



## Text Clustering Algorithm Using Fuzzy Whale Optimization Algorithm

Jagatheeshkumar Gopal<sup>1\*</sup>      Selva Brunda<sup>2</sup>

<sup>1</sup>*Research and Development Centre, Bharathiar University, Coimbatore, India*

<sup>2</sup>*Department of Computer Science and Engineering, Cheran College of Engineering, Karur, India*

\* Corresponding author's Email: jagatheeshkumar.bu@gmail.com

---

**Abstract:** Text clustering is one of the most researchable topics, owing to the massive usage of textual documents. The aim of text clustering algorithms is to group similar textual documents together, which can improve the data organization and makes the process of data analysis simpler. Understanding the benefits, this article intends to present a text clustering algorithm that is based on Fuzzy C Means (FCM) and Whale Optimization Algorithm (WOA). The optimization algorithm (WOA) selects the cluster centre, which helps the FCM to arrive at better textual clusters. The performance of the proposed text clustering algorithm is tested in terms of precision, recall, f-measure, purity and entropy over three different benchmark datasets. The performance of the proposed algorithm is observed to be reasonable when compared to the existing algorithms with an average F-measure of 97.6%.

**Keywords:** Text clustering, FCM, WOA, Optimization algorithm.

---

### 1. Introduction

Data is the quintessential part of all the live transactions. These data are beneficial and are stored for future analysis. Initially, the transactional data are stored in memory for future references, however the data analysis is not simple in this case. The main reasons for the difficult data analysis are the inefficient data storage and retrieval. These drawbacks are addressed rapidly by the modernized storage techniques. Due to the advancement of storage techniques and technology, the knowledge based systems are developed for achieving efficient data storage and retrieval.

The knowledge based systems work intelligently by means of the incorporated techniques. A knowledge based system can achieve two basic tasks, which are clustering and classification. The process of clustering attempts to group similar data being processed and the process of classification aim to distinguish between the entities in the dataset. The clustering and classification activities are called as unsupervised and supervised learning techniques respectively. The process of classification is termed as supervised learning, as the classification system

necessitates the training phase. The objective of training phase is to equip the classifier with the knowledge extracted from the training samples. The classifier is then suitable for classifying between the entities.

On the other hand, the process of clustering does not require any training phase and it works on the go. Hence, the clustering operation is termed as unsupervised and it works without any prior activities. As the related data items are grouped together, the data can be analysed without any hassles. Understanding the advantages of data clustering and the prevalent usage of textual data, this work sets its goal to present a textual data clustering algorithm by combining the Fuzzy C Means (FCM) and Whale Optimization Algorithm (WOA) presented in [1]. FCM algorithm is closely associated with k-means algorithm but is based on the fuzzy theory. Several improvised versions of FCM are proposed in the literature for attaining better clustering results.

The traditional FCM algorithm fixes the count of clusters before the start of the clustering process [2]. In addition to this, a clustering rule that implies the data item with a specific dimension has to be placed

in a specific cluster is fixed. This process can be made better by clubbing the FCM with the global optimization algorithm, such that the clusters are optimized.

With this goal, this article clubs FCM with WOA and the proposed clustering algorithm pre-processing, data similarity calculation and data clustering. The pre-processing step weeds out the needless data from the dataset, in order to reduce the processing time and the computational complexity of clustering algorithm. The data similarity score computes the relationship among the data and the clustering algorithm groups the relevant data together. Some of the work highlights are as follows.

- This work presents optimal clusters and the quality of clusters is proven by analysing the performance of the proposed clustering algorithm with the standard performance metrics such as precision, recall, F-measure, purity and entropy.
- The performance of this work is justified by testing the clustering algorithm over three different textual datasets.
- The combination of FCM and WOA is better with reasonable cluster accuracy rates.

The remainder of this article is organized as follows. Section 2 presents the review of literature with respect to textual clustering and the proposed clustering algorithm is described in section 3. The performance of the proposed clustering algorithm is analysed in section 4 and the conclusions of this work are summarized in section 5.

## 2. Review of literature

This section reviews the state-of-the-art clustering techniques, which are the motivational force of this work.

