A Hybrid Genetic Fuzzy k-modes and Artificial Bee Colony Approach for Clustering YouTube Data

Akash Shrivastava
Computer Science Engineering Department
DIT University, Dehradun

M. L. Garg, PhD
Computer Science Engineering Department
DIT University, Dehradun

ABSTRACT
The genetic fuzzy k-modes algorithm introduced by G.gan, J. Wu and z. Yang. The algorithm is proven to be very effective for cluster structures retrieved from categorical datasets. However, the algorithm is prone to fall into global optima when it is required to be implementing on streaming datasets like social media data. In order to search for suitable approach to overcome the global optimal challenge the hybrid approach of genetic fuzzy k-modes and artificial bee colony is being proposed. Experiments on YouTube datasets are carried out to illustrate the performance of proposed algorithm.

Keywords
Clustering, genetic algorithms, Categorical data, YouTube, artificial bee colony

1. INTRODUCTION
Clustering is the process that is required to classify the data into categories. The category actually depends upon the parameters of similarity and dissimilarity associated with the parameters. As clustering exists in the human life in real time where the history of human development is evident where classification plays an important role [1]. There are existing clustering approaches that have been developed and designed in a manner where data objects are being grouped together with each other. Clustering algorithms ensures that clustering process executed in such a proper manner that objects belong to similar cluster must possess same properties. On the other hand, the data elements which are outside the cluster are not being considered for the clustering process. The unique feature that every clustering algorithm must include is to contain consistently is that cluster possesses only those elements which must have same properties. It not only maintains the clustering basic property but also enable the clustering phenomenon much efficient from the point of computation and efficiency. Generally, clustering divides into two classes [2]: one is hard clustering algorithms and other is fuzzy clustering algorithms. Former is clustering approach where each object supposed to belong to only one cluster uniquely. Earlier hard clustering approaches had been found majorly in the context of clustering. Latter one is just reverse of the former one where every object is permissible to contain membership functions belongs to all cluster rather than bound to belong only one cluster. However, classification system works in two ways: supervise and unsupervised. In supervise classification, the approach works in a manner where set of input data objects is being represented in terms of mathematical function [3] [4] [5]. The set of objects generally contains data elements depending upon the similarity measure of those objects. The similarity measure associated with the data object is actually indicates how close those objects are exists in the cluster. On the other hand unsupervised classification exists which is another class of classification. Unsupervised classification also refers as the term exploratory data analysis [6]. Unsupervised classification actually works on the principle where data available for clustering is unlabeled. The unlabeled means uncertain data which is not being recognized by any of the format but required to be classified [7]. This kind of data contains the information in the insight of data which is needs to be harnessed and transformed into useful format. Unsupervised clustering separates the dataset into classes in a way where unlabeled data which exists in finite number is supposed to be converting into finite set of natural hidden datasets. The goal of unsupervised classification is very clear on the fact that each data object marked unlabeled must be formatted into finite data structure organization.

2. RELATED WORK
Fuzzy clustering algorithm is specifically focusing on the fact that every data object is allowed to have membership function belong to all cluster underlying in the clustering phenomenon. Fuzzy clustering algorithm is also very popular series of clustering algorithm in context to deal with high dimensional data. The traditional clustering algorithm is used to deal with the numerical data whereas fuzzy approach has been developed with the intent to cluster the data with high dimensions. But fuzzy k-means algorithm was proven not enough to contribute much in that direction. However, for numerical datasets fuzzy k-means algorithm is efficiently being implemented to achieve the clustering objectives. In case, categorical datasets are being tested in fuzzy k-means then it is not able to produce significant results. The research focus in this paper is certainly intended to address the challenge to cluster highly categorical dataset which frequently encountered on web-space especially through social medium.

In the same direction, fuzzy K-means algorithm is being evolved [8] which was intent to integrate fuzzy concept with the K-means algorithm which is very popular clustering approach [9]. Fuzzy K-means algorithm is only worked for numerical datasets and as per the huge requirement to deal with data having categorical attributes the extension of fuzzy k-means is being proposed by Huang & Ng in [10] as fuzzy K-modes algorithm. Fuzzy K-modes algorithm is being developed to particularly focus on the dataset which is highly categorical in nature. The categorical data is very difficult to cluster due to its high dimensionality and uncertainty. It became very recognizable fact regarding the fuzzy k-modes algorithm that it certainly able to deal with categorical attributes but like fuzzy k-means it also contains one limitation which is that these both algorithms are prone to fall into local optima. These local optima are associated with the optimization problem as the function involved in the algorithm is non-convex in nature. The problem to get global solution of the optimization problem is evident from past years. Genetic algorithms (GAs) [11] and Tabu search based techniques [12] has been applied to resolve the challenge that provide global solution. In order to determine global solution few existing algorithms have been integrated. Like, genetic approach is being combined with the k-means algorithm [13].

