Improved and Effective Artificial Bee Colony Clustering Algorithm for Social Media Data (I-ABC)

1Akash Shrivastava, 2Dr. M. L. Garg

Department of Computer Science & Engineering, DIT University, Dehradun

1akash.10may@gmail.com, 2dr.ml.garg@dituniversity.edu.in

Corresponding Author: Akash Shrivastava

akash.10may@gmail.com

https://doi.org/10.26782/jmcms.2019.02.00020

Abstract

Social media data made real world like a web of data which is highly categorical in nature. Data having categorical attributes are omnipresent in existing real world. Clustering is an effective approach to deal with categorical data. However, partitional clustering algorithms are prone to fall into local optima for categorical data. A novel approach of ABC K-modes has been proposed to address this issue but acceleration issue of this algorithm was still a challenge for it. In this paper, we address this challenge to reduce the acceleration factor of algorithm and proposing a novel modified ABC K-modes approach which we refer as N-ABC K-modes approach. In our approach, unlike existing ABC K-modes we introduces different attribute matrix for each data sets. In further step, we apply XOR operation to combine the matrix of similar attributes. In last phase, dissimilar data would form a cluster and we apply clustering follow by searching on this cluster. The performance of New ABC K-modes evaluated by a series of tests and experiments over real time streaming social media data like twitter and facebook in comparison with that of other popular algorithms for categorical data.

Keywords: Big data, Twitter, Clustering, Big data Analysis, Artificial Bee Colony (ABC), Data classification.

I. Introduction

In the current world, web is flooded with social media data. Social media become a tool to analyze and generate opinions regarding any product, subject, and political perception including many other domains. Twitter and facebook plays an important role in this emerged practice from past recent years. Twitter becomes a
social site where every popular individual across the globe or any laymen citizen of any nation who has access to internet mostly marks their existence over twitter. It has been reported in [XVII], 140 million plus active users used to publish 400 million 140 characters “Tweets” everyday. This is then quite evident that twitter could be refer as such a social platform which could be able to decide and reflect the flow of sentiments of customers, viewers to the organizations, industries or individual. In addition to twitter platform facebook is also one of the prominent platforms. These social websites now takes analytics domain up to new heights where companies, organizations are looking forward to harness the data furnish from these very social sources. This is the ground over which clustering of data comes into the picture and it has required. Data accumulated on twitter or facebook are in unstructured format which is very tough to process and get transform into knowledge. Here clustering is credible existing approach that could use to deal with this challenge.

Clustering is a special type of classification[XV]. In the prospect of data mining clustering plays an important role to classify the structured or unstructured data in required format [XVII,XIV]. The capability of clustering could be seen as it has been usually used in various fields like information retrieval, social media analysis, image analysis, text analysis and bioinformatics. Clustering targeted to allow the similar objects and dissimilar objects to form a group based on their individual characteristics. The study of literature regarding clustering objectifies the fact that clustering mainly diversified in two types: hierarchical and partitional clustering. Algorithms lies in Hierarchical clustering assigns a group of data objects into a tree of the nested partitions according to agglomerative strategy [II]. On the other hand, algorithms fall under the later one partition a set of data objects into a predefined number of clusters by optimizing an objective cost function.

There are bunch of algorithms have been evolved in clustering over the time. The unique prompt development in this direction had been raised by analyzing the intelligent behavior of social animals like ant, bees, birds for optimizing the problems. It has been observed that the approaches developed in this direction have successfully implemented to clustering. An ant colony clustering algorithm proposed by Shelokar, Jayaraman and kulkarni simulates the same approach which ant used to follow while they are searching for their food source [XVIII]. In the same domain, Kao,Zahara, and Kao united the particle swarm optimization (PSO) approach, it works like the way birds mimic to find the optimal food sources in search space. This algorithm includes k-means procedure and Nelder-Mead simplex search method for effectively improving the performance of clustering [X]. Wan et al. spawn a clustering algorithm which is based on the optimization property of bacterial foraging behavior [XXI].

