ENHANCED LION OPTIMIZATION ALGORITHM BASED ON FUZZY C MEAN FOR TEXTUAL DATA CLUSTERING ALGORITHMS

Dr. Gopal Jagatheeshkumar[1], Dr. S. Selva Brunda[2]

[1]Assistant Professor, Department of Computer Science, KSG College of Arts and Science, Coimbatore.
[2]Professor, Department of CSE, Cheran Engineering College, Karur.

Abstract: *Data is a vital component of the world activity. The dynamic representation of data is wanted to organizes and retrieve most effectively. Information is needed to format some meaningful format then only retrieving of information is knowable. Text mining is concept also called as text analysis process which extracts the related, non related information and knowledge from the unstructured text data. Applications of clustering environment are retrieval, data analysis, statistics, machine learning and linguistics. Similar group of related data item joined together is called clustering. Clustering is evergreen research area in text mining under the data mining domain. The partition cluster analysis Fuzzy C Mean algorithm Combined with Lion Optimization Algorithm found that Enhanced Lion Optimization Algorithm based on Fuzzy C Mean for Textual Data Clustering Algorithms (ELFCM). FCM is better clustering algorithm, some time it will take more time for pick the initial cluster centre point. It may be take more iteration for selecting centre point of cluster. For clustering operation FCM is very best. So overcome this issues pick the initial point from LOA and cluster operation performed by FMC. The proposal shows that very clearly for ELFCM is better clustering algorithm from others. Implement this algorithm in Java and compare with some existing algorithm utilize three different Dataset.*

Keywords: *Information retrieval, Text mining, clustering, Optimization algorithm, Dataset*

1. INTRODUCTION

Text mining encompasses many text processing and classification techniques, such as text categorization, clustering and retrieval and extraction of information. Traditional information management methods are woefully inadequate for the enormous amount of text data. In general, only a very small percentage of easily available document will be applicable to a specific individual of user. And without understanding of what has been in the documents it is difficult to formulate effective requests to analysis and extract useful information from the data. The users need to apply techniques to distinguish between different documents and categorize the importance and significance of documents or find patterns.

Thus, the text mining approach has become important in recent days. Clustering is an unsupervised learning without any prior identified cluster operation work on the go. As the related data items are joined together, the data can be analysed without any disturb. Understanding the advantages of data clustering and prevalent usage of textual data, this work sets its goal to present a textual data clustering algorithm by clubbing FCM and LOA.The enormous amount of unstructured data is being transferred into the information network. The greatest challenge is to organize and extract new information or knowledge from this large, unstructured text, leading towards the concept of text data mining or document grouping.

The aim of this work is to improve the performance of current clustering documents. The aim is to offer an efficient, accurate and scalable, better quality clustering solution. Textual data mining is focus two main techniques. First one supervised learning and second one unsupervised learning. Under the surveillance of teacher student learning in the class room. It is the example for supervised learning. Without a teacher, instead refers to unsupervised learning. Some in which training dataset is not labelled and the typical goal of which of which is to find natural cluster of the patterns specified that is cluster.

2. RELATED WORKS

Clustering algorithm are arrange into group on the similarity of a collection of patterns. Patterns within a valid cluster are intuitively more similar to each other than to a pattern that belongs to another cluster. But there are few previous information on the dataset available in many of such problems and as few assumptions about the data as possible must be made by the decision maker [3].In [4] Text document clustering was most essential research in recent years. It helps to organize textual information in particular order for retrieving information easy.

Text clustering is most useful for many resent task that is either offline or online work based on textual information. It is an extensively spread sub-set of data clustering that uses concepts from all domain. Based on Lion nature clustering operations are made. In [5] this proposal an approach based on ensemble methods for a non-parametric document classification. Non – parametric document clustering can be defined as the process of grouping similar document with requiring either the number of categories of the document or an accurate start of the process.

In [6] multiple label text document categorization technique based on fuzzy relevance clustering. This work clusters the content by considering must link

and cannot_link constraints. The must link constraint works on document with greater semantic information and the cannot_link constrains is meant for information with minimal similarity. In [7] work employs a fuzzy relevance measure for transforming the high dimensional document to the low dimensional fuzzy relevance vector.

