A Modified K-means Clustering Technique to find Initial Cluster Center Using Genetic Algorithm

Himanshu Zankat[1],Prof.Daxa Vekariya[2]
M.E. Student, Department of Computer Engineering, Noble Group of Institutions,Junagadh,Gujarat,India[1]
Assistant professor,Department of Computer Engineering, Noble Group of Institutions,Junagadh,Gujarat,India[2]

ABSTRACT: Genetic algorithm is used to find initial cluster center for K-means. Genetic algorithm is better option to solve the local minima problem of K-Means and will give a better initial cluster center. But Genetic algorithm also has a local minima problem and also easy to trap in blind search within a search area. In this work modified genetic algorithm with modified cuckoo search and try to get better initial cluster center than genetic algorithm for K-Means initial cluster center.

KEYWORDS: Initial cluster center, Genetic Algorithm, Cuckoo search.

I. INTROCTION

Clustering using K-Means, it's result is depending on initial cluster center. Here we try to find more global optimal initial cluster center using modified genetic algorithm. It will also improve clustering results. We try to get initial cluster center which is very close to final initial cluster center. This will make a good cluster. Select a bad initial cluster center may lead to different clustering result. In this paper we present a modified genetic algorithm approach to find a better initial cluster center and try to solve local minima problem. Cluster has a suboptimal solution when center is not closed to final cluster center. Genetic algorithm is used for different optimized problems.

II. GENETIC ALGORITHM

Genetic algorithm produces a search result based on the natural selection and genetics. Genetic algorithm is used to get the optimized result. It is a heuristic algorithm based n natural selection and genetics. It has a local minima problem.
It has mainly three operators.
1. Selection.
2. CrossOver.
3. Mutation.
Selection: Chromosomes are selected from the initial population of n chromosomes.
Crossover: Crossover the genes of two parents chromosomes to produce a better string.
Mutation: After crossover, if needed Apply mutation which is changing 0 to 1 or 1 to 0.
Fitness Function: It is used to check the fitness of a String.

Genetic Algorithm
Step 1: Generate a population of n chromosomes.
Step 2: Crete a new string by repeating following steps.
Step 3: select two parent chromosomes.
Step 4: Crossover.
Step 5: Mutation.
Step 6: Fitness Function
(If required fitness got then stop, otherwise goto step 2)
Step 7: End

III. CUCKOO SEARCH

Cuckoo search algorithm is based on behavior of cuckoo. In this algorithm Cuckoo egg is consider as a new solution and egg in a nest consider as a solution. If Cuckoo egg is a better solution then old egg, replace old egg with new egg.

The Pseudo-code of Cuckoo Search:
Begin
Objective function f(x),x =(x1,x2,x3……xd)
Generate initial population of n host nests xi(i=1,2..N)
While (t< Max Generation) or (stop criterion)
Get a cuckoo randomly by Levy Flights
Evaluate its fitness Fi
Choose a nest among n (say j) randomly

If (Fi>Fj)
Replace j by the new solution;
End If
Fractions (pa) of worse nests are abandoned and new ones are built. Keep the best solutions or nests with quality solutions; Rank the solutions and find the current best
End while
Post process results and visualization
End

IV.PROPOSED ALGORITHM

To solve the local minima problem of Genetic algorithm ,we proposed an algorithm which crossover all higher bit of second chromosome to all lower bit of first chromosome and produce target string from first Chromosome. This change of crossover technique is used to produce more optimal target string. Generate a random cuckoo string, crossover all higher bit of cuckoo string to all lower bit of produced target string. This crossover produce more optimal target string than previous target string. On this way we get better initial cluster center.