In [3], a semi-supervised heterogenous fusion based multimedia data co-clustering technique is proposed. This work is meant for web metadata that involves annotation or description. This work employs the generalized version of Heterogenous Fusion Adaptive Resonance Theory (GHF-ART) for performing the web documents co-clustering. This algorithm is proven to be faster than the existing algorithms, yet this algorithm involves computational complexity.

A document clustering technique is proposed for performing forensic analysis in [4]. This work considers the text clustering algorithms such as k-means, k-medoids, single link, complete link, average link and so on. The conclusion of this work is that the performances of average and complete link algorithms are better than the rest of the

algorithms. A fuzzy relational clustering algorithm is proposed in [5], which is meant for clustering text based on sentence level. This work processes the text data, which is organized in the pattern of square matrix with pairwise similarity. The data is represented in graphs and the expectation-maximization framework is utilised. The time complexity of this work is greater.

In [6], a clustering technique based on evolutionary computing is proposed for Brazilian Portuguese text. This work involves two modules, in which the first module is meant for extracting the important terms followed by indexing them. The second module utilizes the so computed index to group the text under a specific topic. However, the time complexity of this work is more. In [7], a technique to detect and recognize text on videos is proposed. This work is based on corner response feature map and transferred deep convolutional neural networks. The corner response feature map is utilized detecting the text and are segregated into text lines. The classification networks are then established and the text layer is obtained by FCM. However, the clustering performance of the work can still be enhanced.

A Summation bAsed Incremental Learning (SAIL) algorithm is proposed for information-theoretic text clustering in [8]. This algorithm employs an objective function that computes KL divergence by Shannon entropy. The quality of clusters is further improved by variable neighbourhood search scheme accelerated by V-sail algorithm. This work enhances the quality of clusters but suffers from computational complexity. In [9], the multi-topic structures are identified by employing word association in the textual documents. The multi-topic structures are identified by associative gravity, which works on the basis of word association by humans. This work claims itself with better accuracy rates, yet the computational complexity of this work is greater.

A multi-label text categorization technique based on fuzzy relevance clustering is proposed in [10]. This work employs a fuzzy relevance measure for transforming the high dimensional documents to the low dimensional fuzzy relevance vectors. The clustering approach then partitions the relevance space into several regions. This work consumes more time for clustering and involves computational complexity as well. In [11], an efficient semi-supervised MEDLINE document clustering is proposed with the help of Medical Subject Heading (MeSH) semantic and global content constraints. This work considers local, global and MeSH semantic information. This work clusters the content

by considering must-link and cannot-link constraints. The must-link constraint works on documents with greater MeSH semantic information and the cannot-link constraint is meant for information with minimal similarity.

In [12], the information clustering technique with manifold based optimization of Bag-of-Features (BoF) is proposed. This work utilizes the data manifold and the clustering results are formed. A histogram space is built for constructing the dictionary, in which the BoF histograms are formed. This work does not utilize the spectral space, but allots the data item directly to the optimized histogram space.

Certain other clustering algorithms employ evolutionary algorithms for performing clustering. The performance of FCM is enhanced by several optimization algorithms, so as to support the FCM from convergence at local minima. In [13], the FCM is combined with Artificial Bee Colony (ABC) optimization algorithm for data clustering. In [14], the data clustering is performed by combining FCM and invasive weed optimization algorithm. The investigations on bio-inspired algorithms are unstoppable and certain combinations of algorithm prove better performance.

Motivated by these works, this article attempts to combine FCM with WOA for attaining better quality clusters. The detailed description of the proposed work is presented in the following section.

### 3. Proposed text clustering algorithm

The proposed text clustering algorithm is elaborated along with the overview of the algorithm in this section.

#### 3.1 Outline of the work

This article presents a clustering algorithm, which is meant for textual entities. The objective of

this work is to ensure better quality of text clusters and the goal is achieved by combining FCM and WOA. The proposed clustering algorithm enhances the performance of FCM by employing WOA algorithm. WOA imitates the behaviour of whales and the merits of this algorithm are limited control parameters. As the parameters are limited, the algorithm is easy to implement and the flexibility is better.