The integration done on the basis of merits associated with the algorithms which can be utilized to find the global optimal solution. Genetic algorithm evolved in a way that it executes on the coding of set of parameter instead of individual parameter. The set of parameter coded in a prescribed format evolved from the genetic algorithm. This set of coded parameter is actually referred as solutions or chromosomes. The algorithm works in that manner that the objective function involved in the approach is happened to be executed in a manner that its value at solution is just same as its value on the corresponding parameter.

3. GENETIC FUZZY K-MODES

Genetic algorithm as mentioned in above section implement the set of coded parameter which consider the value of objective function at solution is the one that corresponds to the same value on individual parameter. The fix number of population which is also referring as population size is being utilized in the genetic algorithm which belongs to the solutions. The Genetic algorithm has designed and evolved in a way which produces generations. These are the generations which produce new population from the current population. The new population generated from the current population is a resultant produced from series of functions which includes natural selection, crossover and mutation. Genetic fuzzy k-modes are the inclusion of fuzzy k-modes merits with the integration of genetic algorithm. Generally, genetic algorithm includes string representation which extracted from coding, population initialization, selection, crossover and mutation. The experiment performs on the genetic fuzzy k-modes shows the results which evidently proven to exhibits low convergence process. The limitation of genetic algorithm is being resolved by integrating genetic algorithm with fuzzy k-modes algorithm instead of crossover process. Crossover process is being omitted from the genetic algorithm with the fuzzy k-modes which has been added to enhance the convergence process. The one-step fuzzy k-modes have been integrated in the genetic k-modes algorithm to speed up the convergence process. The effectiveness of algorithm has increased by integrating two approaches underlying in the clustering process. The algorithm now framed in a way which includes the following five steps of GA for fuzzy k-modes clustering which is intended to apply on the categorical dataset.

Step 1: Coding has been done to represent string which represented in the form of matrix P of a x b where a is the number of objects and k represent the k fuzzy membership of the first data point.

Step 2: This step is initialization phase where size of population is M. The chromosome (n1, n2, n3……n_{a-b}) has been generated using the method in [14].

for i = 1 to m do

Generate k random numbers $x_{i1}, x_{i2}, \ldots\ldots x_{ik}$ from [0, 1]

for the ith point of the chromosome;

Calculate $n_{(j-1)*m+I} = x_{ij} / \sum_{l=1}^{k} x_{il}$ for j = 1, 2…, k;

end for

if the generated chromosomes satisfies consecutive membership matrix then next set of chromosomes is being generated otherwise repeat the above step.

Step 3: In this step fitness of chromosome has been calculated using rank based evaluation function as $F(g_i) = \alpha(1-\alpha)^{r_i-1}$, where g_i is the ith chromosome in the population, r_i is the rank of g_i, and $\alpha \in [0,1]$ is a parameter which indicates the selective pressure of the algorithm. The selection process is based on [14]. Pj be the cumulative probabilities defined as

$$P_j = \begin{cases} 0 & for\ j = 0; \\ \frac{\sum_{i=1}^{j} F(gi)}{\sum_{i=1}^{N} F(gi)} & for\ j = 1,2 \ldots\ldots, N \end{cases}$$

Then the new population is generated as follows:

for i = 1 to m do

Generate a random real number x from [0, 1];

if $P_{j-1} < x < P_j$ then

select g_i;

end if

end for

Step 4: One step fuzzy k-modes algorithm has been implemented in this step which actually replaced crossover process. As per the theorem defined in [10]

for t = 1 to N do

Let P_t be the fuzzy membership matrix represented by g_t;

new set of cluster centers Z_t has been obtained as per P_t

the fuzzy membership matrix P_t given has been obtained as per Z_t

Replace g_t with the chromosome representing P_t

end for

Step 5: In this final step mutation of chromosomes have been takes place as follows:

for i = 1 to M do

Here $(y_1, y_2, \ldots\ldots y_{n-k})$ indicates g_t

for i = 1 to m do

Generate a random number $x \in [0, 1]$;

if $x \leq P_m$ then

Generate k random numbers $x_{i1}, x_{i2}, \ldots\ldots x_{ik}$ from [0, 1] for the i^{th} point of the chromosome;

Replace $n_{(j-1)*m+I}$ with $x_{ij} / \sum_{l=1}^{k} x_{il}$ for j = 1, 2…, k;

end if

end for

end for

4. ARTIFICIAL BEE COLONY ALGORITHM

In [15], the artificial bee colony (ABC) had been proposed on the basis of foraging behavior of bees. As per karaboga and bastruk their proposed approach is one of the simplest and efficient algorithms. Their approach is proven to be robust which may be applicable to various datasets across different platforms. The key observation regarding the artificial bee colony algorithm is the fact that it has been utilized for numerical datasets. The categorical data is very difficult to be observed and experiment with the evolved approach. ABC algorithm involves three types of bees:

1. *Employed bees*: The bees which are responsible to take the required food source and ensure the information sharing among other bees involved in the process.
2. *Scout*: The bee refers as a scout which is in search for a new food source in the scope of search space.
3. *Onlookers*: The information shared by employed bees is being utilized by the bees and find a new food source. These bees are called as onlookers.