Step 1: Generate a population of n chromosomes.
Step 2: Crete a new string by repeating following steps.
Step 3: Select two parent chromosomes.
Step 4: Crossover all higher bit of second parent chromosome to all lower bit of first parent chromosome and form a new more optimal target string from first chromosome.
Step 5: generate a random cuckoo string and crossover all higher bit of cuckoo string to all lower bit of target string. This change makes form a more optimal target string than previous target string
Step 6: Mutation.
Step 7: Fitness Function.
(If required fitness is there, then stop, otherwise goto step 2)
Step 7: End

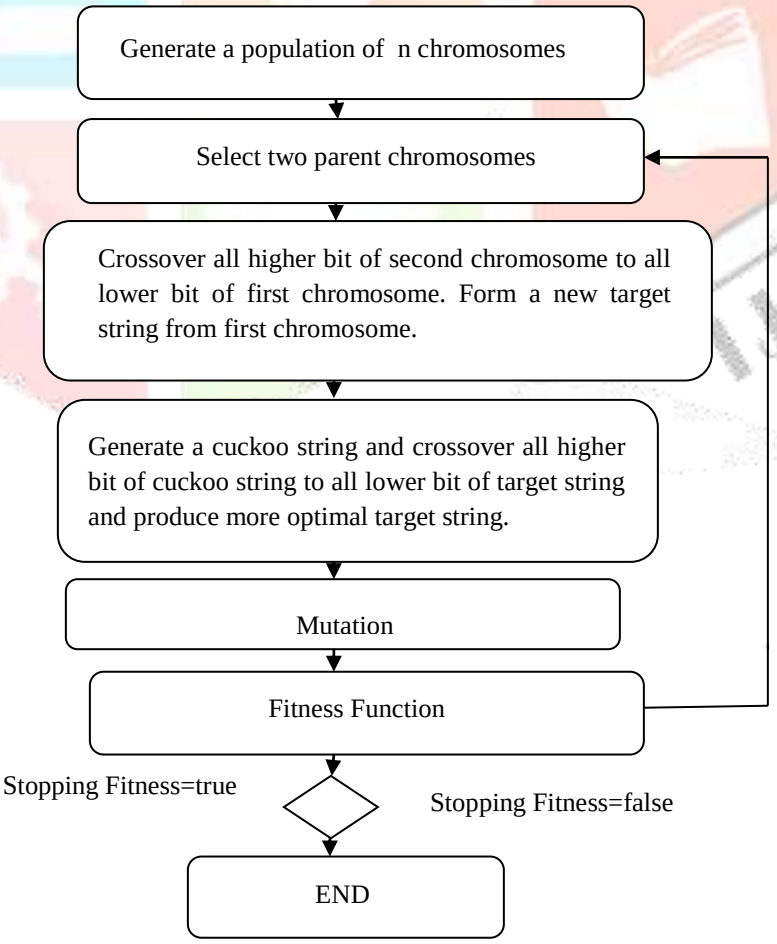

Figure 1:Flow chart of proposed algorithm

V. RESULT ANALYSIS

In order to compare the results of Genetic Algorithm and proposed algorithm, we tested both algorithms on following dataset which takes from the UCI machine learning repository.

1.2014 And 2015 CSM dataset.
2. Tennis-Major-Tornament-Match- Statistics.
(Wimbeldon-men-2013)
3. Default of credit clients.
4. Sales-Transactions-Dataset-Weekly.

Following are the graphs of initial cluster center and searching time of four datasets.Here we are using binary encoding technique to implement proposed algorithm.

Dataset: 2014 and 2015 CSM dataset.

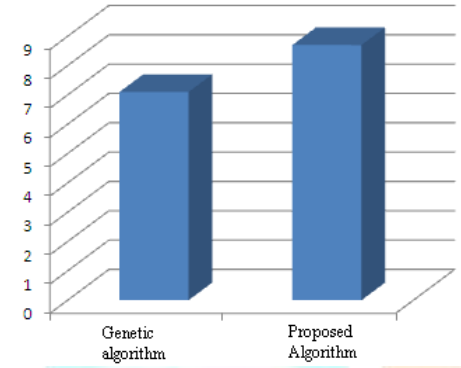

Figure 2: Initial cluster center of CSM dataset

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Initial cluster center based on Ratings of CSM dataset.
Graph in figure 2 displayed the initial cluster center result.