The clustering algorithm involves three important phases as mentioned earlier, which are pre-processing, text similarity computation and clustering. The pre-processing activity aims to remove meaningless data from the dataset, which speeds up the process of clustering. The text similarity computation is attained by cosine similarity measure and the final clustering is performed by FCM-WOA algorithm. The WOA supports the FCM to perform better and the proposed approach yields better results. The following section presents the background of the proposed text clustering algorithm.

#### 3.2 Outline of the work

The underlying pillars of the proposed work, FCM and WOA algorithms are summarized in this subsection.

##### 3.2.1. FCM algorithm

FCM algorithm collects similar data items by considering the similarity between the data items. In FCM, a single data entity can be part of one or more clusters and the relationship between the data entities is computed by means of the membership function. The standard FCM algorithm is presented as follows.

---

#### FCM Algorithm

---

Input – Count of clusters, initial cluster centre, termination criterion

Output – Text clusters

Begin

Fix the objective function as by  $F_n = \sum_{i=1}^r \sum_{j=1}^c \mu_{ij}^f \left\| x_i - y_j \right\|^2$ ;

Compute the centre point of the cluster by  $y_j = \frac{\sum_{i=1}^r \mu_{ij}^{ff} x_i}{\sum_{i=1}^r \mu_{ij}^{ff}}$

Compute fuzzy membership by  $\mu_{ij} = \frac{1}{\sum_{b=1}^c \left( \frac{\left\| x_i - y_b \right\|^2}{\left\| x_i - y_j \right\|^2} \right)^{\frac{2}{f-1}}}$

End;

---

In the above algorithm,  $F_n$  is the objective function,  $ff$  is the fuzzy factor and  $\mu$  is the fuzzy membership value.  $x_i$  is the data item and  $y_j$  is the centre point of the  $j^{th}$  cluster.  $\|x_i - y_j\|^2$  is the Euclidean distance between the data entities, which is computed by

$$\|x_i - y_j\|^2 = \sqrt{\sum_{i=1}^r ((x_i - y_j)^2)}^2 \quad (1)$$

This process of clustering is performed until the termination criterion is satisfied. The following section presents the overview of WOA.

### 3.2.2. WOA algorithm

WOA is an optimization algorithm that imitates the whales [1]. By nature, the whales attack a shoal of fish that swim around the whale in the scope of a circle by forming bubbles. The shoal of fish is enclosed by the bubbles and the attack is performed in two phases. They are exploitation and exploration phases. The exploitation phase is meant for enclosing the prey within a circle and the exploration phase searches the fish in a random fashion. The behaviour of whales is modelled as follows.

$$K = |CV_1 \cdot \vec{A}^*(i) - \vec{A}(i)| \quad (2)$$

$$\vec{A}(i+1) = \vec{A}^*(i) - \vec{CV}_2 \cdot K \quad (3)$$

In the above equations,  $i$  denotes the current iteration,  $A^*$  is the best result attained and the position vector is represented by  $A$ . The absolute value is represented by  $\|$  and dot product ( $\cdot$ ) is performed for all elements.  $CV_1$  and  $CV_2$  are the coefficient vectors, which are calculated as follows.

$$\vec{CV}_2 = 2\vec{cv}_2 \cdot \vec{rv} - \vec{cv}_2 \quad (4)$$

$$\vec{CV}_1 = 2 \cdot \vec{rv} \quad (5)$$

In the above equations, the value of  $\vec{cv}_2$  decreases with iterations and  $\vec{rv}$  is the random vector produced uniformly at the range of [0,1]. The whales change their position by considering the position of food based on the Eq. (3) and the position is managed by  $\vec{CV}_1$  and  $\vec{CV}_2$  respectively. The food of whale is enclosed by the circle by minimizing the value of  $\vec{cv}_2$ , as presented in Eq. (6).

$$cv = 2 - i \frac{2}{Max_i} \quad (6)$$

In Eq. (6),  $i$  is the count of iteration and  $Max_i$  is the maximum count of iterations. The position of the nearby whale is computed for computing the distance between the whale  $x$  and whale  $y$  by the following equation.

$$\vec{A}(i+1) = K' \cdot dis^{cr} \cdot \cos(2\pi r) + \vec{A} \times (i) \quad (7)$$