The artificial bee colony executed on two parts: In the first part, employed bees are being activated and second half belongs to onlookers. Following is the steps involved in the procedure are given as follows:

Step 1: Initialize the population associated with the food sources.

Step 2: Evaluate the corresponding nectar amounts by sending the employed bees over food sources.

Step 3: Probabilities of all food sources are being evaluated by choosing the onlooker bees, and value of probability linked with each food source is identified by its nectar amount. It is to be proposed that as bigger as the nectar amount of the food sources it results to higher probability value;

Step 4: Onlookers bees are allowed to send to food sources: The calculated probabilities in above step decide the food source to be chosen by onlooker. Then the food source is being exploited and the nectar amount has been evaluated. Finally, greedy selection process is being implemented;

Step 5: In case, if the food source consumed by the employed bee is being exhausted then the same employed bee becomes scout bee;

Step 6: These scout bees then send into the search space for randomly finding the new source to be utilized;

Step 7: The identified best food source so far is being memorized;

Step 8: If the required food source obtained then output the best food source otherwise go to step 2.

5. PROPOSED HYBRID GENETIC FUZZY K-MODES AND ARTIFICIAL BEE COLONY ALGORITHM

The proposed algorithm consists of merits of the two above described algorithms. The fuzzy approach of the genetic fuzzy k-modes algorithms is combined with the artificial bee colony algorithm. The foraging behavior of bees are being observed and applied for clustering. The clustering approach developed in a way that it allows to implement over real-time streaming data. Earlier, it applied majorly for numerical dataset. The approach is presented for categorical dataset of high dimensions.

Proposed approach:

Input: N is the size of bee colony, number of clusters k, and L.

Output: The best food source.

1. Initialize the population of food sources randomly as $F_p = \{p_1, p_2, \ldots\ldots\ldots, p_H\}$.
2. For each food source, select k data objects randomly from the dataset X as the mode of cluster;
3. Nectar amount is considered as $NA(p_i)$ which is calculated as

$$NA\ (p_i) = \frac{1}{E(p_i)+1}$$

4. For each employed bee

a. Generate a new food source p_j from the current food source p_i by using initialization step used in genetic fuzzy k-modes
for i = 1 to n do
Generate k random numbers $x_{i1}, x_{i2}, \ldots\ldots x_{ik}$ from [0, 1]
for the ith point of the chromosome;
Calculate $n_{(j-1)*n+I} = x_{ij} / \sum_{l=1}^{k} x_{il}$ for j = 1, 2…, k;
end for

b. Evaluate the nectar $NA\ (p_j)$ for the food source p_j as per the step 3.
c. If $NA\ (p_j) > NA\ (p_i)$, then new food source replace the current food source.

5. Probability of food source q_i is being evaluated as

$$q_i = \frac{NA(p_i)}{\sum_1^H NA(p_i)}$$

6. For each onlooker bee
   a. As per the evaluated probabilities one food source p_i as the current food source.
   b. Generate a new food source p_k from the current food source p_i by using initialization step used in genetic fuzzy k-modes
   for i = 1 to n do
   Generate k random numbers $x_{i1}, x_{i2}, \ldots\ldots x_{ik}$ from [0, 1]
   for the ith point of the chromosome;
   Calculate $n_{(j-1)*n+I} = x_{ij} / \sum_{l=1}^{k} x_{il}$ for j = 1, 2…, k;
   end for
   c. Evaluate the nectar amount of p_k, i.e. $NA\ (p_k)$
   d. If $NA\ (p_k) > NA\ (p_i)$, the current food source p_i is replaced by the new food source p_k
   e. Otherwise the current food source p_i is retained;
   f. Update the probability q_i.
7. For each food source p_i, if the exploitation number E_{ni} is no less than L, this food source is abandoned, and the corresponding employed bee becomes a scout.
8. If there exists an abandoned food source p_i.
   a. Send the scout in the search space to find T candidate food source $\{p_i^1, p_i^2, \ldots\ldots\ldots\ldots\ldots p_i^T\}$.
   b. Evaluate the nectar amounts $\{N(p_i^1), NA\ (p_i^2), NA\ (p_i^3), \ldots\ldots\ldots\ldots NA\ (p_i^T))\}$
   c. Choose the food source with the highest nectar amount as the new food source p_j.
   d. If $NA\ (p_j) > NA\ (p_i)$, the current food source p_i is replace by the new food source p_j;
   e. Otherwise the current food p_j is retained