In few recent years, honeybees behavior had been investigated which includes learning, memorizing, and information sharing mechanism, has emerged as
an interesting research direction in swarm intelligence [XXIV]. There are various engineering and management problems in real world has been dealt with bee colony optimization heuristic inspired by foraging behavior of bee swarms introduced by Lucic and Teodorovic [XIX]. Numerical optimization problem are used be a challenge to manage in due course of this domain Karaboga and Basturk presented an artificial bee colony to address this challenge [XI]. Using the ABC optimization strategy, Karaboga and Ozturk proposed a clustering approach of artificial bee colony[XII]. We are including the overviews and some highlights of existing clustering algorithms but most of these approaches designed by keeping in concern about numerical data. They are not fitting in the domain of categorical data. As mentioned the ubiquity of categorical data in current scenario and real world applications it is now required to develop an ABC based clustering approach for categorical data. In this direction, Wei Pang, Zhewang proposed a novel artificial bee colony clustering approach for categorical data. In this algorithm they firstly introduced the one-step k-modes procedure and then amalgamate this procedure with the artificial bee colony heuristic to cluster categorical data. But there is acceleration issue still there in this algorithm. The second aspect which we found is that this approach had never implemented over social media data like twitter which is highly categorical in nature.

In this paper, we propose an effective approach as new novel artificial bee colony clustering algorithm. We proposed much needed measures in the existing artificial bee colony algorithm which effectively implemented to reduce the acceleration issue stated for the algorithm. We successfully able to reduce the acceleration issue by analyzing our approach which has been demonstrated by applying across various platforms of social media particularly Twitter, Facebook and YouTube. Result produced justifies the effectiveness of algorithm. It has been proven to be a novel and unique approach of clustering applies over social media data.

The rest of the paper would involve the review of related work. This description of related work is then followed by the presentation of our proposed approach. Then, we present our experimental results and analysis of our work which we carried on desired categorical data sets; it highlights the effectiveness of proposed method comparatively with existing one. In the final phase of our paper, we draw conclusions and elaborate future work.

II. Related work

In the first section we primarily review the k-means and k-modes algorithm, then the idea of artificial bee colony optimization described.
K-means algorithm

K-means algorithm [XVI] is known for its simplicity, efficiency and low cost of computation approach for clustering. Considering this fact it is also being observed that functions involved in the approach of clustering most of the functions are non-convex and non-linear, categorically standard k-means algorithm is very reactive to initializations and also prone to be trapped in local optimal solution. As the result of gradual increase in number and dimensions of data sets, finding solutions to functions involved has become an NP-hard problem. However, variants have been proposed to standard k-means method to provide strategy to solve this problem [I, IX]. Since importance and significance of clustering approaches could be evident in many fields, global optimization method[IV, XX, XIII, XVIII], such as genetic algorithms (GA), ant colony optimization (ACO) and particle swarm optimization (PSO), have been applied to solve clustering problems [IV, XX, XIII, XVIII, III, XXII]. In case to solve clustering problems, these mentioned algorithms start from an initial population and explore the solution space through number of iterations to reach a near optimal solution.

K-modes algorithm

The K-modes algorithm Huang in [V] was first introduced the k-modes algorithm which was effectively developed for clustering categorical data. Let $P = \{ P_1, P_2, P_3, \ldots, P_x \}$ denote a dataset consisting of $x$ data objects and $P_i$ ($1 \leq i \leq x$) be a data object characterized by $m$ categorical attributes $Q_1, Q_2, Q_3, \ldots, Q_m$. Each categorical attribute $Q_j$ has a domain of values denoted by $\text{Dom}(Q_j) = \{ q_{1j}, q_{2j}, q_{3j}, \ldots, q_{tj} \}$, where $t$ is the number of categorical values for the attribute $Q_j$. A data object $P_i$ is generally represented in the form of a vector $[P_{i1}, P_{i2}, P_{i3}, \ldots, P_{im}]$