The information clustering technique with manifold based optimization of Bag-og-Features (BoF) is proposed [8]. This proposed utilizes the data manifold and clustering results are formed. In [9] attempted to present another text clustering technique by employing genetic algorithm. The claimed that performance of harmony search algorithm is better than the performance of genetic algorithm on combine with k-means algorithm. The performance of Fuzzy C mean is enhanced by several optimization algorithm, so as to support the FCM from Convergence at local minima [10].

3. PROPOSED ELFCM ALGORITHM

The proposed text clustering algorithm is form a group with similar together that could organize text data item. This algorithm is clubbing Lion optimization algorithm with Fuzzy mean algorithm. The optimization algorithm for choose initial point and FCM for cluster operation. LOA imitates the character of Lion. The merits with this algorithm restricted control parameters. There are three important steps involve ELFCM.

- Pre-processing
- Text Similarity Computation
- Cluster Operation
-

The pre-processes are removing unwanted meaningless data through source dataset. To eliminate mask that is unwanted product. ELFCM us used vector space method for making pre-process. The product of pre-process is called Bag of word. It represented as follows

$$t = (td_{i1}, td_2, \dots\dots\dots, td_n) \quad (1)$$

$$d_x = (W_{x_1}, W_{x_2}, \dots\dots\dots, W_{x_k}) \quad (2)$$

Based on weight of the documents are assigned, Before cluster operation on dataset. Pre-process refine some stops and removing stem words.
Similarity computing utilized to measure distance from document to document. It predict whether text are similar or dissimilar based their distance. The clustering is to grouping of data item into particular group. To calculates similarity between data items internal or externally. Euclidean distance is extract suitable for calculate distance between textual data item. Here implement MVS with Euclidean distance. MVS is used for calculate distance in multiple distance vector space.

$$sim(d_i, d_j) = \frac{1}{n-n_r} \sum d_n \in S \; sim(d_i - d_{\square}, d_{\square} - d_j) \quad (3)$$

Lion optimization algorithm is better to pick the initial point of cluster. Based on living nature it divided into two type as single and group such as nomad and pride. It is decide to hunting behaviour and success rate of capture prey. Some principle apply for dataset and get the initial point. To calculate it quality and fixed as a initial point of cluster.

$$Lion[fitness] = \frac{f_i}{\sum_{i=1}^{n} f_n} \quad (4)$$

Fuzzy C-Mean algorithm described by Fuzzy Matrix μ with n rows and C columns. The n is number of data objects and c is the number of cluster. μ_{ij} the element in the $i^{t\square}$ row and $j^{t\square}$ column in μ.

$$F_n = \sum_{i=1}^{r} \sum_{j=1}^{c} \mu_{ij}^{ff} \left\| x_i - x_j \right\|^2 \quad (5)$$

ff is the fuzzy factor and μ is the fuzzy membership value.

$$X = (x_1, x_2, \dots\dots, x_c) \quad (6)$$

$$Y = (y_1, y_{2,\dots\dots\dots} y_r) \quad (7)$$

Cluster centres of cluster s(r) can have w features.

$$C_i = (C_{i,1}, C_{i,2} \dots\dots\dots C_{i,w}, C_i, r \times (w-1) + 2 \dots\dots (i.r \times w) \quad (8)$$

Proposed ELFCM Algorithm

Input : Text dataset(N)
Output: Text cluster(C)
Begin
For all
Step 1: Start Pre-Processing;
Step 2: Initialize the population of Lion;
Selecting random Nomad and Pride;
Step 3: Distribute the prey;
For each pride of nomad
Compute fitness of lion;
Compute the success rate ;
Step 4: if (fitness(nomad lion)>fitness(pride lion)
Modify the living location of lion;
Verify the search space of lion;
Alter the search space;
End if
Step 5: Apply FCM for cluster operation;
Calculate Membership factor;
Create μ Matrix;
End if
Step 6: Store the optimum solution;
End for
End
Step 7: Some step involving in Textual dataset;

4. EXPERIMENTAL SETPU AND RESULTS

In this section evaluated the proposed text cluster algorithm with existing approach. A new proposal

ELFCM evaluated by standard performance metrics such as Precision, Recall, F-Score, Purity, Entropy. This work utilized some of popular dataset through the respectively their links. The performance associated with the text clustering algorithm is implemented in Core Java with 12 GB RAM. The following table represent about the datasets.