In the above equation,  $K' = |\vec{A}^*(i) - \vec{A}(t)|$  represents the distance between the  $n^{th}$  whale and the best food source detected,  $c$  is the constant which indicates the shape of the curve and  $r$  is the random value between -1 and 1. The food source enclosure and the path determination are represented with a probability ( $pr$ ) of 0.5, which is presented as follows.

$$\vec{A}(i+1) = \begin{cases} \text{Enclosing food source eqn. (3) when } pr < 0.5 \\ \text{path determination eqn. (7) when } pr \geq 0.5 \end{cases} \quad (8)$$

In the above equation,  $pr$  is the random value ranging from 0 to 1.

In the exploration phase, the whales are chosen in a random fashion to perform better search of the food source. The vector  $\vec{CV}_1$  with random numbers aim to search for the best known neighbourhood whale as represented in the following equations.

$$\vec{K} = |\vec{CV}_1 \cdot \vec{A}_r - \vec{A}| \quad (9)$$

$$\vec{A}(i+1) = \vec{A}_r - \vec{CV}_2 \cdot \vec{K} \quad (10)$$

$\vec{A}_r$  is the whale being chosen in a random fashion.

Hence, the WOA algorithm is presented and the proposed text clustering algorithm is presented in the following sub-section.

### 3.2.3. Proposed clustering algorithm

The proposed text clustering algorithm involves three significant phases, which are data pre-processing, similarity computation and text clustering. This work is related to the text clustering algorithm proposed in [15] and so, the data pre-processing and the similarity computation is utilized as such. This section presents the proposed text clustering algorithm alone and the intention of the

proposed algorithm is to improvise the performance of FCM algorithm by introducing WOA into it.

The optimization algorithm WOA supports in improving the performance of the FCM clustering algorithm. This proposed FCMWOA algorithm clusters the data items by fuzzy approach. Initially, the cluster centre is identified by means of the solution obtained from the WOA and the FCM algorithm is executed by modifying the cluster centres and the process is terminated after reaching the termination criterion. The fitness function of the proposed algorithm is presented as follows.

$$Fn = \sum_{i=1}^r \sum_{j=1}^c \mu_{ij}^{ff} \|x_i - y_j\|^2 \quad (11)$$

In the above equation,  $ff$  is the fuzzy factor and  $\mu$  is the fuzzy membership value.  $x_i$  is the data item and  $y_j$  is the centre point of the  $j^{th}$  cluster.  $X = (x_1, x_2, \dots, x_c)$  is the total count of features of a data item and  $Y = (y_1, y_2, \dots, y_r)$  is the cluster centre points. The outcome is considered as a whale in the algorithm. A single solution may present in one or

more clusters. Hence, the cluster centres of the clusters ( $r$ ) can have  $w$  features, which means that the length of the optimization entity can be  $r \times w$ . This can be represented by the following Eq. (12).

$$Z_i = (z_{i,1}, z_{i,2}, \dots, z_{i,w}, z_{i,r \times (w-1)+1}, z_{i,r \times (w-1)+2}, \dots, z_{i,r \times w}) \quad (12)$$

In Eq. (12), the first appearing  $w$  of  $Z_i$  vector is meant for the first cluster centre and the second one is for the second cluster centre and this gets incremented. This process makes the algorithm to converge quicker, yet obtains the better global optimal fitness. This leads to better quality clusters and the process is continued for all the iterations until the termination condition is met. When the termination condition is met, the best solution is obtained as the cluster centre point and the  $\mu_{i,j}$  is computed for the cluster centre. The proposed clustering algorithm is presented as follows.