The K-modes algorithm objective is to divide a dataset $M$ into $n$ clusters by minimizing the following cost function:

$$E(T, R_l) = \sum_{i=1}^{n} \sum_{j=1}^{x} \text{dis} (m_i, R_{lj})$$

(1)

Here $R_l$ is the set of the most frequent value for each attribute in a cluster $l$, and it is called the mode of the cluster $l$; $\text{til}(0 < \text{til} < 1)$ is an element of the partition matrix $T_{nxk}$; $n$ is the number of clusters, and $\text{dis}(p_i, R_l)$ is the distance measure as given below:

$$\text{dis}(p_i, R_l) = \sum_{i=1}^{m} \alpha (p_{ij}, r_{lj})$$

(2)

In Eq (2), $\alpha (p_{ij}, r_{lj})$ is defined as:
\[ \alpha(p_{ij}, r_{lj}) = f(x) = \begin{cases} 
0, & \text{if } p_{ij} = r_{lj}, \\
1, & \text{if } p_{ij} \neq r_{lj}, 
\end{cases} \]  
(3)

Where \( r_{lj} \) is the most frequent value of the \( j \)th categorical attribute in the cluster \( l \). The process of the k-modes algorithm is depicted as follows:

Step 1. In the initial modes of clusters, \( d \) data objects are being randomly picked up from \( Y \) dataset.

Step 2. In terms of Eq. (2), assign data object to the cluster on the basis of the mode of which is the nearest one to this data object as compared with the other modes of clusters. The mode of all clusters must get updated after clusters have been assigned to all data objects.

Step 3. When all the data objects have been assigned to clusters, the dissimilarity which is underlying in between the current modes and data objects has been re-evaluated.

Step 4. In case, if it is found that nearest mode of assigned data object belongs to another cluster instead of the current one, then this data object will be reassign to that cluster and the modes of both clusters have been updated.

Step 5. Finally if none of data objects have changed the clusters after a full circle test of \( Y \), terminate the algorithm; otherwise go to step 3.

The artificial bee colony algorithm

Karaboga and Basturk [XII] proposed the artificial bee colony (ABC) algorithm which is well known for its simplicity and robustness. This algorithm is being popular for optimizing numeric problems. In the ABC algorithm, three types of bees are there in artificial bee swarm: employed bee, onlookers and scouts. The employed bee approach a specific food source which it wants to exploit then information about the food source shares with the onlookers in the nest; scout job is to search new food source in the search space, and rest onlookers waits in the nest and through the information shared by employee bees finds a food source. In ABC algorithm, the exploration and exploitation phenomenon are performed simultaneously. In this algorithm the task has been specified for employed bees, onlookers and scouts. Exploitation process is being carried out specifically by employed bees and onlookers whereas scouts execute the exploration process. The bee colony explores and exploits the food sources in a way to enhance the nectar being stored in the nest. From the perspective of optimistic problem if this algorithm would be observed then the interpretation is like that where food source means a possible solution, the nectar amount of a food source measures the quality of the
corresponding solution and the goal is to obtain the optimal value of the objective function.

Step 1. Population of food sources has been initialized.

Step 2. Nectar amounts (NA) is to be evaluated by sending employed bees onto the food sources.

Step 3. Nectar amount (NA) is responsible for determining the probability value of each food source, then the probabilities of all the food sources which is to be chosen by the onlooker bees is being evaluated. The fact is as much as the nectar amount of food source is big, higher the value of probability.

Step 4. The probabilities calculated from previous step i.e. step 3, is being consider for onlooker to choose its food source for which onlookers send onto food sources.

Step 5. In case, if the food source of employed bee becomes exhausted then terminates its exploitation process. Then this employed bee becomes a scout bee.

Step 6. Then new food sources are randomly to be finding into the search space that has been considered by sending the scouts into it.

Step 7. The best food source found so far considered memorizing, if the requirements are met, output the found best food source; otherwise going to step 2.

