IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

An Improved K-Lion Optimization Algorithm With Feature Selection Methods for Text Document Cluster

Jagatheeshkumar. G[1*], S. Selva Brunda[2]

[1] R&D center, Bharathiar University,Coimbatore,Tamilnadu, India
[2] Department of CSE, Cheran College of Engineering, Karur, Tamilnadu, India

Available online at: www.ijcseonline.org

Accepted: 20/July/2018, Published: 31/July/2018

Abstract- Growth of Computer applications in most of the people and companies are wanted to work through computers. They mostly use computer to store and retrieve information. Data mining is organizing and retrieving information from large data set. Now a day's dataset may be dynamic. Text Document clustering is a passion or an interested area of data mining. Many of the clustering method needed for a new one requires better clustering approaches. A new proposal is an improved KLOA with feature selection method for text mining that is Improved KLOA. K-means is one of the active algorithms for wider application of clustering technique. But it has some inconvenience to form a cluster in the initial point. A novel KLOA algorithm is refined and enhanced by k-means algorithm. This is used to pick the initial point and perform well when some think is rendered. To implement Feature selection method is to find subset and improve the process of cluster. Using Feature selection method is to improve the quality of cluster and find intrinsic properties of dataset. In this new article using wrapper technique of feature selection method is implemented and produces high quality of text clusters, with more accuracy and performance.

Keywords: Clustering Technique, Data mining, Feature selection, Optimization, Text Clustering.

I. INTRODUCTION

Data plays an important role in this activity. All the domain data is used for storing and retrieving. The dynamic purpose of data is overwhelming and in some cases inconvenient to maintain. It may be difficult to store and as well as retrieve data from dataset. The removing hassle must be ensured, such that the needed data can be obtained as and when required. This is more Strule for large dynamic data set. Cluster is a concept that forms a group which is related from one to another. This makes measuring similar or dissimilar. Similarity is measured by document vector space or distance based. Hence the new proposal is a text mining based on k-mean and Lion optimization algorithm.

Problem optimizations are usually very complicated to find solution and many applications have to be solved with these complex problems. Some traditional optimization concepts do not provide a better solution for them. The entire work can be arranged into three different stages. Such as pre-process, similarity measures, clustering operations. Prepositions, pronouns, articles, and irrelevant document have been removed. Choose the initial cluster point and finally associate K-mean with LOA. Feature selection is an interesting field that has been discovers unhidden area in research and redevelopment. Arrange information into a subset of feature or variable. Aim to find data in order to obtain more essential and compact representation of the available data set. Feature selection compares to other that have considered finding the fitness of subset of attributes with respect to a particular data mining task. Adding feature selection, to get a subset of features and clustering. To simultaneously find a subset and cluster defined by features. The high quality of dataset needed to more efficiency and effectiveness of feature selection method for discover cluster. Feature selection decrease memory, eliminate unwanted information or noise. It will be improve the quality of data mining algorithm.

Based on weights, feature extraction is made. Finally the data to new one if it have high quality of text cluster and it will become significant. The semantic work both in terms of the optimal solutions of the optimization and the quality of the solutions obtained from the LOA. Formulate K-mean, a new feature selection method for clustering operation. The propose method to initialized cluster point and choose optimal subset from features selection method. The remainder of this work is as follows. I. Related Work with appropriate text clustering algorithm. II.The Proposed text clustering algorithm. III. Evaluating new algorithm with existing algorithm. Finally concluded with performance, efficiency of proposal and enhance further work.

II. RELATED WORKS

To derive new solutions and Territorial Defense and Territorial Takeover, while intending to find and replace worst solution ,to the new in the best solution, Lion pride optimizer [1].The LOA is meta-heuristic algorithm that describes the character of lions[2][3].The resident lions resides by forming a group called pride. Each pride consists of four to five lionesses. When it comes to hunting, certain number of lionesses surrounds the prey. The Binary Particle Swarm optimization (BPSO) is a popular swarm intelligence optimization method to perform discrete optimization problem such as feature selection[4].