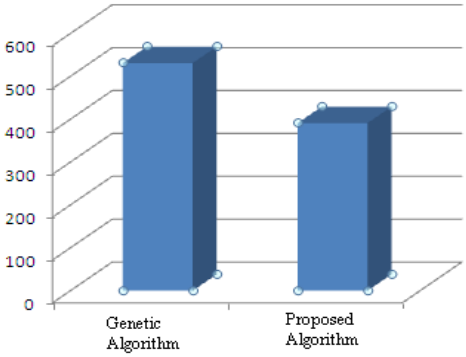

Figure 3: Searching time to search Initial cluster center of CSM dataset

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Searching time to find Initial cluster center in Milliseconds.
Graph in figure 3 displayed searching time result.
Dataset: Tennis-Major-Tournament-Match- Statistics (wimbeldon-men-2013).

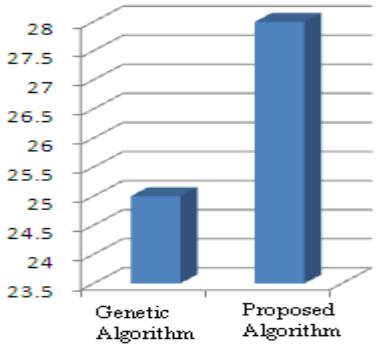

Figure 4: Initial cluster center of dataset Tennis-Major-Tournament-Match-Statistics(wimbeldon-men-2013).

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Initial cluster center based on ACE of wimbeldon-men-2013 dataset.
Graph in figure 4 displayed the initial cluster center result.

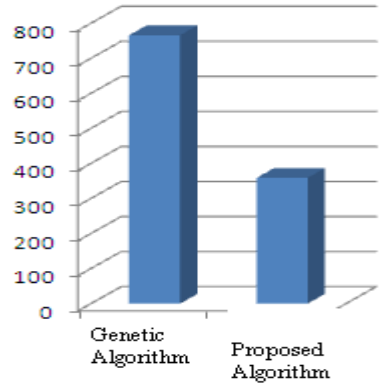

Figure 5: Searching time to search Initial cluster center of Tennis-Major-Tournament-Match-Statistics (wimbeldon-men-2013).

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Searching time to find Initial cluster center in Milliseconds.
Graph in figure 5 displayed searching time result.
Dataset: Default of credit clients.

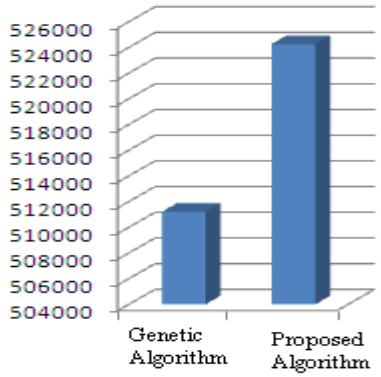

Figure 6: Initial cluster center of dataset Default of credit clients

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Initial cluster center based on LIMIT_BAL of Default of credit clients dataset.
Graph in figure 6 displayed the initial cluster center result.

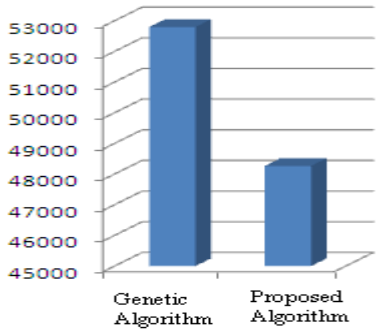

Figure 7: Searching time to search Initial cluster center of Default of credit clients

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Searching time to find Initial cluster center in Milliseconds.
Graph in figure 7 displayed searching time result.
Dataset: Sales-Transactions-Dataset-Weekly.