**Novel Effective ABC Clustering algorithm**

An effective Novel ABC clustering algorithm has been proposed by W pang, Y Zhang [VIII] on the basis of artificial bee colony and k-modes approach. The complexity and convergence analysis of the algorithm have been effectively shown. This algorithm has been tested on biological data which have been really proven to be much effective. But it is not relevant and proven to be effective for categorical sets of social media, in fact it have been strongly observed that the acceleration issue of this algorithm is major challenge. The approach of algorithm found to be very efficient to deal with unstructured data and hence is useful for clustering. The proposed approach required some sort of changes and updation which has been successfully experimented in this paper. This clustering algorithm is being considered as one of the remarkable research done in context of dynamic clustering. Generally, clustering algorithm is used to apply over numerical datasets but it is derived to be utilized for categorical dataset in this approach. The research carried out in the way that it inspires to take bio-inspired clustering for the categorization of unstructured data.

**III. The proposed approach of Improved ABC Clustering Algorithm for Social Media Data**

In this proposed approach, it is likely to be considering the philosophy of Novel ABC clustering algorithm in our I-ABC clustering algorithm. Artificial
honeybees are of three types as explained above. These three are employed bees, onlookers and scouts. The optimization has been done for the food source which is essentially corresponds to a feasible solution of the problem. Another aspect is the nectar amount associated with the food source which is responsible to characterize the quality of the corresponding solution. In the process of clustering, it is observed that clustering results depend upon the cluster centers. The clustering results are being determined in the condition when cluster centers are fixed. This is the reason that the issue of clustering can be observed as the optimization of the cluster centers. The possible solutions have been derived from set of cluster centers.

For categorical data clustering, let \( g_i = \{ S_1, S_2, \ldots, S_k \} \) denote a food source, where \( S_l \) is the mode of cluster \( l \). \( E(g_i) = E(T, g_i) \) is the objective cost function, and \( F(T, g_i) = \sum \sum t_i \text{dis}(x_i, S_l) \), where the symbols have the same meaning as in Eq. (1). Then, the nectar amount of a food source \( f_i \) is given by:

\[
\text{NA}(g_i) = \frac{1}{E(g_i) + 1} \quad (4)
\]

As elaborated in ABC approach, in this proposed also colony of artificial bees has divided into two parts:

Employed bees are the first half of the artificial bees and onlookers are the second half of the artificial bees. For a food source only one employed bees has been existed and it is also considered that number of employed bees are actually equal to the number of solutions identified in the population.

Let \( Q_{g_i} = \{ g_1, g_2, \ldots, g_k \} \) denote the population of food sources, where \( K \) is the number of the food sources, and \( g_i \) is the ith food source. Then the probability of the ith food source being picked up by an onlooker is given by:

\[
\text{pro}_{i} = \frac{\text{NA}(g_i)}{\sum_{i=1}^{K} \text{NA}(g_i)} \quad (5)
\]

The candidate food source required to be derived from the current one in memory. The required process to be carried out by introducing one step K-modes procedure referred as OKM. This is one iteration step process essentially in the search process of k-modes algorithm. It is being utilized to find the neighbor food source which is actually based on the current food source. This is being carried out in the exploitation process which is performed by employed bees and onlookers.

Let \( g_i \) be the current food source, then the OKM consists of the following two steps:
1. Cluster with the nearest mode is being allocated to each of the data object.

Partition matrix \( W \) has been then formed. If the \( i \)th data object which consider to be belongs to the \( l \)th cluster \( w_{il} = 1 \); otherwise \( w_{il} = 0 \), where \( w_{il} \) is one element of \( W \);

2. The new modes hence then be calculated on the basis of the partition matrix \( W \) and eventually candidate food source \( f'_i = \{W'_1, W'_2, \ldots, W'_k\} \)

An employed bee becomes a scout for the colony of bees, when its food source is exhausted. To avoid abandonment of the food source, \( B \) parameter has been adopted which has been considered as a number of trials which is in a range of predetermined number. In case, \( B \) trials are failed to improve the food source then food source is treated as an abandoned. However, Employed bee becomes a scout.