---

#### *Proposed FCMWOA algorithm*

---

*Input:* Textual data ( $d$ )

*Output:* Text clusters

*Begin*

*For all d*

*Pre-process the data;*

*Initialize the population of whales and other parameters;*

$A^*$  = Best whale

*Distribute the prey;*

*For each whale*

*Compute the fitness of all the solutions by eqn.11;*

*While* ( $i < \max_i$ )

*For each solution*

*Upgrade*  $cv_2, CV_2, CV_1, r, pr$

*if* ( $pr < 0.5 \&\& |CV_2| < 1$ )

*Modify the location of the solution by eqn.3;*

*if* ( $pr < 0.5 \&\& |CV_2| > 1$ )

*Choose a whale in a random fashion;*

*Modify the location of whale by eqn.10;*

*End if;*

*if* ( $p \geq 0.5$ )

*Modify the location of the search by eqn. 7;*

*End if;*

*End for;*

*Verify the search space of the solution;*

*Alter the search space;*

---

---

```

Modify A* if better solution i= i + 1
End while;
Return A*;
End;
End While;

```

---

The idea of this work is that the traditional FCM computes the cluster centre points by calculating  $\mu_{ij}$  in the first step. This  $\mu_{ij}$  is fixed in a random fashion in the initial step. However, the proposed clustering algorithm the optimal cluster centres obtained from WOA are fed to FCM, which results in better clusters. The performance of the proposed algorithm is evaluated in the following section.

#### 4. Performance evaluation

The efficiency of the proposed clustering algorithm is implemented in **Java** and assessed with respect to precision, recall, F-measure, purity and entropy over three different datasets that are present in [16-18]. The benchmark datasets being utilized are as webkb, re0 and 20Newsgroups. The total iteration of this work is 100 and the initial population is set as 50.

The performance of the proposed approach is measured in terms of precision ( $P_r$ ), recall ( $\mathbb{R}$ ), purity (P), entropy (E) and F-measure (F). All the stated performance metrics are based on the true positive( $T_p$ ), true negative ( $T_n$ ), false positive ( $F_p$ ), and false negative ( $F_n$ ) rates.  $T_p$  is the rate that is computed by the correctly clustered documents to the entire number of documents available in the dataset.  $T_n$  rates are the documents that are correctly distinguished as they are not the candidates of the particular cluster.  $F_p$  rate is the fraction of the documents that are wrongly included as a candidate of wrong cluster. Likewise,  $F_n$  is computed by considering the documents that are wrongly clustered, where that particular document a specific class is declared not as a candidate of the class.

Based on the values of  $T_p$ ,  $T_n$ ,  $F_p$ , and  $F_n$ , the metrics such as  $P_r$ ,  $\mathbb{R}$ , P, E and F are computed. The clustering algorithm can attain maximum precision value, when the  $F_p$  values are minimal. The  $F_p$  and  $F_n$  rates must be as minimal as possible, such that the precision and recall rates can be improved. A good clustering algorithm should prove maximum precision and recall rates.

$$P_r = \frac{T_p}{T_p + F_p} \quad (13)$$

$$\mathbb{R} = \frac{T_p}{T_p + F_n} \quad (14)$$

$$F = \frac{2(P_r \times \mathbb{R})}{P_r + \mathbb{R}} \quad (15)$$

The purity of a cluster must be greater such that it is ensured that the quality is cluster is better. The cluster purity takes the cluster size and the candidate documents in the cluster into account. The purity of a cluster is computed by

$$P(cl_i) = \frac{1}{ndc_i} \max_c ndc_i^c \quad (16)$$

In the above equation,  $\max_c ndc_i^c$  is the overall count of textual data that belongs to a particular class  $ndc_i$ .  $ndc_i^c$  is the count of textual documents of a cluster  $ndc_i$ . Entropy is a metric that computes the overall document distribution with respect to the class in a cluster and is measured by

$$E(cl_i) = -\frac{1}{\log cd} \sum_{c=1}^m \frac{ndc_i^c}{ndc_i} \log \left( \frac{ndc_i^c}{ndc_i} \right) \quad (17)$$

$$E = \sum_{c=1}^m \frac{ndc_i^c}{ndc_i} E(cl_i) \quad (18)$$

In Eq. (17),  $cd$  is the count of classes being present in the dataset.  $ndc_i^c$  is the number of textual documents in the class  $c$ , which is under the cluster  $cl_i$ . Eq. (18) represents the overall entropy value of all the clusters. The lesser the entropy, the better is the quality of the clusters. In order to justify the combination of FCM and WOA, both the algorithms are analysed separately and the results are as follows.