An improved k-means clustering algorithm for text is proposed in[6]. The disadvantage of the k-means is selecting an initial cluster point. This is overcome by computing density of all the data items in the dataset and the data items with maximum density are measured [5]. The MVSC-I_R , MVSC-Iv are criterion functions. These are calculated by intra or inter cluster similarity based on it average cluster size weight. The I_R, I_V is the simple form of criterion functions. It shows how to optimize clustering solution in text[6]. The distance based similarity measures Fuzzy sets that are considered to have a high importance in reasoning methods handling sparse Fuzzy rules bases [7].

Feature in basically raw dataset have the necessary information but it is not in suitable for the data mining. In this feature constructed out of the entire feature can be more used that original feature [8].Once a cluster solution is generated, interpretation is best motivated by ignoring the components and folding back in the original features along with any additional descriptive information not directly used in the solution. A few heuristics are the best guides to qualitative insight. It can be as easy generating a spreadsheet that our clusters based on average or medians for each feature [9].

To experimentally characterize the behavior of these feature selection algorithm, specified the structure that can and can discover performance of find subset. The results are produce approximation methods have cluster and features [10].

Efficiency stands required time to find a subset of features, and the effectiveness belongs to good quality of the subset of features. High dimensionality of data takes over efficiency and effectiveness point of view in feature selection algorithm [11].A cluster based approach for good feature selection evaluated using minimum variance method. The purpose of the this method is to reduce the computational complexity, reduce the number of initial features and increase the classification accuracy of the selected subset. The main idea to find the dependent attributes for the cluster and removes the other members in the cluster [12].Feature subset is identified good features and it will produce quality of cluster with compare target classes. Using this method for high dimensional dataset. Partitioning of minimum spanning tree and best features are selected clusters of spanning tree [13].

Ranking methods may filter features to deduce dimensionality to the feature space. Any inherent feature selections built in, such as the NNM or some types of neural networks. The number of step needed to find the optimal feature subset is very small comparing to the cost of wrapper approach for large number of features [14].

The collective and social behavior of living creatures motivated researchers to undertake the study is Swarm Intelligence. Particle Swarm Optimization (PSO) is initially introduce for simulating human social behaviors.PSO performs global search and K-mean is responsible for local search [15].

The Ant Colony Algorithm (ACS) based on natural system are successful sample of solving combined optimization problems. It has introduced new context in problem solving methods. ACS is creating an intelligent form of movement for ants in the standard and-clustering algorithm [16].

The FLA algorithm adapted with the clustering procedure. The iteration is continued till the tolerance is achievable only the best solution point is chosen as the optimal centroid. Multi kernel-based distance measurement can be included for finding the fitness of the cluster process. Further Strengthened with different optimization theory is defined by solution [17].

Ant Lion Optimization algorithm describe the hunting mechanism of ant lions in nature. Five main steps hunting prey such as the random walk of ants, building traps, entrapment of ants in traps, catching preys, and re-building traps are implemented. It can be concluded that the proposed algorithm benefits from high exploitation and convergence rate. ALO can be concluded from the results of multimodal and composite test functions which is due the employed random walk and roulette wheel selection mechanisms.

III. PROPOSED METHODOLOGY

Basically all the sub-phases are involved in the text clustering algorithm. Our Novel's proposed approach aim is to find an optimum solution for text clustering algorithm based on k-mean and Lion optimization Algorithm with feature selection. There are three important sub-phases that are eliminates the stop words which are preprocessing, similarity between intra or inter clustering is carried out cosine to similarity between the dataset. Finally the new proposal is achieved through textual data clustering. This clustering mechanism is designed for document cluster. The input dataset process Improved KLOA with Feature selection. New proposal find initial point using fitness function. After that implemented Feature selection method find optimal subset of given attribute. Fitness function used not only fitness of data, as well as quality of data. It improves and enhances the cluster performance. Using Lion optimization algorithm find the initial point that centroid of the cluster. Select best optimal point from selecting attribute Using Feature selection method. Apply k-mean for perform document cluster. Similar content place into one group (fit), and dissimilar group of data placed into one group

(unfit).This article based on behavior of Lion to Improve KLOA algorithm for document clustering algorithm.