Let the abandoned food source be \( g_i \), and then the search operation of a scout finding a new food source is given by:

\[
g'_i = \text{Rand}(\text{Dom}(Y)),
\]

where \( I \in \{1, 2, \ldots, F\} \) and \( \text{Rand}(\text{Dom}(Y)) \) is the operation of randomly selecting \( d \) data objects from the data set \( X \). The idea of multisource search inspired from the batch processing carried out in [VII] is being utilized in this proposed approach. The method implemented just to accelerate the convergence of proposed algorithm. Multi search method works as \( V \) candidate food has been search by the scout bees first and then the best food source has been picked up in the process.

The detailed calculation formula has been successfully introduced for relevant variables. The proposed Improved ABC clustering algorithm (I-ABC) for categorical social media (twitter) data is given as follows:

Input: The size of bee colony \( S \), the maximum cycle number \( \text{MCN} \), the number of clusters \( c \), and number of predetermined trials as \( B \) to control abandonment of a food source.

Output: Best food source.

1. Initially, population of food sources \( P = \{g_1, g_2, \ldots, g_H\} \) is randomly being initialized; specifically, for each food source, select \( k \) data objects randomly from the dataset \( X \) as the modes of clusters, set the exploitation numbers of food

\[
\text{En}_1 = 0, \text{En}_2 = 0, \ldots, \text{En}_H = 0.
\]

2. The nectar amounts of the food sources \( NA(g_1), NA(g_2), \ldots, NA(g_H) \), is being evaluated according to Eq(4);

3. Set CN(the cycles number) to 1;

4. For each employed bee {
a. Generate a new food source \( f_i' \) from the current food source \( f_i \) by using the one step k-modes procedure OKM, and set \( E_n = E_n + 1 \);

b. Evaluate the nectar amount \( NA(g_i') \) for the food source \( g_i \) according to Eq (4);

c. If \( NA(g_i') > NA(g_i) \), the current food source \( f_i \) is replaced by the new food sources; otherwise the current food source \( g_i \) is retained.

5. Evaluate the probability \( p_{ni} \) for each food source \( g_i \) according to Eq (5);

6. For each onlooker bee {
   a. Create a data matrix, with the column size of number of attributes.
   b. Perform an XOR operation for each data set with each other data set.
   c. If any data column has 0 in it, combine the two data sets and continue until the XOR operation yield 1 for every data set. }

7. If there exists an abandoned food source \( g_i \),
   a. Send the scout in the search space to find \( V \) candidate food sources \( \{ g_1^i, g_2^i, \ldots, g_V^i \} \) according to Eq (6);
   b. Evaluate the nectar amount \( \{ NA(g_1^i), NA(g_2^i), \ldots, NA(g_V^i) \} \) of the food sources \( \{ g_1^i, g_2^i, \ldots, g_V^i \} \);
   c. Choose the food source with the highest nectar amount as the new food source \( g_i \), and set \( E_n = 0 \);
   d. If \( NA(g_i') > NA(g_i) \) the current food source \( g_i \) is replaced by the new food source \( g_i \); otherwise the current food source \( g_i \) is retained.

8. \( CN = CN + 1 \);