9. The number of cycles is being updated with one.
10. If number of cycle reach to maximum number of cycle then terminate the algorithm and output the best food source;
11. Otherwise go to step 4.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the performance of our proposed hybrid genetic fuzzy k-modes and artificial bee colony clustering algorithm is being experimented and evaluated; the proposed algorithm has been carried out by executing on YouTube real time streaming categorical dataset. Yangs accuracy measure [16] has been adopted to validate the experiment carried out in the research. 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 performance analysis, the proposed hybrid genetic fuzzy k-modes and artificial bee colony clustering algorithm implemented on real time streaming YouTube datasets. Then the clustering results of the proposed algorithm is being compared with that of the other two algorithms i.e. Genetic Fuzzy k-modes and artificial bee colony algorithms in terms of the best (Best), average (Avg.), and standard deviation of AC, PR, RE, and RI. All algorithms are implemented in python language and executed on Intel core i7, 3.9 GHz, 64GB RAM computer. In all experiments, the parameters of the proposed clustering algorithm are set as follows: S=20, MCN=1000, which are typical values used in the original ABC algorithm [17]; B=5 and V=5 are set by the rule of thumb. In all three algorithms, the cluster number k is set as per the number of classes provided by the class information of the dataset. It has been observed 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 and credibility involved in the methods and variables used are being maintained up to larger extent. The unstructured and categorical social media data must undergo the hybrid approach for better clustering.

Table 1 The AC of the three algorithms on the YouTube dataset

Algorithms	AC		
	Best	Avg	Std
Hybrid Genetic Fuzzy k-modes and ABC	0.9205	0.9123	0.0102
Genetic fuzzy k-modes	0.9103	0.8967	0.0060
ABC	0.8434	0.8924	0.0203

Table 2 The PR of the three algorithms on the YouTube dataset

Algorithms	PR		
	Best	Avg	Std
Hybrid Genetic Fuzzy k-modes and ABC	0.9012	0.8792	0.0093
Genetic fuzzy k-modes	0.8312	0.8691	0.0095
ABC	0.8135	0.8671	0.0122

Table 3 The RE of the three algorithms on the YouTube dataset

Algorithms	RE		
	Best	Avg	Std
Hybrid Genetic Fuzzy k-modes and ABC	0.8275	0.8151	0.0064
Genetic fuzzy k-modes	0.7971	0.8111	0.0075
ABC	0.7959	0.8105	0.0083

The implementation of the algorithm over the retrieved YouTube data sets shows a significant change in the best, average and lower standard values in AC, PR, RE and RI and therefore the algorithm justified to be a better computability than the rest of the two algorithms used for comparison. The Proposed algorithm working is essentially based on the principle of addition of a new layer along with the approach of combination of global search and local search. The social media datasets may be structured through the proposed method.

7. CONCLUSION

The existing clustering algorithms have majorly seen to be applied for numerical data. The categorical data is frequently encountered which need to be cluster and classify in structured format. The genetic fuzzy k-modes and artificial bee colony algorithms combined to form a hybrid approach which is being experimented to generate better results for clustering the categorical datasets. The other clustering algorithms are proposed to be applying for other social media datasets in future.

8. REFERENCES

[1] M. Anderberg, Cluster Analysis for Applications. New York: Academic, 1973.

[2] Jain A., & Dubes, R. (1988). Algorithms for clustering data. Englewood Cliffs, New Jersey: Prentice Hall.

[3] C. Bishop, Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995

[4] V. Cherkassky and F. Mulier, Learning From Data: Concepts, Theory, and Methods. New York: Wiley, 1998.

[5] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed. New York: Wiley, 2001.

[6] B. Everitt, S. Landau, and M. Leese, Cluster Analysis. London: Arnold, 2001.

[7] A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," ACM Comput. Surv., vol. 31, no. 3, pp. 264–323, 1999.

[8] Dunn, J. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 3(3), 32–57.

[9] Ruspini, E. (1969). A new approach to clustering. Information and Control, 15, 22–32.

[10] Huang, Z., & Ng, M. (1999). A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, 7(4), 446–452.

[11] Davis, L. (1991). Handbook of genetic algorithms. New York, USA: van Nostrand Reinhold.

[12] Glover, F., & Laguna, M. (1997). Tabu search. Boston: Kluwer Academic Publishers.

[13] Krishna, K., & Narasimha, M. (1999). Genetic k-means algorithm. IEEE Transactions on Systems, Man and Cybernetics, Part B, 29(3), 433–439.

[14] Zhao, L., Tsujimura, Y., & Gen, M. (1996). Genetic algorithm for fuzzy clustering. In Proceedings of IEEE International Conference on Evolutionary Computation, 1996 (pp. 716–719). Nagoya Japan: IEEE.

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

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

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