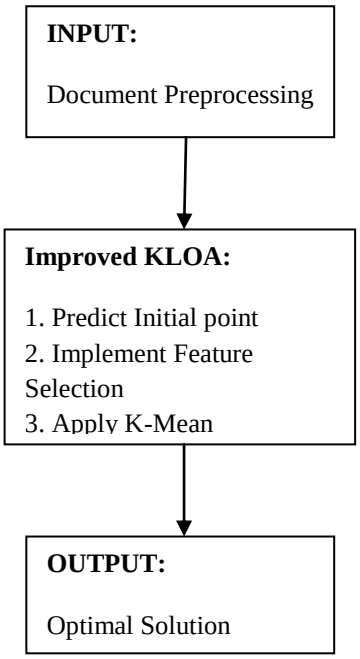

Figure 1: Overview of Improve KLOA

A. Preprocessing the text data

Preprocessing is an initial step in text mining. To remove unwanted mask of data item, mostly the stop words means on its own. Basically stop words are used to form a sentence or for grammar. To eliminate the stop words results in conversation time. Some examples for stop words are given below.

Table 1: Stop words

A	And	End	The
Of	Above	so	Because
About	Then	Also	Any
From	Beside	But	Beyond
After	that	Follow	They
Until	Un	Let	Be
All	Bellow	Such	Yet

These are the sample of stop words. It is not used to cluster operation. Only it associates with other words providing meaning. It may consist of pronouns, vowels, conjunction, and preposition. After preprocessing, the input data is ready for standardizing the data format. The standardize input data is the vector space or Bag of Word. Document vectors space is represented along calculate with term

T = ($td_1, td_2, \dots\dots\dots\dots, td_n$)

Associated weight$d_x = (W_{x1}, W_{x2}, \dots\dots, W_{xk})$

To eliminate stop words using feature extracting method.

B. Similarity Measurement

Mostly similarity is measured in single angle, now we are using a multiple angle that is MVS with Kullback-Leibler Divergence. In order to calculates the difference between probability distributions. Widely Cosine similarity is used to find similarity between document vector space. It has a popular similarity score in text mining.

Sim(di,dj) = cos(di,dj) =$d_i^t d_j$ (1)

With KL divergence similarity can express

Sim(d_i, d_j) = $D_{AvgKL}(d_i - 0, d_j - 0\ )$ (2)

C. KLOA text clustering Algorithm

K-mean algorithm is an efficient partition of clustering algorithm for text mining. It has some disadvantages in predicting the initial point of K. It may be an average value. So it is necessary that we need different initial points to find different clusters.

$SSE = \sum_{i=1}^{n}(\mathrm{d}x_i - \mathrm{d}\bar{x})^2$ (3)

dx,$d\bar{x}$ is either 0 or 1.k-means producing tighter cluster.

Lion optimization algorithm is a met heuristic algorithm [2]. It denotes lion characteristic and nature of lions. Basically lion live in two social debuts such as pride and nomadic. The hunting capability decides to fitness. Check it is very effective for hunting or not. It becomes effective and leads the pride to go and hunt. Otherwise it is not fit. There is a chance to attack nomadic male lion in the pride. The fitness of the lion is computed and the probability of success in hunting is measured. Similarly nomadic lion has the prey generated and the fitness is compute. The lion that is low in fitness is removed. This principle is applied to text document and hence the text clusters are formed. The initial process are initialized and followed by the computation of the similarity between the documents using cosine similarity.