9. If \( CN = MCN \), terminate the algorithm and output the best food source; otherwise go to step 4).

IV. Complexity Analysis

In this section, the complexity of the proposed I-ABC clustering approach is being explored and discussed. In the proposed method the time complexity contains five parts: the initialization, employed bees search operation, probability of food sources calculation, and the search operation of scouts and onlookers. The computational cost of these five parts are \( O(Hkmn) \), \( O(H(nKm + nkC)) \), \( O(Vnkkm) \), and \( O(H(H+(nmk+kn)/e + nk)/e) \) where \( e \) is and dynamic value generated from the number of subsets of data each iteration in step 6 in above proposed approach. Here \( n \) is the number of data objects in the dataset \( Y \); \( m \) is the number of attributes; \( k \) is the number of clusters; \( H \) is the number of employed bees or food sources; \( C= \) is the total number of categories for all attributes. Therefore, the overall time complexity of the proposed approach is \( O(Hkmn + s(Vnkkm + H(H + nk + nkC))) \). Here, \( s \) is the number of iterations. For space complexity, it requires \( O(mn) \) to store the dataset \( Y \), \( O((H+V)km) \) to store the food sources, and \( O(nk) \) to store the partition matrix. Thus, the overall space complexity of our methodology is \( O(mn + (H+T)km + nk) \). The time
complexity and the space complexity of the k-modes algorithm are \( O(nkm + (nkC + nk)) \) and \( O(m(n+k)+km) \), respectively.

Now with the increase in value of \( m \) the complexity is affected vastly for a \( m > t \) where \( t \) is number greater than 10k. So a dynamic factor \( e \) is introduced as a counter to increment rate of \( m \), making it \( m/e \), where \( e = \frac{\sum_{i=0}^{n} \sum_{j=0}^{m} \text{dis}(k, m, i)}{k} \).

V. Experimental Results and Discussion

In this section, the performance of our proposed clustering algorithm Improved ABC clustering is being demonstrated and evaluated; the proposed algorithm has been carried out by executing on TWITTER real time streaming categorical dataset. In this research, Yang's accuracy measure [XXIII] has been adopted and the Rand Index [XVI] is being utilized to assess the clustering results that has obtained. In Yang's method, the definitions of accuracy (AC), precision (PR), and recall(RE) are given as follows:

\[
AC = \frac{\sum_{i=1}^{k} a_i}{n} \quad (7)
\]

\[
PR = \frac{\sum_{i=1}^{k} \frac{a_i}{a_i + b_i}}{k} \quad (8)
\]

\[
RE = \frac{\sum_{i=1}^{k} \frac{a_i}{a_i + c_i}}{k} \quad (9)
\]

Where,

- \( a_i \) = The number of data objects that are correctly allocated to class \( C_i \),
- \( b_i \) = The number of data objects that are incorrectly allocated to class \( C_i \),
- \( c_i \) = The number of data objects that are incorrectly denied from class \( C_i \),
- \( k \) = The total number of class contained in a dataset, and
- \( n \) = The total number of data objects in a dataset,

In the above measures the AC has the same meaning as the clustering accuracy \( r \) defined in [VI]. Given a dataset \( X = \{x_1, x_2, \ldots, x_n\} \) as well as two partitions of this dataset: \( Y = \{y_1, y_2, y_i\} \) and \( Y' = \{y_1', y_2', \ldots, y_{i'}\} \), the Rand Index (RI) [XVI] is given by

\[
RI = \frac{\sum_{i=1}^{n} \sum_{j=2<i<j}^{n} \frac{a_{ij}}{\binom{n}{2}}} \quad (10)
\]
Where \( \alpha_{ij} = \begin{cases} 1, & \text{if there exist } t \text{ and } t' \text{ such that both } x_i \text{ and } x_j \text{ are in both } y_t \text{ and } y_{t'}, \\ 1, & \text{if there exist } t \text{ and } t' \text{ such that } x_i \text{ is in both } y_t \text{ and } y_{t'}, \\ \text{while } x_j \text{ is in neither } y_t \text{ or } y_{t'}, \\ 0, & \text{otherwise} \end{cases} \)