Table No :1

Dataset

Data	Source	Classes	Number Document	Pages
Hitext	TREC	17	2301	13170
Reuter7	Reuters	07	2500	4977
WebKb	WebAce	20	2340	13859
Re0	Reuters	13	1504	2886
20 News Group	WebAce	20	2200	20000

T_P, T_N, F_P, F_N based on these value performance metric are computed

Table No: 2

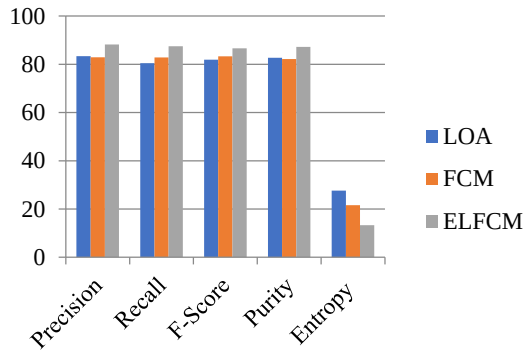
Performance Metrics

	Precision (%)	Recall (%)	F-Score (%)	Purity (%)	Entropy (%)
LOA	83.36	80.46	81.9	82.7	27.62
FCM	82.9	82.82	83.3	82.2	21.6
ELFCM	88.2	87.5	86.6	87.2	13.3

The above table describe the details about the performance analysis of ELFCM. The graphical representation of ELFCM is follows

Figure no: 1

Performance metric of ELFCM



The proposed algorithm ELFCM compare with existing algorithm it clearly proven works better. Based on the Precision rate of the proposal are greater than existing employed individually. Recall rate can price are greater. F-Score analysis demonstrably enhanced. The entropy of this proposal approach is smaller compare with LOA and FCM. Performance metrics are shows that ELFCM very efficient textual cluster algorithm.

5. CONCLUSION

In this paper, we proposed a new textual clustering algorithm, named as ELFCM. It would utilize the merits of Lion Optimization Algorithm and Fuzzy C Mean. We are proven ELFCM attain highly effective as well as efficiency based on empirical evidence and theoretical analysis. The key concept of this approach is initial point of cluster from LOA and cluster operation performed by FCM. The formed clusters are measured with performance metrics. In future this work can be improve by including semantic based analytics for cluster documents.

6. REFERENCES

[1] Hang Jia et al, "Unsupervised Feature Selection with Feature Clusters", IEEE Digital Library, May 2013.

[2] G.Jagatheeshkumar and Dr.S.Selva Brunda, "An Improved K-Lion Optimization Algorithm with Feature Selection Methods for Text Document Cluster",IJCSE,Vol(6), Issue(7), July 2018.

[3] A.H & Xu D Tan, "Semi-supervised heterogeneous fusion for multimedia data co-clustering", IEEE transactions on Knowledge and Data Engineering, 26(9), 2014.

[4] Yazdani and et all, "LOA: a nature inspired metaheuristic algorithm", JCDE, 3(1),24-36, 2016.

[5] Novakovic , J, "Toward optimal feature selection using ranking methods and classification algorithm", Yugoslav Journal of operation Research, 21(1), 2016.

[6] Jiang J.Y, " Multilabel text categorization based on Fuzzy relevance clustering", IEEE transaction on Fuzzy systems, 22(6),2014.

[7] Wu.J, & Xiong.H, "Summation-based incremental learning for information – theoretic text clustering", IEEE Transactions on Cybernetics, 43(2), 2013.

[8] Tefas & Passalis.N, " Information clustering using manifold-based optimization of the bag-of-features representation", IEEE transaction on cybernetic, 48(1).2018.

[9] L.M.Q Abuligah, "Feature selection and Enhanced Krill Herd Algorithm for Text document clustering", Springer, 2019.

[10] G. Jagatheeshkumar and S.Selvabrunda, "Text clustering Algorithm using Fuzzy Whale optimization Algorithm", IJIES, Vol(12), Issue (2), 2019.