$f_i = \sum_{i=1}^{c}\sum_{d_m} \epsilon c_n \|d_m - c_{cen}\|^2$ (4)

$s_{pr} = \frac{f_i}{\sum_{i=1}^{n} f_n}$ (5)

The success rate is decided to the fitness value of data. This is the centroids total count of cluster. The fitness value 0 to 1.It show similarity between the documents. The k-means algorithm applied to, after the cluster is formed has some iterative process. Similar process is done in text document and compute text cluster. The distance between the text data and the cluster centre is found out and the data with the minimal distance must be considered as a part of the cluster. The dataset are allotted to the most relevant data cluster. The performance of the proposal is evaluated to next section.

D.Improved KLOA with Feature Selection Method

Feature selection method is normally used to finding the subset and evaluating each one. To increase efficiency of feature using wrapper technique for find intrinsic properties of dataset. Using Improved KLOA algorithm find cluster and arrange an order. After that implement feature selection technique which wrapper method to find best cluster. Using

this technique to improve the quality and accuracy of cluster is very high.

Lion [fitness] = $\alpha\gamma_R(D) + \beta\frac{|R|}{|C|}$ (6)

R- Length of Pride or nomads
D- Decision
C- Total Number of features in the Pride or Nomads
A and β – Two parameters quality & subset length.

Initially feature selection methods are applied on the original feature set. Features are ranked in ascending order based on their metrics. Similarly using improved KLOA to arrange dataset into subset and using feature selection technique find fitness based rank list and finally get effective text mining algorithm. The enhanced method of KLOA is improved KLOA. In this proposal to enhance feature selection method for finding best attribute. There are two type of feature selection method filter another wrapper. Using wrapper technique, find best solution from group of attributes. Generally text mining based on vector space model. The binary representation of document is usually calculated by term frequency (TE/ID).

$$w_{ij} = tf_{ij}\log\frac{N}{df_i} \quad (7)$$

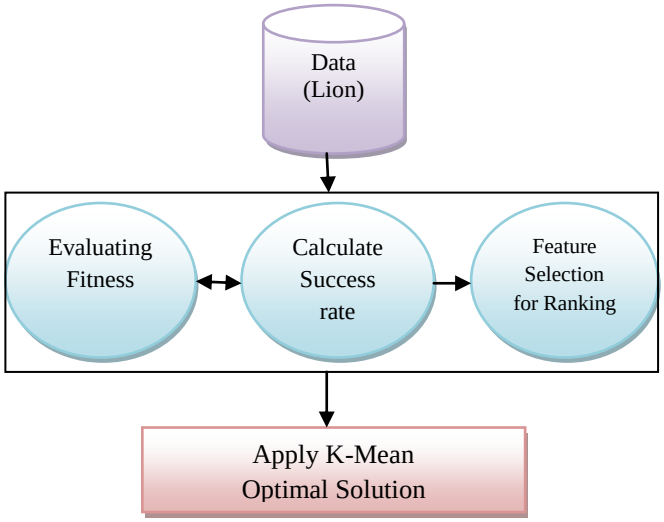

Figure 2: Improved KLOA

Proposed improved KLOA algorithm

Input: N Number of Text Document
Output: Optimum Text Cluster
Begin
For all
Step 1: Start Pre-processing;
Step 2: Initialize the population of lions;
Selecting randomly pride and Nomad;
Step 3: For each pride and nomad of lions
Randomly select a lion for hunting;
Randomly select a Lioness for hunting;
Calculate fitness of lion and lioness;
Compute the success rate;
Post ranking using Feature Selection method;
Step 4: Apply K-mean operation and find optimal solution;
End for
Step 5: if(fitness(nomadic lion) < fitness(pride lion)
Interchange the lion to pride;
Apply feature selection method;
Remove unfitness lion and lioness
Step 6: Apply K-mean and produce optimum solution
Calculate fitness of lion or lioness;
Store the optimum solution;
End if
End for
End
Step 7: Produce cluster in Text document using the same step (step 1 to step6) until better cluster form.