The RI is evaluated by utilizing the true clustering and the clustering that is obtained from a clustering algorithm. According to these mentioned measures evaluation, it has been known that a better clustering In the performance analysis, the proposed hybrid FCM and fuzzy K-modes algorithm, implemented on real time streaming twitter datasets. Then the clustering results of the proposed hybrid FCM and fuzzy K-modes algorithm is being compared with that of the other two algorithms i.e. fuzzy C-means and fuzzy k-modes in terms of the best (Best), average (Avg.), and standard deviation of AC, PR, RE, and RI. All algorithms are implemented in Java language and executed on Intel core i7, 3.9 GHz, 32GB RAM computer. In all experiments, the parameters of the proposed I-ABC clustering algorithm are set as follows: \( S=20, MCN=1000, \) which are typical values used in the original ABC algorithm [XII]; \( B=5 \) and \( V=5 \) are set by the rule of thumb. The cluster number \( k \) in all four algorithms is set according to the number of classes provided by the class information of the dataset. It has been remarked and pointed out that other class information is not used in the clustering process apart from the number of classes. The other parameters for other prescribed algorithms are set the same as those stated in their original papers. The authenticity of methods and variables used are being tried to maintain up to larger extent. The unstructured and categorical data must undergo this phenomenon for effective clustering.

Table 1: The AC of the four algorithms on the Twitter dataset

| Algorithms           | AC   |       |       |
|----------------------|------|-------|-------|
|                      | Best | Avg   | Std   |
| I-ABC Clustering     | 0.9305 | 0.9133 | 0.0109  |
| Novel ABC-K-Modes    | 0.9103 | 0.8997 | 0.0110  |
| K-Means              | 0.8834 | 0.8934 | 0.0503  |
| K-modes              | 0.8826 | 0.8371 | 0.0490  |
**Fig.1.** Graph for the value of AC of the four algorithms on the Twitter Dataset

| Algorithms          | PR       |
|---------------------|----------|
|                     | Best     | Avg     | Std      |
| I-ABC Clustering    | 0.9088   | 0.8794  | 0.0091   |
| Novel ABC-K-Modes   | 0.8913   | 0.8692  | 0.0099   |
| K-Means             | 0.8834   | 0.8678  | 0.0131   |
| K-Modes             | 0.8795   | 0.8330  | 0.0451   |

Table 2: The PR of the four algorithms on the Twitter dataset
Fig.2. Graph for the value of AC of the Four algorithms on the Twitter Dataset

Table 3: The RE of the four algorithms on the Twitter dataset

| Algorithms          | RE      |
|---------------------|---------|
|                     | Best    | Avg    | Std    |
| I-ABC Clustering    | 0.8283  | 0.8141 | 0.0078 |
| Novel ABC-K-Modes   | 0.7981  | 0.8121 | 0.0089 |
| K-Modes             | 0.7979  | 0.8118 | 0.0082 |
| K-modes             | 0.8142  | 0.6122 | 0.1035 |

Fig.3. Graph for the value of AC of the Four algorithms on the Twitter Dataset
The Application of the Algorithm on the proposed data sets shows a significant improvement in the best, average and lower standard values in AC, PR, RE and RI and therefore the algorithm provides a better computability than the rest of the four algorithms used for comparison. The significance of the algorithm evident in the graph plotted which contains all four algorithms. The Proposed algorithm works on the principle of addition of a new layer along with the previous concept of combination of global search local search. The global and local search combination has achieved by use of OKM operator and ABC optimization framework and along with these the new layer which makes the process of searching in general dynamic which in hence controls the acceleration issue.

The result set shows the application of algorithm on various data set which are taken as classless and unorganized data set values.

VI. Conclusion

In this research we proposed I-ABC clustering approach which is the efficient form of existing novel ABC-K-Modes algorithm. It has been observed that novel ABC-K-Modes algorithm is based on the traditional ABC K-modes algorithm and ABC optimization approach. In novel ABC-K-Modes algorithm the challenge is the acceleration issue related to it and it is also observed that it mostly applied to numerical attributes. In proposed algorithm, the implementation is being carried out over categorical attributes by applying it on Twitter social media datasets. It has been analyzed that its potential that is being seen when it’s applied to social media data is highly effective. The acceleration issue that existed in ABC-K-Modes algorithm has been successfully reduced. The experimental results demonstrated that the proposed algorithm was superior to other two well-known algorithms according to evaluation measures AC, PR, RE and RI respectively.

