Novel Hybrid Clustering Approach Based on Black Hole Algorithm for Document Clustering", *IEEE Access*, Vol. 10, pp. 97310-97326, 2022.

[7] M.I. Nadeem, K. Ahmed, D. Li, Z. Zheng, H. Naheed, A.Y. Muaad, A. Alqarafi, and H. Abdel Hameed, "SHO-CNN: A metaheuristic optimization of a convolutional neural network for multi-label news classification", *Electronics*, Vol. 12, No. 1, p. 113, 2022.

[8] B. Diallo, J. Hu, T. Li, G.A. Khan, and A.S. Hussein, "Multi-view document clustering based on geometrical similarity measurement", *International Journal of Machine Learning and Cybernetics*, Vol. 13, pp. 663-675, 2022.

[9] E.K. Jasila, N. Saleena, and K.A. Abdul Nazeer, "An Efficient Document Clustering Approach for Devising Semantic Clusters", *Cybernetics and Systems*, 2023.

[10] I.H. Hassan, A. Mohammed, Y.S. Ali, I. Jeremiah, and S.A. Abdulraheem, "Metaheuristic algorithms in text clustering," *Comprehensive Metaheuristics*, pp. 131-152, 2023.

[11] G. Chandwani, A. Ahlawat, and G. Dubey, "An approach for document retrieval using cluster-based inverted indexing", *Journal of Information Science*, Vol. 49, No. 3, pp. 726-739, 2023.

[12] M. Shahroz, M.F. Mushtaq, R. Majeed, A. Samad, Z. Mushtaq, and U. Akram, "Feature discrimination of news based on canopy and KMGC-search clustering", *IEEE Access*, Vol. 10, pp. 26307-26319, 2022.

[13] D. Mustafi, A. Mustafi, and G. Sahoo, "A novel approach to text clustering using genetic algorithm based on the nearest neighbour heuristic", *International Journal of Computers and Applications*, Vol. 44, No. 3, pp. 291-303, 2022.

[14] L. Huang, P. Shi, H. Zhu, and T. Chen, "Early detection of emergency events from social media: A new text clustering approach", *Natural Hazards*, Vol. 111, No. 1, pp. 851-875, 2022.

[15] M. Ponnusamy, P. Bedi, T. Suresh, A. Alagarsamy, R. Manikandan, and N. Yuvaraj, "Design and analysis of text document clustering using salp swarm algorithm", *The Journal of Supercomputing*, Vol. 78, No. 14, pp. 16197-16213, 2022.

[16] M. Yarlagadda, K.G. Rao, and A. Srikrishna, "Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering", *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, No. 4, pp. 1098-1109, 2022.

[17] N.V. Chandran, V.S. Anoop, and S. Asharaf, "Topicstriker: A topic kernels-powered approach for text classification", *Results in Engineering*, Vol. 17, p. 100949, 2023.

[18] K. Berahmand, F. Daneshfar, M. Dorosti, and M.J. Aghajani, "An Improved Deep Text Clustering via Local Manifold of an Autoencoder Embedding", 2022.

[19] S. Jung, and S. Ka, "GAE-Based Document Embedding Method for Clustering", *IEEE Access*, Vol. 10, pp. 130089-130096, 2022.

[20] P.D. Kusuma, and A. Dinimaharawati, "Extended stochastic coati optimizer", *International Journal of Intelligent Engineering and Systems*, Vol. 16, No. 3, pp. 482-494, 2023, doi: 10.22266/ijies2023.0630.38.

[21] P.D. Kusuma, and A.L. Prasasti, "Walk-Spread Algorithm: A Fast and Superior Stochastic Optimization", *International Journal of Intelligent Engineering & Systems*, Vol. 16, No.

5, pp. 275-288, 2023, doi: 10.22266/ijies2023.1031.24.

[22] Reuters-21578 dataset link: https://www.kaggle.com/datasets/thedevastator/uncovering-financial-insights-with-the-reuters-2.

[23] 20Newsgroup dataset link: https://www.kaggle.com/datasets/crawford/20-newsgroups.

[24] BBC sports database link: https://www.kaggle.com/datasets/maneesh99/sports-datasetbbc.

[25] N. Garg, and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data", *International Journal of Electrical & Computer Engineering (2088-8708)*, Vol. 12, No. 1, pp. 776-784, 2022.

[26] J.C. Costa, T. Roxo, J.B. Sequeiros, H. Proenca, and P.R. Inacio, "Predicting cvss metric via description interpretation", *IEEE Access*, Vol. 10, pp. 59125-59134, 2022.

[27] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification", *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, No. 8, pp. 5059-5074, 2022.