IV. EXPERIMENTAL RESULTS

This section has been evaluated and compared with existing clustering method with new proposal such as Improved KLOA. This work has utilized some of the popular datasets such as hitech, reuters7, k1b, webkb, re0, news3,sports,wap,tr23, la2 and reviews. There are downloads from the links[18-20] respectively. The performance of the text clustering algorithm is implemented in Core Java with 12 GB RAM.The below table describes the characteristics of dataset. The standard preprocess made such as removing stop-words, frequent occurrence words, and stemming.

Table: 2: Text Document Dataset

Data	Source	c	n	m	Balance
hitech	TREC	17	2301	13170	0.191
reuters7	Reuters	7	2500	4977	0.082
K1b	WebACE	6	2340	13859	0.043
webkb	WebACE	20	2340	13859	0.192
Re0	Reuters	13	1504	2886	0.018
Sports	TREC	7	8580	18324	0.036
Wap	WebACE	20	1560	8440	0.015
Tr23	TREC	6	204	5831	0.066
La2	TREC	6	6279	21604	0.282
Reviews	TREC	5	4069	23220	0.099

c:No of classes, n:No of documents, m:No of words, Balance = (smallest class size)/(largest class size)

A. Performance analysis w.r.t Clustering Algorithm

The performance analysis measured by some performance metrics. Such as Precision(P_r), Recall®, Entropy(E), F-Score (F) and Purity(P). These are compute by T_P, T_n, F_p, F_n rates.
T_p – Correctly clustered document to the Entire Number of document in the dataset
T_n – Not Candidates of the particular cluster
F_p – Fraction of the document, wrongly included as a candidate of wrong cluster
F_n – Wrongly clustered, not as a candidate of cluster
The performance metrics computed based on these value such as T_p, T_n, F_p, F_n.. the following table to describe average percentage of performance metrics for table 2 data set

Table 3: Performance metrics

	Precision (%)	Recall (%)	F-Measure (%)	Purity (%)	Entropy (%)
LOA	77.59	73.03	75.24	74.96	36.23
KLOA	84.2	78.8	81.4	83.06	27.03
IKLOA	96.3	96.6	96.6	94.73	11.53

Precision

The clustering algorithm reach high Precision rate, F_P values are minimal. A best clustering algorithm should prove high precision rate.

$$P_r = \frac{T_p}{T_p+F_p} \quad (8)$$

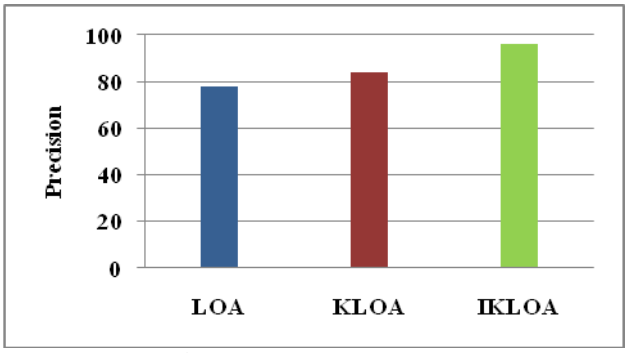

Figure3: Precision Metrics

Figure shows that comparison of the new proposal with existing methods using precision metric. The x axis is algorithms plotted. The y axis is precision rate plotted from zero the hundred. Improved KLOA shows higher precision rate (96.3). This new prototype provides more precision value than existing such as LOA and KLOA.

Recall

The recall is viewed to retrieve relevant documents on a set of documents. The recall is compute quality of text cluster. It divides total number of false negative by total number of relevant elements. The recall value is calculated following formula.

$$R = \frac{T_p}{T_p+T_n} \quad (9)$$

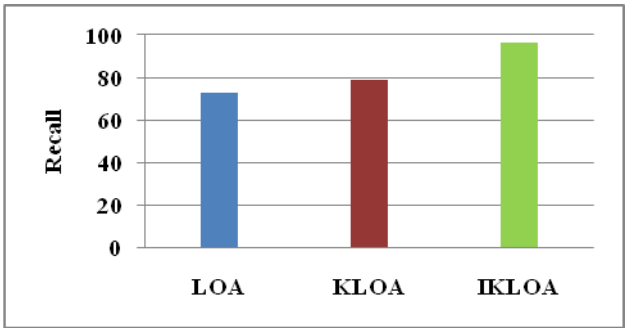

Figure 4: Recall Metrics

The above figure shows that comparison of new prototype to existing algorithms. It clearly proves Improved KLOA is providing higher recall metrics value as 96.6 %.