In near future, this algorithm could be explored to other social media platforms like facebook, YouTube to make clustering efficient for big data spreading across web. Furthermore, the proposed approach would strengthen the thought to encourage development of clustering algorithm which must be suitable for unstructured data with categorical attributes as well as mixed data.
References

I.  Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In N. Bansal, K. Pruhs, & C. Stein (Eds.), Proc. of the eighteenth annual ACM-SIAM symposium on discrete algorithms, SODA (pp. 1027–1035).

II. Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd ed. Waltham: Morgan Kaufmann; 2012.

III. Handl, J., Knowles, J., & Dorigo, M. (2006). Ant-based clustering and topographic mapping. Artificial Life, 12(1), 35–62.

IV. Hruschka, E., Campello, R., & de Castro, L. (2006). Evolving clusters in gene-expression data. Information Sciences, 176(13), 1898–1927.

V. Huang Z. Clustering large data sets with mixed numeric and categorical values. In the first Pacific-Asia Conference on Knowledge Discovery and Data Mining. 1997; pp. 21–34.

VI. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery. 1998; 2: 283–304.

VII. Ikura Y, Gimple M. Efficient scheduling algorithms for a single batch processing machine. Operations Research Letters. 1986; 5: 61–65.

VIII. Ji J, Pang W, Zheng Y, Wang Z, Ma Z (2015) A Novel Artificial Bee Colony Based Clustering Algorithm for Categorical Data. PLoS ONE 10(5): e0127125. doi:10.1371/journal.pone.0127125

IX. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2004). A local search approximation algorithm for k-means clustering. Computational Geometry, 28(2–3), 89–112.

X. Kao Y-T, Zahara E, Kao IW. A hybridized approach to data clustering. Expert Systems with Applications. 2008; 34: 1754–1762.

XI. Karaboga D, Basturk B. On the performance of artificial bee colony (ABC) algorithm. Applied Soft Computing. 2008; 8: 687–697.

XII. Karaboga D, Ozturk C. A novel clustering approach: artificial bee colony (ABC) algorithm. Applied Soft Computing. 2011; 11: 652–657.

XIII. Li, L., Yang, Y., Peng, H., & Wang, X. (2006). An optimization method inspired by chaotic ant behavior. International Journal of Bifurcation and Chaos, 16, 2351–2364.

XIV. Luo C, Pang W, Wang Z (2014) Semi-Supervised clustering on heterogeneous information networks. In: Proceedings of 18th Pacific Asia Conference of Knowledge Discovery and Data Mining (PAKDD’14). Taiwan, pp 548-559.
XV. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley symposium on mathematical statistics and probability (pp. 281–297).

XVI. Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. 1971; 66: 846–850

XVII. Shamanth Kumar, Fred Morastatter, Huan Liu, Twitter Data analytics, Springer, Aug 19, 2013.

XVIII. Shelokar PS, Jayaraman VK, Kulkarni BD. An ant colony approach for clustering. Analytica Chimica Acta. 2004; 509: 187–195.

XIX. Teodorović D. Bee Colony Optimization (BCO). In: Lim C, Jain L, Dehuri S, editors. Innovations in Swarm Intelligence. Berlin: Springer-Verlag; 2009. pp. 39–60.

XX. Van der Merwe, D. W., & Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. In Proceedings of IEEE congress on evolutionary computation (pp. 215–220).

XXI. Wan M, Li L, Xiao J, Wang C, Yang Y. Data clustering using bacterial foraging optimization. Journal of Intelligent Information Systems. 2012; 38: 321–341.

XXII. Wan, M., Li, L., Xiao, J., Yang, Y., Wang, C., & Guo, X. (2010). CAS based clustering algorithm for web users. Nonlinear Dynamics, 61(3), 347–361.

XXIII. Yang Y. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval. 1999; 1: 67–88.

XXIV. Zhang C, Ouyang D, Ning J. An artificial bee colony approach for clustering. Expert Systems with Applications. 2010; 37: 4761–4767.