##### 4.1 Performance analysis w.r.t clustering algorithms

Initially, the performance of the proposed clustering algorithm is justified by analysing the performance of the FCM and WOA algorithms separately followed by its combination. The optimal cluster centres are computed by WOA and the FCM employs it to attain better quality clusters. The experimental results by varying the clustering algorithms are presented as follows.

Dataset/Perf metrics	Precision(%)			Recall (%)			F-measure (%)		
	FCM	WOA	FCM WOA	FCM	WOA	FCM WOA	FCM	WOA	FCM WOA
<b>Webkb</b>	81.2	86.8	<b>98.2</b>	82.3	84.9	<b>97.6</b>	81.7	85.8	<b>97.8</b>
<b>Re0</b>	82.3	88.1	<b>96.3</b>	83.4	86.7	<b>98.4</b>	82.8	87.3	<b>97.3</b>
<b>20Newsgroup</b>	84.9	87.9	<b>97.3</b>	81.7	85.5	<b>98.2</b>	83.2	86.6	<b>97.7</b>
<b>Average</b>	82.8	87.6	<b>97.26</b>	82.4	85.7	<b>98.06</b>	82.56	86.5	<b>97.6</b>

Dataset/Perf metrics	Purity (%)			Entropy (%)		
	FCM	WOA	FCM WOA	FCM	WOA	FCM WOA
<b>Webkb</b>	82.7	87.9	<b>97.6</b>	35.4	26.3	<b>9.6</b>
<b>Re0</b>	81.4	86.3	<b>98.4</b>	33.9	24.7	<b>10.2</b>
<b>20Newsgroup</b>	83.2	88.4	<b>96.7</b>	30.2	23.2	<b>11.8</b>
<b>Average</b>	82.4	87.5	<b>97.5</b>	33.1	24.7	<b>10.5</b>

From the experimental results, it is clearly evident that the performance of the combination of FCM and WOA improves the quality of cluster with better precision and recall rates. Better precision and recall rates, obviously result in better F-measure rates. Apart from this, the proposed clustering algorithm shows reasonable purity and entropy rates. Though the algorithms cannot perform better individually, the performance is proven when the algorithms are combined. Hence, the reason of clubbing these algorithms is justified.

#### 4.2 Performance comparison of FCMWOA with the existing algorithms

The performance of the proposed algorithm is analysed by considering two state-of-the-art clustering algorithms such as presented in [14, 15]. The performance analysis is carried out with the help of the same performance metrics and datasets and the experimental results are presented in Figs. 1 and 2.

The presented results of Figs. 1 and 2 are meant for webkb and Re0 datasets. The description of all the three datasets is presented in [15]. The proposed algorithm proves its efficiency, when compared to the other two existing algorithms and the main

reason for better performance is the cluster centre choice by WOA. Though k-means algorithm performs well, the inclusion of fuzzy concept shoots up the efficiency of the algorithm. The following Fig. 3 shows the experimental results of the proposed approach over 20NewsGroup dataset.

The main reason for the better performance of the proposed approach is that the cluster centres are optimized before the cluster centres are treated by FCM. This reduces the computational complexity, while increasing the efficiency of the clustering algorithm. On the other hand, the combination of FCM and PSO does not work well, while following the same strategy. WOA performs better than PSO with minimal time consumption. On the other hand, kLoA follows the same strategy as that of the proposed approach. However, FCM performs better than k-means algorithm, as the FCM is based on the fuzzy membership and allows the data item to be a part of more than a cluster. This concept improves the clustering performance of the algorithm.

From the experimental results, it is clearly evident that the performance of the proposed clustering algorithm is better in terms of precision, recall, f-measure, purity and entropy. A better clustering algorithm has to prove maximal precision,