F-Measure

F-Measure is single measurement of Precision and Recall. It tests accuracy. It calculates effectiveness of retrieval speed. It describes the documents are labels. Assuming a one – to – compare. Both true positives and true negatives among the total number of class examined. It calculating formula as given bellow

$$F = \frac{2(P_r \times R)}{P_r+R} \quad (10)$$

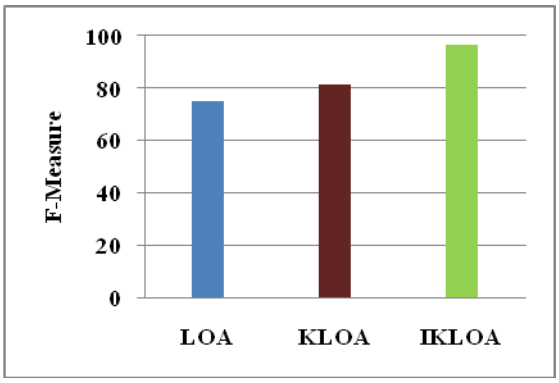

Figure 5: F-Measure Metrics

The above figure shows that F-Score Analysis of Improved KLOA with existing methods. The y axis is plotted F-measure ratio and x-axis is plotted methods. The new proposal provides high F-measure rate that 96.6% and prove more efficiency of cluster accuracy.

Purity

The quality of cluster based on purity of cluster. This compute by

$$P(cl_i) = \frac{1}{ndc_i} {}^{max}_{c} ndc_i^c \quad (11)$$

Purity is an external evaluation quality of cluster criterion. The following figure to shows that Purity comparison. The new prototype compare with existing prototype. The x-axis is plotted methods which are comparing purity. The y-axis is plotted value of purity. The Improved KLOA is providing more purity value that 94.73%.

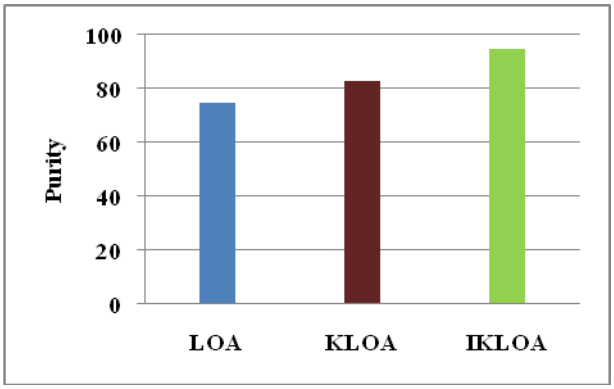

Figure 6: Purity Metrics

Entropy

Entropy value is less mean the quality of cluster is high. The above equitation is calculating Entropy.