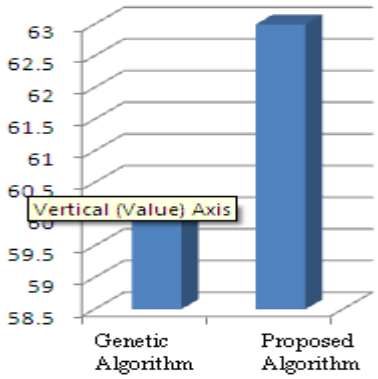

Figure 8: Initial cluster center of dataset Sales-Transactions-Dataset-Weekly

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Initial cluster center based on MAX sale of Sales-Transactions-Dataset-Weekly.
Graph in figure 8 displayed the initial cluster center result.

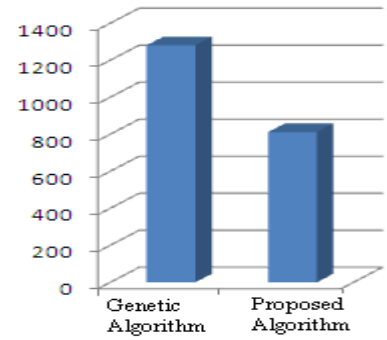

Figure 9: Searching time to search Initial cluster center of Sales-Transactions-Dataset-Weekly

X-axis:-Genetic algorithm and proposed algorithm
Y-axis:-Searching time to find Initial cluster center in Milliseconds.
Graph in figure 9 displayed searching time result.

V. CONCLUSION AND FUTURE WORK

Here we tested our proposed algorithm on related 4 data set. Proposed algorithm implements with the use of genetic binary encoding technique. We observed from graph results that proposed algorithm will give a better optimized initial cluster center than genetic algorithm. It also takes less time to search optimized initial cluster center than genetic algorithm. Proposed algorithm is also solved the local minima problem of genetic algorithm. Here Output of initial cluster center and searching time is depending on chromosomes selection, Genes selection and cuckoo String selection. Searching speed also depends on hardware configuration and code optimization. Future work is to improve the proposed algorithm which will give better initial cluster center with less searching time than proposed algorithm.

REFERENCES

1.David E.Goldberg "Genetic Algorithms in search,optimization and machine learning",Pearson
2. https://en.wikipedia.org/wiki/Genetic_algorithm
3. https://en.wikipedia.org/wiki/Cuckoo_search.
4. Li Jun Tao,Liu Yin Hong,Hao Yan. "The improvement of a K-Means Clustering Algorithm".2016 International Conferrence of Cloud computing and Big Data Analysis.
5. Zhongxiang Fan,Sun Yan."Clustering of College Students Based on Improved K-Means Algorithm" 2016 International Computer Symposium.
6.Aishwarya Palaiah , Akshata H Prabhu , Reetika Agrawal , Dr.S.Natarajan "Clustering Using Cuckoo Search Levy Flight" 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India.
7. Pushpendra Kumar Yadav1, Dr.N.L.Prajapati " An Overview of Genetic Algorithm and Modeling" International Journal of Scientific and Research Publications, Volume 2, Issue 9, September 2012
8. "Hongqing ZHENG,Youngquan ZHOU "A novel Cuckoo Search optimization Algorithm base on Gauss Distribution" Journal of Computational Information System 8:10(2012)
9. "Michael L. Raymer,William F Punch,Eric D Good Man,Anil Jain "Dimensionality Reduction
using Genetic Algorithms.IEEE Transaction on evolutionary computational vol. 4,no 2,july 2000
10. Edwin S.H.,Hong Ren "A Genetic Algorithm for Multiprocessor Scheduling" IEEE Transaction and Distributed System Vol-5.February 1994.
11. Shruti Kapil and Meenu Chawla "Performance Evalution of K-Means Clusetering algorithm with various distance Metrics".1st IEEE International Conference on Power Electronics,Intelligent Control and Energy Systems (ICPEICES-2016)

12. Nourddine Bouhmala,Anders Viken,Jonas Blasas Lonnum " A Multilevel K-Means Algorithm for the Clustering Problem": 2016 International Conference on Cloud Computing and Big Data Analysis.
13. https://archive.ics.uci.edu/ml/datasets.html