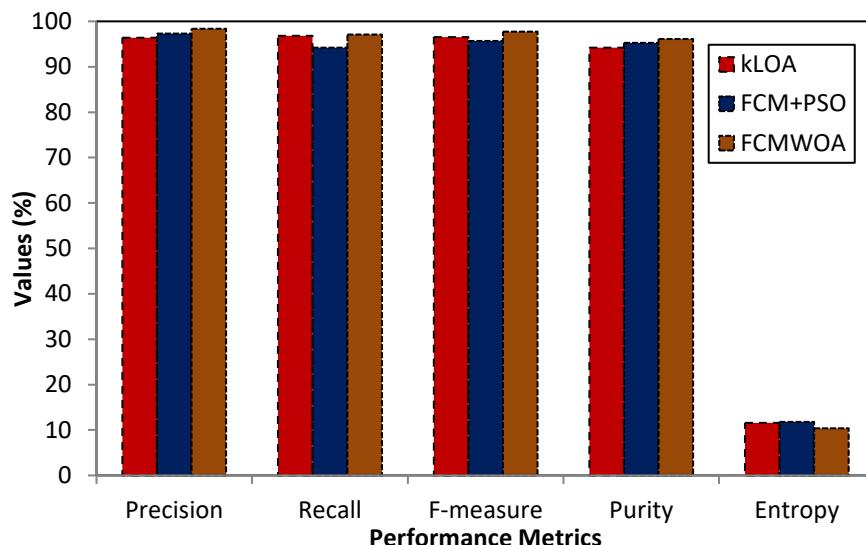


Figure. 1 Performance results on webkb dataset

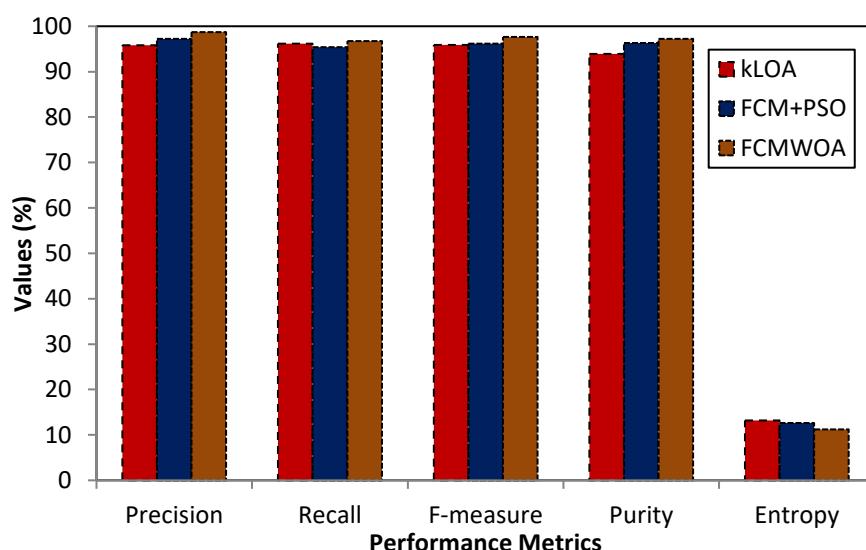


Figure. 2 Performance results on Re0 dataset

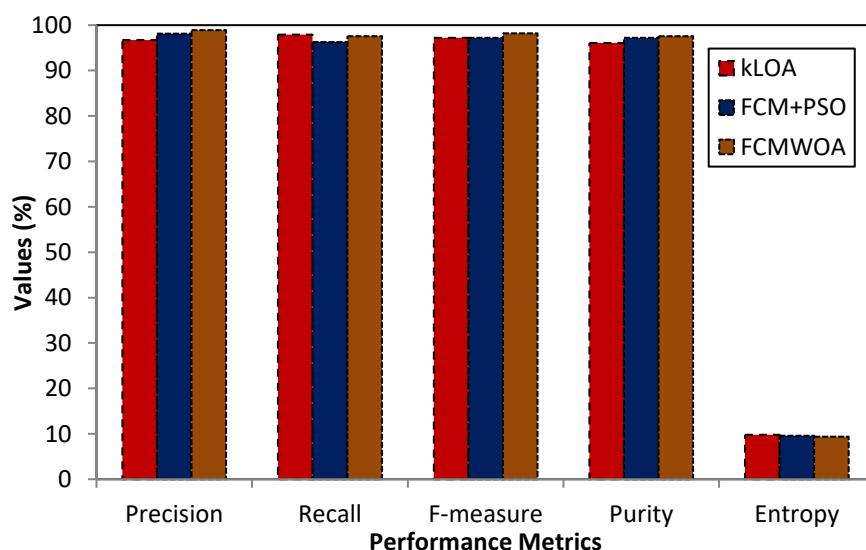


Figure. 3 Performance results on 20NewsGroup dataset

recall, f-measure, purity and minimal entropy. The proposed clustering algorithm satisfies all the conditions and hence, the objective of this work is attained.