$$E=\sum_{c=1}^{m} \frac{ndc_i^c}{ndc_i^c} E(cl_i) \quad (12)$$

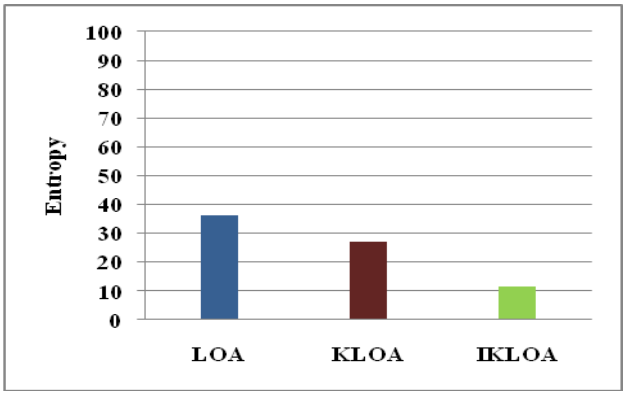

Figure 7: Entropy Metrics

The figure contains the Entropy Metrics of Improved KLOA with Existing methods. X-axis is plotted Entropy ratio and y-axis contains methods. This experimental result IKLOA had less percentage of Entropy value as 11.53. It proves more effectiveness and efficiency of cluster. On analysis the experimental results, It is clearly proven that the combination of Improved KLOA works better. The reason for attaining maximum precision, recall, F-measure and purity is that the initial cluster points are chosen by the LOA and the formed clusters are enhance Feature Selection method and finally implement K-mean Algorithm for cluster operation. This enhancement refines the quality of cluster which results in better results.

B. Performance Comparison with the Existing System

The analysis of proposed algorithm is compared with existing protocol. After that it will proved that the new is better than the older. The above 10 dataset has been compared with existing and was done based on the result in NMI.This is an internal criterion for the quality.

NMI Comparison

NMI Measures the information the true class partition and the cluster assignment share. It measures how much knowing the clusters help us know classes.

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j} log\left(\frac{n n_{i,j}}{n_i n_j}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i log\frac{n_i}{n}\right)\left(\sum_{j=1}^{k} n_j log\frac{n_j}{n}\right)}} \quad (14)$$

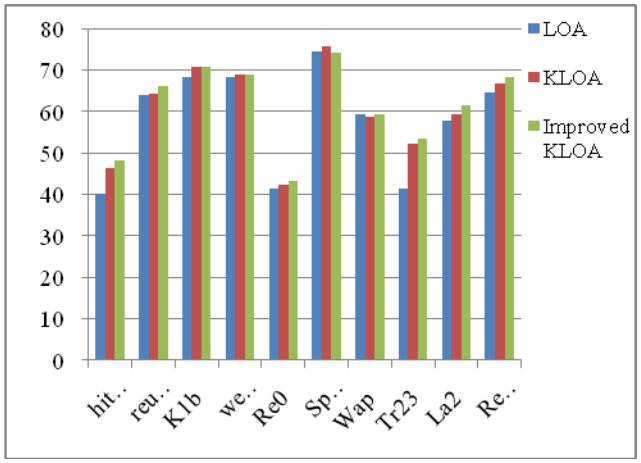

Figure 8: NMI Comparison

Normalized Mutual Information (NMI) comparison to calculate cluster performance. It produces cluster validation of cluster and class size. The above figure 10 describes that NMI comparison of 10 deferent text dateset. The x axis denotes dataset and y axis denote percentage from 0 to 100. The result concluded Improve KLOA more efficient then other optimization algorithm. Thus in the experimental analysis, it is proved that the proposed Improved KLOA algorithm have highest Accuracy, and NMI results. The LOA is chosen to be the best cluster result, as the better features are selected before performing the clustering operation with implement feature selection method. Finally the Improved KLOA proves its efficiency upon other ten different dataset and has shown better cluster of text.

V. CONCLUSION

The Proposal innovate new textual document cluster algorithm, which combines together KLOA with feature selection, where the LOA selects the better cluster center point and k-mean enhances cluster operation. Feature selection method to improve cluster quality. The new Improved KLOA proves its efficiency with some experimental results. Already we have presented its efficiency with various famous algorithms with Improved KLOA. The formed clusters are measured with quality and performance of Improved KLOA to analysis in some of standard performance metrics. The results conclude that the proposal algorithm provides better text document cluster algorithm.

REFERENCE

[1] Leukocyte Saraswat M, Arya KV, Sharma H, segmentation in tissue images using differential evolution algorithm , swarm Evolut, computer 2013:11(0) 31-45.

[2] Fariborz Jolai and Yazdani, "Lion Optimization algorithm:Nature – inspired metaheuristic algorithm.Journal of computational Design and Engineering, Vol.3.16 June 2015.