## 5. Conclusion

This article presents a text clustering algorithm that relies on FCM and WOA. FCM is a standard and promising clustering algorithm still today. However, the efficiency of FCM can be improvised even more, when it is combined with the optimization algorithms. Several existing works have attempted to club FCM with optimization algorithms to enhance the quality of clusters. Following that path, this work intends to present a text clustering algorithm that combines FCM with the recent optimization algorithm WOA. The cluster centres are chosen by the WOA and passed to FCM, which increases the performance of the clustering algorithm. As the optimal cluster centre points are chosen, the clusters are of better quality. The performance of the proposed clustering algorithm is compared with kLoA and FCM+PSO. In future, this work can be extended by including semantic based analytics for clustering documents.

## References

- [1] S. Mirjalili and A. Lewis, "The whale optimization algorithm", *Advances in Engineering Software*, Vol. 95, pp. 51–67, 2016.
- [2] J.C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm", *Computers & Geosciences*, Vol.10, No.2-3, pp.191-203, 1984.
- [3] L. Meng, A. Tan, and D. Xu, "Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol.26, No.9, pp. 2293-2306, 2014.
- [4] L. F.C.Nassif and E.R. Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", *IEEE Transactions on Information Forensics and Security*, Vol. 8, No.1, pp.46-54, 2013.
- [5] A. Skabar and K. Abdalgader, "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", *IEEE Transactions on Knowledge and Data Engineering*, Vol.25, No.1, pp.62-75, 2013.
- [6] A. R. Afonso, "Brazilian Portuguese Text Clustering Based on Evolutionary Computing", *IEEE Latin America Transactions*, Vol.14, No.1, pp. 3370-3377, 2016.
- [7] W. Lu, H. Sun, J. Chu, X. Huang, and J. Yu, "A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network", *IEEE Access*, Vol.6, pp.40198-40211, 2018.
- [8] J. Cao, Z. Wu, J. Wu, and H. Xiong, "SAIL: Summation-bAsed Incremental Learning for Information-Theoretic Text Clustering", *IEEE Transactions on Cybernetics*, Vol. 43, No.2, pp.570-584, 2013.
- [9] A. Klahold, P. Uhr, F. Ansari, and M. Fathi, "Using Word Association to Detect Multitopic Structures in Text Documents", *IEEE Intelligent Systems*, Vol.29, No.5, pp. 40-46, 2014.
- [10] S. Lee and J. Jiang, "Multilabel Text Categorization Based on Fuzzy Relevance Clustering", *IEEE Transactions on Fuzzy Systems*, Vol.22, No.6, pp. 1457-1471, 2014.
- [11] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu, "Efficient Semisupervised MEDLINE Document Clustering with MeSH-Semantic and Global-Content Constraints", *IEEE Transactions on Cybernetics*, Vol.43, No.4, pp. 1265-1276, 2013.
- [12] N. Passalis and A. Tefas, "Information Clustering Using Manifold-Based Optimization of the Bag-of-Features Representation", *IEEE Transactions on Cybernetics*, Vol.48, No.1, pp. 52-63, 2018.
- [13] D. Karaboga and C. Ozturk, "Fuzzy clustering with artificial bee colony algorithm", *Scientific Research and Essays*, Vol.5, No.14, pp.1899–1902, 2010.
- [14] X. Zhu, W. Pedrycz, and Z. Li, "Fuzzy clustering with nonlinearly transformed data", *Applied Soft Computing*, Vol. 61, pp. 364-376, 2017.
- [15] G. Jagatheeeshkumar and S. Selva Brunda, "Lion Optimization Algorithm based k-Means for Textual Data Clustering", *International Journal of Pure and Applied Mathematics*, Vol.117, No.22, pp.167-171, 2017.
- [16] <http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.gtar.gz>
- [17] <http://www.daviddlewis.com/resources/testcollections/reuters21578>.
- [18] <http://qwone.com/~jason/20Newsgroups/>