[3] GB Schaller, The Serengeti Lion: A study of predator relations. Wildlife behavior and ecology series. USA 1972.

[4] Pramod Kumar Singh and Jay Prakash, "Particle swarm optimization with K-means for Simultaneous Feature Selection and Data Clustering",IEEE Digital Library,ISCMI,2015.

[5] Zhen Hua, Caiquan Xiong, Ke Lv, Xuan Li, "An improved k-mean text clustering algorithm by optimizing Initial Cluster Centers", ICOCCBD, Macau, China, 16-18 Nov, 2016.

[6] Lihui Chen and Duc Thang Nguyen and Chee Keong Chan, "Clustering with Multiviewpoint- Based Similarity Measure,IEEE Transactions on Knowledge and Data Engineering, Vol24,No 6,June 2012.

[7] Zsolt csaba Johanyak, Kovacs, "Distance base similarity measures of Fuzzy sets",2015.

[8] Yiu-Ming Cheung, Hong Jia "Unsupervised Feature selection with Feature Clustering" , IEEE Digital Library, May 2013.

[9] Cyprien Gilet, Marie Deprez, "Clustering with feature selection using alternating minimization, Application to Computational biology", Cornell University Library, Dec-2017.

[10] Martin Azizyan, Aarti Singh, And Wei WU2 "Experimental Evaluation of Feature selection methods for clustering", Jan 2014,Garnegie Mellon University.

[11] Khedkar S.A. et al,," A Survey on clustered feature slection algorithm for freature high dimentsional data" volume5(3), 2014,IJST.

[12]Sivakumar Venkataraman, Subitha Sivakumar and Rajalakshmi Selvaraj, "A novel clustering based Feature Subset selection framework for effective Data Classification", Vol9(4), Jan 2016.

[13] Mathuri B Patil, Ani Rao, "A Review on clustering Based Feature subset selection algorithm for high dimensional data", vol4(1) January 2015,IJCSIT.

[14] Jasmina NOVAKOVIC, Perica STRBAC, Dusan Bulatovi, "Toward optimal feature slection using Ranking Methods and claasification" March 2011, Yugoslav Journal of Operations Research.

[15] Sagar Tiwari et al,"Algorithm of Swarm Intelligence using Data Clustering", Vol4($), 2013, IJCSIT.

[16] Kayvan Azaryuon, Babak Fakhar, " A novel Document Clustering Algorithm Based on Ant Colony Optimization algorithm", vol(7), JMCS, 2013.

[17] Sathis Chander et al, "Fractional Lion Algorithm – An Optimization Algorithm for Data Clustering", JCS, Aug 2016.

[18] Sathishkumar.K abd David Otto Arthur, "Clustering Mutual outline for multi Assessment Temporal Data and Cancer Data", vol(6), Issue-1, E-ISSN:23472693, 208

[19] A.K.Sharma and S.K.Patel, "Optimization of Dynamic Resource Scheduling algorithm in Grid Computing Environment", Vol(6), Issues-3,2018

[20] Am Amol D.Potgantwar and S.S.Dhable, "Optimizes NP Problem with Integration of GPU Based Parallel Computing", Vol(5), Issues-3, 2017.

Authors Profiles

Mr.G.Jagatheeshkumar, Received his Master of Computer Science for Bharathiar University, Coimbatore, India. His is currently working as a Assistant Professor at KSG College of Arts and Science, Coimbatore, India. Pursuing Ph.D om Bharathiar University, Coimbatore.His area of Interest Data mining, Cloud Computing, Machine Learning.

Dr.S.Selva brunda, Received Ph.D Degree from Mother Teresa University, Kodaikanal, India in 2012. She has been working as Professor & Head Department of Computer Science and Engineering at Cheran College of Engineering, Karur. Her main area of research interests are data mining, Cloud Computing, Artificial Ingelligence & Game Theory. She received young women scientist award from Dr.Abdul Kalam trust for education for the year 2016.

